



SAPIENZA  
UNIVERSITÀ DI ROMA

# Harnessing Sense-Level Information for Semantically Augmented Knowledge Extraction

Scuola di Dottorato in Informatica

Dottorato di Ricerca in Informatica – XXX Ciclo

Candidate

Claudio Delli Bovi

ID number 1571878

Thesis Advisor

Prof. Roberto Navigli

Thesis defended on 12 February 2018  
in front of a Board of Examiners composed by:  
Prof. Nicola Leone (Università della Calabria)  
Prof. Gianluca Foresti (Università di Udine)  
Prof. Sara Foresti (Università di Milano)

The thesis has been peer-reviewed by:  
Prof. Andreas Vlachos (University of Sheffield)  
Prof. Ido Dagan (Bar-Ilan University)

---

**Harnessing Sense-Level Information for Semantically Augmented Knowledge  
Extraction**

Ph.D. thesis. Sapienza – University of Rome

© 2017 Claudio Delli Bovi. All rights reserved

This thesis has been typeset by L<sup>A</sup>T<sub>E</sub>X and the Sapthesis class.

Author's email: [dellibovi@di.uniroma1.it](mailto:dellibovi@di.uniroma1.it)

*We shall not cease from exploration  
and the end of all our exploring  
will be to arrive where we started  
and know the place for the first time.*

**T.S. Eliot**



## Abstract

Nowadays, building accurate computational models for the semantics of language lies at the very core of Natural Language Processing and Artificial Intelligence. A first and foremost step in this respect consists in moving from word-based to sense-based approaches, in which operating explicitly at the level of word senses enables a model to produce more accurate and unambiguous results. At the same time, word senses create a bridge towards structured lexico-semantic resources, where the vast amount of available machine-readable information can help overcome the shortage of annotated data in many languages and domains of knowledge.

This latter phenomenon, known as the *knowledge acquisition bottleneck*, is a crucial problem that hampers the development of large-scale, data-driven approaches for many Natural Language Processing tasks, especially when lexical semantics is directly involved. One of these tasks is Information Extraction, where an effective model has to cope with data sparsity, as well as with lexical ambiguity that can arise at the level of both arguments and relational phrases. Even in more recent Information Extraction approaches where semantics is implicitly modeled, these issues have not yet been addressed in their entirety. On the other hand, however, having access to explicit sense-level information is a very demanding task on its own, which can rarely be performed with high accuracy on a large scale. With this in mind, in this thesis we will tackle a two-fold objective: our first focus will be on studying fully automatic approaches to obtain high-quality sense-level information from textual corpora; then, we will investigate in depth where and how such sense-level information has the potential to enhance the extraction of knowledge from open text.

In the first part of this work, we will explore three different disambiguation scenarios (semi-structured text, parallel text, and definitional text) and devise automatic disambiguation strategies that are not only capable of scaling to different corpus sizes and different languages, but that actually take advantage of a multilingual and/or heterogeneous setting to improve and refine their performance. As a result, we will obtain three sense-annotated resources that, when tested experimentally with a baseline system in a series of downstream semantic tasks (i.e. Word Sense Disambiguation, Entity Linking, Semantic Similarity), show very competitive performances on standard benchmarks against both manual and semi-automatic competitors.

In the second part we will instead focus on Information Extraction, with an emphasis on Open Information Extraction (OIE), where issues like sparsity and lexical ambiguity are especially critical, and study how to exploit at best sense-level information within the extraction process. We will start by showing that enforcing a deeper semantic analysis in a definitional setting enables a full-fledged extraction pipeline to compete with state-of-the-art approaches based on much larger (but noisier) data. We will then demonstrate how working at the sense level at the end of an extraction pipeline is also beneficial: indeed, by leveraging sense-based techniques, very heterogeneous OIE-derived data can be aligned semantically, and unified with respect to a common sense inventory. Finally, we will briefly shift the focus to the more constrained setting of hypernym discovery, and study a sense-aware supervised framework for the task that is robust and effective, even when trained on heterogeneous OIE-derived hypernymic knowledge.



## Publications

### 2017

- Alessandro Raganato, **Claudio Delli Bovi** and Roberto Navigli. *Neural Sequence Learning Models for Word Sense Disambiguation*. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1167–1178, 7-11 September 2017.
- Simone Papandrea, Alessandro Raganato and **Claudio Delli Bovi**. *SupWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation*. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 103–108, 7-11 September 2017.
- **Claudio Delli Bovi**, José Camacho Collados, Alessandro Raganato and Roberto Navigli. *EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text*. Proceedings of the 55th annual meeting of the Association for Computational Linguistics (ACL), pages 594–600, 30 July-4 August 2017.
- **Claudio Delli Bovi** and Alessandro Raganato. *Sew-Embed at SemEval-2017 Task 2: Language-Independent Concept Representations from a Semantically Enriched Wikipedia*. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 252–257, 30 July-4 August 2017.
- **Claudio Delli Bovi** and Roberto Navigli. *Multilingual semantic dictionaries for natural language processing: The case of BabelNet*. Encyclopedia with Semantic Computing and Robotic Intelligence (ESCRI), vol. 1, no. 1, pages 1630015, 2017.

### 2016

- Luis Espinosa Anke, José Camacho Collados, **Claudio Delli Bovi** and Horacio Saggion. *Supervised Distributional Hypernym Discovery via Domain Adaptation*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 424–435, 1-5 November 2016.
- Alessandro Raganato, **Claudio Delli Bovi** and Roberto Navigli. *Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia*. Proceedings of 25th International Joint Conference on Artificial Intelligence (IJCAI), pages 2894–2900, 9-15 July 2016.
- José Camacho Collados, **Claudio Delli Bovi**, Alessandro Raganato and Roberto Navigli. *A Large-Scale Multilingual Disambiguation of Glosses*. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pages 1701–1708, 23-28 May 2016.

### 2015

- **Claudio Delli Bovi**, Luca Telesca and Roberto Navigli. *Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis*. Transactions of the Association for Computational Linguistics (TACL), vol. 3, pages 529–543, 2015.
- **Claudio Delli Bovi**, Luis Espinosa Anke and Roberto Navigli. *Knowledge Base Unification via Sense Embeddings and Disambiguation*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 726–736, 17-21 September 2015.
- Luis Espinosa Anke, Horacio Saggion and **Claudio Delli Bovi**. *Definition Extraction Using Sense-Based Embeddings*. Proceedings of the 2015 International Workshop on Embeddings and Semantics (IWES), pages 10–15, 15 September 2015.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Focus and Objectives . . . . .	7
1.2	Published Material . . . . .	8
1.2.1	Key Publications . . . . .	9
1.2.2	Publications not Included in this Thesis . . . . .	9
1.3	Contributions . . . . .	10
1.4	Outline of the Thesis . . . . .	11
<b>2</b>	<b>Preliminaries</b>	<b>13</b>
2.1	Knowledge Resources . . . . .	14
2.1.1	WordNet . . . . .	15
2.1.2	Wikipedia . . . . .	16
2.1.3	BabelNet . . . . .	19
2.2	From Words to Senses . . . . .	22
2.2.1	Word Sense Disambiguation . . . . .	23
2.2.1.1	Evaluation and Standard Benchmarks . . . . .	23
2.2.1.2	Approaches to WSD . . . . .	24
2.2.2	Entity Linking . . . . .	26
2.2.2.1	Evaluation and Standard Benchmarks . . . . .	27
2.2.2.2	Approaches to EL . . . . .	28
2.2.2.3	Joint WSD and EL: <b>Babel fy</b> . . . . .	28
2.2.3	Sense-based Vector Representations . . . . .	30
2.2.3.1	Evaluation and Standard Benchmarks . . . . .	31
2.2.3.2	<b>SensEmbed</b> . . . . .	32
2.2.3.3	<b>Nasari</b> . . . . .	33
2.3	Information Extraction . . . . .	34
2.3.1	Traditional Approaches . . . . .	36
2.3.2	Open Information Extraction . . . . .	37
2.3.3	Universal Schemas . . . . .	38
2.4	Nomenclature . . . . .	39
<b>3</b>	<b>Related Work</b>	<b>41</b>
3.1	Constructing Sense-Annotated Corpora . . . . .	42
3.1.1	Manually Curated Corpora . . . . .	43
3.1.1.1	<b>SemCor</b> . . . . .	43
3.1.1.2	The Senseval/SemEval datasets . . . . .	44

3.1.1.3	Other WordNet-annotated corpora . . . . .	45
3.1.1.4	Wikipedia-annotated corpora . . . . .	45
3.1.2	Semi-Automatic Approaches . . . . .	46
3.1.2.1	The Princeton WordNet Gloss Corpus . . . . .	46
3.1.2.2	OMSTI . . . . .	47
3.1.3	Fully Automatic Approaches . . . . .	48
3.1.3.1	WordNet-annotated corpora . . . . .	48
3.1.3.2	Wikipedia-annotated corpora . . . . .	49
3.1.3.3	BabelNet-annotated corpora . . . . .	50
3.2	Semantically Informed Open Information Extraction . . . . .	51
3.2.1	<b>Patty</b> . . . . .	53
3.2.1.1	Methodology . . . . .	54
3.2.1.2	Experimental Evaluation . . . . .	55
3.2.2	<b>WiSeNet</b> . . . . .	56
3.2.2.1	Methodology . . . . .	57
3.2.2.2	Experimental Evaluation . . . . .	58
<b>4</b>	<b>Harvesting Sense Annotations on a Large Scale</b> . . . . .	<b>61</b>
4.1	<b>Sew: A Semantically Enriched Wikipedia</b> . . . . .	63
4.1.1	The Hyperlink Propagation Pipeline . . . . .	65
4.1.1.1	Intra-page Propagation Heuristics . . . . .	66
4.1.1.2	Inter-page Propagation Heuristics . . . . .	67
4.1.2	Statistics . . . . .	68
4.1.3	Experimental Evaluation . . . . .	69
4.1.3.1	Intrinsic Evaluation: Annotation Quality . . . . .	70
4.1.3.2	Extrinsic Evaluation #1: Entity Linking . . . . .	71
4.1.3.3	Extrinsic Evaluation #2: Semantic Similarity . . . . .	71
4.1.4	A Broader Evaluation Study: <b>Sew-Embed</b> . . . . .	73
4.2	<b>EuroSense: Sense Annotations from Parallel Text</b> . . . . .	77
4.2.1	Stage 1: High-Coverage Joint Multilingual Disambiguation . . . . .	79
4.2.2	Stage 2: High-Precision Similarity-Based Refinement . . . . .	80
4.2.3	Statistics . . . . .	82
4.2.4	Experimental Evaluation . . . . .	82
4.2.4.1	Intrinsic Evaluation: Annotation Quality . . . . .	83
4.2.4.2	Extrinsic Evaluation: Word Sense Disambiguation . . . . .	84
4.3	<b>SenseDefs: A Multilingual Disambiguation of Textual Definitions</b> . . . . .	85
4.3.1	Gathering Definitional Knowledge across Resources and Languages . . . . .	87
4.3.2	The Disambiguation Pipeline on a Running Example . . . . .	89
4.3.3	Statistics . . . . .	90
4.3.4	Experimental Evaluation . . . . .	92
4.3.4.1	Intrinsic Evaluation #1: Annotation Quality . . . . .	92
4.3.4.2	Intrinsic Evaluation #2: WordNet Glosses . . . . .	93
4.3.4.3	Extrinsic Evaluation: Sense Clustering . . . . .	93

<b>5</b>	<b>Sense-Aware Extraction of Relational Knowledge</b>	<b>97</b>
5.1	DefIE: Open Information Extraction from Definitions . . . . .	99
5.1.1	Relation Extraction . . . . .	100
5.1.1.1	Constructing Syntactic-Semantic Graphs . . . . .	101
5.1.1.2	Identifying Relation Patterns . . . . .	102
5.1.2	Relation Typing and Scoring . . . . .	103
5.1.3	Relation Taxonomization . . . . .	104
5.1.4	Experimental Evaluation . . . . .	105
5.1.4.1	Quality of the Relations . . . . .	106
5.1.4.2	Quality of the Relation Taxonomy . . . . .	108
5.1.4.3	Quality of Entity Linking and Disambiguation . . . . .	108
5.1.4.4	Impact of Definition Sources . . . . .	109
5.1.4.5	Impact of the Approach vs. Impact of the Data . . . . .	110
5.1.4.6	Preliminary Study: Knowledge Resource Enrichment . . . . .	111
5.1.4.7	DefIE on SenseDefs . . . . .	112
5.2	KB-Unify: Sense-Aware Knowledge Base Unification . . . . .	113
5.2.1	Disambiguating and Unifying Knowledge Bases . . . . .	115
5.2.1.1	Identifying Seed Arguments . . . . .	117
5.2.1.2	Relation Specificity Ranking . . . . .	118
5.2.1.3	Disambiguation with Relation Context . . . . .	119
5.2.1.4	Cross-Resource Relation Alignment . . . . .	119
5.2.2	Experimental Evaluation . . . . .	120
5.2.2.1	Evaluating Knowledge Base Disambiguation . . . . .	121
5.2.2.2	Evaluating Specificity Ranking . . . . .	123
5.2.2.3	Evaluating Relation Alignment . . . . .	124
5.3	TaxoEmbed: Sense-Aware Hypernym Discovery . . . . .	127
5.3.1	The TaxoEmbed pipeline . . . . .	129
5.3.1.1	Domain Clustering . . . . .	130
5.3.1.2	Training Data Expansion . . . . .	130
5.3.1.3	Learning a Hypernym Detection Matrix . . . . .	131
5.3.2	Experimental Evaluation . . . . .	131
5.3.2.1	Evaluating Hypernym Identification . . . . .	131
5.3.2.2	Evaluating Extra Coverage . . . . .	133
<b>6</b>	<b>Release</b>	<b>137</b>
6.1	Sew . . . . .	138
6.2	EuroSense . . . . .	140
6.3	SenseDefs . . . . .	141
6.4	OIE-derived Resources . . . . .	143
<b>7</b>	<b>Conclusion</b>	<b>147</b>
7.1	Wrapping Up . . . . .	149
7.2	Future Work and Perspectives . . . . .	151
	<b>Bibliography</b>	<b>157</b>



# Chapter 1

## Introduction

When I use a word,"  
Humpty Dumpty said in a rather scornful tone,  
it means just what I choose it to mean  
neither more nor less."  
Lewis Carroll

Since the earliest days, encoding and representing the semantics of language with computational models has been the key challenge of Natural Language Processing (NLP) and Artificial Intelligence (AI). Getting a handle on the various phenomena that determine and regulate the meaning of linguistic utterances can pave the way for solving many long-standing and ambitious tasks in the field, from Machine Translation to Question Answering and Information Retrieval.

However, a complete and effective semantic model of language needs first of all reliable building blocks. In the last two decades, research in Lexical Semantics (which focuses on the meaning of individual linguistic elements, i.e. words and expressions), has produced increasingly comprehensive machine-readable resources and dictionaries in multiple languages (Section 2.1): like humans, modern NLP systems can now leverage these sources of lexical knowledge to perform Word Sense Disambiguation (WSD), i.e. to discriminate among various senses of a given lexeme, thereby improving their performance on a series of downstream tasks and applications, including Machine Translation (Chan et al., 2007; Neale et al., 2016; Pu et al., 2017), Information Retrieval (Agirre et al., 2010; Zhong and Ng, 2012), Taxonomy Construction (de Knij et al., 2011; Flati et al., 2016; Espinosa Anke et al., 2016b) and Text Categorization (Hidalgo et al., 2005; Pilehvar et al., 2017). Broadly speaking, the use of lexical knowledge resources encompasses all those NLP tasks in which modeling Lexical Semantics is crucial. Two notable examples, both strictly connected with WSD (Section 2.2), are Entity Linking (Rao et al., 2013), where entity mentions can be highly ambiguous, and Semantic Similarity (Budanitsky and Hirst, 2006; Turney and Pantel, 2010), where word-based models conflated different meanings of an ambiguous word into the same semantic representation.

As a matter of fact, the knowledge-based paradigm<sup>1</sup> has always played a key role in NLP; despite the overwhelming and well-established success of corpus-based approaches (Halevy et al., 2009; Collobert et al., 2011), purely data-driven models, whether supervised or unsupervised, still have limitations in terms of scalability and noise, particularly when dealing with fine-grained lexical distinctions. This is why, even today, the development and widespread application of lexical knowledge resources continues to be an important research thread.

### The Problem of Knowledge Acquisition

Nowadays, the main obstacle to developing lexical knowledge resources with high quality and coverage (and, from these, high-performing knowledge-based models for NLP) lies in the so-called knowledge acquisition bottleneck (Gale et al., 1992a; Buchanan and Wilkins, 1993). In fact, even though resources like WordNet (Section 2.1.1) already encode a wide variety of lexical and semantic relations (hypernymy, meronymy, etc.), knowledge-based algorithms need larger amounts of non-taxonomic relations to achieve state-of-the-art results. These relations are mostly syntagmatic in nature (e.g. car related-to driver, or play related-to game), hence not typically available inside lexico-semantic resources with a taxonomic or ontological structure. This is why the majority of syntagmatic relations have to be acquired from text, encoded suitably and harmonized with the structured information already available, in order for knowledge-based models to exploit them at best.

The challenging task described above is strongly connected to the main goal of knowledge acquisition, i.e. building and enriching knowledge resources on a large scale. As such, it has been addressed by a very broad spectrum of approaches over the last years. A popular strategy consists in starting from existing knowledge and then applying some algorithms to collect new information associated with the concepts already known: approaches of this type are strongly tied to the structure of the resource (e.g. computational lexicon, thesaurus, knowledge graph), and can range from disambiguating the textual definitions associated with those concepts (Mihalcea and Moldovan, 2001; Navigli and Velardi, 2005) to inference-based methods that learn to predict missing links inside a knowledge base represented as a labeled graph (Lao et al., 2011; Gardner et al., 2013; Gardner and Mitchell, 2015). This latter trend, referred to as Knowledge Base Completion (West et al., 2014), has recently attracted the attention of researchers working with semantic representations based on neural networks embeddings, and has led to numerous efforts in trying to learn structured embeddings of knowledge bases (Bordes et al., 2011, 2013; Socher et al., 2013; Neelakantan et al., 2015) to perform this task.

Another strategy consists, instead, in focusing on textual corpora and trying to develop models for the automatic extraction of relation triples with various techniques and degrees of supervision (Zhao and Grishman, 2005; Bunescu and Mooney, 2007; Banko et al., 2007; Kozareva and Hovy, 2010; Carlson et al., 2010). This broad research area, known as Information Extraction (Section 2.3), has received considerable interest over the last two decades, and covers a wide and

---

<sup>1</sup>We use the term knowledge-based to refer to any NLP approach that makes substantial use of lexico-semantic knowledge resources, as opposed to a corpus-based approach that instead relies on textual corpora (regardless of the degree of supervision).

heterogeneous range of approaches, from those targeting a constrained and specific set of predefined relations (e.g. in the biomedical domain) to those in which the goal is the general-purpose, unconstrained extraction of an unspecified and open set of semantic relations (Open Information Extraction, Section 2.3.2). Ultimately, all these efforts are geared towards addressing the knowledge acquisition problem and tackling one of the long-standing challenges of AI: Machine Reading (Mitchell, 2005), i.e., as Tom Mitchell puts it, the capability of automatically reading at least 80% of the factual content across the entire English-speaking web, and placing those facts in a structured knowledge base in such a way that computers would be harvesting in structured form the huge volume of knowledge that millions of humans are entering daily on the web in the form of unstructured text.

Apart from the paradigm and the strategy used, a key issue for all these knowledge acquisition systems is that they should keep pace with the increasingly wide scope of human knowledge: new specialized terms are coined every day as new concepts are discovered or formalized, not to mention all knowledge about people, history and society that is continuously changing and evolving. On top of this, another crucial point to be addressed is multilinguality: the bulk of knowledge acquisition research to date still focuses on English, and even though lexical resources do exist for other languages, in most cases they do not have enough coverage to enable the development of accurate NLP models. This, in turn, prevents effective knowledge acquisition approaches to be implemented, especially for under-resourced languages.

### Collaborative Semi-Structured Resources

Fortunately, the stalemate caused by the knowledge acquisition bottleneck has recently begun to loosen up. A possible way of scaling up semantic knowledge, both in terms of scope and in terms of languages, lies in the so-called semi-structured resources (Hovy et al., 2013), i.e. large-scale collaborative knowledge repositories that provide a convenient middle ground between fully structured resources and unstructured textual corpora. These two extremes are indeed complementary: the former consists of manually-assembled lexicons, thesauri or ontologies which have the highest quality, but require strenuous creation and maintenance effort and hence tend to suffer from coverage problems; the latter consists instead of raw, open and unstructured text, much easier to harvest on a large scale but usually noisy and lacking proper ontological structure. Semi-structured resources seem to take the best of both worlds, insofar as they are kept up to date and multilingual and, at the same time, reliant on human-curated semantic information. Although quality should be intuitively lower when non-experts are involved in the process, it has been shown that the collaborative editing and error correction process (wisdom of the crowd) leads to results of remarkable quality (Giles, 2005).

The most prominent resource of this kind is certainly Wikipedia, the largest and most popular collaborative multilingual encyclopedia of world and linguistic knowledge. Wikipedia features articles in over 250 languages, partially structured with hyperlink connections and categories, and constitutes nowadays an extraordinary resource for innumerable tasks in NLP (Cucerzan, 2007; Gabrilovich and Markovitch, 2007; Wu and Weld, 2010). Among others, Wikipedia's semi-structured corpus of articles played recently a key role in the development of semantically informed

Information Extraction approaches, which will be examined more closely in Section 3.2: these approaches lay down the basis for the sense-aware Information Extraction techniques that will constitute the core of Chapter 5. A great deal of research has also focused on enriching Wikipedia itself, thereby creating taxonomies (Ponzetto and Strube, 2011; Flati et al., 2014) and semantic networks (Navigli and Ponzetto, 2012; Nastase and Strube, 2013). Furthermore, machine-readable resources drawing upon Wikipedia have been continuously developed, including Wikidata (Vrandečić, 2012), YAGO (Mahdisoltani et al., 2015), and DBpedia (Lehmann et al., 2014).

The crucial limitation of semi-structured resources, however, is that they tend to focus only on encyclopedic aspects of knowledge and neglect lexicographic ones (i.e. the knowledge encoded in dictionaries). In some cases this is intentional, since collaborative resources are first of all designed for humans to read. Wikipedia, for instance, provides style guidelines<sup>2</sup> suggesting users to hyperlink a certain concept or entity only when relevant and helpful in the context of the page: this avoids cluttered and less-readable articles, but prevents a lot of common-sense knowledge and basic word senses to be modeled within the Wikipedia structure. This issue, among others, will be tackled in the first part of the thesis: in fact, in Section 4.1 we will show how the structure of Wikipedia can be leveraged effectively to turn Wikipedia itself into a full-edged semantically annotated corpus.

### Linking Knowledge Sources Together

Given the advantages and limitations of both structured and semi-structured resources, devising a way of bringing together the fully structured information on general concepts (from the former) and the up-to-date, wide-ranging world knowledge (from the latter) appears to be the key step towards the ambitious objective of creating a comprehensive lexical resource, capable of covering both encyclopedic and lexicographic information for as many languages as possible. Such a resource would enable NLP applications to integrate information otherwise available only in a multitude of heterogeneous lexical resources, thereby laying a solid foundation for large-scale approaches tackling the knowledge acquisition problem. For instance, let us consider a Question Answering scenario, where an intelligent system needs to know (or infer) that Pink Floyd was a group of people: although Wikipedia can be used to discover that Pink Floyd was indeed a band, having a link from band to its correct sense in, e.g. WordNet, would allow the system to immediately follow a hypernymy chain to organization, whose definition includes a group of people.<sup>3</sup>

Apart from Question Answering, the landscape of NLP applications that a comprehensive, multilingual lexico-semantic resource can potentially enable varies widely. Recent research work has already shown the development of effective knowledge-based strategies for joint Word Sense Disambiguation and Entity Linking in multiple languages (Sections 2.2.1 and 2.2.2), as well as multilingual and cross-lingual sense-aware Semantic Similarity (Section 2.2.3). In line with this trend, we will largely exploit a knowledge resource of this type throughout this thesis: BabelNet (Section 2.1.3). BabelNet (Navigli and Ponzetto, 2012) is a multilingual encyclopedic

<sup>2</sup>[https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style)

<sup>3</sup>Example borrowed (and slightly modified) from Gurevych et al. (2016).

dictionary and semantic network originally designed as the seamless integration of WordNet and Wikipedia, which has now become the largest resource of its kind: 13 million concepts and entities, 380 million semantic relations and 271 languages covered. Using BabelNet's wide-coverage sense inventory and semantic network as backbone will turn out to be a fundamental strategic choice for both harvesting sense annotations on a large scale (Chapter 4) and enabling semantically augmented Information Extraction techniques (Chapter 5).

Crucially, the effectiveness of any downstream application based on a knowledge resource depends strictly on the quality of the resource itself: in fact, seamless integration of heterogeneous knowledge requires accurate methods for linking, or aligning, the entities and concepts across the individual inventories. This task gets increasingly challenging due to the fundamental ever-changing nature of knowledge, but also to the continuous development of new knowledge resources, with their own features and advantages; all these isolated efforts foster the vision of a universal linking machine, where the more new knowledge is integrated, the more confirmation is obtained that the current knowledge is appropriate (or not). On the application side, however, it is also arguable that knowledge-based NLP ultimately needs corpus-based learning approaches to attain outstanding results, especially when dealing with semantics, as various contributions have already shown (Pilehvar and Navigli, 2014; Wang et al., 2014; Aletras and Stevenson, 2015; Toutanova et al., 2015; Camacho Collados et al., 2016c; Mancini et al., 2017).

### Is Explicit Semantics Used/Useful in NLP?

Notwithstanding the central role of semantics within NLP and AI, explicit semantic information is very challenging to extract and utilize. While the recent upsurge of deep learning has fueled the development of powerful data-driven approaches, with impressive results in many areas of AI (e.g. Computer Vision), the state of the art in developing explicit semantic models for language seems to have reached a plateau. In fact, encoding semantic information to train and test these models is a very demanding task, which can rarely be performed with high accuracy on a large scale. In the case of WSD, for instance, high-quality sense-annotated corpora have usually been constructed by relying on human annotators (Section 3.1.1): due to the intrinsic difficulty (and, to a certain extent, subjectivity) of annotating word senses manually, obtaining reliable and coherent sense annotations is highly expensive and especially difficult when fine-grained sense inventories are utilized, or when non-expert annotators are involved (de Lacalle and Agirre, 2015). In addition, as new encyclopedic knowledge about the world is constantly being collected, keeping up using only human annotation is becoming an increasingly expensive endeavor, severely hindered by the knowledge acquisition bottleneck: in fact, annotating word senses and entity mentions manually using large and up-to-date knowledge repositories like, e.g., BabelNet (Section 2.1.3), is not feasible. First of all, the number of items to disambiguate is massive; moreover, as the number of concepts and named entities increases, annotators would have to deal with the added complexity of selecting context-appropriate senses from a prohibitively large sense inventory. In terms of figures, while WordNet 3.0 (Section 2.1.1) comprises 117,659 word senses in total, BabelNet 3.0 covers as many as 13,801,844 concepts and named entities.

Not only obtaining annotations at the level of semantics can be troublesome: their actual usefulness for downstream tasks has been questioned a few times in the past, especially with respect to WSD (Kilgarri, 1997; Carpuat and Wu, 2005; Martín-Wanton et al., 2010). For example, in a very popular downstream application like Machine Translation, attempts to utilize features based on explicit semantics have brought mixed results (Carpuat and Wu, 2005; Chan et al., 2007; Wu and Fung, 2009; Neale et al., 2016): this shows that, at the very least, it is not obvious nor immediate for some NLP models to exploit sense-level information, even when they do yield performance improvements on standard benchmarks.

As a result, a great deal of NLP research is now leaving semantic modeling somehow implicit, often developing end-to-end models directly tailored to their specific tasks. Even in the area of Lexical Semantics, one of the most prominent paradigms today is that of distributional semantics and vector space models (Turney and Pantel, 2010), where words are represented as points in a vector space. The recent advances on neural networks have further increased the popularity of this technique, by allowing words to be ‘embedded’ in low-dimensional vector spaces (Mikolov et al., 2013a; Pennington et al., 2014) where they seem to capture very well a number of syntactic and semantic regularities. Such phenomena suggest that these approaches are very effective in modeling semantics implicitly, and researchers have further verified this empirically by employing them to provide performance boosts in various applications (Zou et al., 2013; Bordes et al., 2014; Weiss et al., 2015).

A major limitation of word-level vector space models is that they do not explicitly address lexical ambiguity: instead, they conflate the different meanings of an ambiguous word into a single vector representation, which might encode multiple senses implicitly if the word is ambiguous (Yaghoobzadeh and Schütze, 2016; Arora et al., 2016). While several works have tackled this issue by automatically inducing word senses from text (Huang et al., 2012; Neelakantan et al., 2014; Tian et al., 2014), sometimes even questioning whether modeling multiple word senses is useful (Li and Jurafsky, 2015), another important research thread focuses on going beyond the word level by explicitly modeling senses instead of words (Chen et al., 2014; Jauhar et al., 2015; Iacobacci et al., 2015; Rothe and Schütze, 2015; Camacho Collados et al., 2016c; Pilehvar and Collier, 2016). While the former approaches are solely based on textual corpora and hence more self-contained and flexible, their induced senses are (1) typically difficult to interpret (Panchenko et al., 2017) and, crucially, (2) not easy to map to lexical knowledge resources, a shortfall that limits their expendability within downstream application.

### On the Side of Knowledge

Broadly speaking, the contraposition between these two tendencies (implicit vs. explicit knowledge) goes far beyond distributional semantics, and it is widespread across the NLP community, constantly feeding the debate. In this thesis we take a stance in favor of sense-level approaches in which semantics is explicitly modeled. In fact, in light of our long-term goal of overcoming the knowledge acquisition bottleneck, we aim at showing that explicit sense-level information is not only useful for downstream applications, but it is actually a key component to enable large-scale NLP models that would not be easily attainable in a standard supervised

way. Among other large-scale endeavors, Open Information Extraction, intrinsically unsupervised, is one of those tasks that directly address the knowledge acquisition problem, and where a grasp of Lexical Semantics, either implicit or explicit, is fundamental. While a few recent approaches (Section 3.2) have started to study how to explicitly model Lexical Semantics in the extraction process, in Chapter 5 we take this semantically-informed approach to the next level and demonstrate the potential of sense-aware methods within a full-edged pipeline (Section 5.1) and after the pipeline (Section 5.2) to align, unify and harmonize the extracted knowledge. In addition, we show in Section 5.3 how this extracted knowledge, once `semantic` properly, can in turn be leveraged to develop a competitive sense-aware framework in the constrained, supervised setting of hypernym discovery.

The proved benefits of explicit sense-level information, however, come at a cost: as previously observed, annotating word senses is in itself a demanding task that suffers from the knowledge acquisition bottleneck, thereby being very difficult to carry out on a large scale. We address this issue in the first part of the thesis (Chapter 4) by investigating how, in various settings, the harvesting of sense annotations can be fully automatized and scaled up to larger corpora, while at the same time retaining a reasonably high quality compared to manual or semi-automatic approaches. Summing up, we believe that explicit semantic modeling is indeed possible, and worth pursuing wherever purely data-driven approaches can be effectively supported and augmented by lexico-semantic knowledge encoded in machine-readable resources. While in the present thesis we make the case for Information Extraction, we argue that sense-aware methods capable of exploiting lexical resources stands as a promising way of overcoming the knowledge acquisition problem in many areas across NLP.

## 1.1 Focus and Objectives

The core objective of this thesis lies in developing a principled approach to open-text knowledge acquisition based on explicit semantic analysis. To this aim, we investigate a series of disambiguation and extraction techniques that leverage sense-level information explicitly, not only to address lexical ambiguity in language, but also to take advantage of the scaffolding of structured lexico-semantic information provided by wide-coverage multilingual knowledge resources like BabelNet.

Given that our main target is general-purpose text in natural language (Open Information Extraction) rather than a partially-populated machine-readable knowledge base to be enriched (Knowledge Base Completion), our methodology requires, first of all, to obtain sense-annotated text on a large scale. While off-the-shelf disambiguation systems surely constitute a viable way for harvesting sense annotations, using them blindly might be suboptimal for a series of reasons (e.g. poor context or structural biases in the disambiguation algorithm). Moreover, there are semi-structured settings, such as that of Wikipedia articles, in which the partial structure of the target corpus already provides valuable information that an external disambiguation/linking system would in principle neglect.

Once equipped with a reliable way of obtaining sense annotations, together with a wide-coverage knowledge resource and a structured sense inventory of concepts and entities, we can reframe the Information Extraction task at the sense level, inves-

investigating where and how sense-aware methods are effective to semantically augment the extraction process. Working with word senses and entity mentions should, on the one hand, enhance the extraction procedure with unambiguous relation triples and, on the other, connect smoothly the extracted information with the structured knowledge that is already available in the underlying lexical resource.

To summarize, in the present thesis we tackle a two-fold objective:

- ^ Developing reliable, fully automatic methods to harvest sense annotations on a large scale : given a target textual corpus, our objective is to cover as many content words as possible by labeling them with concepts and named entities from a reference sense inventory. We aim at effective algorithms, scalable to large amounts of text in different languages, and possibly capable of exploiting multilinguality at best;
- ^ Reframing the task of Open Information Extraction at the sense level : i.e. studying the benefits of sense-aware techniques at every stage of the extraction process. In particular, working at the sense level enables us to extract high-quality unambiguous relation triples and link them to an underlying knowledge resource, where they can exploit the structural properties of the resource itself (e.g. taxonomic information, inter-resource mappings).

In a wider perspective, the long-standing challenge we are facing is that of dealing with lexical ambiguity, at least when it comes to understand the factual content of natural language utterances. This effort is strongly intertwined with overcoming the knowledge acquisition bottleneck: in fact, on the one hand, populating and enriching knowledge resources is a key step towards developing large-scale disambiguation algorithms; on the other, however, extracting information from open text is one of those tasks where facing lexical ambiguity is of the utmost importance.

Finally, in this thesis we put special emphasis on unsupervised and knowledge-based approaches, consistently with the premises laid down throughout this chapter. In fact, as we explain in Section 2.2.1, supervised disambiguation systems are currently less effective and more difficult to scale (especially in a multilingual setting) despite reporting higher accuracy on all standard benchmarks. Although research efforts in this direction are under way (Raganato et al., 2017b), studying extensively the behavior of a supervised model in our setting falls outside the scope of the present work. Nevertheless, most of our findings are general and not tied to an unsupervised or a knowledge-based scenario: on the contrary, the sense-annotated resources we present throughout Chapter 4 can potentially pave the way for large-scale supervised disambiguation systems; by the same token, semantically enriched OIE-derived knowledge is also beneficial within a supervised framework, as we show in Section 5.3 in the context of hypernym discovery.

## 1.2 Published Material

For the major part, the content of this thesis has already been published in top conferences and journals of Natural Language Processing and Artificial Intelligence. In the following sections we list these publications, together with other publications co-authored by the candidate that are not included in the present work. The former

set of publications represent the core of this thesis and their content is covered to a great extent in some chapters and sections, as indicated accordingly below.

### 1.2.1 Key Publications

Published material covered in Chapter 4:

- ^ SEW (Raganato et al., 2016b; Delli Bovi and Raganato, 2017): a semantically enriched version of Wikipedia constructed by solely exploiting its hyperlink structure and the sense inventory of BabelNet (Section 4.1);
- ^ EuroSense (Delli Bovi et al., 2017): a multilingual sense-annotated resource built via the joint disambiguation of the Europarl parallel corpus (Section 4.2);
- ^ SenseDefs (Camacho Collados et al., 2016a): a multilingual sense-annotated corpus of definitional knowledge constructed by jointly disambiguating the whole set of glosses in BabelNet (Section 4.3).

Published material covered in Chapter 5:

- ^ DefIE (Delli Bovi et al., 2015b): a full- edged sense-aware Open Information Extraction pipeline for definitional knowledge (Section 5.1);
- ^ KB-Unify (Delli Bovi et al., 2015a): a Knowledge Base Unification approach based on disambiguation and sense-based relation alignment (Section 5.2);
- ^ TaxoEmbed (Espinosa Anke et al., 2016a): a supervised distributional framework for hypernym discovery at the sense level (Section 5.3).

### 1.2.2 Publications not Included in this Thesis

A number of publications, co-authored by the candidate and strongly connected to the topics treated in this thesis, are not covered in the following chapters and sections. These publications include:

- ^ A sense-aware supervised method to extract definitional knowledge (Espinosa Anke et al., 2015), based on semantic features derived from Entity Linking and Semantic Similarity;
- ^ An experimental study on neural sequence learning models for supervised WSD (Raganato et al., 2017b), where a single all-words model achieves state-of-the-art (or statistically equivalent) results on all standard benchmarks;
- ^ A flexible, open-source Java toolkit and RESTful API for supervised WSD (Papandrea et al., 2017), designed to be modular, fast and scalable for training and testing on large datasets.

### 1.3 Contributions

Consistently with our focus and objectives laid down in Section 1.1, the work presented in this thesis puts forward the following contributions:

- ^ A series of disambiguation techniques to obtain reliable sense annotations on a large scale . Throughout Chapter 4 we deal with three disambiguation scenarios, and show that exploiting at best the structure and the properties of the target corpus is key to harvest high-quality annotations in a self-contained way (Raganato et al., 2016b) and to reduce the structural bias of o -the-shelf disambiguation algorithms (Camacho Collados et al., 2016a; Delli Bovi et al., 2017). The principled multilingual disambiguation technique that we detail in Sections 4.2 and 4.3, based on the synergy between a graph-based disambiguation system and a vector-based representation of concept and entities, is a robust, novel approach potentially capable of accommodating different settings not explored in this thesis (e.g. larger comparable texts, news articles on the same subject);
- ^ Three large-scale semantic resources , all providing an unprecedented amount of sense annotations of concept and named entities from the BabelNet sense inventory (Raganato et al., 2016b; Camacho Collados et al., 2016a; Delli Bovi et al., 2017). We assess the quality of each resource intrinsically and extrinsically, leveraging them in various NLP tasks: Word Sense Disambiguation, Entity Linking, Semantic Similarity, Sense Clustering, Information Extraction. Our experiments show that the results obtained using these resources are comparable or even superior to those obtained using a resource constructed semi-automatically. At the same time, thanks to the wide coverage of BabelNet, these resources take a leap forward in terms of scope and coverage (as they include both encyclopedic and lexicographic knowledge), in terms of languages (as BabelNet covers all the languages available in Wikipedia) and in terms of exhibility (as BabelNet's inter-resource mappings can be used to convert these sense annotations to many individual sense inventories, such as WordNet or Wikipedia). All these features contribute to reshape the landscape of WSD, opening up opportunities for supervised systems to scale to larger training sets while retaining high disambiguation accuracy;
- ^ A full- edged Open Information Extraction pipeline for de nitional knowledge . Among the various strategies presented throughout Chapter 5 to rede ne and study OIE at the sense level, in Section 5.1 we show experimentally that moving an OIE system to the denser, virtually noise-free setting of de nitional text (Delli Bovi et al., 2015b, DefIE ) is bene cial: in fact, a comprehensive semantic analysis yields unambiguous relation triples, as well as `semanti ed' relations that can leverage the taxonomic structure of BabelNet and be arranged in a relation taxonomy. Indeed, working at the sense level is the key feature that puts DefIE in line with state-of-the-art OIE approaches based on much larger datasets. Furthermore, our experimental ndings raise some questions as to where valuable knowledge can be found, and whether just tackling very noisy Web-scale corpora is always the optimal choice;

- ^ An approach to Knowledge Base Unification via sense embeddings and disambiguation . In Section 5.2 we depart from previous literature on the subject, and address the issue of merging and harmonizing OIE-derived knowledge by exploiting sense-aware semantic analysis (Delli Bovi et al., 2015a, KB-Unify ). Again, we show that exploiting the sense inventory of BabelNet as a backbone is extremely effective when interconnecting not only lexical knowledge, but also relational knowledge; to this aim, we devise a disambiguation algorithm ad-hoc for relation triples, and then use these disambiguated triples to align heterogeneous knowledge bases at the sense level. While most research efforts focus on developing new extraction procedures, we instead show that semantic analysis can be used to unify the knowledge already extracted, instead of putting forward yet another isolated, OIE-derived knowledge base. Finally, in Section 5.3 we employ hypernymic relation triples from KB-Unify as training set in a supervised sense-aware framework for hypernym discovery, TaxoEmbed (Espinosa Anke et al., 2016a), further demonstrating the robustness and flexibility of working at the sense level, even in the more constrained scenario of hypernymic relations.

Individual Contributions. Some of the contributions presented in this thesis are the output of a joint work of the candidate with other members of the Linguistic Computing Laboratory,<sup>4</sup> or with other researchers from international institutions. In terms of individual contributions, the candidate had the leading role in designing both the methodological approaches and the experimental evaluations presented in Sections 5.1 and 5.2, with the close supervision of his advisor. As regards the material in Section 5.3, the candidate's main focus, in accordance with the topic of this thesis, was the construction of an OIE-derived training dataset for hypernym discovery, as well as the experimental comparison with DefIE (cf. Section 5.22)<sup>5</sup> In Chapter 4, the candidate contributed in designing some components of the hyperlink propagation pipeline in Sew (cf. Section 4.1.1.2, and Section 4.2 in the paper), and focused on the extrinsic experiment on semantic similarity (Section 4.1.3.3, and Section 6.3 in the paper), including the evaluation study of Section 4.1.4, and the participation to the SemEval-2017 task 2 competition (Delli Bovi and Raganato, 2017). He also designed the first stage of the disambiguation pipeline in Sections 4.2 and 4.3, with their respective context enrichment strategies, and he conducted the extrinsic evaluation of SenseDefs on OIE (Section 5.1.4.7)<sup>6</sup>

## 1.4 Outline of the Thesis

The remainder of the thesis is organized as follows:

- ^ We start by providing a broad overview of the machinery used throughout

<sup>4</sup><http://lcl.uniroma1.it>

<sup>5</sup>Sections 1 to 3, and sections 5 to 8, with the exception of 6.3, in the DefIE paper (Delli Bovi et al., 2015b); sections 3 to 8, with the exception of 5.1, 5.2, and 8.2, in the KB-Unify paper (Delli Bovi et al., 2015a); sections 3 and 5.2 in the TaxoEmbed paper (Espinosa Anke et al., 2016a).

<sup>6</sup>Sections 1 to 5, with the exception of 3.2, in the EuroSense paper (Delli Bovi et al., 2017); sections 1, 3.1, 5.2.1, and 7 in the SenseDefs paper (Camacho Collados et al., 2016a).

the thesis in Chapter 2 : rst of all, a bird's-eye view on the most popular lexico-semantic resources used across the NLP community (Section 2.1), with a special emphasis on BabelNet (Section 2.1.3), extensively used in the core chapters and sections of the thesis. Then, in Section 2.2 we survey the state of the art in the areas of Word Sense Disambiguation (Section 2.2.1), Entity Linking (Section 2.2.2), and sense-based vector representations (Section 2.2.3); in this section we describe some key building blocks of the approaches treated in the later chapters. Finally, Section 2.3 introduces and contextualizes the task of Information Extraction, while Section 2.4 clarifies the nomenclature;

- ^ In Chapter 3 we narrow our focus to the key topics of the thesis. We rst analyze the related work on sense-annotated resources (Section 3.1), from manually curated corpora to semi-automatic and fully automatic approaches; we then move to OIE (Section 3.2) and look at how semantic analysis has been carried out in the published literature on the subject. In particular, we bring into focus two OIE approaches that are very similar in spirit to those treated in the present work: Patty (Section 3.2.1) and WiSeNet (Section 3.2.2). Both contributions have marked a clear turning point in the field by enforcing a deeper semantic analysis of text, laying down the basis and inspiration for the sense-aware techniques presented in the following chapters;
- ^ Chapter 4 constitutes the rst core component of this thesis. Here we address the rst objective described in Section 1.1, i.e. that of developing fully automatic methods to harvest sense annotations on a large scale. As previously explained, we deal with three different disambiguation scenarios: semi-structured text (Section 4.1), parallel text (Section 4.2), and definitional text (Section 4.3). In each case, we develop a disambiguation strategy to exploit at best the structure and characteristics of the target corpus, and produce a sense-annotated resource that is extensively validated on experimental grounds and then released to the research community;
- ^ Chapter 5 constitutes the second core component of our work. Here we address the second objective described in Section 1.1, i.e. that of reframing the task of Open Information Extraction at the sense level. We start by studying a sense-level full-edged OIE pipeline designed for definitional text (Section 5.1), where a denser and noise-free setting enables a comprehensive semantic analysis to produce unambiguous triples, and link them to BabelNet; in Section 5.2 we shift our focus at the end of the extraction process, and utilize an array of sense-aware techniques to disambiguate, align and unify heterogeneous knowledge bases extracted from open text; finally, in Section 5.3 we move from the unconstrained setting of OIE to the constrained setting of hypernym discovery, where we present a robust, sense-level supervised framework that leverages OIE-derived hypernymic knowledge at training time.
- ^ Finally, Chapter 6 showcases all the released data and resources that are associated with the material presented herein, while Chapter 7 concludes the thesis by summarizing its main findings and presenting some medium-term and long-term perspectives of future work.

## Chapter 2

# Preliminaries

If you wish to make an apple pie from scratch  
you must first invent the universe.  
Carl Sagan

In this chapter we provide some important background knowledge, necessary to put the rest of this thesis in context. Most of the following sections are devoted to give an up-to-date view on the landscape of Lexical Semantics. First of all, we go over the most important and widely used knowledge resources in the NLP community, with their differences and commonalities (Section 2.1); we put a special focus on BabelNet (Section 2.1.3), the encyclopedic dictionary and semantic network that serves as a backbone for the core Chapters of this thesis. In Section 2.2 we examine the state of the art in three fields of study that constitute the cornerstones of today's research in Lexical Semantics: Word Sense Disambiguation (Section 2.2.1), Entity Linking (Section 2.2.2) and Semantic Representations for lexical items (Section 2.2.3). Again, we put a special emphasis on some important tools that we utilize as building blocks in the following chapters: Babelfy (Moro et al., 2014b), Nasari (Camacho Collados et al., 2016c) and SensEmbed (Iacobacci et al., 2015).

In Section 2.3 we move to Information Extraction, and survey some milestone contributions in this broad and long-standing field of NLP. We focus in particular on Open Information Extraction (Section 2.3.2), i.e. the unsupervised branch of Information Extraction, overviewing the seminal papers on the subject, as well as more recent advances. We will then return on this topic later on, in Section 3.2, where we describe some important contributions connecting unsupervised extraction with Lexical Semantics (hence much closer in spirit to the present work). Finally, Section 2.4 defines some common and well-established nomenclature, that we subsequently employ in the core Chapters of thesis.

## 2.1 Knowledge Resources

As shown in Chapter 1, lexical knowledge resources are of primary importance in many areas across NLP, given their role of encoding human knowledge of language in machine-readable form. Extensively used by the research community, lexical knowledge resources exist nowadays in many flavors and with different features. At the time of writing, over 2,800 language resources are listed in the META-SHARE repository<sup>1</sup>, while the LRE Map<sup>2</sup> contains almost 4,000 entries, including lexicons, dictionaries, ontologies, and terminologies.

To our aim, a lexical knowledge resource can be operatively defined as a structured or semi-structured resource that contains information on lexical units (words and multi-word expressions) of a particular language or set of languages. This information is expressed with respect to canonical word forms, usually lemmas or lexemes (i.e. lemmas with their associated parts of speech). For instance, *wrote* is the lemma of *wrote*, and *write<sub>v</sub>* is the associated verbal lexeme. Inside a lexical knowledge resource, sense-level information is generally encoded as a set of pairings of lemma and meaning (word senses), which constitute the sense inventory of the resource<sup>3</sup>. The sense inventory associates each word sense with a unique sense identifier, in order to deal with cases where a lemma can have more than one meaning (polysemy). For example, there are two distinct meanings of the verb *to write*, which give rise to two distinct senses: one refers to communicating with someone in writing, and another one refers to producing a literary work. Accordingly, the sense inventory might identify them with two distinct identifiers, e.g. *write01* or *write02*.

Depending on its specific focus, each knowledge resource contains a variety of information (e.g., morphological, syntactic, semantic), which determines its particular internal structure. In this thesis we deal with lexico-semantic knowledge resources, where most of the encoded information consists in semantic relationships interconnecting concepts and named entities. As a result, these resources can also be represented as semantic networks i.e. graphs where nodes are concepts, and edges are semantic relationships between them. On the basis of their production process, knowledge resources can be split into three groups:

- ^ Expert-built knowledge resources : these resources are designed, created and edited by a group of designated experts (e.g. lexicographers, linguists or psycho-linguists). Despite their lower coverage, due to their slow and expensive production and maintenance cycles, they have the highest quality, and often include very specialized aspects of language. Expert-built resources are typically structured (e.g. computational lexicons, machine-readable dictionaries, or full-edged ontologies) and usually biased towards lexicographic knowledge;
- ^ Collaboratively-built knowledge resources : this category includes large-scale semi-structured repositories, typically focused on encyclopedic knowledge,

<sup>1</sup><http://www.meta-share.org>

<sup>2</sup><http://www.resourcebook.eu>

<sup>3</sup>Throughout this thesis we use the terms *meaning* and *concept* interchangeably to refer to the possibly language-independent part of a word sense. We also distinguish the case in which the word or multi-word is a mention of a named entity: in this case we use *named entity mention* in place of *word sense* and *named entity* instead of *concept*.

and constructed via the collaborative effort of a community of users. In fact, a crowd of users, that might include experts along with casual non-expert annotators, can substitute a small and organized group of experts in gathering and editing lexical information. This open approach can handle the otherwise enormous effort of building large-scale multilingual resources, that quickly adapt to new information, and yet maintain a high quality thanks to a continuous revision process (Giles, 2005);

- ^ Linked knowledge resources : this category comprises large-scale resources of a hybrid nature, constructed automatically or semi-automatically by integrating, or mapping, two or more resources from the above two groups. Inter-resource mappings are realized via accurate disambiguation and linking algorithms designed to align (link) named entities and concepts across the different individual inventories. In many cases the individual resources being integrated provide complementary information (e.g. lexicographic vs. encyclopedic knowledge) and enable the creation of wide-coverage repositories and richer knowledge representations (Gurevych et al., 2016).

In the following subsections we examine one representative example for each of the three groups above: WordNet (Section 2.1.1), Wikipedia (Section 2.1.2), and BabelNet (Section 2.1.3), respectively.

### 2.1.1 WordNet

Undoubtedly the most popular and widely used lexical knowledge resource in the area of NLP, the Princeton WordNet of English (Miller et al., 1990; Fellbaum, 1998) is an expert-built computational lexicon based on psycholinguistic principles. A concept in WordNet is represented as a synonym set (synset), i.e. a set of words that share the same meaning. For instance, the concept of play as a dramatic work is expressed by the following synset:

$$\{ \text{play}_n^1, \text{drama}_n^1, \text{dramatic play}_n^1 \}$$

where subscript and superscript of each word denote its part of speech and sense number, respectively. Hence, e.g.  $\text{play}_n^1$  represents the first nominal sense of the word play.<sup>4</sup> Being polysemous, the word play might appear in other synsets, e.g. the one referring to the concept of play as children's activity:

$$\{ \text{play}_n^8, \text{child's play}_n^2 \}$$

Similarly to traditional dictionaries, WordNet provides a textual definition (gloss) and small usage examples for each synset. WordNet synsets are also connected with lexico-semantic relations, including:

- is-a relations such as hypernymy (e.g.  $\text{play}_n^1$  is-a dramatic composition<sub>n</sub><sup>1</sup>) and hyponymy, which express concept generalization and specialization, respectively. These relations are extremely important as they structure the concepts expressed by synsets into a lexicalized taxonomy;

<sup>4</sup>n stands for noun (see Section 2.4).

- instance-of relations denoting set membership between a named entity and the class it belongs to (e.g. Shakespeare<sub>n</sub> is an instance of dramatist<sub>n</sub>). In the majority of NLP applications making use of hypernymic information, this type of relation is generally considered and regarded as a specific form of *is-a*;
- part-of relations expressing the elements of a partition by means of meronymy (e.g. stage direction<sub>n</sub> is a meronym of play<sub>n</sub>) and holonymy (e.g. play<sub>n</sub> is a holonym of stage direction<sub>n</sub>).

As of version 3.0, the Princeton WordNet contains 117,659 synsets for all open-class parts of speech, arranged in a semantic network with 364,569 lexico-semantic relations. Being hand-crafted by experts, WordNet's semantic information is of the highest quality, despite its considerably smaller scale in comparison with the other lexical resources. Various research projects stem from the Princeton WordNet, such as the eXtended WordNet (Mihalcea and Moldovan, 2001), which includes structured semantic information extracted from the textual definitions of senses, WordNet Domains (Bentivogli et al., 2004), and SentiWordNet (Baccianella et al., 2010), which assign domain labels and sentiment scores, respectively, to each synset.

Furthermore, the Princeton WordNet for English has inspired the creation of 'wordnets' in many other languages worldwide, many of which also provide links to the English senses in the Princeton WordNet. Examples include, among others, the Italian WordNet (Toral et al., 2010), the Japanese WordNet (Isahara et al., 2008), and the German WordNet (Hamp and Feldweg, 1997, GermaNet). More recently, all these language-specific efforts have been gathered, normalized and interlinked with the creation of Open Multilingual WordNet (Bond and Foster, 2013).

### 2.1.2 Wikipedia

Already introduced in Chapter 1, Wikipedia is a widely used multilingual Web-based encyclopedia with a prominent role in a great variety of NLP areas. It was conceived as a collaborative open-source medium maintained by volunteers, in order to provide a very large wide-coverage repository of encyclopedic information. Each article in Wikipedia is represented as a page (henceforth, Wikipage) and contains a variety of information about a specific concept (e.g. Play (theatre) ) or named entity (e.g., William Shakespeare ). Although, strictly speaking, Wikipedia does not provide an explicit sense inventory, the pairing of an article and the concept or entity it describes can be interpreted as a word sense. This interpretation actually complies with the bracketed disambiguation policy of Wikipedia, which associates ambiguous word in the title of a page with a parenthesized label specifying its meaning (e.g. Java (programming language) and Java (town) ).

Due to its focus on encyclopedic knowledge, Wikipedia contains almost exclusively nominal senses. However, thanks to its partially structured text, it represents an important source of knowledge from which structured information can be harvested (Hovy et al., 2013). Apart from infoboxes (tables summarizing the most important attributes of an entity such as the birth date and biographical details of William Shakespeare ), Wikipages are connected by means of a number of semantic relations, including:

- Redirect pages which used to express alternative expressions for the same concept or entity (e.g. Stageplay and Theatrical Play redirecting to Play (theatre) ), thus modeling synonymy;
- Internal hyperlinks across the text of a Wikipage, which often refer to concepts or entities related to the one being treated therein (e.g. Play (theatre) linked to Literature and Playwright ), thus representing generic or unspecified semantic relatedness;
- Inter-language links, i.e. connections between the concept or entity described in a Wikipage and its counterparts in other languages (e.g. the English Play (theatre) linked to the Italian Drama and the German Bühnenwerk );
- Categories with which multiple Wikipages are associated, used to encode common topics or features among related concepts or entities (e.g. Play (theatre) categorized as Theatre , Drama and Literature );

Wikipedia is a massive multilingual lexical resource where the number of concepts and entities being described is constantly growing; only the English subset, as of 2015, comprised more than 4.3 million Wikipages and over 71 million internal hyperlinks. Over the years, this enormous amount of information has been exploited in a variety of ways. Depending on the task, Wikipedia can be seen as:

1. A large-scale (partially) sense-annotated corpus : in fact, Wikipages bear textual content and, at the same time, implicitly define a sense inventory (as each Wikipage is unambiguously used to refer to a concept or entity). As a consequence, internal hyperlinks represent, by all rights, sense annotations. Despite not being structurally designed as a sense-annotated corpus (a shortcoming we explore deeply in Section 4.1), Wikipedia has been successfully used as such in a variety of prominent NLP tasks, including Named Entity Disambiguation (Section 2.2.2) and Information Extraction (Section 2.3);
2. A lexicalized semantic network : since each Wikipage identifies a specific concept or entity, and it is connected to related concepts or entities via its internal hyperlinks, the whole Wikipedia can be seen as a full-edged semantic network. Each node in the network encodes a concept that is lexicalized (through the title of the Wikipage and the associated redirections) and possibly language-independent (since inter-language links bring together Wikipages describing the same concept or entity in different languages). Using the hyperlink structure of Wikipedia as a semantic network has proved to be very effective for measuring Semantic Similarity (Section 2.2.3), as well as a key step to construct and refine Wikipedia-based taxonomies and ontologies (de Melo and Weikum, 2010; Ponzetto and Strube, 2011; Nastase and Strube, 2013; Mahdizoltani et al., 2015; Flati et al., 2016).

The fact that Wikipedia can be viewed as a semantic network puts forward a crucial commonality with WordNet (Section 2.1.1): despite the profound structural and conceptual differences between the two resources, they can both be represented as directed graphs. An excerpt of such graphs centered on the synset  $st_{n1}$  and

(a) Excerpt of the WordNet graph centered on the synset  $\text{play}_n^1$ . (b) Excerpt of the Wikipedia graph centered on the page Play (theatre).

Figure 2.1. Excerpts of the WordNet (a) and Wikipedia (b) graphs drawn borrowed from Navigli and Ponzetto (2012). Both resources can be viewed as directed graphs with synsets (Wikipedia pages) as nodes and relations (hyperlinks) as edges.

the Wikipedia page Play (theatre) is given in Figure 2.1(a) and 2.1(b), respectively. While there are nodes corresponding to the same concept (e.g.  $\text{tragedy}_n^2$  and Tragedy), each resource also contains specific knowledge which is missing in the other, both general concepts (for instance no Wikipedia entry corresponding to  $\text{direction}_n^1$ ) and named entities (like Ancient Greece missing in WordNet).

#### Enhancing Wikipedia: Wikidata, Freebase and DBpedia

The central role of Wikipedia in NLP has motivated a series of research efforts targeted at turning it into a fully structured knowledge resource. Among many such efforts, the Wikidata project (Vrandečić, 2012) is arguably the most prominent one. Wikidata is operated directly by the Wikimedia Foundation with the goal of providing a common source of data that can be used by other Wikimedia projects. It is designed as a document-oriented semantic database based on items, each representing a topic and identified by a unique identifier (e.g. the item for Politics is Q7163). Knowledge is encoded with statements in the form of property-value pairs.

Part of the information currently in Wikidata comes from another large-scale collaborative knowledge base Freebase (Bollacker et al., 2008). Freebase was an online collection of structured data harvested from many sources, including individual Wikipedia contributions. In contrast to Wikidata, Freebase used a non-hierarchical graph model where tables and keys were replaced by a set of nodes and a set of links expressing semantic relationships. As of today, the project has been officially discontinued, and most of its data moved into Wikidata.

Another popular Wikipedia-based project is DBpedia (Lehmann et al., 2014), a crowd-sourced community effort to extract structured information from Wikipedia and make it available on the Web by means of an RDF database and ontology accessible through SPARQL queries. Similarly to Wikidata, DBpedia exploits infoboxes as one of the richest sources of information.

As noted in Chapter 1, however, not even a wide-coverage resource like Wikipedia (or, for that matter, Wikipedia-derived resources like the ones described above) works at best for all application scenarios. Instead, the optimal way of making use of

Figure 2.2. An illustrative overview of BabelNet drawn from the original article (Navigli and Ponzetto, 2012). Unlabeled edges come from Wikipedia hyperlinks (e.g. Play (theatre) links to Musical (theatre) ), while labeled edges are drawn from WordNet (e.g.  $\text{play}_n^1$  has-part  $\text{stage\_direction}_n^1$ ).

knowledge appears to be the orchestrated exploitation of multiple, heterogeneous resources (Gurevych et al., 2016). In recent years, a series of successful approaches have demonstrated this: Menta (de Melo and Weikum, 2010), Uby (Gurevych et al., 2012), Yago (Mahdisoltani et al., 2015), and, of course, BabelNet (Navigli and Ponzetto, 2012) which we examine in the following section.

### 2.1.3 BabelNet

The high degree of complementarity between WordNet and Wikipedia, together with the similarity between their internal structures (Figure 2.1), opened the way for an integration that brought together seamlessly lexicographic information organized in a fully structured way (from WordNet) and specialized, up-to-date world knowledge in hundreds of languages (from Wikipedia). The result of this integration is BabelNet (Navigli and Ponzetto, 2012)<sup>5</sup>, a large-scale, multilingual encyclopedic dictionary (i.e. a resource where both lexicographic and encyclopedic knowledge is available in multiple languages) and at the same time a semantic network where all this knowledge is interconnected with several million semantic relations.

Formally, BabelNet is structured as a labeled directed graph  $G = (V; E)$  where  $V$  is the set of nodes i.e. concepts such as play and named entities such as Shakespeare and  $E \subseteq V \times R \times V$  is the set of edges connecting pairs of concepts or entities (e.g. play is-a dramatic composition). Each edge is labeled with a semantic relation from  $R$ , e.g., f is-a, part-of, ..., g, with  $\emptyset$  denoting an unspecified semantic relation. Importantly, each node  $v \in V$  contains a set of lexicalizations of the concept for different languages, e.g.  $\{\text{play}_{en}, \text{Theaterstück}_{de}, \text{dramma}_{it}, \text{obra}_{es}, \dots, \text{pièce de théâtre}_{fr}\}$ . Such multilingually lexicalized concepts are called Babel synsets.

Nodes and edges in BabelNet have been harvested from both WordNet and Wikipedia. In order to construct the BabelNet graph  $G$ , extraction took place at different stages: from WordNet, all available word senses (as nodes) and all the lexico-semantic relations between synsets (as edges); from Wikipedia, all the Wikipages (as nodes) and all their internal hyperlinks (as edges). A graphical overview of BabelNet is given in Figure 2.2. Crucially, the overlap between WordNet and Wikipedia (both in terms of concepts and relations) made the merging between the two resources

<sup>5</sup><http://babelnet.org>

possible and enabled the creation of a unified knowledge resource. After establishing this first Wikipedia-WordNet mapping, multilinguality was achieved by collecting lexical realizations of the available concepts in different languages. Finally, multilingual Babel synsets were connected by establishing semantic relations between them. In short, the construction of BabelNet consisted of three main steps:

1. The integration of WordNet and Wikipedia via an automatic mapping between WordNet senses and Wikipedia pages. This avoided duplicate concepts and allowed heterogeneous sense inventories of concepts to complement each other;
2. The collection of multilingual lexicalizations of the newly-created concepts (Babel synsets) by means of (a) the human-generated translations provided by Wikipedia (via inter-language links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora;
3. The interconnection of Babel synsets by harvesting all the relations in WordNet and in the Wikipedias in the languages of interest, and by subsequently weighting them to quantify the strength of associations between the source and target Babel synsets.

Since its earliest version, based solely on the WordNet-Wikipedia mapping just described, BabelNet has been continuously developed and improved, integrating new resources and pursuing the vision of a unified lexico-semantic repository capable of covering as many languages and areas of knowledge as possible. As of today, BabelNet is the largest interlinked semantic resource available, with 14 different knowledge resources integrated and 13,801,844 entries in 271 languages, interconnected in a semantic network of 380,239,084 lexico-semantic relations<sup>6</sup>. From a lexicographic perspective, BabelNet has been referred to as the dictionary of the future (Steinmetz, 2016), because of its encyclopedic breadth and scope, its organizational structure that favors semantic relatedness (instead of the mere alphabetical order of traditional dictionaries) and its richness of information which comprises, among other features, over 40M textual definitions, 10M images and 2.6M domain labels.

As discussed in Chapter 1, the construction and continuous development of a resource like BabelNet, where both lexicographic and encyclopedic knowledge is available and interconnected in multiple languages, has a great impact on a variety of downstream tasks and applications. What emerges from the NLP approaches directly powered by BabelNet that have been developed over the years, is that the advantage of using such resource is generally two-fold: on the one hand, the unification of lexicographic and encyclopedic information enables NLP systems to perform jointly, and hence with mutual benefit from one another, tasks that were previously conceived as separated; on the other, having language-independent information creates a direct bridge towards multilingual and language-agnostic methods, enabling English-centered models to be directly projected to other languages without modifying their basic structure. In Section 2.2 we examine two important examples of this: multilingual joint Word Sense Disambiguation and Entity Linking (Section 2.2.2.3) and sense-based vector representations of concepts and entities (Section 2.2.3).

<sup>6</sup>Detailed statistics, also across versions, can be found at: <http://babelnet.org/stats>.

## The Quest of Harvesting Semantic Relations

Despite its great potential, BabelNet alone is not enough to overcome the knowledge acquisition bottleneck: even with all the edges derived from WordNet and Wikipedia, the process of connecting concepts or entities in BabelNet with high quality and coverage is today still far from complete, and subject to continuous research effort. In fact, if we consider only WordNet and Wikipedia, the vast majority of edges are drawn from the latter resource; as such, they come with no label or specification, only conveying a generic semantic relatedness. Labeled edges, such as hypernyms and hyponymys, are limited to the (much smaller) lexicographic portion of WordNet.

The case of taxonomic information, indeed, is of paramount importance: the huge amount of specialized knowledge from Wikipedia pages still lacks a proper integration with the general concepts and semantic classes populating WordNet. This is crucial for downstream applications: in the illustrative example of Chapter 1, the word *band* is not hyperlinked in the Wikipedia page *Pink Floyd*. Thus, even if the corresponding Wikipedia concept *Musical ensemble* is correctly mapped to the WordNet *band*, there are no means of establishing the connection with *Pink Floyd* and *Musical ensemble* in the first place (and hence, no hypernymy chain to follow). This shortcoming has motivated, among other efforts, the development of a Wikipedia Bitaxonomy (Flati et al., 2014), i.e. an integrated taxonomy of both Wikipedia pages and categories. Constructing the Wikipedia bitaxonomy involves an iterative process of mutual reinforcement in which complementary information coming from either one of the individual taxonomies is propagated into the other (Figure 2.3). As in the construction of BabelNet, this method exploits at best the graph structure of Wikipedia, not only to extend coverage to as many Wikipedia concepts as possible, but also to project the obtained taxonomies from English (used as pivot language) to an arbitrary language, thereby achieving full multilinguality (Flati et al., 2016).

Integrating the Wikipedia Bitaxonomy into BabelNet has been a major step towards characterizing the deluge of unlabeled semantic relations from Wikipedia, followed up by the integration of Wikidata (Vrandečić, 2012). A lot of work still remains to be done in this respect: indeed, this scenario has contributed to shape part

Figure 2.3. An illustration of the Wikipedia Bitaxonomy borrowed from Flati et al. (2016).

The page taxonomy (left) and category taxonomy (right) are connected with cross-links (dashed lines), i.e. links between Wikipages and the categories they belong to.

of the motivation and the focus of this thesis. As we discuss in Section 3.2, approaches based on extracting semantic relations from open text have been investigated and are currently under investigation; however, they are somehow still suboptimal in utilizing the sense-level information that a resource like BabelNet would put at their disposal. In Chapter 5 we study how to bring these approaches to the next level.

## 2.2 From Words to Senses

Studying how to move from words to senses means entering the domain of Lexical Semantics (Cruse, 1986). Lexical Semantics is the subfield of Linguistics concerned with establishing the meaning of lexical items, the building blocks that make up the catalogue of words in a given language (lexicon). In order to deal with Lexical Semantics computationally, we need first of all a few basic operative assumptions:

1. Words in isolation do not have meaning, and a sentence acquires its meaning by virtue of the words that compose it and the manner of their combination (Cruse, 1986; Miller, 1999);
2. For every word  $w$  in a language  $L$ , having multiple senses  $s_1; s_2; \dots; s_n$ , the lexicon expresses these senses as a sorted list of form-meaning associations, i.e. word senses (Pustejovsky, 1991);
3. Senses refer to language-independent conceptual primitives, which are then expressed (lexicalized) in each specific language (Mandler, 1996; Wierzbicka, 1996).

The second assumption, defined by Pustejovsky (1991) as the sense enumeration lexicon, is based on the fact that all form-meaning associations (word senses) in a language can be listed within the lexicon. In this setting, a word form is said to be polysemous when its lexical entry includes multiple distinct word senses. By the same token, a word meaning can be expressed (lexicalized) by one or more word forms, which are said to be synonyms. Despite the many controversies as to whether both the first and the second assumption are legitimate (Pustejovsky, 1991; Hanks, 2000), a list-based lexicon provides a clear computational framework in which Lexical Semantics remains separate and independent from syntactic knowledge: in fact, WordNet (Section 2.1.1) is the prototypical example of such a framework. The third assumption, apart from being a long-standing theoretical controversy in Cognitive Semantics, stands as the foundational hypothesis of multilingual lexical resources like BabelNet (Section 2.1.3), where concepts and entities are encoded as multilingual sets of synonyms representing their language-specific lexicalizations.

Given the above assumptions, a computational model of Lexical Semantics needs to address two fundamental issues: on the one hand, how to represent the semantics of lexical items computationally; on the other, how to establish a mapping from a given lexical item to an appropriate word sense in the lexicon, inside the context of a language utterance. In both cases, computational approaches are based on the basic, well-known principle that, in order to understand the meaning of words, we should look at the various contexts in which they occur:

You shall know a word by the company it keeps (Firth, 1957)

### 2.2.1 Word Sense Disambiguation

The issue of establishing a mapping from lexical items in context to senses in a sense inventory corresponds, in computational terms, to performing Word Sense Disambiguation. The literature on WSD is broad and comprehensive (Agirre et al., 2009; Navigli, 2009, 2012): new models are continuously being developed and tested over a variety of standard benchmarks (Edmonds and Cotton, 2001; Snyder and Palmer, 2004; Pradhan et al., 2007; Navigli et al., 2007, 2013; Moro and Navigli, 2015). Moreover, the field has been explored in depth from different angles by means of extensive empirical studies and evaluation frameworks (Pilehvar and Navigli, 2014; Iacobacci et al., 2016; McCarthy et al., 2016; Raganato et al., 2017a).

Following Navigli (2009), we define WSD as the task of computationally determining which sense of a word is activated by its use in a particular context. It is usually performed on one or more texts (although in principle any collection of naturally occurring words might be employed), which we can represent as a sequence of words  $T = (w_1; w_2; \dots; w_n)$ . Given  $T$ , WSD can be formally described as the task of assigning the appropriate sense(s) to all or some of the words  $w_i$ , that is, to identify a mapping  $A$  from word to senses, such that  $A(i) = S(w_i)$ , where  $S(w_i)$  is the set of senses encoded in a sense inventory for word  $w_i$  (cf. Section 2.1) and  $A(i)$  is the subset of the senses  $w_i$  which are appropriate in the context  $T$ .<sup>7</sup>

WSD can certainly be (and has been) viewed as a classification task, with words in context being the input instances, and word senses being the classes; a classifier can be trained to assign each occurrence of a word to one or more classes based on the evidence from the context  $T$ , as well as from external knowledge resources. However, an important difference between WSD and other typical classification tasks studied throughout NLP (e.g. part-of-speech tagging, named entity recognition, text categorization) is that the latter use a single predefined set of classes (e.g. parts of speech, categories), whereas in the former the set of classes changes depending on the word to be classified. In this respect, WSD actually comprises  $n$  classification tasks, where  $n$  is the size of the lexicon.

#### 2.2.1.1 Evaluation and Standard Benchmarks

We can distinguish two variants of the generic WSD task: *Lexical Sample WSD* (Kilgarri, 2001; Mihalcea et al., 2004), where a system is required to disambiguate a restricted set of target words, usually occurring one per sentence, and *All-words WSD*, where systems are expected to disambiguate all open-class words in a text (i.e. nouns, verbs, adjectives, and adverbs). This latter task is typically harder, as it requires wide-coverage systems capable of dealing with data sparsity issues (including the knowledge acquisition bottleneck discussed in Chapter 1), and it has been proposed in a number of varieties. For instance, depending on the granularity of the sense inventory employed, it has been referred to as *fine-grained WSD*, when WordNet is used as it is as sense inventory, *coarse-grained WSD* when based on a reduced set of coarser senses obtained by clustering the original WordNet sense inventory as in, e.g., Navigli et al. (2007). In fact, it has been shown that

<sup>7</sup>The mapping  $A$  can assign more than one sense to each  $w_i \in T$ , although typically only the most appropriate sense is selected, that is,  $\sum_j |A(i)_j| = 1$ .

sense granularity is key when developing and utilizing WSD models (Edmonds and Kilgarri, 2002; Navigli, 2006; McCarthy et al., 2016), especially when the sense distinctions encoded in reference sense inventories like WordNet are too subtle even for human annotators. This phenomenon reflects on the typical upper bound considered in WSD, i.e. the inter-annotator agreement (ITA): on coarse-grained sense inventories the ITA is calculated around 90% (Gale et al., 1992b; Navigli et al., 2007), whereas on fine-grained WordNet-style sense inventory the ITA is estimated between 67% and 80% (Snyder and Palmer, 2004; Palmer et al., 2007).

In the all-words WSD task, a typical baseline is represented by the most likely sense for each word regardless of context, which can be computed as the most frequent sense (MFS), or most common sense (MCS), of that word inside a reference corpus, or as the first sense provided for that word by the sense inventory.<sup>8</sup> Initially conceived as a lower bound, this baseline has been shown to pose serious difficulties to WSD systems, and it is often hard to beat (Gale et al., 1992a; Navigli, 2009). Also, due to the skewness of sense distributions inside many sense-annotated corpora, the MFS represent a strong bias not only for supervised systems (Postma et al., 2016), but also for knowledge-based systems (Calvo and Gelbukh, 2015), being correlated with the number of semantic connections inside WordNet.

In Chapter 4 we use the all-words WSD task as a major testbed for evaluating extrinsically a sense-annotated corpus. In particular, we rely on the standard WSD datasets from the Senseval/SemEval competitions: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 task 1 (Snyder and Palmer, 2004), SemEval-07 task 17 (Pradhan et al., 2007), SemEval-13 task 12 (Navigli et al., 2013) and SemEval-15 task 13 (Moro and Navigli, 2015). Recently, these five datasets have been standardized and unified in the framework of Raganato et al. (2017a), together with some of the sense-annotated corpora described in Sections 3.1.1 and 3.1.2.

Finally, attention is increasingly being paid to performing WSD across languages, a task referred to as cross-lingual WSD (Lefever and Hoste, 2010, 2013): in this setting, an input sentence is provided in a source language (e.g. English) and the WSD system has to provide word senses encoded in a target language (e.g. Italian) using a sense inventory constructed with translations from a parallel corpus. The underlying assumption of this task is that sense distinctions of a word in a source language are determined, at least partially, by its different translations in other languages. Supported by several studies (Gale et al., 1992c; Ide et al., 2002; Ng et al., 2003), this assumption also motivates some of the techniques used in Chapter 4 to obtain sense annotations in a multilingual setting. Alongside cross-lingual WSD, truly multilingual WSD (Navigli et al., 2013; Moro and Navigli, 2015), where the all-words WSD task is entirely redefined with training and test datasets for languages other than English, is another important thread of WSD evaluation, recently enabled by multilingual sense inventories like BabelNet (Section 2.1.3).

### 2.2.1.2 Approaches to WSD

There are three mainstream approaches to WSD, namely:

---

<sup>8</sup>Typically these two baselines coincide: in WordNet, for instance, sense order is determined on the basis of sense-annotated text (cf. Section 3.1.1).

- ^ Supervised WSD , based on the formulation of WSD as a series of classification tasks in which a dedicated classifier (word expert) is trained for each target word using a sense-annotated corpus. Supervised models have been shown to consistently achieve higher performances in all standard benchmarks (Raganato et al., 2017a), at the expense, however, of harder training and limited flexibility. Apart from the fact that obtaining sense-annotated data is a highly expensive endeavor, as discussed throughout Chapter 1, the 'word expert' paradigm is one of the major limitations of these approaches, as it requires framing a separate classification problem for each target word or, in the general case, for each ambiguous entry in the lexicon;
- ^ Knowledge-based WSD , where exploiting the structural properties of knowledge resources (Section 2.1) is key to determine the senses of words in context. Despite lagging behind their supervised counterparts, knowledge-based approaches have the advantage of a wider coverage and increased flexibility, which allows them to scale better in terms of scope and number of languages. Also, in contrast to word experts, knowledge-based systems are not forced to treat each target word in isolation: they usually construct a model based only on the underlying resource, which is then able to handle multiple target words at the same time and disambiguate them jointly. Crucially, however, their performances depend strongly on the richness of the underlying resource (Cuadros and Rigau, 2006; Navigli and Lapata, 2010).
- ^ Unsupervised WSD , or Word Sense Induction , i.e. techniques based on discovering senses automatically from unlabeled corpora. These methods are particularly attractive as they do not require sense-annotated data or a predefined sense inventory: instead, they dynamically induce groups of synonymous words (clusters) based on their occurrences in similar contexts (Schütze, 1998; Brody and Lapata, 2009; Marco and Navigli, 2013). This however makes both comparison and evaluation quite hard, and lexico-semantic relationships between the clusters/word senses (otherwise provided by an external knowledge resource) have to be established in a later phase, either automatically or by manually mapping the clusters to a reference sense inventory.

### The State of the Art

In this section we follow Raganato et al. (2017a), which present an up-to-date comparison and comprehensive analysis of the state of the art in supervised and knowledge-based WSD, and survey briefly some recent contributions in the field.<sup>9</sup>

As regards supervised WSD, traditional approaches, based on extracting a set of local features from the target word and the surrounding words (Zhong and Ng, 2010; Shen et al., 2013), are still very competitive. For instance, *It Makes Sense* (Zhong and Ng, 2010, IMS) uses a Support Vector Machine classifier over a set of conventional features (surrounding words, part-of-speech tags, collocations), and it is

<sup>9</sup>Inasmuch as we define WSD with respect to a predefined sense inventory, unsupervised approaches, i.e. Word Sense Induction, fall outside the scope of this analysis and, more generally, of this thesis. Thus, in this and the following sections we will focus solely on supervised and knowledge-based WSD.

widely used today as reference supervised system in many experimental studies and evaluations, including those we perform in Chapter 4. In fact, in latest developments, this basic model has been enhanced with more complex features based on word embeddings (Taghipour and Ng, 2015a; Rothe and Schütze, 2015; Iacobacci et al., 2016). The recent advances of neural networks have also contributed to fuel WSD research: Kågebäck and Salomonsson (2016) present a supervised classifier based on a bidirectional Long Short-Term Memory (LSTM) for the lexical sample task; similar architectures have also been utilized in instance-based approaches (Yuan et al., 2016; Melamud et al., 2016) where a latent representation is obtained for the whole sentence containing a target word  $w$ , and then this representation is compared with those of example sentences annotated with the candidate meanings  $o_i$ .

State-of-the-art knowledge-based systems, instead, are either based on distributional similarity (Basile et al., 2014; Chen et al., 2014; Camacho Collados et al., 2016c) or on the structural properties of lexicalized semantic networks (Agirre et al., 2014; Moro et al., 2014b; Weissenborn et al., 2015; Tripodi and Pelillo, 2017): some of them, such as UKB (Agirre et al., 2014) and Babelify (Moro et al., 2014b) create a graph representation of the input text, and then exploit different graph-based algorithms over the given representation (e.g. PageRank) to perform WSD.

### 2.2.2 Entity Linking

Across the NLP literature, an important task that is usually considered very related to WSD is Entity Linking (Erbs et al., 2011; Rao et al., 2013, EL). The goal of EL is to identify mentions of entities within a text, and then link (disambiguate) them with the most suitable entry in a reference knowledge base. The increasing popularity of EL is connected to the availability of semi-structured resources (cf. Chapter 1), especially Wikipedia to the extent that Wikipedia-based EL is usually known as Wikification (Mihalcea and Csomai, 2007; Milne and Witten, 2008; Ratinov et al., 2011). Originally conceived as one of the fundamental steps within the broader task of Information Extraction (Section 2.3), EL is another long-standing task but, unlike WSD, lacks a standard formal definition of the problem, as well as well-established evaluation benchmarks where EL systems can be meaningfully compared (Ling et al., 2015), despite some recent efforts in this direction (Cornolti et al., 2013; Usbeck et al., 2015). For instance, among the many variants of EL, Wikification aims at linking all kinds of noun phrases to Wikipedia entities, while Named Entity Disambiguation (Bunescu and Paşca, 2006; Cucerzan, 2007) targets only named entities. In this thesis we follow the seminal work of Rao et al. (2013) and consider EL as the better defined problem of solely linking named entities.<sup>10</sup>

The EL can be split in two separate (and typically consequent) subtasks:

1. Mention Extraction, i.e. identifying the boundaries of named entities in a target text. This subtask is closely related to a classical NLP task, Named Entity Recognition (Nadeau and Sekine, 2007, NER), where named entity mentions are extracted and classified into a predefined set of general

<sup>10</sup>As we discuss in Section 2.2.2.3, the task of linking common noun phrases is closely related to WSD, to the extent that they can actually be considered the same task, both formally and operationally.

semantic classes (e.g. PERSON, LOCATION, ORGANIZATION). Seemingly simple, this subtask is actually challenging and controversial, not only because of name variations (e.g. abbreviations, acronyms, or spelling differences) but also because mention boundaries can easily overlap. For instance, the noun phrase *Portland, Oregon* can be considered as a whole, or can be split into two individual mentions, *Portland* for the city and *Oregon* for the city's state, with all three mentions making perfect sense;

2. Entity Disambiguation, i.e. the subtask of appropriately matching a mention to an entry inside a predefined knowledge base, which is often Wikipedia or a Wikipedia-derived knowledge base, such as Yago, Freebase, or DBpedia (cf. Section 2.1.2). The key issue here is, of course, ambiguous mentions, either because of polysemy (e.g. *Washington* being a president, a federal district, or a U.S. state) or metonymy (e.g. *Moscow* referring to the government of Russia rather than the actual city). Another issue is coreference, as two or more mentions can often refer to the same entity (e.g. *Trump*, *D. Trump*, *Mr. President*). Finally, as no standard annotation guidelines exist with respect to EL, a number of structural issues arise (Ling et al., 2015), such as how specific a linked entity should be (e.g. the mention *World Cup* can be legitimately linked to the Wikipage *FIFA World Cup*, as well as to a specific occurrence of the event, say the Wikipage *1998 FIFA World Cup*), or how to deal with entities that are absent from the knowledge base.

### 2.2.2.1 Evaluation and Standard Benchmarks

In contrast to WSD, where well-established benchmarks are provided by the Senseval/SemEval competition series (cf. Section 2.2.1), EL systems are often compared using different datasets (Ling et al., 2015). The most common benchmarks that have been utilized over the years are: the ACE and MSNBC datasets (Cucerzan, 2007; Ratinov et al., 2011), with entity mentions extracted from newswire text and linked to Wikipedia; the TAC-KBP datasets (McNamee et al., 2009), which are however only available to the task participants; and the AIDA-CoNLL test dataset (Hofmann et al., 2011b), the largest publicly available, comprising almost 1,400 English articles and roughly 35,000 entity mentions linked to Yago. Other notable datasets are the Wikipedia-based CSAW (Kulkarni et al., 2009) and AQUAINT (Milne and Witten, 2008), which annotate both concepts and named entities, and KORE50 (Hofmann et al., 2012), a small-size dataset of 50 short English texts annotated using Yago, and built with the idea of testing against a high level of mention ambiguity.

In addition to the many datasets available, a variety of metrics have also been used for evaluation, with little agreement on which ones are best (Ling et al., 2015). The most common ones include: Bag-of-Concept F1 (ACE and MSNBC datasets), where a gold bag of Wikipedia entries is evaluated against a bag of Wikipedia entities provided by the system, micro accuracy (TAC-KBP datasets), which is simply the percentage of correctly predicted links, and NER-style F1 (AIDA-CoNLL), where a link is considered correct only if the mention matches the gold boundary and the linked entity is also correct.

### 2.2.2.2 Approaches to EL

The earliest approaches to EL were Wiki cation approaches (Mihalcea and Csomai, 2007; Cucerzan, 2007), in which the local context surrounding the mentions had a primary role, similarly to supervised WSD (cf. Section 2.2.1). Based on a collective notion of coherence among the selected Wikipages, Milne and Witten (2008) focused instead on analyzing the semantic relations between the candidate entity mentions and the unambiguous context. Despite the crucial dependence on unambiguous words within the input context, their approach started a successful trend of EL models based on both local and global features (Kulkarni et al., 2009; Ratinov et al., 2011; Hoart et al., 2011b; Mendes et al., 2011). In recent years, more sophisticated approaches have been developed, exploiting Integer Linear Programming (Cheng and Roth, 2013), generative models (Han and Sun, 2012), stacked generalization (He et al., 2013b) and deep neural networks (He et al., 2013a). Another extremely promising line of work consists in tackling EL by jointly modeling the disambiguation of entities and closely related tasks, such as NER (Sil and Yates, 2013; Nguyen et al., 2016) and coreference resolution (Hajishirzi et al., 2013; Durrett and Klein, 2014), where generative models are employed to capture inter-task interactions. This key intuition, along with the availability of knowledge resources like BabelNet, has motivated the development of joint WSD and EL approaches, which we examine in Section 2.2.2.3.

### 2.2.2.3 Joint WSD and EL: Babelfy

WSD and EL are undoubtedly similar, as in both cases text fragments have to be disambiguated according to a reference inventory. However, there are two important differences between them: the nature of sense inventory (dictionaries and lexicons for WSD, encyclopedic knowledge bases for EL), and the fact that in EL mentions in context are not guaranteed to be complete but can be (and often are) partial. As a result of these and other discrepancies, the research community has spent a lot of time tackling WSD and EL separately, not only leading to duplicated efforts and results, but also failing to exploit the fact that these two tasks are deeply intertwined. Consider the following example:

He loves driving around with his Mercedes.

where the verb driving should be resolved by a WSD system with the sense of 'operating vehicles', and the partial mention Mercedes should be recognized by an EL system and linked to the automobile brand (Mercedes-Benz). In this setting, the WSD system would clearly benefit from knowing that a brand of vehicles is mentioned in the local context of driving and, at the same time, the EL system would easily take advantage of the sense-level information about driving referring to vehicles when linking the mention Mercedes.

This is where linked lexical resources like BabelNet (Section 2.1.3) play a role: by providing a large-scale encyclopedic dictionary as common ground for WSD and EL, they enable the design and development of unified WSD and EL models. The first of this kind is Babelfy (Moro et al., 2014b), based on BabelNet: Babelfy is a graph-based approach to joint WSD and EL that exploits a loose identification of candidate meanings, and a densest-subgraph algorithm to select high-coherence semantic interpretations. Babelfy disambiguates as follows:

Figure 2.4. Excerpt of the semantic interpretation graph for the example sentence *Thomas and Mario are strikers playing in Munich* borrowed from Moro et al. (2014b). The edges connecting the correct meanings (e.g. *Thomas Müller* for *Thomas* and *Mario Gomez* for *Mario*) are in bold.

1. Given a lexicalized semantic network, such as BabelNet, a semantic signature is computed for each concept or entity. A semantic signature is a set of highly related vertices obtained by performing Random Walks with Restart (RWR) (Tong et al., 2008) for each vertex  $v$  of the semantic network. RWR models the conditional probability  $P(v^0|v)$  associated with an edge  $e(v; v^0)$ :

$$P(v^0|v) = P \frac{\text{weight}(v; v^0)}{\sum_{v^0 \in V} \text{weight}(v; v^0)} \quad (2.1)$$

where  $V$  is the set of vertices in the semantic network and  $\text{weight}(v; v^0)$  is the weight associated with  $e(v; v^0)$ . This is a preliminary step that needs to be performed once and for all, independently of the input text;

2. Given an input text, Babelfy extracts all the linkable fragments and lists all their possible meanings according to the sense inventory. Candidate extraction is a high-coverage procedure based on superstring (instead of exact) matching, and capable of handling partial mentions and overlapping fragments;
3. A graph-based semantic interpretation of the input text is generated using the semantic signatures of all candidate meanings. The Babelfy extracts a dense subgraph of this representation in order to select a coherent set of best candidates for the target mentions. An example of semantic interpretation graph is shown in Figure 2.4. Each candidate in the graph is weighted with a measure that takes into account both semantic and lexical coherence, exploiting graph centrality among the candidates as well as the number of connected fragments. This measure is used in the dense-subgraph algorithm to iteratively remove low-coherence vertices from the semantic graph until convergence.

One of the greatest advantages of Babelfy is exibility: it can be used seamlessly for WSD, EL or even both at the same time. Also, the whole procedure is language-independent, and can be extended to any language for which lexicalizations are available inside the semantic network. In fact, Babelfy can even handle mixed text in which multiple languages are used at the same time, or work without being supplied with information as to which languages the input text contains (language-agnostic setting). On the other hand, as in any knowledge-based approach (Section 2.2.1), the quality of disambiguation depends crucially on the quality of the underlying

resource, and rarely achieves the same results of supervised models (Raganato et al., 2017a). BabelNet and Babelify have inaugurated a new, broader way of looking at disambiguation in Lexical Semantics, which has been further pursued by the research community (Basile et al., 2015; Weissenborn et al., 2015) and has led to the organization of novel shared tasks focused on multilingual WSD and EL as part of the SemEval competition series, namely the SemEval-2013 task 12 on Multilingual Word Sense Disambiguation (Navigli et al., 2013), and the SemEval-2015 task 13 on Multilingual All-words Word Sense Disambiguation and Entity Linking (Moro and Navigli, 2015). Both tasks required participating systems to disambiguate a set of target words and multi-word expressions in a test corpus with the most appropriate sense from the BabelNet sense inventory (or, alternatively, from those of WordNet or Wikipedia) and for 5 and 4 languages, respectively. In particular, the SemEval-2015 task 13 has been the first disambiguation task explicitly oriented to joint WSD and EL, including features of a typical WSD task (i.e. sense annotations for all open-class parts of speech) as well as features of a typical EL task (i.e. annotated named entities and non-specified mention boundaries).

### 2.2.3 Sense-based Vector Representations

While tasks like WSD (Section 2.2.1) and EL (Section 2.2.2) have grown popular across the NLP community mostly over the last two decades, research efforts on semantically representing lexical items dates back to the earlier days of NLP (Harris, 1954; Salton et al., 1975) and have agglutinated a large body of work since then, shaping a field of study now known as Distributional Semantics, in which the meaning of a lexical unit is computed from the distribution of words around it. Stemming from the well-known distributional hypothesis (Firth, 1957), i.e. the fundamental idea that words occurring in the same contexts tend to have similar meanings, the paradigm of vector space models (Turney and Pantel, 2010) took the lead, providing both a theoretical and practical framework in which a word is represented as a vector of numbers in a continuous metric space. Within this framework, linguistic phenomena are framed in terms of mathematical operations and, in particular, semantic similarity and relatedness between two words can be directly expressed in terms of proximity between the corresponding vectors, and computed in a quantifiable way (e.g. using cosine similarity). In recent times, the great success of neural networks and deep learning led to the development of embedded vector spaces (Mikolov et al., 2013c; Pennington et al., 2014), which are compact and fast to compute from unstructured corpora, and at the same time capable of preserving semantic regularities between linguistic items.

However, as discussed in Chapter 1, word representations have a crucial limitation: they tend to encode the different meanings of a word by conflating them into a single vector. A potential way of overcoming this limitation is to move to the sense level and generate representations of word senses, where each distinct meaning of a given word is associated with a distinct vector. Figure 2.5 shows an illustrative example with the word *bank* in a word-level space, the vector for *bank* lies exactly in between two regions that relate to the geographical and financial meanings of *bank*, respectively; this shows that components pertaining to two different semantic areas are inherently mixed up when the word is ambiguous. Instead, in the sense-level

Figure 2.5. Portions of a word-level vector space centered on the word *bank* (left) and a sense-level vector space where two different meanings of *bank* have distinct representations (right). Illustration borrowed from Camacho Collados et al. (2016b).

space, the two regions are neatly separated and the previously conflated meanings of *bank* have their own vectors in the proper semantic areas.

In fact, the representation of individual word senses and concepts has recently become very popular, thanks to several experimental results showing significant performance improvements with respect to word representations (Chen et al., 2014; Jauhar et al., 2015; Iacobacci et al., 2015; Rothe and Schütze, 2015; Camacho Collados et al., 2016c; Pilehvar and Collier, 2016). In this respect, lexical knowledge resources can (and have been) used to construct state-of-the-art models, including WordNet, Wikipedia and BabelNet. Compared to corpus-based approaches, where senses are typically not fine-grained, difficult to evaluate and statistically biased towards frequent meanings, a key advantage of knowledge-based representations is that they are directly linked to existing sense inventories, which makes them readily usable in downstream applications.

### 2.2.3.1 Evaluation and Standard Benchmarks

The most popular benchmark for the evaluation of different semantic representation techniques is semantic similarity, i.e. the task of measuring the semantic closeness of two linguistic items, which is directly computed by comparing the corresponding vector representations. In particular, word similarity is a popular variant of semantic similarity focused on words or multi-word expressions, which provides a series of well-established benchmarks for English: RG-65 (Rubenstein and Goodenough, 1965) and its subset MC-30 (Miller and Charles, 1991), WordSim-353 (Finkelstein et al., 2002), which contains both concepts and named entities, and SimLex-999 (Hill et al., 2014), which is composed of 999 word pairs, 666 of which are noun pairs. As regards other languages, benchmarks for multilingual word similarity and cross-lingual word similarity (where the words or multi-word expressions in a pair belong to different languages) are all constructed on the basis of RG-65 and its translations into German (Gurevych, 2005), French (Joubarne and Inkpen, 2011)

<sup>11</sup> Semantic similarity, which quantifies the likeness of meaning between two linguistic items, is often confused with semantic relatedness which is instead based on any semantic relation between them. For example, *car* is similar to *bus*, but is related (and not similar) to *road* and *driving*.

and Spanish (Camacho Collados et al., 2015a).

In a word similarity benchmark, all word pairs are associated with a similarity score given by a human annotator, and the performance of a system is assessed on the basis of Pearson and Spearman correlations of its similarity scores with human judgment. In the case of sense-based representations, which are defined at either the sense or the concept level, a conventional strategy for word similarity (Resnik, 1995; Budanitsky and Hirst, 2006; Pilehvar et al., 2013; Camacho Collados et al., 2016c) selects, for each pair of words  $w$  and  $w^0$ , the closest pair of candidate senses<sup>12</sup>:

$$\text{sim}(w; w^0) = \max_{v_1 \in L_w; v_2 \in L_{w^0}} VC(v_1; v_2) \quad (2.2)$$

where  $L_w$  represents the set of word senses (or concepts) that contain  $w$  as one of their lexicalizations.  $VC$  denotes the vector comparison measure, typically either standard cosine similarity or Weighted Overlap (Pilehvar et al., 2013), which takes into account the relative ranks of overlapping dimensions between the vectors.

Over the last decade, a broad spectrum of sense-based approaches have been proposed and evaluated experimentally on the semantic similarity task. While a comprehensive survey on sense-based representations is outside the scope of this thesis, in the following sections we focus on two complementary approaches, based on BabelNet (Section 2.1.3), that we utilize as tools for semantic analysis throughout Chapters 4 and 5: SensEmbed (Section 2.2.3.2) and Nasari (Section 2.2.3.3).

### 2.2.3.2 SensEmbed

A possible way of constructing semantic representations of word senses is to leverage existing architectures that already proved effective for word representations (Mikolov et al., 2013a,c), such as the Continuous Bag Of Words (CBOW). CBOW architectures are used to produce continuous vector representations (embeddings) for words based on distributional information from a textual corpus: in particular, they learn to predict a token given its context, which is typically defined by a fixed-size sliding window around the token itself. In order to work at the sense level, the CBOW has to be trained on a sense-annotated corpus, where sense-level information is explicitly attached to words and multi-word expressions; this informs the CBOW that two distinct meanings of the same ambiguous term (e.g. bank) have to be treated as distinct tokens (e.g.  $\text{bank}_1^1$  and  $\text{bank}_1^2$ ) and hence modeled using distinct embeddings. This is the core idea behind SensEmbed (Iacobacci et al., 2015), a technique to obtain continuous representations of word senses (sense embeddings) and use them effectively for word and relational similarity (Medin et al., 1990). SensEmbed relied on a dump of the English Wikipedia automatically disambiguated with Babelfy (Section 2.2.2.3) in order to train a CBOW architecture, obtaining as output latent representations for word senses linked to the sense inventory of BabelNet. By leveraging both distributional knowledge and structured knowledge coming from a lexicalized semantic network, SensEmbed consistently achieved state-of-the-art performances on various similarity benchmark, proving the effectiveness of computing embeddings at the sense level.

<sup>12</sup>Despite being widely used, the strategy of considering only the closest senses has some limitations, as pointed out by Iacobacci et al. (2015), which propose an alternative strategy that takes into account all the different meanings of the two words using a weighted average.

Exploiting the knowledge resource. One of the drawbacks of training embeddings on sense-annotated text, as in SensEmbed, is that it generally requires very large corpora to learn effective representations and, as previously discussed, sense labeling can be considerably expensive on corpora of such a size. An alternative approach consists in exploiting explicitly the features of the underlying knowledge resource that provides the sense inventory. This is the key insight of AutoExtend (Rothe and Schütze, 2015), a method to learn embeddings for sense and synsets that decouples actual embedding learning from their extension based on a lexico-semantic knowledge resource. The rationale of AutoExtend is that the properties and relations of such a resource (e.g. synonymy, hypernymy) can be formalized mathematically as training constraints. Rothe and Schütze (2015) rely on two basic constraints: (i) words representations are expressible as sums of their respective senses and (ii) synset representations are expressible as sums of their respective lexicalizations<sup>13</sup>. The learning process is then carried out using an autoencoder architecture where word embeddings constitute the input and output layers, and hidden layers represent the synset embeddings.

AutoExtend has been proven successful in various similarity tasks, as well as in WSD (Rothe and Schütze, 2015); however, some empirical analysis have shown that it tends to create clusters with a word and all its possible senses when non-predominant senses or less common word types are involved (Mancini et al., 2017).

### 2.2.3.3 Nasari

A major drawback of continuous models, such as word2vec (Mikolov et al., 2013a), SensEmbed (Iacobacci et al., 2015), and AutoExtend (Rothe and Schütze, 2015), is the lack of interpretability: embeddings are compact representations of lexical items where meaning is latent, with no human-readable feature describing their shape and structure. Also, as they are essentially corpus-based techniques, the quality of a word or sense vector depends crucially on the frequency of the corresponding word or word sense inside the training corpus.

To address both issues, Camacho Collados et al. (2016c) propose an alternative vector representation based on BabelNet, named Nasari. Instead of using a sense-annotated corpus, Nasari relies entirely on the lexicalized semantic network of BabelNet to construct a vector representation for each concept or entity (i.e. synset) in the sense inventory for which a Wikipage is available: while SensEmbed (Section 2.2.3.2) learns representations for word senses (hence two synonyms get two different embeddings), Nasari computes a single vector representing a whole Babel synset. This feature, thanks to the multilingual nature of BabelNet, directly translates into comparability across languages and linguistic levels (words, senses and concepts).

The Nasari representation of a given synset  $s$  is computed on the basis of a sub-corpus of contextual information relative to  $s$ , which is obtained as follows: by exploiting BabelNet's inter-resource mappings, Nasari starts from the Wikipage of  $s$  and gathers all Wikipages with an outgoing link to that page, as well as the

<sup>13</sup>For example, the embedding of the word bloom can be expressed as the sum of the embeddings of its two WordNet senses  $\text{bloom}_1^n$  and  $\text{bloom}_2^n$ , while the embedding of the synset containing bloom<sub>n</sub> (11689786r) is given by the sum of the embeddings of its three lexicalizations  $\text{flower}_n^2$ ,  $\text{bloom}_n^2$ , and  $\text{blossom}_n^1$ .

Wikipages of all the synsets connected to  $s$  via taxonomic relations in BabelNet. All content words inside this sub-corpus are then tokenized, lemmatized and weighted according to the source and type of semantic connections to  $s$ ; finally the sub-corpus is turned into a vector using three different techniques that give rise to three different types of representations:

- ^ A lexical representation, i.e. a vector defined in the space of individual words. In this lexical space, dimensions are explicitly associated with words, and the sub-corpus is represented in terms of the relevance of each word inside the text, estimated using lexical specificity (Lafon, 1980), a statistical measure based on the hypergeometric distribution;
- ^ An embedded representation, i.e. a sense embedding in a continuous vector space, obtained from the lexical vector with a two-steps procedure: (1) each dimension (i.e. word) is mapped to its embedded representation learnt from a textual corpus using the CBOW architecture; and (2) these word representations are then combined using a weighted average. The resulting vector is still defined at the concept level but, being based on the same architecture as word2vec, loses its interpretability. As a trade-off, however, it lies in the same semantic space of word embeddings, enabling a direct comparison between words and synsets;
- ^ A unified representation, i.e. a vector defined in the space of Babel synsets. This vector is obtained by clustering the word dimensions of the lexical vector based on whether they have a sense sharing the same hypernym in the BabelNet taxonomy. Clustering sibling words turns a lexical space into a semantic space with multilingual Babel synsets as dimensions; not only does this process provide an implicit disambiguation of ambiguous word dimensions, but it also makes the obtained unified representation language-independent, and hence suitable for cross-lingual applications.

The flexibility of Nasari allowed experimental evaluations on different benchmarks (monolingual and cross-lingual word similarity, sense clustering, WSD), where Nasari reported consistently state-of-the-art performances (Camacho Collados et al., 2016c).

## 2.3 Information Extraction

Information Extraction (Grishman, 1997, IE) is a broad area of NLP that deals with finding and extracting factual information from free text.<sup>14</sup> In other words, an IE system turns unstructured factual information embedded in natural language text into structured data, i.e. facts. A fact can be described in formal terms as a structured object capturing a real-world entity and its attributes mentioned in text, or a real-world event, occurrence, or state, with its argument or actors (e.g. who did what to whom, where and when). Most of the times we can represent such an

<sup>14</sup>IE is often confused with the task of Information Retrieval (IR), which is about selecting, from a collection of textual documents, a subset that is relevant to a particular query. Despite their analogies, the actual goal of IR is that of ranking or selecting documents, rather than deriving structured factual information from unstructured text.

object operatively with one or more triples of the type *entity, relation, entity*. In light of this, the definition of IE can be rephrased as the identification of instances of a particular class of relations in a natural language text, and the extraction of entities that are relevant arguments for that relations (Grishman, 1997).

More concretely, IE can be viewed as an effective way to populate the contents of a relational database, or more generally, of a knowledge resource (Section 2.1). Despite its conceptual simplicity and its targeted nature, the complexity and ambiguity of natural language make IE an extremely challenging task. Factual information can be expressed in multiple equivalent ways, distributed across multiple sentences, or even left implicit, and hence requiring an enormous amount of background knowledge to discern. On the other hand, the scope of IE is typically narrower than the scope of full text understanding (which goes far beyond the strictly factual content of language utterances): this has enabled robust, efficient and high-coverage NLP techniques to tackle many IE problems effectively, even with vast amounts of data.

According to Grishman (1997)'s definition of IE, we can identify two fundamental steps in the pipeline of an IE system: the first step consists in detecting and classifying the named entities occurring in a given text, i.e. performing Named Entity Recognition (Nadeau and Sekine, 2007, NER); the second step, instead, consists in finding and classifying the semantic relations among the entities detected (e.g. *born-in*, *spouse-of*, *works-for*, etc.), a task known as Relation Extraction. This latter step represents the actual gist of IE, and outputs a set of relation triples that can be used to populate a knowledge resource.

### What Kinds of Relations?

Since the core of the IE task is about extracting semantic relationships, designing a set of target relations to be modeled is of primary importance. In fact, in traditional approaches to IE (Section 2.3.1) an inventory of relations of interest is provided as input, so that, once entity mentions are given, the relation extraction step can be modeled as a standard classification task. The variety of relation types modeled depends strictly on the application scenario: in many domain-specific settings (e.g. biomedical documents, or airline routes), the relation inventory is relatively limited and can be hand-crafted by human experts; in a general-purpose IE task, instead, Wikipedia and Wikipedia-derived knowledge resources (Section 2.1.2) offer a large supply of relation types, typically drawn from infoboxes. For example, the Wikipedia infobox for [Stanford University](#) includes structured facts like *location = Stanford, California* and *president = Marc Tessier-Lavigne* that can be turned into relations like *located-in* and *president-of*. Another typical target is the set of ontological relations from WordNet (Section 2.1.1) or WordNet-like ontologies. The prototypical relations of this kind are the *is-a* (hypernymy) relation and the *instance-of* relation, both crucially important under an IE objective of extending an ontology or building it for new languages or domains.

Finally, on the other side of the spectrum are unsupervised approaches (Section 2.3.2), where no relation inventory is provided a priori. In this case the extraction step is completely open unconstrained, and the set of covered relation types emerges as a by-product of the extraction process itself.

### 2.3.1 Traditional Approaches

Earlier IE approaches, based on a fixed set of relations and entities to extract, were either based on purely supervised learning with engineered word-level and syntactic features (Zhao and Grishman, 2005; Mooney and Bunescu, 2006), or weakly supervised multiple-instance learning (Bunescu and Mooney, 2007), where negative examples are automatically generated from non-annotated entity pairs within a sentence. At the same time, given the small size of many annotated datasets for IE, other approaches focused on bootstrapping supervised systems from a high-precision seed patterns (Ravichandran and Hovy, 2002; Carlson and Schafer, 2008; Kozareva and Hovy, 2010): these approaches work by extracting sentences containing the target entity pair, and then generalizing the context around these entities to learn new relation patterns. Some contributions brought this approach to the extreme, with self-training methods that automatically generate their own training data (Agichtein and Gravano, 2000; Etzioni et al., 2005; Rozenfeld and Feldman, 2008; Weld et al., 2009). One of the major issues with semi-supervised approaches, both bootstrapped and self-supervised, is semantic drift, which occurs when erroneous patterns are learnt and lead to erroneous triples which, in turn, generate problematic patterns where the meaning of the original pattern is substantially altered.

One of the most well-known semi-supervised approaches to IE is undoubtedly NELL (Carlson et al., 2010), a Web-scale self-supervised learning system which runs continuously 24 hours a day, presented as a prototype for the never-ending learning paradigm. NELL's sophisticated architecture comprises a pool of extractors, simultaneously trained, that harvest candidate beliefs from the Web with a variety of methods (co-occurrence-based pattern learning, HTML and table mining, etc.). In order to overcome semantic drifts, NELL exploits other modules, together with occasional human supervision, that refine the extracted knowledge into confirmed beliefs subsequently added to NELL's internal knowledge base and training dataset. The whole learning process was bootstrapped with an initial hand-crafted ontology of categories (e.g. person, sportTeam, fruit) and relations (e.g. playsInstrument, playsForTeam), and few seed examples for each category and relation.

In recent times the availability of large-scale knowledge resources (Section 2.1) has enabled IE models to employ a considerably larger amount of examples in place of a handful of seeds, leading to the development of the distant supervision paradigm (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011): distantly supervised systems generate a lot of noisy pattern-based features using triples from (possibly human-contributed) knowledge resources, and then combine all these features using supervised classifiers. The way features are extracted in the first place is similar to that of self-supervised approaches, i.e. based on extracting sentences where the two entities of a given triple occur at the same time. The fundamental assumption underlying this strategy is the following (distant supervision assumption): if two entities participate in a relation, all sentences mentioning these two entities express that relation. Although it is prone to errors, this assumption has been very successfully applied to many IE tasks; furthermore, relaxing it leads to sophisticated approaches based on multi-instance multi-label learning (Surdeanu et al., 2012), where joint graphical models are required to allow relations overlaps: e.g., both founded and CEO-of should be valid relations for the pair (Steve Jobs, Apple).

Finally, the latest IE approaches either make use of the Statistical Relational Learning paradigm to couple actual IE with relational inference over knowledge bases (Wang and Cohen, 2015), or leverage end-to-end deep neural network models to frame the relation extraction task, including Convolutional Neural Networks (Zeng et al., 2015), Long Short-Term Memories (Miwa and Bansal, 2016), selective attention (Lin et al., 2016), and sequence-to-sequence models (Palm et al., 2017).

### 2.3.2 Open Information Extraction

While some of the self-contained IE approaches described in the previous section can legitimately be viewed as unsupervised (Etzioni et al., 2005), as they do not require training data in any form, the Open Information Extraction paradigm (Banko et al., 2007; Banko and Etzioni, 2008; Wu and Weld, 2010, OIE) is based on a way more radical approach: not only is it fully unsupervised, but it does not even rely on a predefined entity or relation inventory at all. In other words, OIE consists in the open and unconstrained extraction of an unspecified set of relations, which is not given as input, but rather obtained as a by-product of the extraction process. The sole input of an OIE system is a large, usually Web-scale, textual corpus.

OIE is typically performed in a single pass over massive amounts of raw text, with no human input of any kind, in order to extract and formalize a large collection of relation triples, or relation instances, where pairs of entity mentions are connected by textual relation phrases (e.g. `is a city in`, `is married to`). The earliest OIE approaches, namely `TextRunner` (Banko et al., 2007), `O-crf` (Banko and Etzioni, 2008), `StatSnowBall` (Zhu et al., 2009) and `Woe` (Wu and Weld, 2010), had two clear limitations hampering their performances: incoherent extractions (i.e. relational phrases with no meaningful interpretation, usually due to sequence labeling errors) and uninformative extractions (i.e. extractions omitting critical information, mostly due to improper handling of light-verb constructions).<sup>15</sup>

In response to these limitations, a second generation (Etzioni et al., 2011) of OIE systems focused on substantially improving both precision and recall of the extraction by imposing a set of generic syntactic and lexical constraints to the identified relation phrases. One of the most popular OIE approach of this type is `ReVerb` (Fader et al., 2011), which implements a general model of verb-based relation phrase identification based on pre-specified syntactic constraints (targeted at avoiding incoherent extractions) and lexical constraints which eliminate very long, rare and over-specified relation phrases by matching them against a dictionary. In contrast to other traditional IE and OIE pipelines, where entities are extracted first, the verb-centric strategy of `ReVerb` starts by identifying valid relation phrases, and then extracts the argument pairs by finding the nearest noun phrases to the left and to the right of the relation phrases. To date, `ReVerb` remains one of the OIE approaches with the highest coverage, with almost 15 million extractions of 664,746 distinct relation phrases obtained from a filtered run on the ClueWeb09 dataset<sup>16</sup>.

A great deal of later OIE approaches have adopted more sophisticated solutions to further improve over `ReVerb` and capture relations not mediated by verbs:

<sup>15</sup>For example, `hHamas claimed responsibility` instead of `Hamas claimed responsibility for the Gaza attack`.

<sup>16</sup><http://lemurproject.org/clueweb09.php>

Ollie (Mausam et al., 2012), based on relation-independent dependency patterns automatically learnt via bootstrapping, KrakeN (Akbik and Löser, 2012), which targets higher-order n-ary extractions, ClausIE (Del Corro and Gemulla, 2013), a clause-based approach that decompose a sentence into basic constituents (used from which relation triples are derived), ReNoun (Yahya et al., 2014), which complements ReVerb by focusing on of nominal attributes, and Stanford's openIE (Angeli et al., 2015), which approaches OIE as entailment, and exploits linguistic structure alongside natural logic inference. Recently, some efforts have also been put in making OIE extractions more compact and precise (Gashteovski et al., 2017).

At the same time, an alternative research thread has looked at the similarities between OIE and another very popular NLP task, i.e. Semantic Role Labeling (SRL), proposing OIE approaches capable of exploiting semantic features derived from SRL (Christensen et al., 2010; Mesquita et al., 2013), or leveraging SRL annotations to construct automatically a large-scale benchmark for OIE (Stanovsky and Dagan, 2016). Rather than being conceived as a downstream task itself, however, OIE has also shown to be effective in producing intermediate structured features for downstream semantics-oriented tasks (Stanovsky et al., 2015), such as word analogy and word similarity, as well as in harvesting semi-structured knowledge for applicative tasks like question answering (Khot et al., 2017).

Finally, as in many related areas of NLP, there has been a growing interest in multilingual and cross-lingual OIE approaches, either based on rules over dependency parses (Gamallo and Garcia, 2015), cross-lingual projection via Machine Translation (Faruqui and Kumar, 2015), SRL-derived predicate-argument analysis (Falke et al., 2016), or even deep sequence-to-sequence models (Zhang et al., 2017).

All these approaches to OIE, despite being effective in their respective settings, tend to have a common limitation in the fact that they do not address Lexical Semantics explicitly. As in many areas of NLP, the current tendency is indeed to model semantic phenomena, such as synonymy or lexical ambiguity, implicitly (cf. Chapter 1). In this thesis we target precisely this issue: first, we examine some notable exceptions of sense-aware OIE approaches in Section 3.2; then, in Chapter 5, we study how to redefine IE, and OIE in particular, at the sense level.

### 2.3.3 Universal Schemas

As examined in the previous sections, both the traditional, or closed (Section 2.3.1, and the open (Section 2.3.2) paradigms of IE have their strengths and weaknesses. The former requires a pre-defined, finite and fixed schema of relation types, as well as training data labelled according to this schema, in order to train one or more extractors with various degrees of supervision; the latter, instead, gains tremendous flexibility by using language itself as source of the schema (which, as a consequence, becomes open and unbounded) but lacks the ability to generalize effectively.

An alternative approach, with the aim of taking the best of both worlds, consists in using a universal schema i.e. the union of all involved schemas (surface-form predicates as in OIE, and relation types from pre-existing databases). By operating simultaneously on relations observed in text and in pre-existing structured resources, the Universal Schema approach enables reasoning about unstructured and structured data in mutually supporting ways. The key underlying idea is focusing on predicting

source data as opposed to modeling semantic equivalence among relations explicitly, which is an arguably elusive matter even with clearly specified cases such as the 'is-a', or hypernymic, relation (Brachman, 1983).

Universal Schemas have been originally proposed by Riedel et al. (2013), where they are designed and evaluated as a Knowledge Base Completion approach based on matrix factorization: the unified, probabilistic knowledge base is represented with a matrix with entity pairs as rows and relations (both surface-form patterns and pre-existing relation types from structured resources) as columns. A series of collaborative filtering models are then exploited to learn lower dimensional manifolds for tuples, relations and entities, and to capture direct correlations between relations (usually asymmetric implicatures). In their experimental evaluations, these models are shown to accurately predict surface-pattern relationships not appearing explicitly in text, substantially improving results over a traditional classifier-based IE approach.

Despite its success and numerous extensions (Rocktäschel et al., 2015; Verga et al., 2016), as well as its relationship with some of the objectives targeted in Chapter 5, this inherently supervised paradigm falls outside of the scope and focus of the present thesis (cf. Section 1.1). However, from the perspective laid out in Chapter 7, it surely opens a number of compelling avenues for future work.

## 2.4 Nomenclature

Throughout this thesis we largely follow the same nomenclature of WordNet and BabelNet with respect to word senses, synsets, concepts and entities (cf. Sections 2.1.1 and 2.1.3). In particular, we use the term *synset* to refer to a specific language-independent meaning encoded as set of synonyms inside a knowledge resource. Synsets may represent concepts (such as the concept of play as a dramatic work, from the example of Section 2.1.1, or the concept of dog as a domestic mammal) or named entities (e.g. Microsoft, World War II, the city of Rome), and are associated with an open-class part-of-speech tag (noun, verb, adjective, adverb<sup>17</sup>) and a unique identifier. Even though a synset is language-independent, it features one or more words and multi-word expressions as external, language-specific representations, which we refer to as *lexicalizations*. For example, the synset representing the concept of play as a dramatic work can be expressed with the English lexicalizations *play*, *drama*, *dramatic play* with the German lexicalizations *Theaterstück*, *Bühnenstück*, *Bühnenwerk* or with the Italian lexicalizations *dramma*, *opera*, *teatrale*. As explained in Section 2.1, lexicalizations are always in a canonical form, i.e. lemma or lexeme.

We refer to the pairing of a lexicalization and its associated meaning as a *word sense* or as a *named entity mention*, depending on whether the latter is a concept or a named entity, respectively. Thus, the word *play* associated with the concept of play as a dramatic work is a word sense, while the word *Washington* associated with the U.S. state of Washington is a named entity mention.<sup>18</sup> With both word senses and named entity mentions we use the following notation (Navigli, 2009):

$$\text{word}_p^n \quad \text{the } n^{\text{th}} \text{ meaning of word with part of speech } p$$

<sup>17</sup>While concepts are not exclusively nominal (e.g. the concept of driving a car, or the concept of being honest), named entities are only associated with nominal synsets.

<sup>18</sup>When a lexicalization is associated with a named entity, we also use the term 'mention'.

The two examples above, for instance, could be represented as  $\langle \text{play}_n^1, \text{Washington}_n^2 \rangle$ .

As regards the Information Extraction domain, we refer to a relation instance  $\langle a_s, r, a_o \rangle$ , or relation triple, as a tuple having the form:

$$t = \langle a_s; r; a_o \rangle \quad (2.3)$$

with  $a_s$  being the subject argument,  $a_o$  being the object argument, and  $r$  being the relation pattern, or relation phrase. Depending on the specific scenario,  $a_s$  and  $a_o$  can be words or multi-word expressions, word senses/named entity mentions, or synsets. For example,  $\langle \text{Seattle}_n^1, \text{located in}, \text{Washington}_n^2 \rangle$  and  $\langle \text{Seattle}_n^1, \text{located in}, \text{Washington}_n^2 \rangle$  are both valid relation instances, the former having lexicalized arguments (words and multi-word expressions) and the latter having sense-level arguments (word senses and named entity mentions). Given a particular relation pattern  $r_k$ , the associated relation  $r$  is identified by the set of all relation instances where  $r = r_k$ .<sup>19</sup> In the above example, the relation pattern 'located in' is associated with the relation 'located-in', which might cover many other relation triples, such as  $\langle \text{Rome}_n^1, \text{located in}, \text{Italy}_n^2 \rangle$ , or  $\langle \text{Melbourne}_n^1, \text{located in}, \text{Australia}_n^2 \rangle$ .

Given a relation  $r$ , we define the set of all subject arguments of its relation instances as the domain  $D$  of  $r$ , i.e.  $D(r) = \{ a \mid a = a_s \wedge \langle a_s; r; a_o \rangle \in rg \}$ . Similarly, we define the set of all object arguments of  $r$  as the range  $G$  of  $r$ , i.e.  $G(r) = \{ a \mid a = a_o \wedge \langle a_s; r; a_o \rangle \in rg \}$ . In some cases, the domain and range of a relation are associated with a type signature, i.e. a semantic class that generalizes all the elements they contain. In the case of the example above, the domain and range of the relation 'located-in' would have these shapes:

$$\begin{aligned} D(\text{located-in}) &= \{ \text{Seattle}, \text{Rome}, \text{Melbourne} \} :: g \\ G(\text{located-in}) &= \{ \text{Washington}, \text{Italy}, \text{Australia} \} :: g \end{aligned}$$

Suitable type signatures for the two sets would be, e.g.  $City$  and  $State$ , respectively.

Finally, we characterize in formal terms a generic knowledge base, or knowledge resource, as a triple  $KB = \langle E; R; T \rangle$  where  $E$  is a set of entities,  $R$  is a set of relation patterns, and  $T$  is a set of relation triples, each defined as in (2.3), where  $a_s, a_o \in E$  and  $r \in R$ . In plain words,  $E$  is the set of all distinct subject and object arguments of all the triples included in the knowledge base (entity inventory),  $R$  is the set of all distinct relation patterns (relation inventory), and  $T$  is the actual content of the knowledge base. From this perspective, if we assume that each argument pair from  $E$  participates in at most one relation instance (i.e. the distant supervision assumption, cf. Section 2.3.2), the relations encoded in a knowledge base define a partition of  $T$ , i.e.  $T = r_1 \cup r_2 \cup \dots \cup r_n$  with  $n = |R|$ <sup>20</sup> and  $r_i \cap r_j = \emptyset$ ; for all  $i$  and  $j$ . Accordingly with the definition of relation instance in (2.3), and depending on the nature of  $KB$ , the entity inventory  $E$  might be composed of elements defined at the lexical level (i.e. words and multi-word expressions), or elements defined at the sense level (i.e. word senses and named entity mentions).

<sup>19</sup>In the particular case in which a relation  $r$  is identified by multiple paraphrastic relation patterns (Section 3.2.1)  $r$  is defined as the set of all relation instances such that  $r \in P_r$ , with  $P_r$  being a set of relation patterns associated with  $r$ .

<sup>20</sup>The strict equality holds under the simplifying assumption that each relation  $r$  is identified by a unique relation pattern; if relations are defined as sets of paraphrastic relation patterns (Section 3.2.1), then  $n < |R|$ .

## Chapter 3

# Related Work

El hecho es que cada escritor crea sus precursores.  
Su labor modifica nuestra concepción del pasado,  
como ha de modificar el futuro  
[The fact is that every author creates his own precursors.  
His work modifies our conception of the past,  
as it will modify the future.]  
Jorge Luis Borges

This chapter is devoted to reviewing in detail some literature work in the areas of Lexical Semantics and Open Information Extraction that is closely related to the contributions presented in this thesis. With respect to Chapter 2, where we broadly introduced the key topics, in the present chapter we narrow our focus on a series of specific approaches and methodologies that have been used in the past to achieve analogous or similar objectives to those outlined in Section 1.4. Most of the research efforts reviewed here constitute a key inspiration for the work presented throughout Chapters 4 and 5, as well as an important reference baseline for comparison in all the experimental evaluations therein.

We start by looking, in Section 3.1, at how sense-annotated resources have been constructed over the years, ranging from high-quality, manually curated corpora (Section 3.1.1) to semi-automatic (Section 3.1.2) and fully automatic (Section 3.1.3) methods targeted at scaling up the annotation process with a variety of techniques. We focus specifically on the latter category, where fully automatic approaches are examined. In the second part of the chapter, instead, we move to OIE (Section 3.2), and examine how semantic analysis at the sense level has been carried out in the published literature on the subject. While a great deal of literature on general OIE has been covered already in Section 2.3.2, in the present chapter we look closely at two semantically informed OIE approaches that are very similar in spirit to the present work, i.e. Patty (Section 3.2.1) and WiSeNet (Section 3.2.2).

### 3.1 Constructing Sense-Annotated Corpora

As discussed extensively in Chapter 1, phenomena like the knowledge acquisition bottleneck make annotating explicit sense-level information across textual data a very expensive and time-consuming endeavor. In fact, matching lexical items to suitable word senses and named entity mentions represents very often a tedious effort for human annotators, which becomes even more vexed when the inventory of concepts and named entities grows very large (as in large-scale knowledge resources like Wikipedia and BabelNet), or when sense distinctions in the lexicon are so fine-grained that telling them apart is problematic and, to a certain extent, subjective. As a result, manually curated corpora with high-quality annotations, such as SemCor (Miller et al., 1993), have limited size and coverage (Section 3.1.1).

To overcome this issue, several works have attempted to construct larger sense-annotated datasets by reducing as much as possible human intervention (Section 3.1.2), or by employing fully automatic disambiguation techniques (Section 3.1.3). Even though these techniques tend to produce noisier annotations, it has been shown that training on them leads to better supervised or semi-supervised models (Taghipour and Ng, 2015a; Raganato et al., 2017a).

While examining these works, we distinguish them not only on the basis of the degree of human supervision (manually curated, semi-automatic, or fully automatic), but also on the basis of the task they are designed for (WSD, EL, or both) and on the basis of the sense inventory they adopt (WordNet, Wikipedia, or BabelNet).<sup>1</sup>

The most prominent sense-annotated corpora covered in this section are summarized in Table 3.1, with some global statistics about the size and coverage of these corpora, together with their main features (nature of the sense annotations, reference sense inventory). Table 3.1 shows that sense-annotated resources can be quite heterogeneous in terms of size (total number of word tokens, and total number of sense annotations) and coverage (total number of distinct concepts and named entities with at least one annotation). As expected, hand-crafted resources tend to be considerably smaller, but their average number of annotations per word token (annotation density) is not necessarily lower compared to semi-automatically constructed corpora. In fact, annotation density depends not only on the degree of human supervision, but also on the nature of the corpus and on the task it is conceived for. Fully automatic approaches, on the other hand, tend to produce corpora with many more sense annotations, especially when they employ high-coverage disambiguation systems like Babelify (Section 2.2.2.3). However, annotation quality aside, automatically obtained resources tend to cover a smaller number of distinct concepts or named entities compared to human-designed resources, especially when the latter are based on definitional knowledge. This is due to the skewed distribution of many word senses and named entity mentions across natural language texts, and to the structural biases affecting the majority of disambiguation algorithms.

---

<sup>1</sup>There have been additional works to provide sense-annotated data based on other inventories, such as the New Oxford American Dictionary (Yuan et al., 2016), or language-specific WordNets (Agirre et al., 2005; Bentivogli and Pianta, 2005; Henrich et al., 2012a). However, for the sake of compactness, in the present section we limit our analysis to the subset of most widely used sense inventories that are considered in Chapters 4 and 5. A comprehensive survey of sense annotated corpora with language-specific WordNets is provided by Petrolito and Bond (2014).

	Sense Inventory	Type	# Annotations	# Senses	# Tokens
SemCor	WordNet	Manual	226,036	33,362	802,443
SemEval (all)	WordNet	Manual	7,253	3,669	25,503
Princeton WN Gloss	WordNet	Semi-automatic	458,825	59,250	1,621,129
OMSTI	WordNet	Semi-automatic	911,134	33,960	30,441,386
Wikipedia	Wikipedia	Collaborative	71,457,658	2,891,660	1,357,105,761
Wikilinks	Wikipedia	Collaborative	40,323,853	2,933,659	N.A.
Babel ed Wikipedia	BabelNet	Automatic	113,896,864	4,239,879	501,862,251
Babel ed MASC	BabelNet	Automatic	286,416	23,175	592,472

Table 3.1. Features and global statistics of some sense-annotated corpora treated in this section, including the reference sense inventory, the total number of sense annotations (# Annotations), the total number of concepts and named entities covered (# Senses), and the total number of word tokens (# Tokens). `Wikipedia` (5th row) refers to the English dump of November 2014, while Semeval (all) (2nd row) refers to the concatenated evaluation dataset from Raganato et al. (2017a).

Wikipedia-derived sense-annotated resources, including Wikipedia itself (cf. Section 2.1.2) represent a special case: despite having undergone a radically different construction process than expert-built resources, they can still technically be considered manually curated, since sense-level information in Wikipedia, encoded in hyperlinks, is always provided by a human editor. However, Wikipedia-based annotations are not aimed at providing training data for automatic methods; thus, only a specific subset of concepts and named entities in the large sense inventory of Wikipedia is actually annotated in text, and, even among those, only a fraction of the total number of corresponding mentions or lexicalizations are explicitly tagged. As a result, both annotation density and coverage of the resulting resources are relatively low, in spite of their larger sizes.

### 3.1.1 Manually Curated Corpora

#### 3.1.1.1 SemCor

The most prominent example of a manually curated resource is arguably the English SemCor corpus (Miller et al., 1993), one of the first sense-annotated corpora produced for English (and, in general for any language). Over many years SemCor stood as the largest textual resource annotated with word senses, and still constitutes one of the most widely used reference datasets across the NLP literature for training supervised disambiguation systems (Raganato et al., 2017a). SemCor has been part-of-speech tagged and sense-annotated manually at Princeton University by the WordNet Project research team, in a very early stage of the WordNet project, and it is currently distributed under the Princeton WordNet license.

SemCor consists of a subset of the Brown Corpus (Francis and Kucera, 1979) with approximately 800,000 words, out of which 200,000 open-class words (or multi-word expressions) have been sense-annotated using the WordNet sense inventory. The corpus is divided into two parts: the first portion (semcor-all) comprises 186 texts with sense annotations from all open-class parts of speech (noun, verb, adjective, and adverb), while in the second portion (semcor-verb) only verbal word senses

are annotated. Given that multi-word expressions such as phrasal verbs (e.g. get up) are also tagged, SemCor's sense annotations are not always continuous spans of text. Of course, closed-class words (such as prepositions and determiners) are only tagged if they are part of a multi-word expression.

The standardized version of SemCor released by Raganato et al. (2017a) comprises a total of 226,036 sense annotations covering 33,362 WordNet synsets, with an annotation density of 0.28 annotations per word token (i.e. approximately one token out of three is sense-tagged). If we consider only the first portion of SemCor, with sense annotations for all open-class parts of speech, the annotation density increases to 0.53, which makes SemCor one of the most densely annotated resources available. However, despite this and the high quality of sense annotations<sup>2</sup>, SemCor has several limitations: first of all, only 16% of the WordNet sense inventory is covered, and the nature of the source corpus makes SemCor's sense-level information somewhat outdated in the context of modern application scenarios<sup>3</sup>.

### 3.1.1.2 The Senseval/SemEval datasets

Another well-known example of corpus manually compiled with sense annotations is given by the training and testing datasets used in the Senseval/SemEval competition series (cf. Section 2.2.1.1). Most of these datasets are extensively used today as evaluation benchmarks for WSD systems. Five of them have been standardized and unified in the framework of Raganato et al. (2017a), including:

- ^ Senseval-2 (Edmonds and Cotton, 2001), the first and largest benchmark dataset, with 5,766 word tokens and 2,282 sense annotations from WordNet 1.7 for all open-class parts of speech;
- ^ Senseval-3 task 1 (Snyder and Palmer, 2004), similar to the Senseval-2 dataset, although slightly smaller (5,541 word tokens and 1,850 sense annotations). The dataset consists of three documents from three different domains (editorial, news story, fiction);
- ^ SemEval-07 task 17 (Pradhan et al., 2007), the smallest benchmark dataset, with 455 sense annotations from WordNet 2.1 for nouns and verbs only. Because of this, this is also the most ambiguous dataset (Raganato et al., 2017a);
- ^ The English portion of the SemEval-13 task 12 (Navigli et al., 2013), a very large dataset comprising thirteen documents from various domains, but including only nominal word senses from WordNet 3.0;
- ^ The English portion of the SemEval-15 task 13 (Moro and Navigli, 2015), the most recent WSD dataset available to date, annotated with WordNet 3.0

<sup>2</sup>Even though the original annotation of SemCor is known to be imperfect Bentivogli and Pianta (2005) estimated that around 2.5% of the sense tags are incorrect. SemCor is considered as one of the sense-annotated corpora with the highest quality, and its sense annotations are generally considered as gold labels when training supervised models.

<sup>3</sup>A typical example is the frequency distribution of word senses associated with the word pipe, where pipe<sub>1</sub> refers to the tobacco pipe and pipe<sub>2</sub> refers to the plastic or metal tubes used to carry water (which is arguably the most common usage of the word pipe nowadays).

and comprising 1,022 sense annotations in four documents from three different domains (biomedical, computing, society).

The concatenation of the above five datasets, used in the empirical comparison of Raganato et al. (2017a), reaches 25,503 word tokens and 7,253 sense annotations from WordNet 3.0. Even if it constitutes the smallest sense-annotated corpus we consider, this concatenation features a very high annotation density (0.28), as all included datasets have been originally designed as benchmarks for all-words WSD.

### 3.1.1.3 Other WordNet-annotated corpora

Beyond SemCor and the Senseval/Semeval datasets, a great deal of corpora annotated with WordNet or WordNet-like inventories have been released over the last two decades, with varying size and features (Petrolito and Bond, 2014). In this section we mention two notable examples, both sense-annotated with respect to the English WordNet, with a prominent role in the NLP literature:

- ^ MASC , Manually Annotated Sub-Corpus (Passonneau et al., 2012; Ide, 2012): an annotated portion of the American National Corpus, released with multiple layers of annotations in a common format that others can leverage to include additional annotations (Ide, 2012). The MASC corpus contains nineteen genres of spoken and written language data in roughly equal amounts, including social media material like tweets and blogs. Beside other annotation layers of various types (token and sentence boundaries, part-of-speech tags and lemma, shallow parse, logical structure), MASC also includes WordNet sense annotations for 1,000 occurrences of a selected set of 100 words and multi-word expressions. The sense-annotated data are distributed separately with links to the original documents in which they appear, without licenses or other restrictions, and they have been either manually produced or automatically produced and then fully hand-validated;
- ^ OntoNotes (Hovy et al., 2006): a collaborative effort among various institutions and universities toward the construction of a large semantically annotated corpus comprising various genres of text (news, conversational speech, weblogs, broadcasts, talk shows). The corpus is annotated with structural information (syntax and predicate-argument structure), and also with lexico-semantic information (word senses and named entity mentions, coreference resolution). As part of the latter, the authors have annotated the most frequent noun and verb senses in a 300,000-words subset of the PropBank corpus, using their multi-stage annotation procedure (Hovy et al., 2006). The released sense annotations are based on coarse-grained clusters of the original WordNet synsets (OntoNotes Sense Groups) and they cover 1.5 million English words.

### 3.1.1.4 Wikipedia-annotated corpora

As discussed in Section 2.1.2, Wikipedia itself can be viewed as a partially sense-annotated corpus which specifies its own encyclopedic sense inventory. In this respect, Wikipedia stands out both in terms of annotation quality (as hyperlinks are first generated and then validated by human editors). and corpus size (the dump of

November 2014, as reported in Table 3.1, is more than two orders of magnitude larger than any WordNet-annotated corpus available). However, given the fundamental nature and structure of Wikipedia, annotation density is extremely low (0.05), and almost half of the sense inventory is not covered at all across the corpus. In fact, exploiting the structure of Wikipedia to turn it into a full-edged sense-annotated corpus is one of the main focuses of Chapter 4.

On the other hand, Wikipedia, and portions of Wikipedia (Brümmer et al., 2016), are not the only Wikipedia-annotated resources available: beside the various training and testing benchmarks available for Entity Linking and Wikification (cf. Section 2.2.2), one of the most prominent resources featuring Wikipedia-based sense annotations is Wikilinks (Singh et al., 2012). The Wikilinks corpus was constructed by crawling the web and collecting hyperlinks (i.e. named entity mentions) linking to Wikipages (i.e. concepts and named entities), together with their surrounding context. With approximately 40 million mentions covering almost 3 million entities, harvested from over 10 million web pages, Wikilinks can be seen as a large-scale, naturally-occurring, crowd-sourced dataset where thousands of human annotators provide gold labels for mentions of interest.

Even though its actual size in terms of number of word tokens is hard to estimate given its heterogeneous composition, Wikilinks technically qualifies as a manually annotated Web-scale corpus, significantly larger than many other Wikipedia-annotated resources. However, Wikilinks does not address the sparsity and coverage issues of Wikipedia and, on the contrary, provides textual data with various kinds of noise, especially due to incoherent contexts (Eshel et al., 2017). While such contextual noise presents an interesting test case supplementing existing datasets (based instead on mostly coherent and well-formed text), it also makes it harder for Wikilinks to be a reliable training set for general-purpose WSD or EL systems, both in terms of annotation coherence and well-formed underlying textual sources.

### 3.1.2 Semi-Automatic Approaches

#### 3.1.2.1 The Princeton WordNet Gloss Corpus

The Princeton WordNet Gloss Corpus<sup>4</sup> is a sense-annotated corpus of textual definitions (glosses) drawn from the synsets of WordNet. The corpus contains 1,621,129 word tokens overall and 458,825 sense annotations, out of which 330,499 (72%) were obtained by manually linking to the context-appropriate sense in WordNet, and the remaining part was automatically disambiguated. Even though composed solely of textual definitions, the WordNet Gloss Corpus is twice as big as SemCor, and features approximately the same annotation density. Moreover, thanks to the nature of definitional text (which is not limited to the most frequently used word types, as in corpora drawn from real-world written or spoken text), the coverage of WordNet synset of the WordNet Gloss Corpus is almost doubled compared to other WordNet-annotated resources, such as SemCor and MSTI (Section 3.1.2.2).

The WordNet Gloss Corpus, a sense-annotated corpus of definitional knowledge, has already been proved useful in various NLP tasks, including semantic similarity (Pilehvar et al., 2013), domain labeling (González et al., 2012) and, of course,

<sup>4</sup><http://wordnet.princeton.edu/glosstag.shtml>

knowledge-based WSD, from earlier definition-based algorithms (Lesk, 1986) to more modern approaches (Baldwin et al., 2008; Agirre and Soroa, 2009; Camacho Collados et al., 2015b). Motivated by the key role of definitional knowledge in WSD and Lexical Semantics in general, we also focus on textual definitions in Section 4.3, where we study a fully automatic algorithm to generate a multilingual sense-annotated corpus of definitional knowledge, and in Section 5.1, where we design a full-edged OIE pipeline targeted at textual definitions.

### 3.1.2.2 OMSTI

An effective way of automatizing the construction of sense-annotated corpora involves the use of parallel text, as in cross-lingual WSD (Section 2.2.1.1). However, in order to obtain reliable sense-level information, human supervision is still needed as, for instance, word alignments might be imperfect and propagate through subsequent stages of the annotation process. This is the case of the One Million Sense-Tagged Instances corpus (Taghipour and Ng, 2015b, OMSTI),<sup>5</sup> a semi-automatically constructed corpus annotated with WordNet synsets. The authors employed a well-known alignment-based WSD approach (Ng et al., 2003; Chan and Ng, 2005) to harvest approximately one million training samples from a large English-Chinese parallel corpus, MultiUN (Eisele and Chen, 2010, MUN). OMSTI has been tested experimentally as training set for supervised WSD and, coupled with SemCor (Section 3.1.1.1), has been widely used after its public release (Taghipour and Ng, 2015a; Iacobacci et al., 2016; Raganato et al., 2017a).

Given a parallel corpus already preprocessed (tokenization, word segmentation) and word-aligned, the semi-automatic annotation procedure carried out in OMSTI (Chan and Ng, 2005) works as follows: for each synset in the WordNet sense inventory associated with an English word  $w_e$ , a hand-crafted list of English-Chinese translations is used to check every lexicalization  $o_f$ ; if a lexicalization matching  $w_c$  (i.e. the Chinese word aligned with  $w_e$ ) is found, then  $w_e$  is tagged with senses. Even though this procedure can generally be very noisy, the authors have manually validated a subset of 1,000 randomly selected sense annotations and estimated an accuracy of 83.7%, which would be reasonably high for fully automatic approaches such as those presented in Section 3.1.3. This shows that semi-automatic methods can often provide a middle ground between the two extremes of full and zero human supervision, at least in terms of annotation quality.

The word types annotated in OMSTI include 649 nouns, 190 verbs, and 219 adjectives selected among the top 60% most frequent word types in the Brown Corpus. The authors have extracted at most 500 random samples per word sense, which have been used to construct a training dataset as balanced as possible. Finally, samples for 28 adverbial word senses have been added from SemCor, together with sense-annotated samples from the DSO corpus (Ng and Lee, 1996) which, however, is proprietary material and was not included in the final release.

Overall, OMSTI is the largest WordNet-annotated resource reported in Table

<sup>5</sup><http://www.comp.nus.edu.sg/~nlp/corpora.html>

<sup>6</sup>Even though the original release of OMSTI features SemCor already included, in the present chapter we follow Raganato et al. (2017a) and use OMSTI when referring to the portion of sense-annotated data from MUN only.

3.1, both in terms of size and number of sense annotations. However, given the high-precision annotation procedure used to construct it, its annotation density is one of the lowest (0.03) and the coverage of concepts and named entities is approximately the same of SemCor, even though OMSTI is almost two orders of magnitude larger.

### 3.1.3 Fully Automatic Approaches

#### 3.1.3.1 WordNet-annotated corpora

The earliest fully automatic approaches to the construction of WordNet-annotated corpora were either based on bootstrapping from one-sense-per-discourse and one-sense-per-collocation heuristics (Yarowsky, 1995), or focused on building rich Web queries in such a way that the words occurring in the retrieved documents are, with some probability, associated with the desired sense (Leacock et al., 1998; Mihalcea and Moldovan, 1999; Agirre and Martínez, 2000). The latter strategy is heavily based on exploiting monosemous words (i.e. words appearing only in one synset, cf. Section 2.2) and their connections to the senses of an ambiguous target word via specific lexico-semantic relations (synonymy, hypernymy, hyponymy), but it can also be extended via topic signatures (Agirre et al., 2000), or seed examples from manually sense-annotated corpora (Mihalcea, 2002). Despite the high precision reported by manual assessments on random samples, larger comparative evaluations (Agirre and Martínez, 2000, 2004) suggested that sense-annotated examples obtained from the Web can be affected by topical biases, especially when a new sense of a given target word is predominant on the web (e.g. mentions like *òasis* and *`nirvana* across the Web are mostly referring to music groups not covered by WordNet). Later approaches have improved by exploiting additional resources, such as the automatic association of Web directories from the Open Directory Project (ODP) to WordNet synsets (Santamaría et al., 2003), or the automatic mapping between WordNet/GermaNet and Wiktionary (Henrich et al., 2012b).

An alternative annotation strategy that has become extremely popular over the years is instead based on exploiting translation correspondences from parallel text, and on projecting them using word alignments or other techniques (Ide et al., 2001; Diab and Resnik, 2002; Ng et al., 2003; Chan and Ng, 2005; Zhong and Ng, 2009; Lefever et al., 2011; Yao et al., 2012; Bonansinga and Bond, 2016). All these approaches are based on the underlying intuition that polysemy can be reduced, at least partially, by looking at the different translations of an ambiguous English word in other languages. As discussed in Section 2.2.1.1, this experimentally verified intuition is the fundamental idea behind cross-lingual WSD (Lefever and Hoste, 2010, 2013), and has demonstrated his effectiveness in producing high-quality sense-annotated data (Chan and Ng, 2005). In recent times, the development of executable knowledge-based WSD models like UKB (Agirre et al., 2014), that can easily be adapted to languages other than English, has led to the fully automatic disambiguation of large-scale parallel corpora, such as Europarl (Koehn, 2005), using off-the-shelf knowledge-based systems (not necessarily based on word alignments) together with a set of language-specific preprocessing pipelines (Otegi et al., 2016). Such an approach demonstrates that not only alignment techniques can be useful for WSD, but also vice versa: multilingual WSD can be exploited to disambiguate

large-scale training datasets for Machine Translation, thereby encouraging machine translation approaches explicitly aware of Lexical Semantics (cf. Chapter 1).

However, regardless of the strategy used, automatic sense-annotation methods have a common issue when dealing with corpora in multiple languages: each language relies on its own sense inventory (e.g. a language-specific WordNet) while an optimal cross-lingual disambiguation approach based on parallel text would require a language-independent annotation framework that goes beyond monolingual WordNet-like sense inventories (Lefever et al., 2011).

### 3.1.3.2 Wikipedia-annotated corpora

In the same way an off-the-shelf WSD system can be used to build a WordNet-annotated corpus automatically, automatic EL approaches (Section 2.2.2) can be used to construct large-scale corpora annotated with Wikipedia or Wikipedia-derived resources with no human supervision, not even collaborative (cf. Section 3.1.1). The most prominent example of this kind is arguably the Freebase annotation of the ClueWeb Corpora (Gabrilovich et al., 2013, FACC), which comprises around 800 million web documents from ClueWeb09 and ClueWeb12 with 11 billion entity mentions automatically disambiguated and linked to the most suitable Freebase entries. The automatic linking procedure of FACC was optimized for precision over recall, and left out many low-confidence annotations. On the basis of manual assessment over a sample of documents, the authors have estimated FACC's overall precision to be about 80-85%, with recall in the range of 70-75%. Similarly to Wikilinks (Singh et al., 2012), another web-scale resource designed to encode Wikipedia-derived information, FACC is arguably among the largest semantic resources available, but annotation density is considerably lower, with an average of approximately 13 sense annotations per document.

The noisy nature of Web-derived textual data, such as those in FACC and Wikilinks, represents a challenge by itself for EL systems, and has recently inspired researchers to investigate the task in a setting where only local and noisy context is available; one of the latest examples is WikilinksNED (Eshel et al., 2017), a large-scale dataset composed of 3.2 million short text fragments from the Web, which is significantly noisier and more challenging than similar annotated corpora for EL. In order to capture the limited and noisy local context surrounding each of the 18,000 mentions inside WikilinksNED, a recurrent neural model was designed and trained with an ad-hoc method for sampling informative negative examples.

Other approaches have been proposed to automatically harvest named entity mentions linked to the DBpedia ontology (Lehmann et al., 2014): the Kaist corpus (Hahm et al., 2014), based on an English Wikipedia dump, comprises 6.8 million sentences and about 157 million word tokens, with more than 98 million mentions linked to DBpedia; the Europarl-QTLeap WSD/NED corpus (Otegi et al., 2016), instead, is based on Europarl coupled with the QTLeap corpus (a collection of 4,000 question-answer pairs in the domain of IT troubleshooting), and includes 3.11 million mentions in six languages linked to DBpedia (but also features NER, WSD, and coreference information). Both contributions either employ a language-independent pipeline (Hahm et al., 2014), or an array of language-specific pipelines (Otegi et al., 2016) to produce large-scale resources annotated with respect to a Wikipedia-derived

(hence language-independent) sense inventory. Even though they do solve some of the issues of WordNet-annotated resources (Section 3.1.3.1), these corpora tend to focus exclusively on named entities and neglect general concepts or non-nominal senses, especially when sense annotations are harvested with automatic EL methods.

### 3.1.3.3 BabelNet-annotated corpora

In recent years the development of multilingual knowledge resources like BabelNet (Section 2.1.3) has marked a clear turning point in the field, introducing comprehensive sense inventories suitable for both WSD and EL at the same time (Section 2.2.2.3). As we discussed throughout this section, WordNet-annotated corpora for WSD have been around for more than two decades (Petrolito and Bond, 2014), and Wikipedia-annotated corpora for EL have also followed the same path: as a result, smaller (but denser) corpora annotated with word senses are available on one side, and larger (but sparser) corpora annotated with named entity mentions are available on the other. Although recent automatic approaches (Otegi et al., 2016) have addressed the issue of building a multilingual sense-annotated corpus suitable for both WSD and EL, a set of different monolingual sense inventories is still required to encode word senses and named entity mentions in many languages.

In contrast, BabelNet enables word senses and named entity mentions in multiple languages to be encoded using a single, unified and language-independent sense inventory. This great advantage, however, comes with a cost: given the size of such an encyclopedic inventory, manual or semi-automatic annotation approaches become prohibitively difficult and time-consuming. This is why, to date, all the research efforts in building BabelNet-annotated corpora have been fully automatic disambiguation approaches based on knowledge-based WSD/EL systems like Babelfy (Moro et al., 2014b). One of the earliest attempts of this kind is the disambiguation of the MASC corpus (Moro et al., 2014a, Babel ed MASC ),<sup>7</sup> a resource that covers different genres and domains, encoded in a convenient unified format that favors the integration of different kinds of semantic annotation (Ide, 2012). The Babel ed MASC corpus comprises 592,472 word tokens and 286,416 sense annotations, and features the highest annotation density (0.48) among all the resources in Table 3.1. However, MASC has a limited size, even compared with earlier WordNet-annotated corpora like SemCor (Section 3.1.1.1), and the accuracy of its sense annotation is estimated to be around 70% (Moro et al., 2014a).

In line with this approach, Babelfy has also been used to disambiguate a large portion of the English and Italian Wikipedias (Scozzafava et al., 2015, Babel ed Wikipedia )<sup>8</sup>, both publicly released in NIF and XML format. In this case, the disambiguation approach took advantage of Wikipedia's internal hyperlinks, which were converted into Babel synsets via BabelNet's inter-resource mappings and used as gold annotations. Similarly to other Wikipedia-based approaches (Hahm et al., 2014), this disambiguation procedure can be easily applied to Wikipedias in other languages, since Babelfy handles all the languages supported by BabelNet. With over 500 million word tokens, the Babel ed Wikipedia is the largest BabelNet-annotated resource reported in Table 3.1, and its annotation density (0.23) is comparable to

<sup>7</sup><http://lcl.uniroma1.it/MASC-NEWS>

<sup>8</sup><http://lcl.uniroma1.it/babelfied-wikipedia>

that of WordNet-annotated corpora; the accuracy of its sense annotations, estimated manually on a sample of 1,000 Wikipages, is similar to the one reported for the Babel ed MASC corpus: 70.5% for English and 72.3% for Italian.

Finally, a Babelnet-based multilingual disambiguation procedure has also been tested experimentally in the history domain: within the semantic indexing pipeline of Raganato et al. (2016a), Babelify has been adapted to disambiguate a version of the Bible translated in four different languages, and enable cross-lingual text retrieval via Babel synsets. Despite the inherent difficulty of the domain, especially for general-purpose disambiguation systems, Babelify outperformed the MCS baseline on a manually-annotated sample of the corpus, with overall precision of 68.8% on the English version, obtained without any domain-specific tuning or prior translations.

All the approaches described in the present section, with their own advantages and limitations, show that a knowledge resource like BabelNet is a key requirement in order to produce large-scale corpora where: 1) word senses and named entity mentions are annotated simultaneously and linked to a unified sense inventory; 2) scaling up to multiple languages does not require ad-hoc specializations of the disambiguation pipeline, or language-specific sense inventories. However, an important issue that still needs to be fully addressed concerns the quality of sense annotations: given the size and scope of BabelNet's sense inventory, obtaining gold (or even silver) labels is unpractical, if not totally unfeasible. Fully automatic disambiguation methods, on the other hand, have to cope with their well-known shortcomings and structural biases (Section 2.2.1) in order to produce high-quality sense annotations that are expendable within downstream tasks and applications.

## 3.2 Semantically Informed Open Information Extraction

As we discussed extensively in Section 2.3, Information Extraction is concerned with harvesting relations between entities and subsequently encoding them as triples of the form  $(\text{entity}, \text{relation}, \text{entity})$  inside a knowledge base. In order to extract relations without pre-defining them, the OIE paradigm was introduced (Section 2.3.2) and became increasingly popular: over the last two decades, the OIE community has witnessed a wide variety of approaches targeted at extracting meaningful and informative relation patterns in textual forms, constantly improving and overcoming the shortfalls of earlier models. Some approaches have even investigated how to exploit deeper semantic features, drawn from Semantic Role Labeling, to extract relation triples with a more solid semantic structure (Christensen et al., 2010; Mesquita et al., 2013; Stanovsky and Dagan, 2016). In fact, recent contributions have looked extensively at the close relationship between OIE and other NLP tasks traditionally associated with a deeper semantic analysis, such as Question Answering and Semantic Parsing (Yao et al., 2014; Khot et al., 2017).

Notwithstanding all the efforts in connecting OIE with semantic analysis to improve the extraction process, none of the techniques mentioned above actually focused on providing a more structured semantic representation of the triples themselves which remain still anchored to surface text: both subject and object arguments are

always represented by noun phrases, while relations are encoded with verb phrases, either raw or lemmatized. In other words, beside some notable exceptions focused on learning latent features for such relations (cf. Section 2.3.3), little or no attention has been paid in making explicit the semantics of the extracted information.

However, taking into account Lexical Semantics is intuitively an important step towards improving the extraction of high-quality relation triples; two of the most prominent linguistic phenomena that cause OIE systems to produce redundant extraction are synonymy and paraphrases. For instance, ReVerb (Fader et al., 2011), one of the most well-known traditional OIE approaches (cf. Section 2.3.2), extracts the following two synonymous relation instances<sup>9</sup>:

Natural Language Processing is a field of, Computer Science  
 Natural Language Processing is an area of Computer Science

In fact, ReVerb enforces purely syntactic constraints on the aspect of a relation phrase (using hand-crafted part-of-speech-based regular expressions), and lexical constraints only on the number of different arguments a relation phrase appears with. In order to reduce this kind of redundancy, relational clustering approaches have been proposed (Kok and Domingos, 2008; Yates and Etzioni, 2009) to identify synonymous phrases like those in the example above. While clustering definitely helps in dealing with paraphrases and synonymy, the issue is only partially solved, as the underlying assumption of such clustering techniques is that a relation phrase can have only one meaning, which limits the number of distinct relation phrases associated with a relation (Yates and Etzioni, 2009). In other words, these approaches fall short in dealing with another major linguistic phenomenon that affects the extraction process: lexical ambiguity. Lexical ambiguity can arise both at the level of arguments and the level of relational phrases. The latter case is the one affecting the example above, with the ambiguous word *field* that could either refer to a piece of land (*field<sub>n</sub><sup>1</sup>* in WordNet) or to a specific branch of knowledge (*field<sub>n</sub><sup>4</sup>*). A viable way of dealing with relation phrase ambiguity is that of ontologizing semantic relations (Pennacchiotti and Pantel, 2006). However, the necessary use of WordNet (Section 2.1.1) to perform this task makes the ontologization step difficult for many domains, because of WordNet's inherent lack of coverage of specialized concepts and named entities.

As for resolving lexically ambiguous arguments, instead, the most suitable solution would consist in performing WSD or EL on the source prior to extracting the relation instance. In fact, while a Named Entity Recognition module can easily get rid of coarse-grained ambiguities, more fine-grained lexical distinctions require a deeper analysis at the sense level, and are often influenced by the semantics of the verb heading the relation phrase (that cannot be captured by a NER module). Let us consider the following example:

Washington is the capital of the United States

In this case, the ambiguous subject argument cannot be resolved by a traditional NER module: even if correctly labeled as *LOCATION*, the mention *Washington* could still refer to the state of Washington, or to the U.S. capital. Identifying the latter as the correct meaning of *Washington* is a WSD problem that, the example above,

<sup>9</sup>Example borrowed from Moro and Navigli (2012).

crucially depends on disambiguating the wordcapital in the relation phrase with, e.g., the WordNet sensecapital<sub>1</sub> (the federal government of the United States).<sup>10</sup>

In this section we examine in detail two OIE approaches, Patty (Section 3.2.1) and WiSeNet (Section 3.2.2), that adopted a radically different strategy, i.e. that of modeling Lexical Semantics explicitly. We refer to this strategy as Semantically Informed Open Information Extraction.

The first and foremost difference with respect to traditional OIE is that, in these approaches, the semantics of subject and object arguments is explicitly modeled by disambiguating their surface-text forms and linking them to knowledge resources like Wikipedia or Yago. This crucial step enables, on the one hand, to resolve ambiguous relation patterns (e.g., in the first example above, knowing that Natural Language Processing is an academic discipline activates the meaning of field that refers to a branch of knowledge); on the other hand, disambiguated arguments can be exploited to identify synonymous relation patterns (such as as is a field of and is an area of) or, in some cases, taxonomize relation patterns by discovering subsumption relationships between the corresponding domains and ranges (e.g. knows subsumes is dating). All these techniques enable Patty and WiSeNet to extract more accurate and structured relation triples, and at the same time generalize better than traditional OIE techniques based only on surface-text analysis.

### 3.2.1 Patty

Motivated by the goal of extending WordNet's taxonomic structure to OIE-derived relation patterns, Patty (Nakashole et al., 2012)<sup>11</sup> puts forward a large-scale OIE approach to systematically harvest relation pattern from Web-scale textual corpora, and to impose a semantically typed structure on them in order to construct a WordNet-style subsumption taxonomy of binary relations. Similarly to Nell's type signatures (Carlson et al., 2010), which however are pre-specified manually, Patty aims at extracting typed patterns such as hSINGERsingshSONG generalizing them with respect to syntactic variations (e.g. sings [PRP]hSONG in place of sings her hSONG and sings his hSONG), and finally discovering pattern subsumptions (e.g. hSINGERsingshSONG being subsumed by hMUSICIANperforms or hCOMPOSITION

Patty addresses many of the issues pointed out in the previous section. In fact, while generalizing patterns helps dealing with synonymy and paraphrasing, typing patterns with semantic signatures is a way of overcoming lexical ambiguity. The resulting new type of relation patterns, not tied to surface text as in traditional OIE, is referred to as syntactic-ontological-lexical (SOL) pattern model. An example of SOL pattern, borrowed from Nakashole et al. (2012), is the following:

hPERSONs [ADJ] voice \* hSONG

This pattern matches can be extracted, for instance, from sentences like:

- (a) Amy Winehouse's soft voice in 'Rehab'
- (b) Elvis Presley's solid voice in his song 'All shook up'

<sup>10</sup>In fact, some recent work on fine-grained NER (Ling and Weld, 2012; Shimaoka et al., 2017) has shown the positive impact of a more diverse and specific set of entity types, thereby gradually blurring the boundaries between NER and actual WSD.

<sup>11</sup><http://www.mpi-inf.mpg.de/yago-naga/patty>

In marked contrast with surface-form OIE patterns, the SOL pattern above not only drops ReVerb's verb-centric assumption, as both (a) and (b) are noun phrases, but includes lexical word features (e.g. voice), together with syntactic generalization based on part-of-speech tags and wildcards (e.g. ADJ or \*) and, crucially, ontological type signatures represented by a pair of entity placeholders, such as PERSON SONG in the example above. Nakashole et al. (2012) define the support set of a given SOL pattern as the set of argument pairs appearing in place of the entity placeholders in the extracted relation instances. Also, they define a given SOL pattern  $p_a$  as syntactically more general than another SOL pattern  $p_b$  when every surface-text phrase that matches  $p_b$  also matches  $p_a$ . Similarly,  $p_a$  is semantically more general than  $p_b$  when the support set of  $p_a$  is a superset of the support set of  $p_b$ . If this relationship holds in both ways, i.e.  $p_a$  is semantically more general than  $p_b$  but at the same time  $p_b$  is semantically more general than  $p_a$ , then the two SOL patterns are synonymous and can be grouped together in a pattern synset.

### 3.2.1.1 Methodology

Patty's pipeline takes as input a textual corpus and comprises three stages:

1. **Pattern Extraction** : in order to obtain a first set of surface-text patterns from the input corpus, Patty performs syntactic dependency parsing (de Marneffe et al., 2006) on each sentence of the corpus to produce a word-level dependency graph. At the same time, named entity mentions across the sentence are detected and linked to the sense inventory of *Yago2* (Hort et al., 2011a) using a disambiguation procedure based on a context-similarity prior (Suchanek et al., 2009). Then, given two disambiguated entity mentions, the dependency graph of the sentence is traversed to get the shortest path between them.<sup>12</sup> To obtain the final textual pattern, the shortest path is subsequently expanded with adverbial and adjectival modifiers;
2. **Pattern Generalization** : the extracted patterns are turned into SOL patterns and generalized syntactically, by replacing less-frequent n-grams with wildcards and part-of-speech placeholders, and semantically, by generalizing their semantic types. In order to avoid too abstract and meaningless patterns, the generalization is stopped when a SOL pattern subsumes multiple patterns with disjoint support sets. In this phase, the statistical strength of a SOL pattern is quantified by associating each pattern  $p$  with a confidence value, computed as the ratio of the support-set sizes of  $p$  and  $p^t$  (an untyped variant of  $p$  where type signatures are replaced by the generic *Yago* type ENTITY);
3. **Taxonomy Construction** : in the third stage, the generated SOL patterns are arranged in a subsumption taxonomy. Since support sets may contain noise in terms of spurious or incomplete entity pairs, pattern subsumption is based on a probabilistic soft set inclusion procedure, where a certain set can be a subset of another set to a certain degree. Also, instead of comparing

<sup>12</sup>In order to deal with noisy extraction, a set of syntactic constraints is used to capture only relations that refer to subject-relation-object triples: for instance, only shortest paths starting with subject-like dependencies (nsubj, rcmmod, partmod) are considered.

every SOL pattern pairwise and check whether subsumption holds, Nakashole et al. (2012)'s approach is based on constructing a pre x-tree for frequent patterns, which is then used to mine subsumptions and semantic equivalences (i.e. synonymy) across patterns. Finally, the obtained taxonomy is refined in order to obtain a directed acyclic graph defined over pattern synsets: the Patty taxonomy.

### 3.2.1.2 Experimental Evaluation

Patty's pipeline was evaluated experimentally on two different input corpora: the New York Times archive (NYT), which includes 1.8 million articles from the years 1987 to 2007, and a June 2011 dump of the English Wikipedia (WKP), featuring about 3.8 million articles. Entity disambiguation and typing was based on two underlying sense inventories, Yago2 (Hofmann et al., 2011a) and Freebase (Bollacker et al., 2008). After being run on both corpora, the Yago2-based extraction pipeline produced 86,982 SOL patterns from the NYT corpus and 360,562 SOL patterns from the WKP corpus, while the Freebase-based variant, which relied on Freebase's coarse-grained categories as semantic types, produced 809,091 and 1,631,531 SOL patterns, respectively.

In order to evaluate the quality of the extracted patterns, Nakashole et al. (2012) employed a manual evaluation based on several human judges, which were shown a sampled pattern synset, its type signature, a few example relation instances, and then asked to state whether the pattern synset indicated a valid semantic relation. This assessment was performed both on the top 100 most confident pattern synsets, as well as on a random sample with the same size, and showed an average precision in the range 87%-95% on the top 100 sample, and in the range 71%-85% on the random sample. These figures demonstrate that dealing explicitly with linguistic phenomena like synonymy and ambiguity enables the extraction of high-quality relation instances, which have shown to be useful also for extrinsic tasks like relation paraphrasing with respect to DBpedia and Yago2 (Nakashole et al., 2012).

On the other hand, most of the downsides of Patty concern the taxonomy construction step. In fact, the best experimental configuration reported by Nakashole et al. (2012), i.e. the Yago2-based pipeline run on the WKP corpus, yields a relatively sparse subsumption taxonomy, composed of 8,162 hypernymy edges with a manually estimated precision of 75% (on the random sample) and 83% (on the top 100 sample). Even though many interesting subsumptions were discovered (e.g. `!PERSONwinner of !AWARD` being subsumed by `!PERSONwas nominated for !AWARD` or `!PERSON's wife !PERSON` being subsumed by `!PERSON's widow !PERSON`) the reported figures suggest that there is still room for improvement, especially in terms of coverage. Indeed, a few follow-up contributions have addressed this issue, and improved over the original subsumption taxonomy of Nakashole et al. (2012). For instance, Harpy (Grycner and Weikum, 2014) puts forward a graph-based alignment algorithm which exploits random walks to link Patty's relation patterns to verb senses in WordNet, obtaining a larger pattern taxonomy and, as a by-product, fine-grained lexical types for the arguments of WordNet's verb senses. Similarly, the method adopted by Relly (Grycner et al., 2015) builds upon Patty and leverages collective probabilistic programming techniques to construct

Figure 3.1. Excerpts of the WiSeNet semantic network before(a) and after (b) the relation ontologization stage. Figure borrowed from Moro and Navigli (2012).

a high-coverage, high-precision taxonomy of about 20,000 relation patterns with 35,000 hypernymy links, while retaining (or even improving) the manually assessed accuracy of Patty 's hypernymy edges.

### 3.2.2 WiSeNet

Among other findings, the experimental evaluation of Patty (Section 3.2.1) has demonstrated the advantages of using Wikipedia, a large-scale general-purpose encyclopedic resource, over the noisier news-based data of the New York Times archive; as observed by Nakashole et al. (2012), some portions of the corpus (e.g. news about the stock market) do not express actual relational information. However, despite being a semantically informed approach Patty is not specifically designed for Wikipedia, and hence does not take into account the semantic information that is already available and encoded within the structure of Wikipedia, such as internal hyperlinks or Wikipedia categories (Section 2.1.2).

In contrast, WiSeNet (Moro and Navigli, 2012, 2013)<sup>13</sup> focuses on combining the advantages of a semi-structured knowledge resource like Wikipedia and the large-scale harvesting techniques of traditional OIE systems, with the goal of building a Wikipedia-based semantic network. Similarly to Patty, WiSeNet explicitly addresses linguistic phenomena like synonymy and polysemy, but instead of formalizing enhanced relation patterns that are subsequently generalized with wildcards or coarser semantic types, relies on Wikipedia's internal hyperlinks to extract non-ambiguous argument pairs, and on Wikipedia categories to generate semantic type signatures for its relation patterns. At the same time, WiSeNet is also able to identify synonymous relation phrases and cluster them into ontologized relation synsets. Following the line of similar approaches (Nastase and Strube, 2013), WiSeNet aims at turning Wikipedia into a full-edged semantic network; however, instead of considering a pre-specified set of infobox-based semantic relations, OIE techniques are leveraged to discover these relations automatically. Figure 3.1 exemplifies the final output of WiSeNet 's extraction pipeline: not only the semantic connections among Wikipages, previously defined by unspecified hyperlinks, are labeled with a suitable relation phrase (Figure 3.1a), but ambiguous relation phrases are then replaced by ontologized synsets of synonymous phrases (Figure 3.1b).

<sup>13</sup><http://lcl.uniroma1.it/wisenet>

### 3.2.2.1 Methodology

WiSeNet's pipeline (Moro and Navigli, 2012) is based on two successive stages:

1. **Relation Extraction** : the objective of this first stage is that of extracting a set of OIE-style relation instances from the input corpus (an English Wikipedia dump) such that, for each relation instance, the left and right arguments are disambiguated entity mentions linked to suitable Wikipages. While the output of this stage is analogous to that of Patty's pattern extraction step, the methodology is substantially different: instead of using a probabilistic WSD technique, WiSeNet identifies pairs of hyperlinked mentions inside a Wikipage; then, instead of applying syntactic analysis, the corresponding relation phrase is obtained by just considering the span of text between the two mentions, if it comprises at least one verb. The output of this step is a shallow semantic network, as displayed in Figure 3.1a;
2. **Relation Ontologization** : this second and final stage, instead, is focused on ontologizing the extracted relation instances, thereby dealing with synonymy and polysemy explicitly. This process is carried out in three steps:
  - ^ Clustering of synonymous relation phrases by means of a distributional method based on defining a measure of semantic similarity between two given relation phrases and  $\rho$ . This method consists in constructing a vector representation for the left and right arguments of  $\rho$  and  $\rho'$ , and then computing the harmonic mean between the cosine similarity of the corresponding left and right vector pairs. This step generates a set of relation synsets (e.g.  $f$  is a field of, is an area of, is studied in) from the extracted relation instances, similarly to the pattern synsets in Patty;
  - ^ Semantic labeling of relation synsets based on the identification of a set of Wikipedia categories describing their arguments. This step is carried out with a depth-first-search exploration of the Wikipedia category hierarchy up to a fixed depth, followed by a ranking of such categories based on the number of visits. This ranking is then used to extend the set of categories that are originally associated with the Wikipages representing the left and right arguments of each relation instance. As output of this step relation synsets are ontologized, i.e. they feature disambiguated arguments identified by one or more Wikipedia categories;
  - ^ Disambiguation of relation instances, the last ontologization step, which deals with lexically ambiguous relation phrases with an explicit disambiguation procedure. Given a relation instance  $\rho = \langle p_1; \rho; p_2 \rangle$ , this procedure disambiguates  $\rho$  with the most suitable relation synset  $R$ , (among all relation synsets containing  $\rho$ ) by maximizing the intersection of common Wikipedia categories between  $\rho$  and  $R$ . For instance, given the triple  $\langle \text{Natural Language Processing}; \rho; \text{Computer Science} \rangle$ , the disambiguation procedure should associate  $\rho$  with the relation synset  $R_1 = \langle f \text{ is a field of, is an area of, is studied in} \rangle$  instead of the relation synset  $R_2 = \langle f \text{ is a field of, is cultivated with, where grows} \rangle$ , since  $R_1$  would be identified by Wikipedia categories related to academic disciplines.

Figure 3.2. The syntactic constraint introduced by Moro and Navigli (2013) with two example dependency trees: one for the artificial phrase  $x$  is located in  $y(a)$ , and another one for the artificial phrase  $\bar{x}$  and  $x$  located in  $y(b)$ . The latter parse, with  $x$  being connected with a `conj` dependency to the head verb, is filtered out.

The overall output of this stage is a semantic network of Wikispaces interconnected with semantically typed relation synsets (Figure 3.1b).

The extraction process just described (Moro and Navigli, 2012) manages to obtain semantically informed relation instances with high coverage, but the accuracy of relation phrases is hampered by two main issues: first of all, since only shallow syntactic analysis is performed at extraction time, over-specific and noisy phrases can be retained (e.g. `is the name Gulliver gives his nurse in Book 11 of but then lost to`); furthermore, measuring the similarity of relation phrases by solely exploiting the left and right arguments might generate many false positives, as the same arguments can be related by multiple semantic relations (e.g. `married to`, `is a friend of`, `started a company with`). In order to overcome these issues, an enhanced version of the WiSeNet pipeline, based on a deeper syntactic and semantic analysis, is proposed by Moro and Navigli (2013), with the following important improvements:

- ^ At extraction time, a syntactic constraint based on a computationally efficient test is used to filter out ill-formed relation phrases: given a relation phrase  $\phi$ , an artificial phrase is constructed by concatenating the symbol  $x$ ,  $\phi$  and the symbol  $y$ ; then a dependency parser is applied and, if  $x$  and  $y$  are marked as subject and object, respectively, in the resulting dependency graph, then  $\phi$  is retained (Figure 3.2). This constraint, in addition to a threshold on the minimum number of relation instances extracted for each relation phrase, helps reducing the amount of noisy extractions to a large extent;
- ^ At ontologization time, a sophisticated soft clustering technique, based on a shortest-path dependency kernel and on a distributed kernel-based K-medoids algorithm (Zhang and Rudnicky, 2002), is used to synergistically cluster synonymous relation phrases, while at the same time letting polysemous relation phrases belong to more than one cluster. Crucially, the kernel-based similarity measure introduced by Moro and Navigli (2013) considers three different aspects of each relation phrase: its dependency structure, the distributional semantics of its words, and the semantics of its arguments.

#### 3.2.2.2 Experimental Evaluation

Both versions of WiSeNet have been evaluated experimentally using an English Wikipedia dump of late 2012. The enhanced version (Moro and Navigli, 2013)

additionally employed the Stanford Parser (de Marne e et al., 2006) for dependency parsing, and the Gigaword corpus<sup>14</sup> to compute distributional vectors. While the earlier version extracted as many as 16,344,622 relation instances with 10,863,122 distinct relation phrases, the enhanced pipeline reduced the number of extractions to 2,271,807 relation instances and 245,935 distinct relation phrases. These figures suggest that the conservative strategy based on syntactic and frequency constraints adopted by Moro and Navigli (2013) helps dealing with data sparsity by cutting a long tail of infrequent and possibly over-specific relation phrases: in fact, its average number of extractions per relation phrase increases from 1.50 to 9.24.

As regards the accuracy of the extracted information, Moro and Navigli (2013) carried out a manual assessment based on Amazon Mechanical Turk, in which both versions of WiSeNet were compared. The pipeline was evaluated at four different levels, each time with a different sample of 2,000 randomly extracted items:

- ^ Level 1 (relation instances) , where human judges were presented with a relation phrase and with the two Wikipages corresponding to its left and right arguments, and asked whether the relation instance was correct. At this level, the enhanced version of WiSeNet reported an accuracy of 91.8%, with an improvement of 9% with respect to the earlier version;
- ^ Level 2 (relation phrases) , in which the judges were presented with a relation phrase and asked if they could think of a subject and object that would fit the phrase. Consistently with the previous evaluation, the enhanced version of WiSeNet achieved 94.5% accuracy on relation phrases, improving over 14% with respect to the earlier version;
- ^ Level 3 (relation synsets) , where the judges were asked to examine two synonymous relation phrases for each relation synset, and state if they could be exchanged with each other to express the same semantic relation. In this case both the enhanced and the original version of WiSeNet achieved comparable results, with 85% and 82.1% accuracy, respectively;
- ^ Level 4 (disambiguated relation instances) , where, similarly to the first evaluation, the judges were asked to examine individual relation instances. However, in this case, all the synonymous relation phrases of the disambiguated relation synset were shown together with the Wikipages associated with the left and right arguments. In this setting, the accuracy decreased to 88.6% (for the enhanced version) and 76.7% (for the original version).

Overall, WiSeNet 's experimental results, together with those of Patty . demonstrate that the choice of modeling Lexical Semantics explicitly is beneficial (especially in the context of a semi-structured resource), as it boosts large-scale OIE approaches targeted to general-purpose encyclopedic text, such as Wikipedia, and enables the extraction of high-quality relation instances from these corpora; being anchored to an underlying knowledge resource, relation instances can leverage their explicit semantic characterization to overcome many limitations of traditional OIE approaches, typically caused by linguistic phenomena like synonymy and polysemy.

<sup>14</sup><https://catalog.ldc.upenn.edu/ldc2011t07>



## Chapter 4

# Harvesting Sense Annotations on a Large Scale

Take care of the sense  
and the sounds will take care of themselves.  
Lewis Carroll

This chapter tackles the first objective of this thesis, outlined in Section 1.1: that of developing robust and reliable methods to harvest sense annotations automatically on a large scale. These methods should be flexible and scalable enough, especially in terms of number of languages; at the same time, they should be capable of retaining a high annotation quality, comparable or possibly higher than previous automatic or even semi-automatic approaches.

As we discussed throughout Section 3.1, a key step towards scalability lies in using BabelNet (Section 2.1.3) as reference sense inventory. By bringing together lexicographic and encyclopedic knowledge, BabelNet enables us to annotate both named entities and concepts using a common reference inventory which not only improves the disambiguation process (in particular, it allows us to utilize joint WSD/EL approaches, as shown in Section 3.1.3.3, without forcing us to treat WSD and EL as separate annotation tasks), but also results in sense-annotated resources not dependent on an array of separate, stand-alone inventories. This is especially critical with multilingual corpora (Bentivogli and Pianta, 2005; Otegi et al., 2016), where each language relies on its specific monolingual sense inventory. The BabelNet sense inventory, instead, is inherently multilingual: beside being practical when utilizing the final resource, this feature can be leveraged at disambiguation time to enforce cross-language coherence among sense annotations, as we show in Sections 4.2 and 4.3. Furthermore, using BabelNet is also advantageous for flexibility purposes: in fact, by being a merger of all the most widely-used knowledge resources in the NLP community, BabelNet provides inter-resource mappings to most individual

sense inventories that might be used in a specific application scenario. This means that BabelNet-annotated corpora can be straightforwardly converted into WordNet-annotated corpora, or Wikipedia-annotated corpora, and vice versa<sup>1</sup>.

Leveraging BabelNet as reference knowledge resource, however, comes at a cost: the size of its encyclopedic sense inventory is prohibitively large to rely on human supervision, not even to a limited extent, as in semi-automatic approaches (Section 3.1.2): hence, fully automatic disambiguation strategies represent the only viable option. While previous work on automatically constructing BabelNet-annotated corpora, discussed in Section 3.1.3.3, has demonstrated the effectiveness of exploiting an off-the-shelf state-of-the-art WSD/EL system to disambiguate on a large scale, the proposed approaches are still suboptimal, for two main reasons:

1. Even a state-of-the-art disambiguation system like Babelfy (Section 2.2.2.3) is affected by a structural bias towards the most connected senses inside the underlying semantic network, which is typical of knowledge-based approaches (Calvo and Gelbukh, 2015), and limits the accuracy of its disambiguation output;
2. None of the proposed approaches is designed to fully exploit the structure and features of the target corpus. For instance, Wikipedia provides a semi-structured scaffolding with categories and internal hyperlinks, both providing important semantic information. Parallel corpora, on the other hand, include useful sentence-level alignments that are neglected when disambiguating each language separately.

In this chapter we address the two limitations above, both by studying how the structure of the target corpus (coupled with the features of the BabelNet sense inventory) can be exploited to improve the disambiguation process, as well as by investigating fully automatic disambiguation strategies where the structural bias discussed above can be fully or partially recovered. To this aim, we focus on three disambiguation scenarios:

- ^ Wikipedia, i.e. the most popular and widely-used semi-structured resource of encyclopedic knowledge (Section 2.1.2). Beside its central role in many NLP areas, Wikipedia provides a large-scale general-purpose textual corpus which covers a wide variety of knowledge domains, while being less noisy compared to the majority of news-based corpora. As such, Wikipedia has proven to be a convenient target corpus both for automatic sense-annotation approaches (Section 3.1.3.3) and for semantically informed OIE systems (Section 3.2). Wikipedia is hence the target of the approach presented in Section 4.1;
- ^ A large parallel text, i.e. the Europarl corpus (Koehn, 2005). Europarl is by far the most popular multilingual corpus used for Machine Translation (MT): in fact, it was originally designed to provide a large sentence-aligned training benchmark for MT systems. Over the years, it has been used widely across other NLP areas, including cross-lingual WSD (Lefever and Hoste, 2010,

<sup>1</sup>Since the BabelNet sense inventory is a superset of, e.g., the WordNet sense inventory, mapping a set of sense annotations from the former to the latter might of course reduce the number of valid annotations, as there might be a word sense or named entity mentions not covered by WordNet. With the reverse procedure, instead, all sense annotations are always retained.

2013), and also as source of sense annotations (Otegi et al., 2016). Europarl is our case study for the approach presented in Section 4.2;

- ^ A corpus of definitional knowledge, i.e. the whole set of textual definitions drawn from BabelNet in all the available languages. In fact, definitional knowledge constitutes not only a convenient target for harvesting sense annotations (as shown in Section 3.1.2.1), but also a fundamental resource for many sense-level approaches (Baldwin et al., 2008; Pilehvar et al., 2013; Camacho Collados et al., 2015b). Furthermore, it brings together features of the first scenario (less noisy text of encyclopedic nature) and of the second scenario (multilinguality and sentence-level alignments), with the added difficulty of having shorter sentences, and hence less context for disambiguation (cf. Section 4.3.4.2), which balances the well-formed nature of definitional text. In Section 4.3 we target this corpus using the same approach of Section 4.2 adapted to definitional knowledge;

In each scenario we adopt a similar methodological approach: we investigate a disambiguation strategy suitable for the target corpus, and then we apply it to produce a sense-annotated resource, which is then publicly released to the community.

For each resource we carry out an extensive experimental evaluation, comprising both intrinsic experiments (typically based on manual assessment over a random sample of sense annotations) and extrinsic experiments (where we use our sense annotations as training or development data for a variety of downstream NLP tasks). In each specific scenario we compare our sense-annotation strategy, both intrinsically and extrinsically, with the closest automatic or semi-automatic approaches (some of which have been treated already in Section 3.1).

## 4.1 Sew: A Semantically Enriched Wikipedia

As discussed extensively through Chapters 1 and 2, Wikipedia represents an extraordinary source of semantic information for innumerable tasks in NLP. In particular, the internal hyperlinks spread out across the textual content of over 4 million Wikipages constitute Wikipedia's fundamental backbone: on one hand, hyperlinks work as semantic connections between the entities described by the source and target Wikipages, framing the whole Wikipedia as a large-scale lexicalized semantic network; on the other, they also provide several million sense annotations of Wikipedia entities in context. Both aspects of Wikipedia have been extensively exploited in many NLP areas (Section 2.1.2) and, in particular, hyperlinks have also played an important role in the automatic construction of sense-annotated corpora (Section 3.1.3.2) and in the development of semantically informed OIE approaches (Section 3.2.2).

Unfortunately, if we consider Wikipedia as it is, the sense-level information available as sense-annotated textual mentions is partial and incomplete, since only a fraction of linkable mentions in Wikipedia are in fact hyperlinked: out of over 580 million noun lemmas across the whole corpus<sup>3</sup> those covered by internal hyperlinks amount to 116 millions (19%). Hyperlink sparseness is partly intentional: the

<sup>2</sup>We detail the structure and format of each released resource in Chapter 6.

<sup>3</sup>Estimated from the Wikipedia dump of November 2014 (Raganato et al., 2016b).

Wikipedia style guidelines suggest to link each concept at most once within a page, and only when it is relevant and helpful in the context<sup>4</sup>. While this is advisable from the perspective of human readers, as too much hyperlinked text would make Wikipages less readable, it also prevents a lot of basic concepts and entities to be modeled within the Wikipedia structure.

Being able to link and disambiguate appropriately every linkable mention across Wikipedia would be a major step for bridging this gap and turning Wikipedia into a full-edged sense-annotated corpus. In the NLP community, this task of automatic identification and linking of referenced Wikipedia entities across text is commonly referred to as *Wikification*, and it has been addressed in various ways (cf. Section 2.2.2). One of the key challenges of *Wikification* lies in resolving mention ambiguity: in Section 3.1.3.3, indeed, we examined some approaches based on off-the-shelf WSD/EL systems with a Wikipedia-based (or BabelNet-based) sense inventory that have been used to this purpose (Hahm et al., 2014; Scozzafava et al., 2015). However, these systems are designed for general text and, although enriching Wikipedia can be seen as the special case of 'wikifying' Wikipedia articles, a general-text system does not take full advantage of the existing Wikipedia structure.

In light of all this, our objective in the present section is that of augmenting Wikipedia with as much sense-level information as possible, by recovering potentially linkable mentions to concepts or named entities that are not covered by original hyperlinks. Although a few previous approaches have addressed this specific task of detecting and annotating potentially linkable mentions in Wikipedia, mainly using *Wikification* (West et al., 2015) or classifiers with Wikipedia-specific features (Noraset et al., 2014), none of these strategies fit our needs: in fact, we aim at a fully automatic and self-contained approach, without employing human intervention or overly tuned supervised systems. Also, we aim at covering as many mentions as possible across the corpus, whereas Noraset et al. (2014) enforce a high-precision setting, and West et al. (2015) focus only on hyperlinks that increase Wikipedia navigability.

Instead, in marked contrast with previous approaches, we rely solely on the structure of Wikipedia itself, with no off-the-shelf disambiguation system. Specifically, we exploit direct connections among Wikipedia articles and Wikipedia categories to propagate hyperlink information across the corpus. Importantly, as we stated at the beginning of the chapter, we use BabelNet as reference sense inventory, and we leverage the BabelNet semantic network (Section 2.1.3) to connect pages across Wikipedias in different languages, as well as across different lexicographic and encyclopedic resources. As a result of our hyperlink propagation pipeline, we obtain a *Semantically Enriched Wikipedia*, or *Sew* (Raganato et al., 2016b)<sup>5</sup>, i.e. a large-scale Wikipedia-based sense-annotated corpus with more than 200 million sense annotations of over 4 million different concepts and named entities drawn from BabelNet. *Sew* covers almost 40% of the nouns in Wikipedia (compared to less than 20% covered by original hyperlinks) and also includes verbs, adjectives and adverbs.

To the best of our knowledge, *Sew* constitutes today the largest resource available comprising word senses and named entity mentions together, annotated using the

<sup>4</sup>[https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style#Links](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style#Links)

<sup>5</sup><http://lcl.uniroma1.it/sew>

uni ed sense inventory of BabelNet. This makes it a suitable dataset for various di erent tasks, e.g. Entity Linking and Semantic Similarity, that usually require dedicated training sets. Throughout the following sections, we proceed as follows: we rst give the details of our disambiguation strategy and examine every stage of our hyperlink propagation pipeline (Section 4.1.1); we then look at some global statistics of the corpus in Section 4.1.2, showing that Sew stands as a prominent sense-annotated resource not only in terms of size (i.e. number of sense annotations and coverage) but also in terms of scope (i.e. lexicographic and encyclopedic knowledge from a wide-coverage sense inventory); in our experimental evaluation of Section 4.1.3, instead, we assess the quality of Sew's annotations intrinsically (on a test set of hand-labeled hyperlinks) and extrinsically in two ways:

- ^ Using Sew as a training set for EL with IMS (Zhong and Ng, 2010), an open-source supervised WSD system, showing that it leads to performances in line with the state of the art on standard benchmarks (Section 4.1.3.2);
- ^ Leveraging propagated hyperlinks to generate two simple, yet e ective, Wikipedia-based language-independent vector representations that achieve competitive results on semantic similarity and sense clustering (Section 4.1.3.3).

Finally, we expand the latter experiment with a broader evaluation study on semantic similarity (Section 4.1.4), where we investigate an embedded augmentation of Sew's original vector representations, Sew-Embed (Delli Bovi and Raganato, 2017).

#### 4.1.1 The Hyperlink Propagation Pipeline

Sew's hyperlink propagation pipeline takes as input a Wikipedia dump and outputs a sense-annotated corpus, built upon the original textual content of Wikipedia, where word senses and named entity mentions are linked to the sense inventory of BabelNet. Some standard preprocessing is applied to the input corpus in the rst place, including tokenization, part-of-speech tagging and lemmatization. At this preliminary stage we also discard disambiguation pages, List of ' articles and pages of common surnames<sup>6</sup>, as they typically contain few lines of meaningful text and tend to introduce noise in the propagation process. After preprocessing, we apply a cascade of hyperlink propagation heuristics to each Wikipage in the input corpus. Each propagation heuristic, when applied, identifies a list of BabelNet synsets  $S^p$  to be propagated across a given Wikipage  $p$ ; then, for each synset  $s \in S^p$ , occurrences of any lexicalization of  $s$  are detected, annotated with  $s$ , and added as new hyperlinks for  $p$ .<sup>7</sup> All propagation heuristics share a common assumption: given an ambiguous mention  $m$  within a Wikipage  $p$ , every occurrence of  $m$  across  $p$  refers to the same sense of  $m$  (one-sense-per-page assumption) and hence it can be annotated using the same synset. This assumption is a Wikipedia-specific version of the one-sense-per-discourse assumption (Yarowsky, 1995) and, albeit simple, tends

<sup>6</sup>[https://en.wikipedia.org/wiki/Lists\\_of\\_most\\_common\\_surnames](https://en.wikipedia.org/wiki/Lists_of_most_common_surnames)

<sup>7</sup>Thanks to BabelNet's inter-resource mappings, each hyperlinked Wikipage can be unambiguously mapped to the corresponding Babel synset, and vice versa. Thus, in the present section we use the terms 'propagated hyperlink' and 'sense annotation' interchangeably.

	Symbol	Type	Scope
Original Hyperlink	HL	-	Wikipedia
Surface Mention Propagation	SP	Intra-page	Wikipedia
Lemmatized Mention Propagation	LP	Intra-page	Wikipedia
Person Mention Propagation	PP	Intra-page	Wikipedia
Wikipedia Inlink Propagation	WIL	Inter-page	Wikipedia
BabelNet Inlink Propagation	BIL	Inter-page	BabelNet
Category Propagation	CP	Inter-page	Wikipedia
Monosemous Content Word	MP	-	BabelNet

Table 4.1. Summary of the hyperlink propagation heuristics used in Sew.

to be surprisingly accurate given the nature and structure of Wikipedia.<sup>8</sup>

As we apply a heuristic  $h$  to a given Wikipage  $p$ , we characterize  $h$  as being either intra-page (when it propagates synsets that already occur as hyperlinks within  $p$  itself) or inter-page (when it exploits the connection of  $p$  with other Wikipages or categories). Also, we refer to the scope of  $h$  as either Wikipedia (when all synsets propagated by  $h$  identify a specific Wikipedia page) or BabelNet (when  $h$  propagates synsets that may not have an associated Wikipedia page).

After all heuristics have been applied we enforce a conservative policy to remove overlapping mentions and duplicates (i.e. multiple sense annotations associated with the exact same fragment of text). We deal with overlaps by penalizing inter-page annotations in favor of intra-page ones, and by preferring the longest match in case of overlapping annotations of the same type. Similarly, we deal with duplicates by preferring intra-page annotations over inter-page ones, consistently with the one-sense-per-page assumption. Finally, if the mention is still ambiguous, all its sense annotations are discarded. All the propagation heuristics composing the pipeline of Sew are summarized in Table 4.1. Most of them are based on methods that proved to be robust and effective in previous works for a variety of different purposes: a one-sense-per-page assumption is used by Wu and Giles (2015) to develop sense-aware Wikipedia-based word representations; Wikipedia categories have been exploited for propagating semantic relations (Nastase and Strube, 2008), learning topic hierarchies (Hu et al., 2015) and building taxonomies (Flati et al., 2014); finally, ingoing links to Wikipedia pages played a key role in the semantic representations of Nasari (Section 2.2.3.3).

#### 4.1.1.1 Intra-page Propagation Heuristics

Intra-page propagation heuristics collect a list of synsets  $S^p$  from the original hyperlinks occurring in Wikipage  $p$  (including the synset associated with  $p$  itself) and then propagate  $S^p$  by looking for potential mentions matching any lexicalization of a synset in  $S^p$ . Every mention discovered this way is then added to the list of propagated hyperlinks for  $p$  if part-of-speech tags are consistent. However,

<sup>8</sup>98% of the Wikipedia pages support the one-sense-per-page assumption, according to the estimation of Wu and Giles (2015).

as potential mentions may contain punctuation or occur in some inflected form, propagation is performed as a two-pass procedure: surface mention propagation (SP) over the original text of  $p$  before preprocessing, and lemmatized mention propagation (LP) over tokenized and lemmatized text.<sup>9</sup>

Moreover, we designed a specific heuristic to propagate person mentions (PP). This heuristic can be seen as a specialized version of coreference resolution restricted to person entities: if a synset  $s \in S^p$  identifies a person according to the BabelNet entity typing, we allow potential mentions to match lexicalizations of  $s$  partially (i.e. only first name, or only last name). Each partial mention is then validated by checking its surrounding word tokens against a pre-computed set of first and last names, drawn from Wikipedia itself,<sup>10</sup> and added as sense annotation only if surrounding tokens do not match any person name. This prevents us from annotating false positives (e.g. siblings of the person identified by  $s$ ).

#### 4.1.1.2 Inter-page Propagation Heuristics

Inter-page heuristics exploit the connections of  $p$  inside Wikipedia and BabelNet. Once synsets to be propagated are collected in  $S^p$ , we apply the same propagation procedure described in the previous section for intra-page heuristics. We exploited three inter-page heuristics:

- ^ The Wikipedia Inlink Propagation (WIL) heuristic collects ingoing links to  $p$  inside Wikipedia, that is other Wikipages where  $p$  is mentioned and hyperlinked, and adds the corresponding BabelNet synsets to  $S^p$ ;
- ^ The BabelNet Inlink Propagation (BIL) heuristic, similarly to WIL, leverages ingoing links to the synsets  $s_p$  that identifies  $p$  in the BabelNet semantic network. These might include, in particular, hyperlinks inside Wikipedias in languages other than English, as well as connections of  $s_p$  drawn from other resources integrated in BabelNet (cf. Section 2.1.3);
- ^ The Category Propagation (CP) heuristic propagates hyperlinks across Wikipages that belong to the same Wikipedia categories of  $p$ . Intuitively, Wikipages belonging to the same categories tend to mention the same entities. This heuristic is based on three successive steps:

1. Given a Wikipedia category  $c$ , CP harvests all hyperlinks appearing in all Wikipages associated with  $c$  at least twice, collects them into the set the set  $S^c$ , and then ranks them by frequency count;
2. In order to filter out categories that are too broad or uninformative (e.g. Living people) CP associates with each category  $c$  a probability distribution over hyperlinks  $f^c$ , and computes the entropy  $H(c)$  of such distribution as:

$$H(c) = - \sum_{h \in S^c} f^c(h) \log_2 f^c(h) \quad (4.1)$$

<sup>9</sup>A common example is the mention  $m = \text{'United States of America'}$  since only shallow preprocessing is applied to the input text (and, in particular, no NER) a lemmatization step would reduce  $m$  to  $\text{'unite state of America'}$ , which is not a valid lexicalization of the corresponding Babel synset. Similar observations apply for song, book or movie titles.

<sup>10</sup>[https://en.wikipedia.org/wiki/List\\_of\\_most\\_popular\\_given\\_names](https://en.wikipedia.org/wiki/List_of_most_popular_given_names)

	# Annotations	# Senses	# Documents	Sense Inventory
Wikipedia	71,457,658	2,898,503	4 313,373	Wikipedia
Sew (all)	250,325,257	4,098,049	4 313,373	BabelNet
Sew	206,475,360	4,071,902	4 313,373	BabelNet
WordNet	116,079,163	67,774	4 313,373	WordNet
Wikipedia	162,614,753	4,020,979	4 313,373	Wikipedia
Wikilinks	40,323,863	2,933,659	10,893,248	Wikipedia
FACC1	11,240,817,829	5,114,077	1,104,053,884	Freebase
OMSTI	1,357,922	31,956	62,815	WordNet
MASC	286,416	23,175	392	BabelNet

Table 4.2. Global statistics of Sew in comparison with other sense-annotated corpora. 'Wikipedia' (rst row) refers to the English dump of November 2014, while Sew (all)' (second row) refers to the corpus before applying the conservative policy.

where  $f^c(h)$  is computed as the normalized frequency count of  $h$  in  $S^c$ . Ranking categories by their entropy values allows to discriminate between broader categories, where a large number of less related hyperlinks appear with relatively small counts (hence higher  $H$ ), and more specific categories, where fewer related hyperlinks occur with relatively higher counts (and lower  $H$ );

3. Finally, given a Wikipedia  $p$ , CP considers each category  $c_p$  associated with  $p$  where  $H(c_p)$  is below a predefined threshold  $H_{th}$ <sup>11</sup> and adds to  $S^p$  all the synsets that identify hyperlinks in  $S^{c_p}$ .

In the last stage of the pipeline, after both intra-page and inter-page heuristics have been applied, we additionally exploit a Monosemous Content Word (MP) heuristic to propagate verb, adjective and adverb senses that are monosemous according to the sense inventory.

#### 4.1.2 Statistics

The experimental setup described in Raganato et al. (2016b) includes the English Wikipedia dump of November 2014 as input corpus, and the Stanford CoreNLP pipeline<sup>12</sup> for preprocessing. Table 4.2 reports some global statistics: compared to the original Wikipedia, Sew achieves 3.5 times the amount of annotations (58.03 average annotations per page against 16.57 of the original Wikipedia) and adds 1,199,546 new concepts and entities not covered by the original hyperlinks. 17.5% ambiguous annotations are removed by the conservative policy, but the overall coverage of senses remains almost unchanged. Table 4.2 also includes two reduced versions of Sew with only Wikipedia (fth row) or WordNet (fourth row) as sense inventories, respectively. The bottom rows of Table 4.2 report global statistics on other sense-tagged corpora mentioned in Section 3.1: Wikilinks (Singh et al., 2012), FACC1 (Gabrilovich et al., 2013), OMSTI (Taghipour and Ng, 2015b) and

<sup>11</sup>Raganato et al. (2016b) uses a fixed  $H_{th} = 0.5$ , empirically validated on a small set of held-out Wikipages, for all the experimental evaluations.

<sup>12</sup><http://stanfordnlp.github.io/CoreNLP>

	Sew (%)	Only HL (%)
Nouns	227,326,282 (38.75%)	116,342,382 (19.83%)
Verbs	8,080,280 (6.71%)	1,799,680 (0.82%)
Adjectives	33,402,556 (27.87%)	9,913,634 (8.27%)
Adverbs	17,163,713 (33.95%)	245,468 (0.49%)
Total	285,972,831 (29.26%)	128,301,164 (13.13%)

Table 4.3. Coverage of content words by part of speech.

	HL	SP	LP	PP	WIL	BIL	CP	MP
Sew (all)	71,457,020	33,780,057	24,510,995	6,735,336	7,237,505	32,713,194	25,650,945	48,240,205
Sew	71,457,020	33,589,710	14,936,540	6,411,877	2,174,818	19,850,111	14,271,461	43,783,185

Table 4.4. Sense annotations by heuristic type. Sew (all)' (rst row) refers to the corpus before applying the conservative policy.

the sense-tagged MASC corpus (Moro et al., 2014a). Compared to Wikilinks, the Wikipedia portion of Sew adds 122M annotations and 1,087,320 covered senses. FACC1 is considerably larger than any other reported corpus and features 1.12G annotations, which are however drawn from 1.1G documents (with an average of 10.18 annotations per document) and restricted to named entities in Freebase. Finally, compared to OMSTI, the WordNet portion of Sew adds over 114M sense annotations and 35,818 covered senses.

Table 4.3 reports the coverage of Sew at the lemma level. Out of 977,203,946 lemmas in total, Sew annotates 38.75% of the nouns, 6.71% of the verbs, 27.87% of the adjectives, and 33.95% of the adverbs. In comparison, the original Wikipedia hyperlinks cover 19.83% of the nouns, 8.27% of the adjectives, and less than 1% of verbs and adverbs. Overall, Sew achieves almost 30% coverage, improving more than 16% with respect to the original Wikipedia (13.3%) and extending coverage to non-nominal content words (verbs, adverbs, adjectives). Finally, Table 4.4 shows the number of sense annotations by heuristic type. Each heuristic is identified by the corresponding symbol in Table 4.1. Apart from original hyperlinks (which provide 28.55% of the annotations) and monosemous mentions (19.27%), the SP and BIL heuristics provide 13.49% and 13.07% of annotations respectively, followed by the CP heuristic with 10.25%. The annotations discarded after applying the conservative policy are mostly derived from inter-page heuristics (WIL, BIL, CP), which open up to a broader context with respect to intra-page ones (being therefore prone to noisier propagations).

#### 4.1.3 Experimental Evaluation

We evaluated Sew with an intrinsic and an extrinsic evaluation. In the former (Section 4.1.3.1) we compared Sew's sense annotations against those discovered by 3W (Noraset et al., 2014), a Wikipedia-specific classifier designed to add automatically high-precision hyperlinks to Wikipages; in the latter we used Sew as

	Precision	Recall	F-score
Sew	0.934	0.468	0.623
Sew w/o SP	0.907	0.409	0.564
Sew w/o LP	0.914	0.456	0.608
Sew w/o PP	0.916	0.457	0.610
Sew w/o WIL	0.917	0.453	0.607
Sew w/o BIL	0.907	0.413	0.567
Sew w/o CP	0.916	0.415	0.571
Sew w/o MP	0.945	0.458	0.617
3W	0.989	0.310	0.471

Table 4.5. Performance on the hand-labeled evaluation set of Noraset et al. (2014).

a training set for Entity Linking (Section 4.1.3.2) and we exploited it to develop Wikipedia-based language-independent vector representations for semantic similarity (Section 4.1.3.3), comparing Sew against a baseline given by the original Wikipedia.

#### 4.1.3.1 Intrinsic Evaluation: Annotation Quality

We assessed the quality of Sew's sense annotations on a hand-labeled evaluation set of 2,000 randomly selected Wikipages, described by Noraset et al. (2014) and used for training, validating and testing 3W. We first ran the hyperlink propagation pipeline on those Wikipages and then, following Noraset et al. (2014), we checked the 1,530 solvable mentions against the gold standard. Results are reported in Table 4.5 and compared against 3W<sup>13</sup>. While obtaining a substantially higher recall, Sew manages to keep precision above 93% and achieves an F-score of 62.3% against 47.1% of 3W. It is also worth noting that gold standard mentions, being labeled with Wikipages, do not take parts of speech into account and hence include several adjective mentions (e.g. American, German) labeled as nouns (United States, Germany) whereas Sew annotates them with the corresponding adjectival senses (American<sub>a</sub>, German<sub>a</sub>). If we account for these cases Sew achieves 64.4% F-score, showing a precision (96.5%) comparable to a supervised system tuned for high precision, while at the same time granting a much higher coverage, with an average of 31.3 new annotations per page (Section 4.1.2) against an estimate of 7 added by 3W. It is worth noting that the purpose of this evaluation is not that of overcoming 3W (which could easily be tuned to work at a lower precision and boost its recall, cf. Figure 1 by Noraset et al. (2014)) but rather that of showing how a self-contained vanilla approach behaves against a supervised high-precision upper bound.

We used the same gold standard to perform an ablation test on our propagation heuristics: for each heuristic, we discarded the hyperlinks propagated by it and then repeated the experiment. Results (Table 4.5) show that significant contributions in terms of F-score come from both intra-page propagations (SP, +5.89%) and inter-page ones (BIL and CP, +5.2% and +5.3% respectively).

<sup>13</sup>We used the recommended setting of 3W with threshold at 0.934.

	SemEval-2013	SemEval-2015	MSNBC	AIDA-CoNLL
IMS+ Sew	0.810	0.882	0.789	0.726
IMS+HL	0.775	0.758	0.695	0.712
MFS	0.802	0.857	0.620	0.535
UMCC-DLSI	0.548	-	-	-
Babelfy	0.874	-	-	-
DFKI	-	0.889	-	-
SUDOKU	-	0.870	-	-
Wiki er	-	-	0.812	0.724
M&W	-	-	0.685	0.823

Table 4.6. Results in terms of F-score on various standard benchmarks for WSD and EL.

#### 4.1.3.2 Extrinsic Evaluation #1: Entity Linking

In the first extrinsic experiment we used Sew as Entity Linking training set for It Makes Sense (Zhong and Ng, 2010, IMS), a supervised system originally designed for all-words WSD, and based on Support Vector Machines. As a baseline, we considered IMS with the same features and parameters, but trained only on the original Wikipedia hyperlinks. Results are shown in Table 4.6 in terms of F-score: IMS+ Sew and IMS+HL represent IMS trained on Sew and its baseline, respectively. We included for each dataset a Most Frequent Sense (MFS) baseline, as well as the results reported by other state-of-the-art EL systems in the literature: Babelfy (Section 2.2.2.3) and the best performing system reported by Navigli et al. (2013) for SemEval-2013; the two best performing systems reported by Moro and Navigli (2015) for SemEval-2015; finally, Wiki er (Cheng and Roth, 2013) and Wikipedia Miner (Milne and Witten, 2008, M&W) for MSNBC and AIDA-CoNLL.

In each dataset, IMS trained on Sew consistently outperforms its baseline version, suggesting that our propagated hyperlinks lead to more accurate supervised models. Furthermore, the IMS model trained on Sew outperforms the best and second-best systems reported in the SemEval 2013 and 2015 tasks respectively, putting IMS (a WSD model based on local features, that is not even designed for EL) in line with more recent EL approaches, significantly outperformed only by systems that are specifically designed to exploit Wikipedia information (Wiki er, M&W).

#### 4.1.3.3 Extrinsic Evaluation #2: Semantic Similarity

Another interesting test-bed for Sew is provided by word similarity, where several successful approaches make explicit use of Wikipedia, such as ~~asari~~ (Section 2.2.3.3). Others, like SensEmbed (Section 2.2.3.2), report state-of-the-art results when trained on an automatically disambiguated version of Wikipedia. In order to test experimentally whether Sew constitutes a preferable starting point than the original Wikipedia, with its increased hyperlink connections (in the former case) and its increased sense-tagged mentions (in the latter case), we designed two sense-based explicit vector representations for nominal concepts and entities, built upon Sew:

- ^ A Wikipage-based representation (WB-Sew) where each synset is in the sense

		WB-Sew		SB-Sew		WB-HL		SB-HL	
		RC	LS	RC	LS	RC	LS	RC	LS
WS-Sim	r	0.65	0.64	0.50	0.57	0.58	0.58	0.53	0.52
		0.69	0.70	0.56	0.57	0.59	0.61	0.49	0.51
SL-666	r	0.38	0.38	0.26	0.34	0.32	0.32	0.28	0.31
		0.40	0.41	0.33	0.36	0.31	0.32	0.27	0.27

Table 4.7. Results on word similarity in terms of Pearson ( $r$ ) and Spearman ( $\rho$ ) correlation.

		WB-Sew		WB-HL		Word2Vec		Polyglot
		RC	LS	RC	LS	original	retro tted	
EN	r	0.673	0.674	0.619	0.614	-	-	0.51
		0.608	0.620	0.592	0.592	0.73	0.77	0.55
FR	r	0.808	0.811	0.773	0.778	-	-	0.38
		0.755	0.759	0.693	0.681	0.47	0.61	0.35
DE	r	0.639	0.639	0.584	0.580	-	-	0.18
		0.689	0.695	0.637	0.615	0.53	0.6	0.15
ES	r	0.811	0.804	0.757	0.740	-	-	0.51
		0.815	0.812	0.764	0.759	-	-	0.56

Table 4.8. Pearson ( $r$ ) and Spearman ( $\rho$ ) correlation results for multilingual word similarity on the RG-65 dataset.

inventory is represented by a vector  $v_s$  where dimensions are Wikipages. We computed, for each Wikipage  $p$ , the corresponding component of  $v_s$  as the frequency of  $s$  appearing as hyperlink in  $p$ ;

- ^ A synset-based representation (SB-Sew) where each Wikipage  $p$  is represented by a vector  $v_p$  where dimensions are Babel synsets. For each synset  $s$ , the corresponding component of  $v_p$  is computed as the frequency of  $s$  appearing as hyperlink in  $p$ .

We estimated frequencies using both raw counts (RC) and lexical specificity (LS), as in Camacho Collados et al. (2016c), and we considered the two largest word similarity benchmarks (cf. Section 2.2.3) for testing: the similarity portion of WordSim-353 (WS-Sim) and the noun subset of SimLex-999 (SL-666). In both cases we used Weighted Overlap (Pilehvar et al., 2013) as similarity measure. Table 4.7 reports the performance of WB-Sew and SB-Sew in comparison with baseline vectors (WB-HL and SB-HL) computed using only the original Wikipedia hyperlinks, and shows a consistent improvement over the baseline in both datasets. On WordSim-353, in particular, WB-Sew obtains higher correlation figures than those reported by state-of-the-art approaches like ADW (Pilehvar et al., 2013) ( $r = 0.63$  and  $\rho = 0.67$ ) and ESA (Gabrilovich and Markovitch, 2007) ( $r = 0.40$  and  $\rho = 0.47$ ). On the other hand, both approaches lag behind the state of the art on the noun portion of Simlex-999 (Camacho Collados et al., 2016c). Finally, WB-Sew consistently improves over SB-Sew in both our benchmarks, suggesting that a synset-based vector space might be affected by mappings errors across BabelNet.

	WB-Sew		SB-Sew		WB-HL		SB-HL	
	RC	LS	RC	LS	RC	LS	RC	LS
500-pair	0.67	0.67	0.71	0.67	0.67	0.65	0.23	0.19
SemEval	0.63	0.64	0.63	0.64	0.56	0.56	0.29	0.24

Table 4.9. F-score results on Wikipedia sense clustering.

Since both WB-Sew and SB-Sew are defined with respect to a multilingual sense inventory, and hence are inherently language-independent, we also tested our best performing model (WB-Sew) on multilingual word similarity. As reported in Table 4.8, WB-Sew consistently beats the baseline and shows a considerable improvement on French, German and Spanish over Word2Vec (Mikolov et al., 2013a), both the original model<sup>14</sup> and the model retrofitted into WordNet (Faruqui et al., 2015), as well as over the language-specific pre-trained embedding models from the Polyglot project (Al-Rfou' et al., 2013). It is worth noting that the availability of training data is a crucial factor in the evaluation of the data-driven models in Table 4.8: this is why Word2Vec still manages to keep the lead on English, while WB-Sew cannot take direct advantage of the increased underlying data.

Finally, we tested WB-Sew and SB-Sew on the Wikipedia sense clustering task (Dandala et al., 2013), considering both benchmark datasets (500-pair and SemEval). We followed the same clustering approach proposed by Camacho Collados et al. (2016c), with empirically validated thresholds of 0:1 (WB-Sew) and 0:5 (SB-Sew). F-score results reported in Table 4.9 are consistent with the experiment on word similarity (Table 4.7) and show that both vector representations improve consistently over their baseline counterparts, with F-scores close to the state of the art reported by Nasari (72% on 500-pair and 64.2% on SemEval).

#### 4.1.4 A Broader Evaluation Study: Sew-Embed

Both WB-Sew and SB-Sew, developed for the second extrinsic evaluation (Section 4.1.3.3), consist of high-dimensional sparse vectors, not immediately comparable with many existing approaches, especially those based on word embeddings, and less flexible to use within downstream applications. This is why we broadened the scope of the experiment by participating in the SemEval 2017 task 2 on multilingual and cross-lingual word similarity (Camacho Collados et al., 2017), where we studied an alternative low-dimensional representation based on Sew. Specifically, we considered WB-Sew, and designed an embedded augmentation of its explicit high-dimensional vectors, obtained by plugging in an arbitrary word (or sense) embedding model, and computing a weighted average in the continuous vector space. Regardless of the particular model used, the resulting vector representation, Sew-Embed (Delli Bovi and Raganato, 2017), is still defined at the concept level, and hence immediately expendable in a multilingual or cross-lingual setting. The work flow of our procedure to generate Sew-Embed is depicted in Figure 4.1 with an illustrative example.

<sup>14</sup>Following Camacho Collados et al. (2016c), we consider the pre-trained Word2Vec vectors obtained from the Google News corpus (EN), and from a 1-billion-token sample of Wikipedia (DE

Figure 4.1. Illustrative example, borrowed from Delli Bovi and Raganato (2017), of Sew-Embed's embedded representation  $\mathbf{b}$ ) for the BabelNet entity Lorenzo de Medici (bn:00052034n) obtained from the corresponding explicit representation  $\mathbf{a}$ ).

**Methodology** In order to compute the embedded augmentation of an explicit WB-Sew vector  $\mathbf{v}_s$  (Figure 4.1a), we followed Camacho Collados et al. (2016c) and exploited the compositionality of word embeddings (Mikolov et al., 2013c), i.e. the fact that the representation of an arbitrary compositional phrase can be expressed as the combination (typically the average) of its constituents' representations. In particular, we considered each dimension  $p$  (i.e. Wikipage, cf. Section 4.1.1) of  $\mathbf{v}_s$  and mapped it to the embedding space  $\mathcal{E}$  provided by an external pre-trained model to obtain an embedded vector  $\mathbf{e}_p$ . The way this mapping was carried out depended on the specific external model utilized:

- In case of a word embedding model we considered the Wikipage title as lexicalization of  $p$ , and then retrieved the associated pre-trained embedding. If the title is a multi-word expression, and no embedding is available for the whole expression, we exploited compositionality again and averaged the embedding vectors of its individual tokens;
- In case of a sense or concept embedding model we instead exploited BabelNet's inter-resource mappings, and mapped  $p$  to the target sense inventory of  $\mathcal{E}$ , for which the corresponding embedding vector could be retrieved.

The embedded representation  $\mathbf{e}_s$  of  $s$  (Figure 4.1b) was then computed as the weighted average over all the embedded vectors  $\mathbf{e}_p$  associated with the dimensions of  $\mathbf{v}_s$ :

$$\mathbf{e}_s = \frac{\sum_p \mathbf{v}_s(p) \mathbf{e}_p}{\sum_p \mathbf{v}_s(p)} \quad (4.2)$$

where  $\mathbf{v}_s(p)$  is the lexical specificity weight of dimension  $p$ . In contrast to a simple average, in (4.2) we exploited the ranking of each dimension  $p$  (represented by  $\mathbf{v}_s(p)$ ) and hence gave more importance to the higher weighted dimensions of  $\mathbf{v}_s$ .

**Experimental Setup** In our experimental setup, i.e. the monolingual and cross-lingual benchmark of the Semeval 2017 Task 2 (Camacho Collados et al., 2017), we

---

and FR).

	EN			FA			DE			IT			ES		
	r	Mean		r	Mean		r	Mean		r	Mean		r	Mean	
Sew-Embed <sub>w2v</sub>	0.56	0.58	0.57	0.38	0.40	0.39	0.45	0.45	0.45	0.57	0.57	0.57	0.61	0.62	0.62
Sew-Embed <sub>Nasari</sub>	0.57	0.61	0.59	0.30	0.40	0.34	0.38	0.45	0.42	0.56	0.62	0.59	0.59	0.64	0.62
WB-Sew	0.61	0.67	0.64	0.51	0.56	0.53	0.51	0.53	0.52	0.63	0.70	0.66	0.60	0.66	0.63
Nasari	0.68	0.68	0.68	0.41	0.40	0.41	0.51	0.51	0.51	0.60	0.59	0.60	0.60	0.60	0.60

Table 4.10. Results on the multilingual word similarity benchmarks (subtask 1) of Semeval 2017 task 2, in terms of Pearson correlation  $r()$ , Spearman correlation ( $\rho$ ), and the harmonic mean of  $r$  and  $\rho$ .

considered two versions of Sew-Embed: one based on the pre-trained English word embeddings of Word2Vec<sup>15</sup> used as comparison system in Section 4.1.3.3 (Sew-Embed<sub>w2v</sub>), and another one based on the embedded concept vectors of Nasari (Sew-Embed<sub>Nasari</sub>). Both versions relied on a back-off similarity value of 0.5 (i.e. the middle point in the similarity scale) when no candidate sense is found for either one of the two target words. In both benchmarks we compared Sew-Embed against the explicit vectors of WB-Sew and by Nasari.<sup>16</sup>

**Evaluation** Table 4.10 shows the overall performance on multilingual word similarity for each monolingual dataset. Both Sew-Embed<sub>w2v</sub> and Sew-Embed<sub>Nasari</sub> show correlation figures in the same ballpark as the Nasari baseline for Italian, Farsi, and Spanish; instead, they lag behind in English and German. Most surprisingly, however, the explicit representations based on Sew reach the best result overall in 4 out of 5 benchmarks: this might suggest that many word pairs across the test sets are actually being associated with synsets that are well connected in Sew, and hence the corresponding sparse vectors are representative enough to provide meaningful comparisons. In general, the performance decrease on German and Farsi for all comparison systems is connected to the lack of coverage: both Sew and Sew-Embed use the back-off strategy 70 times for Farsi (14%) and 54 times (10.8%) for German.

Table 4.11 reports the overall performance on cross-lingual word similarity for each language pair. All approaches based on Sew seem to perform globally better in a cross-lingual setting: on average, the harmonic mean of  $r$  and  $\rho$  is 2.2 points below the Nasari baseline (compared to 3.2 points in Table 4.10). This suggests the potential of Wikipedia as a bridge to multilinguality: in fact, even though Sew was constructed automatically on the English Wikipedia, semantic information transfers rather well via inter-language links and has a considerable impact on the cross-lingual performance. Again, the best figures are consistently achieved by WB-Sew: the improvement in terms of harmonic mean of  $r$  and  $\rho$  is especially notable in benchmarks that include a less-resourced language such as Farsi (+11.75% on average compared to the Nasari baseline). This improvement does not occur with Sew-Embed, since in that case sparse vectors are eventually mapped to an embedding space trained specifically on an English corpus.

Overall, Sew-Embed reached the 4th and 3rd positions in the global rankings

<sup>15</sup> <https://code.google.com/archive/p/word2vec>

<sup>16</sup> For an extensive comparison including all participating systems in the task, the interested reader is referred to the task description paper (Camacho Collados et al., 2017).

	DE-ES			DE-FA			DE-IT			EN-DE			EN-ES		
	r	Mean		r	Mean		r	Mean		r	Mean		r	Mean	
Sew-Embed <sub>w2v</sub>	0.52	0.54	0.53	0.42	0.44	0.43	0.52	0.52	0.52	0.50	0.53	0.51	0.59	0.60	0.59
Sew-Embed <sub>Nasari</sub>	0.47	0.55	0.51	0.35	0.45	0.39	0.47	0.55	0.51	0.46	0.55	0.50	0.59	0.63	0.61
WB-Sew	0.57	0.61	0.59	0.53	0.58	0.56	0.59	0.64	0.61	0.58	0.62	0.60	0.61	0.63	0.61
Nasari	0.55	0.55	0.55	0.46	0.45	0.46	0.56	0.56	0.56	0.60	0.59	0.60	0.64	0.63	0.63

  

	EN-FA			EN-IT			ES-FA			ES-IT			IT-FA		
	r	Mean		r	Mean		r	Mean		r	Mean		r	Mean	
Sew-Embed <sub>w2v</sub>	0.46	0.49	0.48	0.58	0.60	0.59	0.50	0.53	0.52	0.59	0.60	0.60	0.48	0.50	0.49
Sew-Embed <sub>Nasari</sub>	0.41	0.52	0.46	0.59	0.65	0.62	0.44	0.54	0.48	0.58	0.64	0.61	0.42	0.52	0.47
WB-Sew	0.58	0.63	0.61	0.64	0.71	0.68	0.59	0.65	0.62	0.63	0.70	0.66	0.59	0.65	0.62
Nasari	0.52	0.49	0.51	0.65	0.65	0.65	0.49	0.47	0.48	0.60	0.59	0.60	0.50	0.48	0.49

Table 4.11. Results on the cross-lingual word similarity benchmarks (subtask 2) of Semeval 2017 task 2, in terms of Pearson correlation ( $r$ ), Spearman correlation ( $s$ ), and the harmonic mean of  $r$  and  $s$ .

of subtask 1 and 2 respectively (with scores 0.552 and 0.558, not including the Nasari baseline). Thus, perhaps surprisingly, the embedded augmentation yielded a considerable decrease in terms of global performance in both subtasks, where the original explicit representations of WB-Sew achieved a global score of 0.615 in subtask 1, and a global score of 0.63 in subtask 2<sup>17</sup>. Intuitively, multiple factors might have influenced this negative result:

- ^ Dimensionality reduction : converting an explicit vector (with around 4 million dimensions) into a latent vector of a few hundred dimensions leads inevitably to losing some valuable information, and hence to a decrease in the representational power of the model. Such a phenomenon was also shown by Camacho Collados et al. (2016c), where the lexical and unified representations of Nasari tend to outperform the embedded representation on several word similarity and sense clustering benchmarks;
- ^ Lexical ambiguity: while the original concept vectors of Sew are defined in the unambiguous semantic space of Wikipedia pages, we constructed their embedded counterparts via the word-level representations of their lexicalized dimensions; hence, when moving to the word level, Sew-Embed conflates the different meanings of an ambiguous word or expression;
- ^ Non-compositionality . the compositional properties of word embeddings fall short in many cases, such as idiomatic expressions or named entity mentions (e.g. Wall Street or New York). The explicit vectors of Sew, instead, do not require the compositional assumption and always consider a multi-word expression as a whole.

Apart from the points above, multiple other factors (e.g. design choices, hyperparameters) should be taken into account when dealing with embedded representations, as they can greatly influence their performances on distributional similarity (Levy

<sup>17</sup>The global score is computed as the average harmonic mean of Pearson and Spearman correlation on the best four (subtask 1) and six (subtask 2) individual benchmarks (Camacho Collados et al., 2017).

et al., 2015a). Either way, even though the embedded representations of Sew did not match up to the accuracy of explicit ones on experimental benchmarks, they still constitute a convenient alternative in terms of compactness and flexibility (thanks to their reduced dimensionality), and also in terms of comparability, as they are defined in the same vector space of many popular Word2Vec-based representations.

**Final Remarks.** From the comprehensive experimental evaluation we carried out for Sew, two important points emerge: (1) when resources like Wikipedia are leveraged to harvest sense annotations, the semi-structured knowledge they already encode, either implicitly or explicitly, is extremely valuable to the extent that it can substitute off-the-shelf disambiguation systems when cleverly used; (2) The availability of large amounts of sense level information can greatly boost both performance and flexibility, enabling vanilla approaches, like those in Sections 4.1.3.3 and 4.1.4, to compete with more sophisticated state-of-the-art systems. These vanilla approaches are not meant to overcome full-edged models, of course, but to show how using Sew we can set robust performance baselines for multiple tasks and datasets, from Entity Linking to Word Similarity.

## 4.2 EuroSense: Sense Annotations from Parallel Text

With the automatic construction and evaluation of a high-quality Wikipedia corpus, i.e. Sew, Section 4.1 demonstrated how a disambiguation strategy can greatly take advantage of semi-structured knowledge encoded in the target corpus. However, despite the popularity and wide use of Wikipedia, such a semi-structured corpus represent a special, isolated case. In our second disambiguation scenario, we shift to a different setting: a parallel corpus, i.e. a corpus available in multiple languages, where the various translations of its textual content are aligned pairwise at the sentence level (text). Although explicit semantic information is in this case absent from the corpus structure, solely composed of unstructured text, parallel corpora have a key feature: manually established sentence alignments, by means of which equivalent language-specific translations are related. This is why, apart from its prominent role in MT, parallel data have been exploited widely across the NLP community to, e.g., perform cross-lingual WSD (Lefever and Hoste, 2010, 2013; Gonen and Goldberg, 2016), develop cross-lingual word embeddings (Hermann and Blunsom, 2014; Gouws et al., 2015; Coulmance et al., 2015; Vyas and Carpuat, 2016; Vulić and Korhonen, 2016; Artetxe et al., 2016) and multi-sense embeddings (Ettinger et al., 2016; Huster et al., 2016), and also harvest sense annotations (Section 3.1).

Given their extensive use across various NLP areas, parallel corpora exist in many flavors, covering multiple topics and comprising textual content of different natures (Tiedemann, 2012; Steinberger et al., 2014; Lison and Tiedemann, 2016). As stated at the beginning of the present chapter, here we focus on Europarl (Koehn, 2005)<sup>18</sup>, one of the largest and most popular resources, as well as a reference training dataset in the area of MT. Extracted from the proceedings of the European Parliament, the latest release of the Europarl corpus comprises parallel text for 21 European languages, with more than 743 million tokens overall.

<sup>18</sup><http://opus.lingfil.uu.se/Europarl.php>

Consistently with the key objective of this chapter, our aim is to augment Europarl with sense-level information for multiple languages, thereby constructing a large-scale sense-annotated multilingual corpus that would constitute a valuable resource for both WSD and MT. However, in marked contrast with previous cross-lingual disambiguation approaches (cf. Sections 3.1.2.2 and 3.1.3.1), we do not rely on pre-computed word alignments against a pivot language, as that would require us to integrate an additional external module into the pipeline, with the consequent increase of preprocessing errors propagating and affecting the disambiguation process (Taghipour and Ng, 2015b). Instead, we consider all available languages at the same time in a joint disambiguation procedure, that is subsequently refined using distributional similarity. This disambiguation strategy is substantially different from that of Section 4.1: in this case we do not have semi-structured semantic information at our disposal, and a cascade of simple propagation heuristics would not be sufficient to disambiguate with reasonably high quality. This is why, in this scenario, we follow previous approaches (Moro et al., 2014a; Scozzafava et al., 2015) and exploit Babelify to harvest as many sense annotations as possible. The way we integrate Babelify into our disambiguation pipeline, however, differs from previous work in two important respects:

- ^ We leverage parallel data to implicitly enforce cross-lingual semantic coherence throughout the disambiguation process (Section 4.2.1). Crucially, this is made possible by the multilingual sense inventory of BabelNet, where synsets are lexicalized in multiple languages;
- ^ We design a refinement procedure, based on distributional semantic similarity, in order to contrast the structural bias of Babelify towards the MFS (Section 4.2.2). This refinement step increases the accuracy of sense annotations at the expense of a reduced coverage, since sense annotations that are less semantically related with the global semantics of a target sentence are discarded;

By applying the disambiguation pipeline described above to the Europarl corpus, we obtain as a result EuroSense (Delli Bovi et al., 2017),<sup>19</sup> a multilingual sense-annotated corpus with almost 123 million sense annotations of more than 155 thousand distinct concepts and named entities drawn from the multilingual sense inventory of BabelNet, and covering all the 21 languages of the Europarl corpus.

Our methodological approach is analogous to that of Section 4.1. We first detail the two stages of EuroSense's disambiguation pipeline in Sections 4.2.1 and 4.2.2; as output of the former we obtain a first, high-coverage variant of the corpus, while the latter generates the final, refined version of EuroSense, more suitable for high precision applications. Then, in Section 4.2.3, we look at some global statistics about the corpus, and finally Section 4.2.4 presents its experimental evaluation: as with the previous disambiguation scenario, in this case we also evaluate EuroSense intrinsically (with a manual assessment on a randomly extracted sample of sentences) and extrinsically (as training set for all-words Word Sense Disambiguation).

<sup>19</sup><http://lcl.uniroma1.it/eurosense>

Figure 4.2. Illustrative example of EuroSense 's disambiguation strategy on a target set of aligned sentences.

#### 4.2.1 Stage 1: High-Coverage Joint Multilingual Disambiguation

The objective of this first stage is to obtain an intermediate high-coverage version of EuroSense , where we harvest as many sense annotations as possible by using Babelfy on all the available translations of the Europarl corpus. The resulting sense annotations, which we use as input for the subsequent stage of the disambiguation pipeline (Section 4.2.2), are also publicly released, together with the final version of EuroSense :<sup>20</sup> in fact, in a number of downstream high-recall applications, such as general-purpose Open Information Extraction (cf. Section 3.2), covering a large number of word senses and named entity mentions could be a key requirement.

**Gathering Multilingual Text.** As a preprocessing step, we part-of-speech tag and lemmatize each monolingual version of Europarl using TreeTagger (Schmid, 1995)<sup>21</sup>. At both stages of the pipeline, we aim at performing disambiguation at the sentence level. However, instead of considering each sentence in isolation, language by language, we first identify all available translations of a given sentence and then gather these together into a single multilingual text. To this aim, we utilize Europarl's sentence-aligned bitexts, relying on English as pivot language: our incremental procedure considers each bitext and, whenever two sentences of different languages are associated with the same English translation, they are put together and aligned. As a result, we reshape the Europarl corpus and turn it into a single multilingual text, where each English sentence is directly aligned to all its available translations.

<sup>20</sup>We detail structure and format of all the released data in Section 6.

<sup>21</sup>Pre-trained TreeTagger models are released for a wide variety of languages, and cover already X of the 21 languages of Europarl. We instead rely on the internal preprocessing pipeline of Babelfy for those languages not supported by TreeTagger.

**Joint Disambiguation.** We then disambiguate this multilingual text jointly using Babelify. Our underlying idea is based on the fact that knowledge-based disambiguation systems like Babelify work better with richer context, even when they use no supervision: at disambiguation time, Babelify considers the content words across the target text in order to construct an associated semantic graph, whose richness in terms of nodes and edges depends strictly on the number of content words (cf. Section 2.2.2.3). Thus, given that Babelify is capable of handling text with multiple languages at the same time, this multilingual extension effectively increases the amount of context for each sentence, and directly helps in dealing with highly ambiguous words in any particular language, as the translations of these words may be less ambiguous in some different language. Moreover, given the multilingual nature of our sense inventory, Babelify's approach based on semantic coherence favors naturally sense assignments that are consistent across languages (i.e. those having fewer distinct senses shared by more translations of the same sentence<sup>22</sup>).

This process is depicted in the illustrative example of Figure 4.2, where, for instance, the Babel synset representing the State of the Union address (bn:14473459n) occurs in the majority of sentences, with different language-specific lexicalizations (state of the Union, État de l'Union, Estado de la Unión). For those languages where a lexicalization of the State of the Union synset is not available (German and Italian in the example of Figure 4.2), Babelify disambiguates only a part of the mention, but still selects a context-relevant meaning (i.e. 'Union' as the association of Northern American states, rather than the abstract concept of 'union' as a generic collection of entities). This procedure, however, is not perfect, and Babelify's structural bias towards the Most Frequent sense might also affect sense assignments that are coherent across languages: e.g., in Figure 4.2, the synset of climate intended as the weather situation (bn:00019780n) is incorrect despite occurring in every sentence.

#### 4.2.2 Stage 2: High-Precision Similarity-Based Refinement

At this stage we aim at improving the sense annotations obtained in the previous step (Section 4.2.1). In order to get a handle on Babelify's MFS bias and improve disambiguation accuracy we adopt a refinement based on distributional similarity, which is not affected by the MFS. This refinement allows us to discard low-confidence sense annotations, and to correct the output of Babelify in a number of cases. As a result of this final stage, we obtain the refined high-precision version of EuroSense.

**Isolating Low-Confidence Disambiguations.** Let  $D$  be the set of word senses and named entity mentions connected to the corresponding Babel synset  $d$  (disambiguated instances henceforth) in a target multilingual sentence. First of all, for each disambiguated instance  $d \in D$  we compute a coherence score  $C(d)$ . The coherence score of  $d$  is given by the number of semantic connections between the synset associated with  $d$  and the synset associated with any other disambiguated instance  $i \in D$ ,

<sup>22</sup>This is due to the fact that each target content word, regardless of the language, is included in the same graph-based representation of the sentence.

normalized by the total number of disambiguated instances:

$$C(d) = \frac{|\text{Disambiguated instances connected to } d|}{|\text{Disambiguated instances}|} \quad (4.3)$$

We set a coherence score threshold to 0.125 (i.e. one semantic connection out of eight disambiguated instances) using a held-out validation set of manually annotated sentences, and identify  $L_{\text{low}} \subseteq D$  as the set of disambiguated instances below this threshold (namely the low-coherence disambiguations).<sup>23</sup>

**Similarity-Based Refinement.** In order to refine the disambiguated instances in  $L$ , we exploit the embedded vector representations of Nasari (Section 2.2.3.3), and associate an additional score (Nasari score) with all those instances in  $L$  for which a Nasari vector can be retrieved.<sup>24</sup> First, we calculate the centroid  $\mu$  of all the Nasari vectors associated with the disambiguation instances in  $H = D \setminus L$  (i.e. the high-coherence disambiguations):

$$\mu = \frac{\sum_{d \in H} \mathbf{d}}{|H|} \quad (4.4)$$

where  $\mathbf{d}$  is the Nasari vector associated with a disambiguated instance  $d$ .  $\mu$  represents the vector of maximum coherence, as it corresponds to the point in the vector space which is closer to all synsets associated with  $H$  on average. Once we have  $\mu$ , we consider each disambiguated instance  $d \in L$ , retrieve all the candidate senses of its surface form, and calculate a Nasari score for each candidate sense. The Nasari score  $N(s)$  of a candidate sense  $s$  is given by the cosine similarity between its associated Nasari vector  $\mathbf{s}$  and the centroid  $\mu$ :

$$N(s) = \cos(\mathbf{s}; \mu) \quad (4.5)$$

As with the coherence score, we empirically set a Nasari score threshold to 0.75 (i.e. the upper quarter of the similarity scale). Each  $d \in L$  is then re-disambiguated with the sense  $s$  obtaining the highest  $N(s)$ , provided that  $N(s)$  exceeds the threshold:

$$\hat{s} = \operatorname{argmax}_{s \in S_d} N_s \quad (4.6)$$

where  $S_d$  is the set containing all the candidate senses for  $d$ . If no candidate sense  $s \in S_d$  achieves a value of  $N(s)$  beyond the threshold, we discard  $d$  as a whole.

In the example of Figure 4.2, the synset of climate intended as the weather situation (bn:00019780n), incorrectly selected by Babelify in the previous stage, is now replaced with the synset of climate intended metaphorically as mood of a situation or event (bn:00019781n). At the same time, the synset of place intended as a physical or geographical location (bn:00019780n) is discarded, as no alternative sense of place and lugar is found to be close enough to  $\mu$ .

<sup>23</sup> $L$  includes also those instances for which Babelify did not provide a disambiguation. In fact, Babelify associates with each disambiguated instance an internal coherence score (Babelify score): when this score goes below 0.7, an MFS back-off strategy is activated by default for that instance, replacing the original output of Babelify.

<sup>24</sup>Nasari computes a vector for each Babel synset that includes a Wikipage (cf. Section 2.2.3.3): hence we can retrieve a Nasari vector with virtually all nominal disambiguated instances in  $L$ .

		Total	EN	FR	DE	ES
Full	# Annotations	215,877,109	26,455,574	22,214,996	16,888,108	21,486,532
	# Lemma Types	567,378	60,853	30,474	66,762	43,892
	# Senses	247,706	138,115	65,301	75,008	74,214
	Average coherence score	0.19	0.19	0.18	0.18	0.18
Re ned	# Annotations	122,963,111	15,441,667	12,955,469	9,165,112	12,193,260
	# Lemma Types	453,063	42,947	23,603	50,681	31,980
	# Senses	155,904	86,881	49,189	52,425	52,859
	Average coherence score	0.29	0.28	0.25	0.28	0.27

Table 4.12. Global statistics on EuroSense before (full) and after re nement (re ned) for all the 21 languages. Language-speci c gures are also reported for the 4 languages of the intrinsic evaluation (Section 4.2.4.1).

### 4.2.3 Statistics

Table 4.12 reports some global statistics on EuroSense regarding both its high-coverage (cf. Section 4.2.1) and high-precision (cf. Section 4.2.2) versions. Joint multilingual disambiguation with Babelify generated more than 215M sense annotations of 247k distinct concepts and entities, while similarity-based re nement retained almost 123M high-con dence instances (56.96% of the total), covering almost 156k distinct concepts and entities. 42.40% of these retained annotations were corrected or validated using distributional similarity. As expected, the distribution over parts of speech is skewed towards nominal senses (64.79% before re nement and 81.79% after re nement) followed by verbs (19.26% and 12.22%), adjectives (11.46% and 5.24%) and adverbs (4.48% and 0.73%). We note that the average coherence score increases from 0.19 to 0.29 after re nement, suggesting that distributional similarity tends to favor sense annotations that are also consistent across di erent languages. Table 4.12 also includes language-speci c statistics on the 4 languages of the intrinsic evaluation, where the average lexical ambiguity ranges from 1.12 senses per lemma (German) to 2.26 (English).

Interestingly enough, if we consider all the 21 languages, the total number of distinct lemmas covered is more than twice the total number of distinct senses: this is a direct consequence of having a uni ed, language-independent sense inventory (BabelNet), a feature that sets EuroSense apart from previous multilingual sense-annotated corpora (Otegi et al., 2016). Finally we note from the global gures on the number of covered senses that 109 591 senses (44.2% of the total) are not covered by the English sense annotations: this suggests that EuroSense relies heavily on multilinguality in integrating concepts or named entities that are tied to speci c social or cultural aspects of a given language (and hence would be under-represented in an English-speci c sense inventory).

### 4.2.4 Experimental Evaluation

As in the previous disambiguation scenario (Section 4.1) we assessed the quality of EuroSense 's sense annotations both intrinsically, by means of a manual evaluation on four samples of randomly extracted sentences in di erent languages (Section 4.2.4.1), as well as extrinsically, by augmenting the training set of a supervised

all-words WSD system (Zhong and Ng, 2010) and showing that it leads to consistent performance improvements over two standard WSD benchmarks (Section 4.2.4.2).

#### 4.2.4.1 Intrinsic Evaluation: Annotation Quality

We carried out a manual evaluation on 4 different languages (English, French, German and Spanish) with 2 human annotators per language. We sampled 50 random sentences across the subset of sentences EuroSense featuring a translation in all 4 languages, totaling 200 sentences overall. For each sentence, we evaluated all sense annotations both before and after the refinement stage, along with the sense annotations obtained by a baseline that disambiguates each sentence in isolation with Babelify . Overall, 5818 sense annotations were manually verified across the three configurations (1518 in English, 1564 in French, 1093 in German and 1643 in Spanish). In every language the two judges agreed in more than 85% of the cases, with an inter-annotator agreement in terms of Cohen's kappa (Cohen, 1960) above 60% in all evaluations (67.7% on average).

**Evaluation Setup.** For each sentence in the sample, each annotator was shown the text of the sentence, together with every sense annotation paired with the corresponding BabelNet synset. The annotator had to decide independently, for each sense annotation, whether it was correct (score of 1), or incorrect (score of 0). The disambiguation source (i.e. whether the annotation came from Babelify , Nasari , or the Babelify baseline) was not shown. In some special cases where a certain sense annotation was acceptable but a more suitable synset was available, a score of 0.5 was allowed. One recurrent example of these indecisive annotations occurred on multi-word expressions: being designed as a high-coverage all-word disambiguation strategy, Babelify can output disambiguation decisions over overlapping mentions when confronted with fragments of text having more than one acceptable disambiguation. For instance, the multi-word expression 'Commission of the European Union' can be interpreted both as a single mention, referring to the specific sense European Commission (executive body of the European Union), and as two mentions, one (Commission) referring to the sense Parliamentary committee (a subordinate deliberative assembly), and the other (European Union) referring to the the sense European Union (the international organization of European countries). In all cases where one part of a certain multi-word expression was tagged with an acceptable meaning, but a more accurate annotation would have been the one associated with the whole multi-word expression, we allowed annotators to assign a score of 0.5 to valid annotations of nested mentions and a score of 1 only to the complete and correct multi-word annotation. Another controversial example of indecision is connected to semantic shifts due to Wikipedia redirections, which lead to sense annotations that are lexically acceptable but wrong from the point of view of semantic roles. For instance, the term painter inside Wikipedia redirects to the Wikipeage Painting , while the term Basketball player redirects to the Wikipeage Basketball . These redirections are also exploited by Babelify as acceptable disambiguation decisions and, as such, they are also allowed a score of 0.5.

<sup>25</sup>This policy is very often used in Entity Linking and Wikification (cf. Section 2.2.2).

	EN		FR		DE		ES	
	Prec.	Cov.	Prec.	Cov.	Prec.	Cov.	Prec.	Cov.
Babelify	76.1	100.0	59.1	100.0	80.4	100.0	67.5	100.0
EuroSense (full)	80.3	100.0	67.9	100.0	84.6	100.0	76.7	100.0
EuroSense (re ned)	81.5	75.0	71.8	63.5	89.3	53.8	82.5	62.9

Table 4.13. Precision (Prec.) and coverage (Cov.) percentage (%) of EuroSense, manually evaluated on a random sample in 4 languages. Precision is averaged between the two judges, and coverage is computed assuming each content word in the sense inventory to be a valid disambiguation target.

	SemEval-2013	SemEval-2015
IMS <sub>SemCor</sub>	65.3	69.3
IMS <sub>OMSTI</sub>	65.0	69.1
IMS <sub>EuroSense</sub>	66.4	69.5
UKB	62.9	63.3
Babelify	66.4	70.3
SOTA	67.3	71.9
MFS	63.0	67.8

Table 4.14. Results in terms of F-score on all-words WSD.

Table 4.13 shows that joint multilingual disambiguation improves precision consistently over the baseline, from an increase of 4.2% in English to 9.2% in Spanish. Even if the nature of source text (i.e. parliamentary proceedings) played a significant role, this strategy proved effective in improving the disambiguation performance of Babelify compared to previously reported results (cf. Section 3.1.3.3). The similarity-based refinement boosts precision even further (with a 3.9% average increase), at the expense of a reduced coverage, which drops by 36.2% on average. Over the 4 languages, sense annotations appear to be most reliable for German, consistently with its lower lexical ambiguity on the corpus (cf. Section 4.2.3).

#### 4.2.4.2 Extrinsic Evaluation: Word Sense Disambiguation

We carried out an extrinsic evaluation of EuroSense by mapping its refined sense annotations for English to WordNet, and using them as a training set for the same supervised WSD system used in Section 4.1.3.2: It Makes Sense (Zhong and Ng, 2010, IMS). Following Taghipour and Ng (2015b), we started with SemCor (Section 3.1.1.1) as initial training dataset, and then performed a subsampling of EuroSense up to 500 additional training examples per word sense. Crucially, instead of sampling randomly as in Taghipour and Ng (2015b), we sorted sense annotations by decreasing coherence score, and considered the top occurrences of each word sense. We then

<sup>26</sup> Both Babelify and the baseline always attempt an answer for every possible disambiguation target, hence they achieve maximum coverage in each configuration. Note that in Table 4.13 we consider coverage (i.e. number of content words covered) in place of recall, since the number of 'correct' answers is not clearly defined in many cases, e.g. with overlapping mentions (as discussed in Section 4.2.4.1).

trained IMS on this augmented training set and tested on the two most recent standard benchmarks for all-words WSD: SemEval-2013 and SemEval-2015, from the standardized framework of Raganato et al. (2017a). As baselines we considered IMS trained on SemCor only and on OMSTI (Section 3.1.2.2). As shown in Table 4.14, where we also include two knowledge-based systems Babelfy and UKB (Agirre et al., 2014), the MFS baseline, and the current state of the art (SOTA) on both datasets (Raganato et al., 2017a), IMS trained on the EuroSense -augmented training set consistently outperforms all baseline models, showing competitive results even against IMS trained on semi-automatic sense annotations (Taghipour and Ng, 2015b). Even though the F-score increase is not statistically significant in these specific benchmarks, it demonstrates that our fully automatic method can perform on par with semi-automatic approaches in extracting high-quality sense annotations.

**Final Remarks.** Our experimental evaluation shows, once again, that exploiting at best the features of the target text is crucial to achieve high-quality disambiguation in a fully automatic fashion. Specifically, with EuroSense we explored the effectiveness of multilinguality in the disambiguation process: instead of relying on external translations or pre-computed alignments, however, we let semantic coherence across languages emerge naturally at disambiguation time, thanks to the flexibility of a language-independent sense inventory and its multilingual lexicalizations. In contrast to the disambiguation pipeline of Section 4.1, building EuroSense required using two external tools, Babelfy and Nasari, and a structured pipeline to cope with their respective shortcomings. The proved benefits of this solution are: (1) the release of two different versions of EuroSense, complementary with respect to the downstream applications they are most suitable for; (2) the fact that each sense annotation is associated with multiple confidence scores (Section 4.2.2) enabling to further tune EuroSense for a specific task, application, or use.

### 4.3 SenseDefs: A Multilingual Disambiguation of Textual Definitions

In this third and final disambiguation scenario our target is definitional text. We focus on a large definitional corpus that shares some features with the Wikipedia corpus of Section 4.1 (i.e. the encyclopedic nature), as well as some features with the parallel corpus of Section 4.2 (i.e. equivalent sentences in multiple languages), with, however, an important difference: the short and concise nature of definitions.

**Why Definitions?** In addition to lexicography, where their use is of paramount importance, textual definitions (or glosses) drawn from dictionaries or encyclopedias have been widely used in various NLP tasks and applications. Definitional knowledge is effective inasmuch as it conveys the crucial semantic information and the distinguishing features of a given subject (deniendum): this means that, on the one hand, a definition often provides a fair amount of discriminative power that can be leveraged to automatically represent and disambiguate the deniendum; on the other, definitions are usually concise and encode dense, virtually noise-free information that can be best exploited with knowledge acquisition techniques. To date, some of

the areas where the use of de nitional knowledge has proved to be key in achieving state-of-the-art results are Word Sense Disambiguation (Lesk, 1986; Banerjee and Pedersen, 2003; Navigli and Velardi, 2005; Agirre and Soroa, 2009; Faralli and Navigli, 2012; Fernandez-Ordonez et al., 2012; Chen et al., 2014; Basile et al., 2014; Camacho Collados et al., 2015b), Taxonomy and Ontology Learning (Velardi et al., 2013; Flati et al., 2016; Espinosa Anke et al., 2016c), Information Extraction (Richardson et al., 1998; Delli Bovi et al., 2015b), Plagiarism Detection (Franco-Salvador et al., 2016), and Question Answering (Hill et al., 2016). In fact, textual de nitions are today widely available in knowledge resources of various kinds, ranging from lexicons and dictionaries, such as WordNet (Section 2.1.1) or Wiktionary, to encyclopedic Wikipedia-derived knowledge bases (Section 2.1.2). Interestingly enough, sources of de nitional knowledge also include Wikipedia: despite its purely encyclopedic nature, and although the format of a Wikipedia article does not include an explicit gloss or de nition, the rst sentence of each article is generally regarded as the de nition of its subject.

**Related Work.** Disambiguating de nitions has attracted a considerable amount of interest over the years. Among others, WordNet has de nitely been the most popular and the most exploited target resource in this respect, as WordNet glosses have still been used successfully in recent work (Khan et al., 2013; Chen et al., 2015). A rst attempt to disambiguate WordNet glosses automatically was proposed as part of the eXtended WordNet project (Novischi, 2002).<sup>27</sup> However, this attempt's estimated coverage did not reach 6% of the total number of sense-annotated instances. Moldovan and Novischi (2004) proposed an alternative disambiguation approach, speci cally targeted at the WordNet sense inventory and based on a supervised model trained on SemCor (Section 3.1.1.1); another disambiguation task focused on WordNet glosses was presented as part of the Senseval-3 workshop (Litkowski, 2004). However, the best reported system obtained precision and recall gures below 70%, which is arguably not enough to provide high-quality sense-annotated data for current state-of-the-art NLP systems. In addition to annotation reliability, another issue that arises when producing a corpus of textual de nitions is coverage. In fact, reliable corpora of sense-annotated de nitions produced to date, such as the Princeton WordNet Gloss Corpus (Section 3.1.2.1), have usually been obtained employing human annotators and, we discussed extensively in previous sections, human supervision is increasingly expensive and time-consuming as the size of the sense inventory grows larger. Furthermore, new encyclopedic knowledge about the world is constantly being harvested, and WordNet's de nitions fail to capture many up-to-date concepts and entities. With a view to tackling this problem, a great deal of research has recently focused on the automatic extraction of de nitions from unstructured text (Navigli and Velardi, 2010; Benedictis et al., 2013; Espinosa Anke and Saggion, 2014; Espinosa Anke et al., 2015; Dalvi et al., 2015); as a consequence, disambiguating de nitional text has to be framed necessarily as a large-scale task.

**Motivation.** Irrespective of the nature of the knowledge source, an accurate semantic analysis of textual de nitions is made di cult by the short and concise

<sup>27</sup><http://www.hlt.utdallas.edu/~xwn>

nature of definitional text, a crucial issue for automatic disambiguation systems that rely heavily on local context. Furthermore, the majority of approaches making use of definitions are restricted to corpora where each concept or entity is associated with a single definition; instead, definitions coming from different resources are often complementary and might give different perspectives on the denotandum. Moreover, equivalent definitions of the same concept or entity may vary substantially according to the language, and be more precise or self-explanatory in some languages than others. In fact, the way a certain concept or entity is defined in a given language is sometimes strictly connected to the social, cultural and historical background associated with that language, a phenomenon that also affects the lexical ambiguity of the definition itself. This difference in the degree of ambiguity when moving across languages is especially valuable in the context of disambiguation, as we demonstrated in the previous disambiguation scenario (Section 4.2).

In light of this, in the present section we adapt the disambiguation pipeline designed for EuroSense to a definitional setting. The underlying disambiguation idea is, indeed, almost the same: bringing together definitions drawn from different resources and different languages, and exploiting their cross-lingual and cross-resource complementarities at disambiguation time. As in the case of EuroSense, a large-scale high-quality disambiguation requires us to use off-the-shelf techniques which, for flexibility and scalability purposes, are based on a single multilingual disambiguation model. In fact, while language- and resource-specific techniques can certainly be used for disambiguation, the number of models required would add up to the order of hundreds, without even considering the need for large amounts of sense-annotated data for each language and resource. Therefore, we first gather a target corpus of textual definitions in multiple languages from BabelNet (section 4.3.1); then we apply the two-stage disambiguation pipeline described in Sections 4.2.1 and 4.2.2 to each group of definitions referring to the same denotandum (Section 4.3.2). As a result we obtain SenseDefs (Camacho Collados et al., 2016a)<sup>28</sup> a multilingual corpus of textual definitions featuring over 38 million definitions in 263 languages, with almost 250 million sense annotations for both concepts and named entities drawn from the BabelNet sense inventory. Following the same methodology of Sections 4.1 and 4.2, we examine some global statistics about the corpus in Section 4.3.3, and then we carry out an experimental evaluation in Section 4.3.4, including both intrinsic and extrinsic experiments.

#### 4.3.1 Gathering Definitional Knowledge across Resources and Languages

We construct a target corpus of definitional knowledge by collecting all textual definitions associated with every concept or named entity inside BabelNet, for all the languages available. Being a merger of various different knowledge resources (cf. Section 2.1.3), BabelNet provides a very heterogeneous set of definitions. Specifically, the definitional knowledge inside BabelNet comes from the following sources:

- ^ WordNet : being hand-crafted by expert annotators, definitional knowledge

<sup>28</sup> <http://lcl.uniroma1.it/disambiguated-glosses>

from WordNet is among the most accurate available and includes non-nominal parts of speech rarely covered by other resources (e.g. adjectives and adverbs). However, given its considerably smaller scale, WordNet provides less than 1% of the overall number of definitions in BabelNet, and covers only the English language;

- ^ **Wikipedia** : Wikipages do not provide explicit glosses or definitions, however, according to the style guidelines of Wikipedia<sup>29</sup> a Wikipage should begin with a short declarative sentence defining what (or who) the subject is and why it is notable. Following previous literature, we also consider the first sentence of a Wikipage as a valid definition of the corresponding concept or entity. Furthermore, text snippets drawn from the associated disambiguation pages can also be regarded as definitions<sup>30</sup>. Wikipedia provides the largest proportion of definitional knowledge by far ( 77%), including many definitions in languages other than English;
- ^ **Wikidata** : Wikidata is the second largest individual contribution to SenseDefs (more than 8 million items and 22% of the total), even though, given its strictly computational nature, it often provides minimal definition phrases containing only the superclass of the deniendum.
- ^ **Wiktionary, OmegaWiki** : beyond WordNet, Wikipedia and Wikidata, the remaining definitions ( 1% of the total) are provided by two collaborative multilingual dictionaries: Wiktionary and OmegaWiki. Wiktionary<sup>31</sup> is a Wikimedia project designed to represent lexicographic knowledge that would not be well suited for an encyclopedia (e.g. verbal and adverbial senses). It is available for over 500 languages typically with a very high coverage, including domain-specific terms and descriptions that are not found in WordNet. Similar to Wiktionary, OmegaWiki<sup>32</sup> is a large multilingual dictionary based on a relational database, designed with the aim of unifying the various language-specific Wiktionaries into a unified lexical repository.

Overall, the corpus of definitional knowledge obtained from BabelNet comprises more than 38 million definitions associated with more than 8 million synsets, both concepts and named entities (see Section 4.3.3). The key feature of this corpus, that we will leverage at disambiguation time, is the fact that BabelNet's inter-resource and inter-language mappings enable us to combine multiple definitions (drawn from different resources and in different languages) of the same concept or named entity. Thus, if we re-arrange the corpus by grouping all the definitions by deniendum, we can view it as a collection of around 8 million multilingual definitional texts.

<sup>29</sup> [https://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style)

<sup>30</sup> The release format of SenseDefs (cf. Section 6.3) specifies two distinct attribute values for definitions extracted from the first sentence of Wikipedia articles ( WIKI) and definitions extracted from disambiguation pages (WIKIDIS).

<sup>31</sup> <https://www.wiktionary.org>

<sup>32</sup> <http://www.omegawiki.org>

### 4.3.2 The Disambiguation Pipeline on a Running Example

As highlighted at the beginning of the section, definitional knowledge is not easy to analyze automatically at the sense level. Since many definitions are short and concise, the lack of sufficient and/or meaningful context might negatively affect the performance of an off-the-shelf disambiguation system that works at the sentence level (i.e. targeting individual definitions one by one). In light of this, we consider the input corpus arranged as described in Section 4.3.1; while with EuroSense we considered, for each English sentence, a multilingual text given by all its translations, in this case we consider, for each definition, a multilingual text given by all its definitions. In this way, we can associate a much richer context with each target definition, and this context is semantically meaningful (since it is composed of other definitions describing the same concept or entity).

Following the EuroSense pipeline, we perform an initial preprocessing step on all definitions, which consists of tokenization, part-of-speech tagging and lemmatization for a subset of languages using standard NLP tools (Camacho Collados et al., 2016a). Then we apply stage 1 (Section 4.2.1) and stage 2 (Section 4.2.2) of the pipeline to obtain the full, high-coverage version and the refined, high-precision version of SenseDefs. Let us go through the stages of the disambiguation pipeline using the running example from Camacho Collados et al. (2016a).

**Running Example.** As an example, consider the following definition of castling in chess ( $\text{castling}_h^1$ ) as provided by WordNet:

Interchanging the positions of the king and a rook. (4.7)

The context in this example is limited and it might not be obvious for an automatic disambiguation system that the concept being defined relates to chess: for instance, an alternative definition of  $\text{castling}_h^1$  where `chess` is explicitly mentioned would definitely help the disambiguation process. When provided solely with the English WordNet definition of (4.7), Babelify disambiguates rook incorrectly as rookie, inexperienced youth ( $\text{rook}_h^7$ ). Instead, as additional definitions from other resources and languages are included Babelify exploits the added context to construct a richer semantic graph, and disambiguates rook with its correct chess-related sense ( $\text{rook}_h^1$ ) in the first stage of the pipeline. Multilingual joint disambiguation, however, is still not enough to provide a high-confidence disambiguation for the word king, which was then incorrectly disambiguated using the MCS back-off strategy ( $\text{king}_h^1$ ). This error is subsequently corrected with the refinement step, as the chess-related sense of king ( $\text{king}_h^8$ ) achieves higher semantic similarity with the disambiguated instances in  $H$  (cf. Section 4.2.2) compared to its predominant sense. In fact, thanks to the augmented context in the first stage, many chess-related senses have been disambiguated with high confidence, including  $\text{rook}_h^1$ , but also, e.g.,  $\text{enroque}_h^1$  from the Spanish Wikipedia definition, or  $\text{Schach}_h^1$  from the German Wikidata definition (both of which are, incidentally, monosemic cases).

	# Definitions		# Sense Annotations	
	Full	Re ned	Full	Re ned
Wikipedia	29,792,245	28,904,602	223,802,767	143,927,150
Wikidata	8,484,267	8,002,375	22,769,436	17,504,023
Wiktionary	281,756	187,755	1,384,127	693,597
OmegaWiki	115,828	106,994	744,496	415,631
WordNet	146,018	133,089	843,882	488,730
Total	38,820,114	37,334,815	249,544,708	163,029,131

Table 4.15. Total number of definitions and sense annotations by knowledge resource in the full and re ned versions of SenseDefs .

### 4.3.3 Statistics

Table 4.15 shows some global statistics of the full and re ned versions of SenseDefs ,<sup>33</sup> divided by resource. The output of the full version is a corpus of 38,820,114 disambiguated glosses, corresponding to 8,665,300 BabelNet synsets and covering 263 languages and 5 different resources (Wiktionary, WordNet, Wikidata, Wikipedia and OmegaWiki). It includes 249,544,708 sense annotations (6.4 annotations per definition on average). The re ned version of the resource includes fewer, but more reliable, sense annotations and a slightly reduced number of glosses containing at least one sense annotation. As noted in Section 4.3.1, Wikipedia is the resource with by far the largest number of definitions and sense annotations, including almost 30 million definitions and over 140 million sense annotations in both versions of the corpus. Additionally, Wikipedia also features textual definitions for the largest number of languages (over 200).

Statistics by language. Figure 4.3 displays the number of definitions and sense annotations, respectively, divided by language. As expected, English provides the largest contribution (5.8 million glosses and 37.9 million sense annotations in the re ned version), followed by German and French. Even though the majority of sense annotations overall concern resource-rich languages, the language rankings in Figure 4.3a and 4.3b do not coincide exactly: this suggests, on the one hand, that some languages (such as Vietnamese and Spanish, both with higher positions in Figure 4.3b compared to Figure 4.3a) actually benefit from a cross-lingual disambiguation strategy; on the other hand, it also suggests that there is still room for improvement, especially for some other languages (such as Swedish or Russian) where the tendency is reversed, and the number of annotations is lower compared to the amount of definitional knowledge available.

Table 4.16 shows the number of annotations divided by part-of-speech tag and disambiguation source. In particular, the full version comprises two disambiguation sources: Babelfy and the MFS back-off (used for low-confidence annotations). The re ned version, instead, removes the MCS back-off, either by discarding or correcting

<sup>33</sup> Consistently with Section 4.2, we refer to the output of the first stage of the pipeline as the full version of SenseDefs, and to the final output of the pipeline as the re ned version.

Figure 4.3. Total number of textual definitions (a) and sense annotations (b) in SenseDefs by language (top 15 languages).

the annotation with Nasari (cf. Section 4.2.2). Additionally, 17% of the sense annotations obtained by Babelfy (without the MFS back-off) are also corrected or discarded. Assuming the coverage of the full version to be 100%, as in Section 4.2.4.1, the coverage of our system after the refinement step is estimated to be 65.3%. As shown in Table 4.16, discarded annotations mostly consist of verbs, adjectives and adverbs, which are often harder to disambiguate as they are very frequently not directly related to the denotandum. In fact, the coverage figure on noun instances is estimated to be 73.9% after refinement.

		All	Nouns	Verbs	Adjectives	Adverbs
Full	Babelfy	174,256,335	158,310,414	4,368,488	10,646,921	930,512
	MFS	75,288,373	56,231,910	8,344,930	9,256,497	1,455,036
	Total	249,544,708	214,542,324	12,713,418	19,903,418	2,385,548
Refined	Babelfy	144,637,032	140,111,921	1,326,947	3,064,416	133,748
	Nasari	18,392,099	18,392,099	-	-	-
	Total	163,029,131	158,504,020	1,326,947	3,064,416	133,748

Table 4.16. Total number of definitions and sense annotations by part-of-speech tag (columns) and by source (rows) in the full and refined versions of SenseDefs.

		#Ann.	Prec.	Rec.	F1	IAA ROA	
EN	Babelfy	671	84.3	69.6	76.1	94.6	71.7
	SenseDefs <sub>full</sub>	714	80.0	70.2	74.8	94.2	70.1
	SenseDefs <sub>re ned</sub>	745	83.1	76.1	79.5	95.3	71.9
ES	Babelfy	678	85.8	59.3	70.2	91.4	51.1
	SenseDefs <sub>full</sub>	737	82.6	62.1	70.9	92.4	66.2
	SenseDefs <sub>re ned</sub>	725	86.6	64.0	73.6	95.1	63.3
FR	Babelfy	516	84.3	49.8	62.6	97.2	85.7
	SenseDefs <sub>full</sub>	568	81.3	52.8	64.0	96.7	86.4
	SenseDefs <sub>re ned</sub>	579	87.1	57.7	69.4	95.1	65.8
IT	Babelfy	540	81.7	53.5	64.7	94.5	74.3
	SenseDefs <sub>full</sub>	609	73.9	54.5	62.8	92.4	78.0
	SenseDefs <sub>re ned</sub>	618	77.5	58.1	66.4	94.7	83.0

Table 4.17. Precision (Prec.) and recall (Rec.) percentage (%) of SenseDefs, manually evaluated on random samples of 120 textual definitions in 4 languages (English, Spanish, French, and Italian). Precision is averaged between the two judges, and recall is computed assuming each content word in a sentence should be associated with a distinct sense.

#### 4.3.4 Experimental Evaluation

In line with the previous sections, we carried out both an intrinsic and an extrinsic evaluation of SenseDefs. The former consists of two experiments: a manual assessment on four samples of randomly extracted definitions in different languages (Section 4.3.4.1), and an automatic evaluation on the manually annotated portion of the Princeton Gloss Corpus (Section 4.3.4.2). The latter, instead, evaluates SenseDefs on the Wikipedia sense clustering task (Section 4.3.4.3).

##### 4.3.4.1 Intrinsic Evaluation #1: Annotation Quality

We evaluated sense annotation quality in SenseDefs on four different languages: English, French, Italian and Spanish. To this end, we first randomly sampled 120 definitions for each language. Then, two annotators validated the sense annotations given by SenseDefs (both full and re ned) and by the same Babelfy baseline used in Section 4.2.4.1. The evaluation setup is the same as the one in Section 4.2.4.1. However, in this case we excluded those sense annotations coming from the MFS back-off, in order to assess explicitly the output of our disambiguation pipeline. We also calculated the Inter Annotator Agreement (IAA) between the two annotators of each language by means of Relative Observed Agreement (ROA), i.e. the proportion of equal answers, and Cohen's kappa (Cohen, 1960). Finally, the two annotators in each language adjudicated the answers which were judged with opposite values.

Table 4.17 shows the results of this manual evaluation. In the four languages, our re ned version of the corpus achieved the best overall results. SenseDefs achieved over 80% precision in three of the four considered languages, both in its full and re ned versions. In the Italian sample the precision dropped to 73.9% and 77.5%, respectively, probably due to lower coverage in BabelNet. Finally, it is worth noting that, for all the examined languages, both the full and re ned versions of SenseDefs provided more annotations than using the baseline on isolated definitions.

	#WN Annot.	Prec.	MFS-Prec.
SenseDefs <sub>full</sub>	162,819	76.4	66.1
SenseDefs <sub>re ned</sub>	169,696	76.4	65.2
Babelify	130,236	69.1	65.6
IMS	275,893	56.1	55.2

Table 4.18. Overall precision (Prec.) percentage (%) and number of compared WordNet annotations (#WN Annot.) on the Princeton Gloss Corpus (Section 3.1.2.1). On the rightmost column, precision of the MFS baseline (MFS-Prec.) on the same sample.

#### 4.3.4.2 Intrinsic Evaluation #2: WordNet Glosses

We performed an additional intrinsic evaluation where we compared the WordNet annotations given by SenseDefs with the manually-crafted annotations of the disambiguated glosses from the Princeton Gloss Corpus (Section 3.1.2.1). Similarly to the previous manual evaluation, we included a baseline based on Babelify disambiguating the definitions sentence-wise in isolation, and a supervised baseline based on the pre-trained models of IMS (Zhong and Ng, 2010) or MSTI.<sup>34</sup> As in our previous experiment, we did not consider the sense annotations for which the MFS back-off strategy was activated on any of the comparison systems. Finally, we included the MFS result on each of the subsets of sense annotations provided the systems. Table 4.18 shows the precision of SenseDefs, Babelify and IMS on the Princeton Gloss Corpus. SenseDefs achieved a precision of 76.4% in both versions. Even though results are not directly comparable,<sup>35</sup> IMS reported a considerably lower precision than our pipeline's, and also lower compared to its performance on standard benchmarks (Raganato et al., 2017a). This result highlights the difficulty of dealing with definitional text, even for supervised systems: in fact, most definitions do not provide enough local context for an accurate disambiguation at the sentence level.

#### 4.3.4.3 Extrinsic Evaluation: Sense Clustering

In our extrinsic experiment we evaluated the re ned version of SenseDefs on the Wikipedia sense clustering task (Dandala et al., 2013). Specifically, we exploited SenseDefs to enhance the vectorial representations of Nasari, by enriching the semantic network used in the original implementation. Since re ned sense annotations tend to identify synsets that are highly semantically related to the denendum, they can actually be viewed as semantic connections between these synsets and the synset identified by the denendum, and hence utilized as additional edges. We performed this enrichment step, ran again the original Nasari pipeline to generate the vectors, and then evaluated these on the Wikipedia sense clustering task, following the original experiment by Camacho Collados et al. (2016c).

Table 4.19 shows the accuracy and F-score results of our enhanced version of Nasari (Nasari + SenseDefs). As a comparison we included the Support Vector

<sup>34</sup> <http://www.comp.nus.edu.sg/~nlp/corpora.html#onemilwsd>

<sup>35</sup> Since our pipeline annotates with BabelNet synsets, the set of candidate senses is often larger than IMS and the MFS baseline (both based on WordNet).

	500-pair		SemEval	
	Accuracy	F-score	Accuracy	F-score
Nasari + SenseDefs	86.0	74.8	88.1	64.7
Nasari	81.6	65.4	85.7	57.4
SB-Sew <sub>best</sub>	-	71.0	-	64.0
SVM-monolingual	77.4	-	83.5	-
SVM-multilingual	84.4	-	85.5	-
Baseline	28.6	44.5	17.5	29.8

Table 4.19. Accuracy and F-score results on Wikipedia sense clustering.

Machine classifier of Dandala et al. (2013), which exploits information from Wikipedia in English (SVM-monolingual) and in four different languages (SVM-multilingual), together with a naive baseline that clusters every Wikipedia pair. We also report the results obtained by the original English lexical vectors of Nasari, and those obtained by the best configuration of SB-Sew (cf. Section 4.1.3.3). As shown in Table 4.19, the enrichment produced by SenseDefs proved to be highly beneficial, with a significant improvement on the original results reported by Camacho Collados et al. (2016c), and the best overall performance on the task.

**Final Remarks.** With this experimental evaluation we assessed the flexibility and effectiveness of the disambiguation pipeline we designed in Section 4.2 on a heterogeneous multilingual corpus of definitional text. With the broadened intrinsic evaluation on the WordNet glosses (Section 4.3.4.2), in particular, we saw that the extremely limited local context of most definitions is a crucial problem also for trained and tuned supervised systems in English. This suggests once again that, when adding multiple languages into the picture (including languages for which sense-annotated data are not available), using an array of language-specific supervised models to carry out a reliable disambiguation procedure on each monolingual subset of the corpus becomes unpractical. The pipeline we propose, instead, employs a single model for which adding additional languages contributes to creating a richer context for disambiguation. Differently from the parallel text scenario of Section 4.2, in this case a further advantage is given by the fact that multiple definitions of a given synset can be put together from different resources even when a single language is considered (as in the example of Section 4.3.2). In general, while parallel text is useful to enforce cross-language semantic coherence (but new translations of the same sentences are less likely to provide novel and complementary information), in the present case additional definitions from other languages might be completely different in describing the denotandum. As a result, the precision figures reported by our pipeline in the intrinsic evaluation, consistently with the previous case (Section 4.2.4.1), are higher on average compared with those estimated in previous literature for fully automatic systems, which very rarely go beyond 75% (cf. Section 3.1.3.3); moreover, both SenseDefs and EuroSense can be further tuned using the confidence scores associated with each sense annotation.

	# Languages	# Annotations	# Senses	# Tokens	Accuracy
Sew	1	206,475,360	4,071,902	1,357,105,761	93.4%
EuroSense	21	122,963,111	155,904	48,274,313	81.5%
SenseDefs	263	163,029,131	10,870,032	71,109,002	79.7%
Babel ed Wikipedia	3	113,896,864	4,239,879	501,862,251	70.5%
Babel ed MASC	1	286,416	23,175	592,472	72.4%

Table 4.20. Global statistics on the sense-annotated corpora treated in this section, including the number of languages covered, the total number of sense annotations, the total number of concepts and named entities covered, the total number of word tokens, and the estimated accuracy of sense annotations for English.

As a general summary, Table 4.20 puts together some global statistics about the three sense-annotated resources presented in this section, i.e. Sew, EuroSense, and SenseDefs,<sup>36</sup> and compares them with the two BabelNet-annotated corpora examined in Section 3.1.3.3, i.e. the Babel ed Wikipedia and the Babel ed MASC. As we already discussed, all the three resources provide sense annotations with higher quality compared to previous approaches, as estimated in their respective intrinsic evaluations.<sup>37</sup> Among the resources presented in this chapter, Sew stands out in terms of size and total number of annotations, being constructed from the largest source corpus (a Wikipedia dump). However, annotation density (0.15) is lower than EuroSense (2.55) and SenseDefs (2.29), and the sense-annotated corpus is currently available only for English. As regards EuroSense and SenseDefs, instead, they both represent the largest available resources of their kinds (parallel text and definitional text, respectively) providing sense annotations for concepts and named entities in multiple languages. Since both of them have been constructed using the same disambiguation pipeline, they show comparable accuracy and coverage. In both cases, however, there is still room for improvement: in Section 7 we come back to the disambiguation strategies presented in this chapter, and discuss some open problems and perspective of future work to further improve their performances.

<sup>36</sup>We considered the final version of the three resources at the end of their disambiguation pipelines: the refined versions of EuroSense and SenseDefs, and Sew after applying the conservative policy.

<sup>37</sup>In the case of SenseDefs we averaged the precision figures obtained in the two intrinsic experiments of Sections 4.3.4.1 and 4.3.4.2.



## Chapter 5

# Sense-Aware Extraction of Relational Knowledge

Any fool can know.  
The point is to understand.  
Albert Einstein

In this chapter we address the second objective outlined in Section 1.1: that of reframing the task of Open Information Extraction at the sense level, and exploring the benefits of sense-aware techniques at the various stages of the extraction process. As anticipated in Chapter 1, our focus is on Open Information Extraction, rather than traditional (closed) Information Extraction. The motivation for this choice is two-fold: on the one hand, being completely unsupervised, OIE tackles explicitly issues like the knowledge acquisition bottleneck, and complies perfectly with the long-term goal of the present work; on the other, the fact that semantic relations modeled by OIE are not pre-specified or encoded formally in a database, but instead bound to their surface-text realizations (Section 2.3.2), makes them particularly susceptible to many linguistic phenomena studied in Lexical Semantics (e.g. polysemy, synonymy). Thus, OIE is one of the NLP areas where sense-level information appears to have greater impact and really make a substantial difference.

In fact, we examined in Section 3.2 some recent approaches that have started moving in this direction: Patty (Section 3.2.1) and WiSeNet (Section 3.2.2). These methods demonstrate how the choice of modeling Lexical Semantics explicitly (e.g. with a more structured semantic representation of relation patterns and relation instances) is not only feasible but also tremendously effective, as it enables the extraction of high-quality relation instances on a large scale. Also, being anchored to an underlying knowledge resource, these relation instances can easily leverage their explicit semantic characterization to generalize better and overcome many limitations of traditional approaches.

Even though *Patty* and *WiSeNet* have laid the foundations of sense-aware knowledge extraction, they suffer from a number of shortcomings on a practical ground, mostly connected with the fact that a deeper semantic analysis is made difficult by these systems' attempts to cope with data sparsity and noisy extractions, even with encyclopedic Wikipedia text as target corpus (cf. Sections 3.2.1.2 and 3.2.2.2). For instance, *WiSeNet*'s identification of argument pairs is limited to hyperlinked Wikipedia entities, while relation phrases are clustered but not taxonomized; on the other hand, *Patty*'s subsumption taxonomy for relations is solely based on soft set inclusion principles, and only a relatively small subset of its large collection of relation patterns can be taxonomized with high confidence.

In addition, both *Patty* and *WiSeNet* produce their own, isolated OIE-derived knowledge bases: even if such knowledge bases are equipped with explicitly 'semantic' arguments and a partial ontological structure, there is no way of discovering whether, e.g., they have extracted the same relation triple or, for that matter, they have discovered the same semantic relation. Broadly speaking, any kind of interaction among OIE-derived knowledge bases generally requires manual inspection, even when they have been constructed from the same input corpus.

In the present chapter we address these and other limitations of previous approaches by taking Semantically Informed OIE to the next level, and showing how sense-level information can be leveraged to extract, ontologize, align and unify relational knowledge. Our analysis consists of three parts, organized as follows:

1. In Section 5.1 we investigate how to integrate a sense-aware approach into a full-edged OIE pipeline by moving to the denser, virtually noise-free setting of definitional text. In this scenario, we show that a comprehensive semantic analysis yields unambiguous relation triples, as well as 'semantic' relations that can be effectively arranged in a relation taxonomy;
2. Section 5.2, instead, addresses the issue of merging and harmonizing OIE-derived knowledge bases. We show that a sense-aware semantic analysis enables to interconnect not only lexical knowledge, but also relational knowledge, even when drawn from a set of very heterogeneous resources;
3. Finally, in Section 5.3 we demonstrate that OIE-derived knowledge, when properly 'semantic', can be leveraged in the more constrained IE setting of supervised hypernym discovery. In fact, working at the sense level in this scenario enables very heterogeneous knowledge to be utilized seamlessly as training data.

Throughout the three stages of our analysis, as in Chapter 4, we rely on *BabelNet* (Section 2.1.3) as a fundamental backbone and reference sense inventory. Indeed, we share with Chapter 4 the goal of developing sense-aware approaches that are both flexible and scalable. With the present chapter, not only do we employ *BabelNet* as sense inventory for disambiguation, but we also make explicit use of the structured knowledge it provides in a number of different circumstances: for instance, we exploit taxonomic information for concepts and named entities in Section 5.1, while we take advantage of *BabelNet*'s inter-resource mappings in Sections 5.2 and 5.3. Moreover, we utilize extensively *BabelNet*-powered tools like *Babelify* (Section 2.2.2.3), *Nasari* (Section 2.2.3.3), and *SensEmbed* (Section 2.2.3.2).

## 5.1 DefIE: Open Information Extraction from Definitions

As we discussed in Section 3.2, Semantically Informed OIE has shown that integrating a deeper linguistic analysis into an OIE pipeline, traditionally limited to surface-text dependencies, is key for obtaining high-quality extractions. Indeed, relation triples with explicit semantic information are able to generalize over synonymous relation phrases, as well as to reduce lexical ambiguities. However, ambiguity issues have not yet been addressed in their entirety. While arguments are typed and linked in both Patty (Section 3.2.1) and WiSeNet (Section 3.2.2), relation phrases are still bound to surface text and lack actual semantic content. Furthermore, attaching a clear ontological structure to a set of extracted patterns is not trivial, and typically requires additional processing steps, such as pattern subsumption mining (Nakashole et al., 2012), statistical inference mapping (Dutta et al., 2014), graph-based alignment (Grycner and Weikum, 2014), or collective probabilistic programming (Grycner et al., 2015), in order to obtain satisfactory results.

A limiting factor for the performance of Semantically Informed OIE approaches, emerged throughout Section 3.2, is that noise and sparsity in the input text make it difficult to enforce a comprehensive semantic analysis at both extraction and ontologization time. In fact, in most cases, the semantic characterization of a relation (or relation synset) is completely dependent on the semantics of its argument set, and only a sufficient number of extractions would provide reliable semantic types. An appropriate modeling of semantic types (e.g. selectional preferences) constitutes a line of research by itself, rooted in earlier works (Resnik, 1996) and focused on either class-based (Clark and Weir, 2002), or similarity-based (Erk, 2007), approaches. However, these methods do not fit our needs, as they model the semantics of verbs rather than arbitrary patterns. More recently some strategies based on topic modeling have also been proposed, either to infer latent relation semantic types from OIE relations (Ritter et al., 2010), or to directly learn an ontological structure from a starting set of relation instances (Movshovitz-Attias and Cohen, 2015). However, the knowledge they generate is often hard to interpret and integrate with existing knowledge resources without human intervention (Ritter et al., 2010).

In light of all the above, our strategy is to leverage a full syntactic and semantic analysis, similarly to previous Semantically Informed OIE approaches, while moving from large-scale open and noisy texts to smaller corpora of dense definitional knowledge. In this setting, which is virtually noise-free and mostly composed of concise prescriptive text, we are not forced to impose a series of constraints to cope with noisy data or difficult extractions (due to, e.g., relative clauses or coreference), and we are able to design a full-edged OIE pipeline aimed at extracting as much information as possible by unifying syntactic analysis and joint WSD/EL on textual definitions. As a trade-off, such a system is quasi-OIE, as it is limited to text having definitional nature. In fact, this definition-specific sense-aware quasi-OIE approach, named DefIE (Delli Bovi et al., 2015b),<sup>1</sup> takes as input a corpus of textual definitions and harvests fully disambiguated relation instances (i.e. relation instances where both the argument pairs and the relation phrases include sense-level information), which

---

<sup>1</sup><http://lcl.uniroma1.it/defie>

are then integrated automatically into a high-quality taxonomy of semantic relations. By running DefIE on the same definitional corpus we built for SenseDefs in Section 4.3.1, which comprises 4.3 million textual definitions, we obtain a large-scale OIE-derived knowledge base with over 20 million relation instances, 250,000 distinct relations and almost 2.4 million concepts and entities involved, showing very competitive accuracy and coverage in comparison with state-of-the-art OIE systems based on much larger corpora, including *Patty* and *WiSeNet*.

The following sections are organized as follows: we first give the details of DefIE pipeline, which comprises three successive stages: relation extraction (Section 5.1.1), relation refinement via typing and scoring (Section 5.1.2), and relation taxonomization (Section 5.1.3). We then carry out an extensive experimental evaluation of DefIE (Section 5.1.4), where we assess the quality, coverage and novelty of the extracted knowledge, and we study the impact of the various components of the pipeline on the overall performance of DefIE. In Section 5.1.4.6 we explore the effectiveness of DefIE in providing semantic labels to unlabeled edges across the semantic network of BabelNet; finally, we investigate in Section 5.1.4.7 how to further improve DefIE's extractions by utilizing the sense annotations from SenseDefs (Section 4.3), which are computed from the same definitional corpus.

### 5.1.1 Relation Extraction

The first stage of the DefIE pipeline is the extraction stage, where the input corpus is processed definition-wise, and a set of semantic relations is obtained as output. As stated at the beginning of this section, each semantic relation built by DefIE at this stage is composed of fully disambiguated relation instances, i.e. relation instances where both  $a_s$ ;  $a_o$  and (ideally all) the content words appearing in  $r$  identify word senses or named entity mentions linked to the sense inventory of BabelNet.

Compared to the approaches in Section 3.2, where Lexical Semantics is mostly modeled in the ontologization phase, DefIE addresses polysemy and synonymy directly at extraction time, by performing WSD/EL on each target definition. In fact, incorporating explicit sense-level content in the relation patterns makes them less ambiguous without resorting to their arguments' semantics; at the same time, it also generalizes over specific lexicalizations of their content words, merging together many synonymous relation patterns without ad-hoc clustering strategies.

The extraction process is depicted in Figure 5.1. Each definition is first parsed and disambiguated (Figure 5.1a-b), and then syntactic and semantic information is combined into a structured graph representation (Figure 5.1c, Section 5.1.1.1). Rather than using plain syntactic dependencies, DefIE injects explicit semantics into the dependency graph of a target definition, in order to generate a unified syntactic-semantic graph.<sup>3</sup> Finally, this syntactic-semantic graph is used to extract relation patterns as shortest paths between concept or entity pairs (Section 5.1.1.2).

<sup>2</sup>We refer to the definition of relation instance given in equation (2.3) of Section 2.4.

<sup>3</sup>Similar graphs have been proposed for a number of tasks (Lao et al., 2012; Moro et al., 2013), showing the effectiveness of unifying syntactic and semantic information, but, to the best of our knowledge, never applied in an OIE setting. They also share some similarities with the recent Abstract Meaning Representation formalism (Banarescu et al., 2013), which however provides a purely semantic structure abstracting away from many syntactic idiosyncrasies.

Figure 5.1. Example of syntactic-semantic graph construction from the textual definition  $d = \text{Atom Heart Mother is the fifth album by English band Pink Floyd}$ . Semantic nodes and regular syntactic nodes in  $G_d^{\text{sem}}$  are marked in grey and white, respectively.

### 5.1.1.1 Constructing Syntactic-Semantic Graphs

The first and foremost step of the extraction process consists in parsing and disambiguating a given definition  $d$  to obtain syntactic information, i.e. a dependency graph  $G_d$  (Figure 5.1a), and semantic information, i.e. a sense mapping  $S_d$  from surface text to word senses and named entities mentions drawn from BabelNet (Figure 5.1b). In Delli Bovi et al. (2015b) parsing is carried out using C&C (Clark and Curran, 2007), a log-linear parser based on Combinatory Categorical Grammar, or CCG<sup>4</sup>, while disambiguation is based on Babelify (Section 2.2.2.3).

The information extracted by parsing and disambiguating  $d$  is then unified into a syntactic-semantic graph  $G_d^{\text{sem}}$  where concepts and named entities identified and are arranged in a graph structure encoding their syntactic dependencies (Figure 5.1c). Given a dependency graph  $G_d$  for  $d$ , semantic information from the sense mappings  $S_d$  could be incorporated directly in the vertices of  $G_d$  by attaching available matches between words and senses to the corresponding vertices. Dependency graphs, however, encode dependencies solely on a word basis, while our sense mappings may include multi-word expressions (e.g.  $\text{Pink Floyd}_h^1$ ,  $\text{Atom Heart Mother}_h^1$ ). In order to extract consistent information, subsets of vertices referring to the same concept or entity are merged to a single semantic node  $e$  which replaces the subgraph covered in the original dependency structure. In Figure 5.1,  $\text{Pink Floyd}_h^1$  covers two distinct and connected vertices in the dependency graph  $G_d$ , one for the noun  $\text{Floyd}$  and one for its modifier  $\text{Pink}$ , and in the actual semantics of the sentence, encoded in  $G_d^{\text{sem}}$ , these two vertices are merged to a single node referring to  $\text{Pink Floyd}_h^1$  (the English rock band), instead of being assigned single-word interpretations.

Practically speaking, the procedure for building  $G_d^{\text{sem}}$  takes as input a typed dependency graph  $G_d$  and a sense mapping  $S_d$ , both extracted from a given definition  $d$ .  $G_d^{\text{sem}}$  is first populated with the vertices of  $G_d$  referring to disambiguated content words, merging those vertices covered by the same sense  $s \in S_d$  into a single node (e.g.  $\text{Pink Floyd}_h^1$  and  $\text{Atom Heart Mother}_h^1$  in Figure 5.1c). Then, the remaining vertices and edges are added as in  $G_d$ , discarding non-disambiguated adjuncts and modifiers (e.g.  $\text{the}$  and  $\text{fifth}$  in Figure 5.1c).

<sup>4</sup>CCG rules are especially suited to longer definitions and various linguistic phenomena (Steedman, 2000), such as coordinating conjunctions, that appear often across definitional text.

---

**Algorithm 1** Relation Extraction
 

---

```

procedure ExtractRelationsFrom (D)
  1: T := ;
  2: for each d in D do
  3:   Gd := dependencyParse(d)
  4:   Sd := disambiguate(d)
  5:   Gdsem := buildSemanticGraph(Gd; Sd)
  6:   for each hs; sj in Sd do
  7:     hs; rij; sj := shortestPath(si; sj)
  8:     T := T [fh si; rij; sj ig
  9: filterPatterns (T; )
return T;
  
```

---

### 5.1.1.2 Identifying Relation Patterns

After constructing a syntactic-semantic graph  $G_d^{\text{sem}}$  for a definition  $d$ , DefIE considers every pair of identified concepts or named entities across the graph and extract the relation pattern  $r$  between them as the shortest path between the two corresponding vertices in  $G_d^{\text{sem}}$ . This enables us to exclude less relevant information, typically carried by adjuncts or modifiers. The shortest path is computed using the Floyd-Warshall algorithm (Floyd, 1962), and the only syntactic constraint that we enforce on the resulting path is that it must include at least one verb node as in ReVerb (Fader et al., 2011). This condition filters out meaningless single-node patterns (e.g. two concepts connected with a preposition) and, given the prescriptive nature of  $d$ , is unlikely to discard semantically relevant attributes compacted in noun phrases. As an example, consider the two sentences ‘Mutter is the third album by German band Rammstein’ and ‘Atom Heart Mother is the fifth album by English band Pink Floyd’. In both cases, two valid shortest-path patterns are extracted:

$$X \text{ ! is ! album}_h^1 \text{ ! by ! } Y$$

with  $X = \text{Mutter}_n^3$ ,  $Y = \text{Rammstein}_h^1$  in the first sentence and  $X = \text{Atom Heart Mother}_h^1$ ,  $Y = \text{Pink Floyd}_h^1$  in the second one, and:

$$X \text{ ! is ! } Y$$

with  $X = \text{Mutter}_n^3$ ,  $Y = \text{album}_h^1$  in the first sentence and  $X = \text{Atom Heart Mother}_h^1$ ,  $Y = \text{album}_h^1$  in the second one. Thanks to joint WSD and EL (Section 5.1.1.1), DefIE discovers general knowledge (e.g. that  $\text{Mutter}_n^3$  and  $\text{Atom Heart Mother}_h^1$  are instances of the concept  $\text{album}_h^1$ ) and, at the same time, relational facts (encoded in both cases with the relation pattern ‘is album<sub>h</sub><sup>1</sup> by’).

A pseudo-code for DefIE’s extraction stage is shown in Algorithm 1. Each  $d \in D$  is first parsed and disambiguated to produce a syntactic-semantic graph  $G_d^{\text{sem}}$  (Section 5.1.1.1); then all the named entity/concept pairs  $h_s; s_j$  are examined to detect relation instances as shortest paths. Finally, all relations for which the number of extracted instances is below a fixed threshold is filtered out.<sup>5</sup>

<sup>5</sup>In all the experiments of Section 5.1.4 we set  $\tau = 10$ , empirically validated on a small held-out set of manually annotated definitions.

Relation pattern of $r$	$\text{score}(r)$	$H_r$
X directed by Y	4,025.80	1.74
X known for Y	2,590.70	3.65
X is election district <sup>1</sup> of Y	110.49	0.83
X is composed <sup>1</sup> from Y	39.92	2.08
X is street <sup>1</sup> named after Y	1.91	2.24
X is village <sup>2</sup> founded in 1912 in Y	0.91	0.18

Table 5.1. Some examples of relation scoring and corresponding entropy  $H_r$  (third column).

### 5.1.2 Relation Typing and Scoring

After the extraction stage, we further characterize and refine the semantics of DefIE's relations by computing semantic type signatures for each  $r \in T$ , i.e. by attaching a proper semantic class to both its domain and range (cf. Section 2.4). Since every element in the domain and range of  $r$  is disambiguated, we retrieve the corresponding Babel synsets and collect their direct hypernyms from the taxonomy of BabelNet (Section 2.1.3). We then select the hypernym covering the largest subset of arguments as the representative semantic class for the domain (or range) of  $r$ . By leveraging the distribution of direct hypernyms over domain and range arguments of  $r$ , we estimate the quality of  $r$  and associate a confidence value with its relation pattern  $r$ . Intuitively we want to assign higher confidence to relations where the corresponding distributions have low entropy<sup>6</sup>. For each relation  $r$ , we compute:

$$H_r = - \sum_{i=1}^n p(h_i) \log_2 p(h_i) \quad (5.1)$$

where  $h_i$  ( $i = 1; \dots; n$ ) are all the distinct argument hypernyms over the domain and range of  $r$ , and probabilities  $p(h_i)$  are estimated from the proportion of arguments covered in such sets. The lower  $H_r$ , the better semantic types of  $r$  are defined. As a matter of fact, however, some valid but over-general relations (e.g. 'is a', 'is used for') have inherently high values of  $H_r$ . To obtain a balanced score, we therefore consider two additional factors, i.e. the number of extracted instances for  $r$  and the length of the associated pattern  $r$ , obtaining the following empirical measure:

$$\text{score}(r) = \frac{|r|}{(H_r + 1) \text{length}(r)} \quad (5.2)$$

The +1 term accounts for cases where  $H_r = 0$ . As shown in the examples of Table 5.1, relations with rather general patterns (such as 'known for') achieve higher scores compared to very specific ones (e.g. 'village<sup>2</sup> founded in 1912 in') despite higher entropy values. We validated our measure on the samples of Section 5.1.4.1, computing the overall precision for different score thresholds. The monotonic decrease of sample precision in Figure 5.2a shows that our measure captures the quality of extracted patterns better than  $H_r$  (Figure 5.2b).

<sup>6</sup>For instance, if both sets have a single hypernym covering all arguments, then  $r$  arguably captures a well-defined semantic relation and should be assigned high confidence.

Figure 5.2. Average precision vs.  $\text{score}(r)$  (a) and  $H_r$  (b) on the sample of Section 5.1.4.1.

### 5.1.3 Relation Taxonomization

In the last stage of the pipeline, the set of extracted and refined relations is arranged automatically in a relation taxonomy. The process is carried out by comparing relations pairwise, looking for hypernymic relationships between the corresponding patterns; the final taxonomy is then built by connecting with an edge those relation pairs for which such a relationship is found. We adopt two straightforward methods to detect hypernymic relationships, both of which examine noun nodes across each relation pattern  $r$ , and consider for taxonomization only those relations whose patterns are identical except for a single noun node<sup>7</sup>.

**Hypernym Generalization.** A way of identifying hypernym/hyponym noun nodes across relation patterns is to analyze the sense-level information attached to them. Given two relation patterns  $r_i$  and  $r_j$ , differing only in the noun nodes  $n_i$  and  $n_j$ , we retrieve the hypernym sets  $H(c_i)$  and  $H(c_j)$ , of the associated synsets,  $c_i$  and  $c_j$ . Hypernym sets are obtained by iteratively collecting the superclasses of  $c_i$  and  $c_j$  from the semantic network of BabelNet, up to a fixed height. For instance, given  $c_i = \text{album}_h^1$ ,  $H(c_i) = \{ \text{work of art}_h^1, \text{creator}_h^2, \text{artifact}_h^1 \}$ . Once we have  $H(c_i)$  and  $H(c_j)$ , we just check whether  $c_j \in H(c_i)$  or  $c_i \in H(c_j)$  (Figure 5.3a). According to which is the case, we conclude that  $r_j$  is a generalization of  $r_i$ , or vice versa.

<sup>7</sup>The simplifying assumption we exploit here is that two given relation patterns may be in a hypernymy-hyponymy relationship only when their plain syntactic structure is equivalent (e.g.  $\text{` is } N_1 \text{ by'}$  and  $\text{` is } N_2 \text{ by'}$ , with  $N_1$  and  $N_2$  being two distinct noun nodes).

Figure 5.3. Hypernym (a) and substring (b) generalization of relation patterns.

	DefIE	Nell	Patty	ReVerb	WiSeNet	Freebase	DBpedia
# Relations	255,881	298	1,631,531	664,746	245,935	1,894	1,368
Avg. Extractions	81.68	7,013.03	9.68	22.16	9.24	127,727.99	24,451.48
# Rel. Instances	20,352,903	2,089,883	15,802,946	14,728,268	2,271,807	241,897,882	33,449,631
# Senses	2,398,982	1,996,021	1,087,907	3,327,425	1,636,307	66,988,232	10,338,501

Table 5.2. Comparative statistics on the relation extraction process, including the number of distinct relations (# Relations), the average number of extractions per relation (Avg. Extractions), the number of relation instances (# Rel. Instances), and the number of distinct concepts or named entities involved (# Senses).

**Substring Generalization.** The second procedure focuses on the noun (or compound) represented by the node. Given two relation patterns,  $r_i$  and  $r_j$ , we apply the following heuristic: from one of the two nouns, be it  $n_i$ , any adjunct or modifier is removed, retaining the sole head word  $\hat{n}_i$ . Then,  $\hat{n}_i$  is compared with  $n_j$  and, if  $\hat{n}_i = n_j$ , we assume that the relation  $r_j$  is a generalization of  $r_i$  (Figure 5.3b).

#### 5.1.4 Experimental Evaluation

General statistics on DefIE's extraction process are shown in Table 5.2, and compared with other prominent OIE approaches, each of which is considered in the setting detailed below (Experimental Setup). Even though no direct quality comparison is possible at this stage, as these OIE approaches are run on different source corpora and evaluated differently, the reported figures highlight some interesting differences in the nature of each extraction process. In particular, DefIE extracts 20,352,903 relation instances, out of which 13,753,133 feature a disambiguated pattern, with an average of 3.15 disambiguated relation instances extracted from each definition. The resulting knowledge base comprises 255,881 distinct semantic relations, 94% of which also have disambiguated content words in their patterns. DefIE extracts a considerably larger amount of relation instances compared to similar approaches, despite the much smaller amount of text used. For example, we managed to harvest over 5 million relation instances more than Patty, using a much smaller corpus (single sentences as opposed to full Wikipedia articles) and generating a number of distinct relations that was six times less than Patty's. As a result, we obtained an average number of extractions that was substantially higher than those of other OIE methods, which reflects the fact that DefIE, by stripping away syntactic modifiers (Section 5.1.1.1) and replacing synonymous words with their synset identifiers, generalizes over relation patterns. Furthermore, our semantic analysis captured 2,398,982 distinct arguments (either concept or named entities), outperforming almost all open-text systems examined.

**Experimental Setup.** All the manual evaluations carried out in the following sections were based on two human judges, with an inter-annotator agreement, as measured by Cohen's kappa coefficient (Cohen, 1960), above 70% in all cases. In these evaluations we compared DefIE with the following approaches:

- ^ Nell (Carlson et al., 2010) with beliefs updated to November 2014;

	Top 100		Top 250		Rand 100		Rand 250	
DefIE	0:93	0:01	0:91	0:02	0:79	0:02	0:81	0:08
Patty	0:93	0:05	N/A		0:80	0:08	N/A	

Table 5.3. Precision of relation patterns.

- ^ Patty (Nakashole et al., 2012) with Freebase types and pattern synsets from the English Wikipedia dump of June 2011 (cf. Section 3.2.1.2);
- ^ ReVerb (Fader et al., 2011), using the set of normalized relation instances from the ClueWeb09 dataset;
- ^ WiSeNet (Moro and Navigli, 2012, 2013) with relational phrases from the English Wikipedia dump of December 2012 (cf. Section 3.2.2.2).

In addition, we also compared DefIE's knowledge base with human-contributed resources, namely Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2014), both from the dumps of April/May 2014.

#### 5.1.4.1 Quality of the Relations

**Relation Precision.** We first assessed the quality and the semantic consistency of our relations using manual evaluation, along the lines of previous approaches (Sections 3.2.1.2 and 3.2.2.2). We ranked our relations according to their score (Section 5.1.2) and then created two samples (of size 100 and 250 respectively) from the top scoring relations. In order to evaluate the long tail of less confident relations, we created another two samples of the same size with randomly extracted relations. We presented these samples to the human judges, accompanying each relation with a set of 50 argument pairs and the corresponding textual definitions from BabelNet. For each item in the sample we asked whether it represented a meaningful relation and whether the extracted argument pairs were consistent with this relation and the corresponding definitions. If the answer was positive, the relation was considered as correct. Finally we estimated the overall precision of the sample as the proportion of correct items. Results are reported in Table 5.3 and compared to those obtained by our closest competitor, Patty.<sup>8</sup> In Patty the confidence of a given pattern was estimated from its statistical strength (cf. Section 3.2.1). As shown in Table 5.3, DefIE achieved a comparable level of accuracy in every sample. An error analysis identified most errors as related to the vagueness of some short and general patterns, e.g. 'take', 'make'. Others were related to parsing (e.g. in labeling the head word of complex noun phrases) or disambiguation.

**Information Novelty.** We used the same samples to estimate the novelty of the extracted information in comparison to currently available resources. We examined each correct relation pattern and looked manually for an equivalent relation in the knowledge bases of both our OIE competitors and human-contributed resources.

<sup>8</sup>Nakashole et al. (2012) only report Patty's precision figures on the samples of size 100.

	Nell	Patty	ReVerb	WiSeNet	Freebase	DBpedia
Top 100	0.571	0.238	0.214	0.155	0.571	0.461
Rand 100	0.942	0.711	0.596	0.635	0.904	0.880

Table 5.4. Novelty of the extracted information.

Gold Standard	DefIE	WiSeNet	Patty
163	131	129	126
	ReVerb	Freebase	DBpedia
	122	69	39

Table 5.5. Coverage of semantic relations.

For instance, given the relation 'born in', Nell and ReVerb have the equivalent relations 'personborninlocation' and 'is born in', while Freebase and DBpedia have 'Place of birth' and 'birthPlace' respectively. We then computed the proportion of novel relations among those previously labeled as correct by the human judges. Results are shown in Table 5.4 for both the top 100 sample and the random sample of the same size. The high proportion of relations not appearing in existing resources (especially across the random samples) suggests that DefIE is capable of discovering information not obtainable from available knowledge bases, including very specific relations ('is blizzard in', 'is Mayan language spoken by', 'is government-owned corporation in'), as well as general but unusual ones ('used by writer of').

**Relation Coverage.** To assess the coverage of DefIE we first tested our extracted relations on a public dataset described in Nakashole et al. (2012), and consisting of 163 semantic relations manually annotated from five Wikipages about musicians. Following the line of previous works (Nakashole et al., 2012; Moro and Navigli, 2013), for each annotation we sought a relation in DefIE's knowledge base carrying the same semantics. Results are reported in Table 5.5. Consistently with the results in Table 5.4, the proportion of novel information places DefIE in line with its closest competitors, achieving a coverage of 80.3% with respect to the gold standard. Examples of relations not covered by DefIE's competitors are 'hasFatherInLaw' and 'hasDaughterInLaw'. Furthermore, relations holding between entities and general concepts (e.g. 'critizedFor', 'praisedFor', 'sentencedTo'), are captured only by DefIE and ReVerb (which, however, lacks any argument semantics). To complement this experiment, we also assessed manually the coverage of resources based on human-defined semantic relations by extracting three random samples of 100 relations from Freebase, DBpedia and Nell and looking for semantically equivalent relations in our knowledge base. In this setting, DefIE reports a coverage of 83%, 81% and 89% respectively, failing to cover mostly relations that refer to numerical properties (e.g. 'numberOfMembers'). Finally, we tested the coverage over individual relation instances: we selected a random sample of 100 triples from the two closest competitors exploiting textual corpora, i.e. Patty and WiSeNet and, for each selected triple, we sought an equivalent relation instance in DefIE's knowledge

	Hyp. Gen.		Substr. Gen.		Patty (Top)		Patty (Rand)	
Precision	0:87	0:03	0:90	0:02	0:85	0:07	0:62	0:09
# Edges	44,412				20,339			
Density	1:89 $10^{-6}$				7:64 $10^{-9}$			

Table 5.6. Precision and coverage of the relation taxonomy.

base. This experiment showed a coverage greater than 65% over both samples (66% and 69% on `Patty` and `WiSeNet`, respectively) which, given the dramatic reduction of corpus size, indicates that definitional knowledge can be extremely valuable for relation extraction approaches. Of course, these manual assessments are necessarily carried out on a small scale and affected by subjectivity: intuitively, many relations and relation instances still remain out of reach for approaches limited to definitional knowledge. However, these results might suggest that, even in large-scale OIE-derived resources, a substantial amount of knowledge is likely to come from a rather smaller subset of definitional sentences within the source corpus.

#### 5.1.4.2 Quality of the Relation Taxonomy

We evaluated `DefIE`'s relation taxonomy by manually assessing the accuracy of both taxonomization procedures described in Section 5.1.3. We compared the results with those of `Patty`, the only system among our closest competitors that generates a taxonomy of relations. The setting for this evaluation was the same of that of Section 5.1.4.1. However, as we lacked a confidence measure in this case, we just extracted a random sample of 200 hypernym edges for each generalization procedure. We presented these samples to our human judges and, for each hypernym edge, we asked whether the corresponding pair of relations represented a correct generalization. We then estimated the overall precision as the proportion of correct edges. Results are reported in Table 5.6, along with `Patty`'s results; as `Patty`'s edges are ranked by confidence, we considered both its top confident 100 subsumptions and a random sample of the same size. Even though no conclusive quality comparison can be made, as `DefIE` and `Patty` are run on different data, Table 5.6 shows that `DefIE` outperforms `Patty` in terms of precision, and generates more than twice the number of edges overall. As mentioned in Section 3.2.1.2, `Harpy` (Grycner and Weikum, 2014) enriches `Patty`'s taxonomy with 616,792 hypernym edges, but its alignment algorithm also includes transitive edges and still yields a sparser taxonomy compared to ours, with a graph density of  $2:32 \cdot 10^{-7}$ .

#### 5.1.4.3 Quality of Entity Linking and Disambiguation

We evaluated the disambiguation quality of `DefIE` by comparing `Babelfy` against other state-of-the-art EL systems (cf. Section 2.2.2). To set a level playing field, we selected a random sample of 60,000 glosses from the input corpus of textual definitions, and ran the relation extraction step (Section 5.1.1) using a different competitor in the disambiguation step each time. We then used the mappings in `BabelNet` to express each output using a common dictionary and sense inventory.

	# Relations	# Triples	# Entities	Sem. Nodes
Babelify	96,434	233,517	79,998	2.37
TagME 2.0	88,638	226,905	89,318	1.67
WAT	24,083	56,503	38,147	0.39
DBpedia Spotlight	67,377	140,711	38,254	1.45
Wikipedia Miner	39,547	88,777	37,036	0.96

  

	# Relations	# Relation instances
Babelify	82.3%	76.6%
TagME 2.0	76.0%	62.0%
WAT	84.6%	72.6%
DBpedia Spotlight	70.5%	62.6%
Wikipedia Miner	71.7%	56.0%

Table 5.7. Coverage (top) and precision (bottom) for different disambiguation systems.

The coverage obtained by each competitor was assessed by looking at the number of distinct relations extracted in the process, the total number of relation instances extracted, the number of distinct concepts or entities involved, and the average number of semantic nodes within the relation patterns. For each competitor, we also assessed the precision obtained by evaluating the quality and semantic consistency of the relation patterns, in the same manner as in Section 5.1.4.1, both at the level of semantic relations (on the top 150 relation patterns) and at the level of individual relation instances (on a randomly extracted sample of 150 triples). Results are shown in Table 5.7 for Babelify and the following systems: TagME (Ferragina and Scaiella, 2012),<sup>9</sup> WAT (Piccinno and Ferragina, 2014), DBpedia Spotlight (Mendes et al., 2011),<sup>10</sup> and Wikipedia Miner (Milne and Witten, 2013),<sup>11</sup>. As shown in top part of Table 5.7, Babelify outperforms all its competitors in terms of coverage and, due to its unified WSD/EL approach, extracts semantically richer patterns with 2.37 semantic nodes on the average per sentence. This reflects on the quality of semantic relations, reported in the bottom part of Table 5.7, with an overall increase of precision in terms of both relations and relation instances; even though WAT shows slightly higher precision over relations, its considerably lower coverage yields semantically poor patterns (0.39 semantic nodes on the average) and impacts on the overall quality of relations, where some ambiguity is necessarily retained. As an example, the pattern `is station in`, extracted from WAT's disambiguation output, covers both railway stations and radio broadcasts. Babelify produces, instead, two distinct relation patterns for each sense, tagging `station` as `railway station`<sub>1</sub> for the former and `station`<sub>5</sub> for the latter.

#### 5.1.4.4 Impact of Definition Sources

Given the heterogeneous input corpus in our experimental setup, we carried out an empirical analysis to study the impact of each source of textual definitions in isolation. The leftmost column of Table 5.8 shows the composition of the input corpus with

<sup>9</sup>tagme.di.unipi.it

<sup>10</sup>spotlight.dbpedia.org

<sup>11</sup>wikipediadataminer.cms.waikato.ac.nz

	# Definitions	# Relations	# Relation instances	Avg. Extractions
Wikipedia	3,899,087	251,954	19,455,992	77.58
Wikidata	364,484	5,414	1,033,732	191.01
WordNet	41,356	2,260	128,200	56.73
Wiktionary	39,383	2,863	143,990	50.52
OmegaWiki	13,017	1,168	45,818	39.45

Table 5.8. Impact of each definition source on relation extraction.

	# Wikipages	# Sentences	# Extractions	Precision
All	14,072	225,867	39,684	61.8%
Top 100	10,334	161,769	13,687	59.0%

Table 5.9. Extraction results of DefIE over non-definitional text.

respect to each of these definition sources. The distribution is rather skewed, with the vast majority of definitions coming from Wikipedia (almost 90% of the input corpus). We ran the relation extraction step (Section 5.1.1) on each subset of the input corpus. Results, as shown in Table 5.8, are consistent with the composition of the input corpus: by relying solely on Wikipedia's first sentences, the extraction algorithm discovered 98% of all the distinct relations identified across the whole input corpus, and 93% of the total number of extracted instances. Wikidata provides more than 1 million extractions (5% of the total) but definitions are rather short and most of them (44.2%) generate only is-a relation instances. The remaining sources (WordNet, Wiktionary, OmegaWiki) account for less than 2% of the extractions.

#### 5.1.4.5 Impact of the Approach vs. Impact of the Data

DefIE is explicitly designed to target textual definitions. Hence, the result it achieves is due to the mutual contribution of two key features: an OIE approach and the use of definitional data. In order to decouple these two factors and study their respective impacts, we carried out two experiments: in the first we applied DefIE to a sample of non-definitional text; in the second we applied our closest competitor, Patty, on the same definitional corpus used as input for DefIE in the previous experiments.

**Extraction from non-definitional text.** We selected a random sample of Wikipages from the English Wikipedia dump of October 2012. We processed each sentence as in Section 5.1.1, and extracted instances of those relations produced by DefIE in the original definitional setting; we then automatically filtered out those instances where the arguments' hypernyms did not agree with the semantic types of the relation. We evaluated manually the quality of extractions on a sample of 100 items (as in Section 5.1.4.1) for both the full set of extracted instances and for the subset of extractions from the top 100 scoring relations. Results are reported in Table 5.9: in both cases, precision figures show that relation quality drops consistently in comparison to Section 5.1.4.1, suggesting that DefIE by itself is less accurate when moving to more complex sentences (with, e.g., subordinate clauses or coreferences).

	# Relation instances	# Relations	# Edges
Patty (definitions)	3,212,065	41,593	4,785
Patty (Wikipedia)	15,802,946	1,631,531	20,339
DefIE	20,807,732	255,881	44,412

Table 5.10. Performance of Patty on definitional text.

Source	Label	Target
enzyme <sub>n</sub>	catalyzes reaction of	chemical <sub>n</sub>
album <sub>n</sub>	recorded by	rock group <sub>n</sub>
officer <sub>n</sub>	commanded brigade of	army unit <sub>n</sub>
bridge <sub>n</sub>	crosses over	river <sub>n</sub>
academic journal <sub>n</sub>	covers research in	science <sub>n</sub>
organization <sub>n</sub>	has headquarters in	city <sub>n</sub>

Table 5.11. Examples of labeled edges in BabelNet derived from DefIE.

Patty on textual definitions. We implemented a version of Patty based on Babelfy for disambiguation. We then ran it on our corpus of BabelNet definitions and compared the results against those originally obtained by Patty on the entire Wikipedia corpus (cf. Section 3.2.1.2) and those obtained by DefIE. Figures are reported in Table 5.10 in terms of number of extracted relation instances, distinct relations and hypernym edges in the relation taxonomy, show that the dramatic reduction of corpus size affects the support sets of Patty's relations, worsening both coverage and generalization capability.

#### 5.1.4.6 Preliminary Study: Knowledge Resource Enrichment

As a preliminary study, we explored the application of DefIE to the enrichment of existing resources. We focused on BabelNet as a case study. In BabelNet's semantic network, nodes representing concepts and entities are only connected via lexicographic relationships from WordNet (hypernymy, meronymy, etc.), Wikidata relations, or unlabeled edges derived from Wikipedia hyperlinks (cf. Section 2.1.3). DefIE has the potential to provide useful information to both augment unlabeled edges with labels and explicit semantic content, and create additional connections based on novel semantic relations. Some examples are shown in Table 5.11. We carried out a quantitative analysis using all disambiguated relations with at least 10 extracted instances. For each relation pattern  $r$ , we first examined the pair of semantic classes associated with its type signatures and looked in BabelNet for an unlabeled edge connecting the pair. Then we examined the whole set of extracted relation instances in  $r$  and looked in BabelNet for an unlabeled edge connecting the arguments  $a_s$  and  $a_o$ . We found that only 27.7% of the concept pairs representing relation type signatures are connected in BabelNet (as of version 2.5), and most of these connections are unlabeled. By the same token, more than 4 million distinct argument pairs (53.5%) do not share any edge in the semantic network and, among those that do, less than 14% have a labeled relationship. These proportions suggest

that DefIE 's relations (but, more in general, properly 'semantified' OIE-derived relations) are able to enrich substantially the underlying knowledge resource in terms of both connectivity and labeling of existing edges.

#### 5.1.4.7 DefIE on SenseDefs

Since most of the pipeline of DefIE builds upon sense-level information, having high-quality disambiguations when processing a textual definition is of utmost importance: poor context of particularly short definitions may introduce disambiguation errors at preprocessing time, which tend to propagate and then reflect on the extraction of both relations and relation instances. While the relation extraction stage, as described in Section 5.1.1, assumes a generic input corpus of definitional knowledge, and processes it definition by definition, in the experimental evaluation we relied, instead, on the heterogeneous input corpus of definitions drawn from BabelNet that we built to develop SenseDefs (Section 4.3). With SenseDefs we indeed showed that exploiting the nature of the target corpus leads to a more structured and effective disambiguation strategy. In order to investigate the impact of this strategy on the relational knowledge extracted by DefIE, we adapted its pipeline to consider SenseDefs as target corpus, and evaluated the results obtained at the end of the pipeline in terms of quality of relation and relation instances.

We first selected a random sample of 150 textual definitions from the high-coverage version of SenseDefs. We generated a baseline for the experiment by discarding all disambiguated instances from the sample, and treating the sample itself as an unstructured collection of textual definitions which we used as input for DefIE, letting the original pipeline of the system perform the disambiguation step. Then we carried out the same procedure using a modified implementation of DefIE that takes into account SenseDefs 's disambiguated instances of a target definition instead of disambiguating it from scratch. In both cases, we evaluated the output in terms of both relations and relation instances. Following previous experiments (Section 5.1.4), we employed two human judges, and performed the same evaluation procedure described therein over the set of distinct relations extracted from the sample, as well as the set of extracted relation instances.

Results reported in the top part of Table 5.12 show a slight but consistent improvement on coverage that results from using SenseDefs in place of the original corpus of definitions, in terms of extracted relations, extracted triples, and number of glosses with at least one extraction. Similarly, SenseDefs also improves the estimated precision of such extractions, as shown in the bottom part of Table 5.12. The joint disambiguation of glosses across resources and languages enabled the extraction of 6.5% additional instances from the sample (2.26 extractions on the average from each definition) and, at the same time, increased the estimated precision of relation and relation instances over the sample by about 1%.

**Final Remarks.** The gist of DefIE lies in its comprehensive syntactic-semantic analysis targeted to textual definitions. In contrast to many competitors, where syntactic constraints are necessary in order to keep precision high when dealing with noisy data (cf. Section 3.2), DefIE shows comparable (or greater) performances by exploiting a dense, noise-free definitional setting to generate a large OIE-derived

	# De nitions	# Relation Instances	# Relations
DefIE + SenseDefs	150	340	184
DefIE	146	318	171

  

	Relation	Relation Instances
DefIE + SenseDefs	0.872	0.780
DefIE	0.865	0.770

Table 5.12. Extractions (top) and precision (bottom) of DefIE on the evaluation sample.

knowledge base, in line with prominent OIE systems, derived from a much smaller amount of input data. The target corpus of de nitions used by DefIE comprises less than 83 million tokens overall, while other OIE systems exploit massive corpora like Wikipedia (typically more than 1.5 billion tokens), ClueWeb (more than 33 billion tokens), or the Web itself. Crucially, the experiments in Sections 5.1.4.4 and 5.1.4.5 demonstrate that the performances of DefIE result from the interplay between a fully sense-aware quasi-OIE approach and a target text composed of de nitional knowledge: in fact, from the strict point of view of OIE, its extraction pipeline is improvable in many ways. Rather than improving OIE per se however, our objective is that of showing how reframing OIE at the sense level can effectively compensate the unavailability of large amounts of data. A clear example of this is given by relation taxonomization: while the approach of Patty is that of discovering subsumptions between semantic relations by looking at the shape of their support sets (Section 3.2.1), DefIE enforces a very restrictive assumption and focuses on very basic cases where the explicit semantic characterization of a relation pattern can be leveraged (Section 5.1.3). Hence, the accuracy of the former depends crucially on having a sufficient number of extractions for a given relation, whereas the latter works perfectly even with very rare relation patterns and depends, instead, on the quality of disambiguation. On the other hand, however, the latter also relies on the well-formed nature of de nitional text, whereas in open text the restrictiveness of its assumption could significantly hinder recall.

## 5.2 KB-Unify: Sense-Aware Knowledge Base Uni cation

Another important limitation of most OIE systems to date is the lack of interoperability. As we examined in Section 2.3.2, these systems can be very different in nature; still, they have been developed with their own type inventories, and no portable ontological structure. This issue is actually broader than OIE: distantly supervised approaches (Mintz et al., 2009; Riedel et al., 2010, 2013; Surdeanu et al., 2012; Fan et al., 2014), where noisy extractions are complemented with structured knowledge, and systems like Nell (Carlson et al., 2010), which combines a hand-crafted taxonomy of entities and relations with self-supervised large-scale extraction from the Web, require additional processing for linking and integration (Dutta et al., 2014). Even Semantically Informed OIE approaches, like Patty and WiSeNet

(Section 3.2), produce their own, isolated OIE-derived knowledge bases. In order to discover whether two OIE systems are able to extract a specific kind of semantic relation (e.g. `is a', or `located in'), as we did in the experimental evaluation of DefIE (Section 5.1.4.1), manual inspection is required, even if the source textual corpus used as input was the same for the two systems.

This is also why, in recent years, a research thread focused on Knowledge Base Completion (Nickel et al., 2012; Bordes et al., 2013; West et al., 2014) has emerged, where the aim is to integrate new knowledge into an already existing knowledge base (KB). However, beside some notable exceptions (Section 2.3.3), the majority of integration approaches nowadays are not designed to deal with many different resources at the same time.

**Integrating Knowledge Bases.** On the other hand, we discussed in Chapter 1 how the integration of knowledge drawn from different sources has received much attention over the last decade (Gurevych et al., 2016). However, while great effort has been put into aligning knowledge at the concept level, most approaches do not tackle the problem of integrating heterogeneous knowledge at the relation level, nor do they exploit effectively the huge amount of information harvested with OIE systems, even when this information is unambiguously linked to a structured resource (cf. Section 3.2). Yet, as the number of knowledge resources increases, some approaches have started addressing the task of aligning KBs: Dutta et al. (2014) describe a method for linking arguments in NELL triples to DBpedia by combining First Order Logic and Markov Networks; Grycner and Weikum (2014) semantify PATTY's pattern synsets and connect them to WordNet verbs; Lin et al. (2012) propose a method to propagate Freebase types across ReVerb and deal with the problem of unlinkable entities. All these approaches achieve very competitive results in their respective settings but, like KB completion approaches, they limit the task to one-to-one alignments. A few contributions have also tried to broaden the scope and include different resources at the same time: Riedel et al. (2013) propose a universal schema that integrates structured data with OIE data by learning latent feature vectors for entities and relations (Section 2.3.3); Knowledge Vault (Dong et al., 2014) uses a graph-based probabilistic framework where prior knowledge from existing resources (e.g. Freebase) improves Web extractions by predicting their reliability. Finally, a recent trend of research focuses on learning embedding models for structured knowledge and their application to tasks like relation extraction and KB completion (Socher et al., 2013; Weston et al., 2013; Bordes et al., 2013; Neelakantan et al., 2015).

**Motivation.** The latter integration approaches described above are very effective but still unfit to our scenario, as they are inherently based on surface-text techniques; in accordance with the objectives of this thesis, our aim is instead to bring Lexical Semantics into play. In this respect, sense-aware OIE approaches have shown their benefits over surface-text extraction, especially when we restrict the target to well-formed definitional text (Section 5.1). Following this thread, in the present section we address the issue of merging and harmonizing KBs (with a special focus on OIE-derived KBs) at the sense level we aim at showing that a sense-aware strategy

enables to interconnect not only lexical knowledge but also relational knowledge, even when drawn from a set of very heterogeneous KBs. The approach we propose, named KB-Unify (Delli Bovi et al., 2015a),<sup>12</sup> is based on the key idea of bringing together knowledge from an arbitrary number of OIE systems, regardless of whether these systems provide links to some general-purpose inventory, come with their own ad-hoc structure, or have no structure at all. Knowledge from each source, in the form of *hsubject, predicate, object* triples, is disambiguated and linked to the sense inventory of BabelNet (Section 2.1.3). This enables us to discover alignments at the sense level between relations from different KBs, and to generate a unified, fully disambiguated KB of entities and semantic relations. We detail the pipeline of KB-Unify in Section 5.2.1; then, in Section 5.2.2 we test KB-Unify experimentally on a set of four heterogeneous KBs.

### 5.2.1 Disambiguating and Unifying Knowledge Bases

KB-Unify takes as input a set of KBs  $K = \{KB_1, \dots, KB_n\}$  and outputs a single, unified and fully disambiguated KB, denoted as  $K^*$ .<sup>13</sup> Depending on the nature of each  $KB_i$ , entities in  $E_i$  might be disambiguated and linked to an external inventory (e.g. the argument *Washington* linked to the Wikipage *George Washington*), or unlinked and only available as ambiguous mentions. We can thus partition  $K$  into a subset of linked resources  $K_D$ , and one of unlinked resources  $K_U$ . In order to align very different and heterogeneous KBs at the semantic level, KB-Unify exploits:

- ^ A unified sense inventory  $S$ , which acts as a superset for the inventories of individual KBs. We choose BabelNet for this purpose: by merging complementary knowledge from different resources (e.g. Wikipedia, WordNet, Wikidata and Wiktionary, among others), BabelNet provides a wide coverage of entities and concepts whilst at the same time enabling convenient inter-resource mappings for  $KB_i$  in  $K_D$ . For instance, each Wikipage (or Wikidata item) has a corresponding synset in BabelNet, which enables a one-to-one mapping between BabelNet's synsets and entries in, e.g. DBpedia or Freebase (cf. Section 2.1.3);
- ^ A vector space model  $V_S$  that enables a semantic representation for every item in  $S$ . Current distributional models, like word embeddings (Mikolov et al., 2013c), are not suitable to our setting: they are constrained to surface word forms, and hence they inherently retain ambiguity of polysemous words and entity mentions. We thus leverage SensEmbed (Iacobacci et al., 2015), a sense-level approach to embeddings. SensEmbed is trained on a large BabelNet-annotated corpus and produces continuous representations for individual word senses (sense embeddings) according to the sense inventory of BabelNet (cf. Section 2.2.3.2).

Figure 5.4 illustrates the work ow of KB-Unify 's unification approach. Entities coming from any  $KB_i \in K_D$  can be directly (and unambiguously) mapped to the corresponding entries in  $S$  via BabelNet inter-resource linking (Figure 5.4a): in

<sup>12</sup><http://lcl.uniroma1.it/kb-unify>

<sup>13</sup>Throughout this section we refer to the definition of KB specified in Section 2.4.

Figure 5.4. Unification algorithm work ow.

Figure 5.5. Disambiguation algorithm work ow.

the above example, the argument `Washington` linked to the Wikipage `George Washington` is included in the BabelNet synset `bn:00040239n` with the word sense `Washington1`. In contrast, unlinked (and potentially ambiguous) arguments need an explicit disambiguation step (Figure 5.4b) connecting them to appropriate entries, i.e. synsets, in  $S$ : this is the case, in the above example, for the ambiguous argument `Washington` that has to be linked to either the president, the city or the state. Therefore, our approach comprises two successive stages:

- ^ A disambiguation stage, where all  $KB_i \in K$  are linked to  $S$ , either by inter-resource mapping (Figure 5.4a) or disambiguation (Figure 5.4b, Sections 5.2.1.1-5.2.1.3), and all  $E_i$  are merged into a unified set of entities  $E$ . As a result of this process we obtain a set  $K^S$  comprising all the KBs in  $K$  redefined using the common sense inventory  $S$ ;
- ^ An alignment stage (Section 5.2.1.4, Figure 5.4c) where, for each pair of KBs  $KB_i^S, KB_j^S \in K^S$ , we compare every relation pair  $\langle r_i, r_j \rangle$ ,  $r_i \in R_i^S$  and  $r_j \in R_j^S$ , in order to identify cross-resource alignments and merge relations sharing equivalent semantics into relation clusters (relation synsets). This process yields a unified set of relation synsets  $R$ . The overall result is  $KB = \langle E; R; T \rangle$ , where  $T$  is the set of all disambiguated triples redefined over  $E$  and  $R$ .

**Disambiguating a Knowledge Base.** In the disambiguation phase (Figure 5.4a and b), all  $KB_i \in K_U$  are linked to the unified sense inventory  $S$  and added to

Figure 5.6. Example of disambiguation for high-confidence argument pairs with the relation triple  $\langle \text{Armstrong works for, NASA} \rangle$ . For clarity, only the most prominent BabelNet senses for both arguments are shown.

the set of redefined KBs  $K^S$ . As explained before, while each KB in  $K_D$  can be unambiguously redefined via BabelNet inter-resource links and added to  $K^S$ , KBs in  $K_U$  require an explicit disambiguation step. Given  $KB_i \in K_U$ , our disambiguation module (Figure 5.4b) takes as input its set of unlinked triples  $T_i$  and outputs a set  $T_i^S = T_i$  of disambiguated triples with subject-object pairs linked to  $S$ . The triples in  $T_i^S$ , together with their corresponding entity sets and relation sets, constitute the redefined KB  $K_i^S$  which is then added to  $K^S$ . However, disambiguating the content of KB (i.e. a set of relation instances) is not a trivial task: as we show in Sections 4.2 and 4.3, off-the-shelf disambiguation systems, including knowledge-based ones, require a rich and meaningful context to provide high quality disambiguations. Hence, applying a straightforward approach that disambiguates every triple in isolation might lead to very imprecise results, due to the lack of available context for each individual triple. We thus devise a disambiguation strategy, illustrated in Figure 5.5, that comprises three successive steps:

1. We identify a set of high-confidence seeds from  $T_i$  (Section 5.2.1.1), i.e. triples  $\langle e_d; r; e_g \rangle$  where subject  $e_d$  and object  $e_g$  are highly semantically related, and disambiguate them using the senses that maximize their similarity in  $V_S$ ;
2. We use the seeds to generate a ranking of the relations  $r_i$  according to their degree of specificity (Section 5.2.1.2). We represent each  $r_i \in R_i$  in  $V_S$  and assign higher specificity to relations whose arguments are closer in  $V_S$ ;
3. We jointly disambiguate the remaining non-seed triples in  $T_i$  (Section 5.2.1.3) starting from the most specific relations, and jointly using all participating argument pairs as context.

### 5.2.1.1 Identifying Seed Arguments

The first stage of the disambiguation pipeline aims at extracting reliable seeds from  $T_i$ , i.e. triples  $\langle e_d; r; e_g \rangle$  where subject  $e_d$  and object  $e_g$  can be confidently disambiguated without additional context. In order to do this we leverage the embeddings associated with each candidate sense for  $e_d$  and  $e_g$ . We consider all the available senses for both  $e_d$  and  $e_g$  in  $S$ , namely  $s_d = \{s_d^1; \dots; s_d^m\}$  and  $s_g = \{s_g^1; \dots; s_g^o\}$ , and the

corresponding sets of sense embeddings  $v_d = \{v_d^1, \dots, v_d^m\}$  and  $v_g = \{v_g^1, \dots, v_g^m\}$ . We then select, among all possible pairs of senses, the pair  $s_d, s_g$  that maximizes the cosine similarity between the corresponding embeddings  $v_d, v_g$ :

$$h_{v_d, v_g} = \operatorname{argmax}_{v_d \in V_d, v_g \in V_g} \frac{v_d \cdot v_g}{\|v_d\| \|v_g\|} \quad (5.3)$$

For each disambiguated triple  $h_{s_d, r, s_g}$ , the cosine similarity value associated with  $h_{v_d, v_g}$  represents its disambiguation confidence  $c_{dis}$ . We rank all such triples according to their confidence, and select those above a pre-specified confidence threshold  $\tau_{dis}$ . The underlying assumption is that, for high-confidence subject-object pairs, the embeddings associated with the correct senses  $s_d$  and  $s_g$  will be closest in  $V_S$  compared to any other candidate pair. Intuitively, the more the relation  $r$  between  $e_d$  and  $e_g$  is semantically well defined, the more this assumption is justified. As an example, consider the triple  $h_{\text{Armstrong works for}, \text{NASA}}$  in Figure 5.6: among all the possible senses for **Armstrong** (the astronaut Neil Armstrong, the jazz musician Louis Armstrong, the cyclist Lance Armstrong, etc.) and **NASA** (the space agency, the racing organization, the Swedish band, etc.) we expect the vectors corresponding to the astronaut sense of **Armstrong** and to the space agency sense of **NASA** to be closest in the vector space model  $V_S$ .

### 5.2.1.2 Relation Specificity Ranking

The assumption that, given an ambiguous subject-object pair, correct argument senses are the closest pair in the vector space (Section 5.2.1.1) is easily verifiable for general relations (e.g. *is a*, *is part of*). However, as a semantic relation becomes specific, its arguments are less guaranteed to be semantically related (e.g. *is a professor in the university of* and *a disambiguation approach based exclusively on similarity is prone to errors*). On the other hand, specific relations tend to narrow down the scope of possible entity types occurring as subject and object. In the above example, *is a professor in the university of* requires entity pairs with professors as subjects, and cities (or states) as objects. Our disambiguation strategy should thus vary according to the specificity of the relations taken into account. In order to consider this observation in our disambiguation pipeline, we first need to estimate the degree of specificity for each relation in the relation set  $R_i$  of the target KB to be disambiguated. Given  $R_i$  and a set of seeds from the previous step (Section 5.2.1.1), we apply a specificity ranking policy and sort relations in  $R_i$  from the most general to the most specific. We compute the generality  $\text{Gen}(r)$  of a given relation  $r$  by looking at the spatial dispersion of the sense embeddings associated with its seed subjects and objects. Let  $V_D$  ( $V_G$ ) be the set of sense embeddings associated with the domain (range) seed arguments of  $r$ . For both  $V_D$  and  $V_G$ , we compute the corresponding centroid vectors  $v_D$  and  $v_G$  as:

$$v_k = \frac{1}{|V_k|} \sum_{v \in V_k} v; \quad k \in \{D, G\} \quad (5.4)$$

Then, the variances  $\sigma_D^2$  and  $\sigma_G^2$  are given by:

$$\sigma_k^2 = \frac{1}{|V_k|} \sum_{v \in V_k} (1 - \cos(v, v_k))^2; \quad k \in \{D, G\} \quad (5.5)$$

We finally compute  $\text{Gen}(r)$  as the average of  $\frac{2}{D}$  and  $\frac{2}{G}$ . The result of this procedure is a relation specificity ranking that associates each relation  $r$  with its generality score  $\text{Gen}(r)$ . Intuitively, we expect more general relations to show higher variance (hence higher  $\text{Gen}(r)$ ), as their subjects and objects are likely to be rather dispersed throughout the vector space; instead, arguments of very specific relations are more likely to be clustered together in compact regions, yielding lower values of  $\text{Gen}(r)$ .

### 5.2.1.3 Disambiguation with Relation Context

In the third step, both the specificity ranking and the seeds are exploited to disambiguate the remaining triples in  $T_i$ . To do this we leverage `Babelify` (Section 2.2.2.3). As we observed in Section 5.2.1.2, specific relations impose constraints on their subject-object types and tend to show compact domains and ranges in the vector space. Therefore, given a triple  $(e_d; r; e_g)$ , knowing that  $r$  is specific enables us to put together all the triples in  $T_i$  where  $r$  occurs, and use them to provide an enriched and meaningful context for disambiguation. If  $r$  is general, instead, its subject-object types are less constrained, and additional triples do not guarantee to provide semantically related context (on the contrary, they could introduce noise).

At this third and final stage, the disambiguation pipeline takes as input the set of triples  $T_i$ , along with the associated disambiguation seeds (Section 5.2.1.1), the specificity ranking (Section 5.2.1.2), and a specificity threshold  $\tau_{\text{spec}}$ .  $T_i$  is first partitioned into two subsets:  $T_i^{\text{spec}}$ , comprising all the triples for which  $\text{Gen}(r) < \tau_{\text{spec}}$ , and  $T_i^{\text{gen}} = T_i \setminus T_i^{\text{spec}}$ . We then employ two different disambiguation strategies:

- ^ For each distinct relation  $r$  occurring in  $T_i^{\text{spec}}$ , we first retrieve the subset  $T_{i;r}^{\text{spec}} \subseteq T_i^{\text{spec}}$  of triples where  $r$  occurs, and then disambiguate  $T_{i;r}^{\text{spec}}$  as a whole with `Babelify`. For each triple in  $T_{i;r}^{\text{spec}}$ , context is provided by all the remaining triples along with the disambiguated seeds extracted for  $r$ .
- ^ We disambiguate the remaining triples in  $T_i^{\text{gen}}$  one by one in isolation with `Babelify`, providing for each triple only the predicate string  $r$  as additional context.

### 5.2.1.4 Cross-Resource Relation Alignment

After disambiguation (Figure 5.4a and b) each KB in  $K$  is linked to the unified sense inventory  $S$  and added to  $K^S$ . However, each  $KB_i^S \in K^S$  still provides its own relation set  $R_i^S \subseteq R$ . Instead, in the unified KB, relations with equivalent semantics should be considered as part of a single relation synset even when they come from different KBs. Therefore, at this stage, an alignment procedure is applied to identify pairs of relations from different KBs having equivalent semantics. We exploit the fact that each relation  $r$  is now defined over entity pairs linked to  $S$ , and we generate a semantic representation of  $r$  in the vector space  $V_S$  based on the centroid vectors of its domain and range. Due to representing the semantics of relations on this common ground, we can compare them by computing their domain and range similarity in  $V_S$ . We first consider each  $KB_i^S \in K^S$  and, for each relation  $r_i$  in  $R_i^S$ , we compute the corresponding centroid vectors  $r_d^i$  and  $r_g^i$  using formula (5.4). Then, for each pair of KBs  $\{KB_i^S; KB_j^S\} \in K^S \times K^S$ , we compare all

	$K_U$		$K_D$	
	Nell	ReVerb	Patty	WiSeNet
# Relations	298	1,299,844	1,631,531	245,935
# Triples	2,245,050	14,728,268	15,802,946	2,271,807
# Entities	1,996,021	3,327,425	1,087,907	1,636,307

Table 5.13. Statistics on the input KBs.

relation pairs  $\{r_i; r_j\} \in R_i^S \times R_j^S$  by computing the cosine similarity between domain centroids  $s_D$  and between range centroids  $s_G$ :

$$s_k = \frac{\frac{r_i}{k} \cdot \frac{r_j}{k}}{\frac{r_i}{k} \cdot \frac{r_i}{k} + \frac{r_j}{k} \cdot \frac{r_j}{k}} \quad (5.6)$$

where  $\frac{r}{k}$  denotes the centroid associated with relation  $r$  and  $k \in \{D; G\}$ . The average of  $s_D$  and  $s_G$  gives us an alignment confidence  $s_{align}$  for the pair  $\{r_i; r_j\}$ . If confidence is above a given threshold  $align$  then  $r_i$  and  $r_j$  are merged into the same relation synset. Relations for which no alignment is found are turned into singleton relation synsets. As a result of this alignment procedure we obtain the unified set of relations  $R$ .

## 5.2.2 Experimental Evaluation

We carried out an extensive experimental evaluation to assess the effectiveness of KB-Unify's unification pipeline. In particular, we evaluate the disambiguation pipeline in Section 5.2.2.1; then, in Section 5.2.2.2 we test our assumption on relation specificity on a manually verified sample of specificity rankings, and in Section 5.2.2.3 we evaluate the cross-resource alignment step of Section 5.2.1.4. The input set of KBs for this experimental evaluation was the following:

- We selected Patty (Section 3.2.1) and WiSeNet (Section 3.2.2) as linked resources. We used Patty with Freebase types and pattern synsets derived from Wikipedia, and WiSeNet 2.0 with Wikipedia relational phrases;
- We selected Nell (Carlson et al., 2010) and ReVerb (Fader et al., 2011) as unlinked resources. We used KB beliefs updated to November 2014 for the former, and the set of relation instances from ClueWeb09 for the latter.

Comparative statistics in Table 5.13 show that the input KBs are rather different in nature: Nell is based on 298 predefined relations and contains beliefs for about 2 million entities. The distribution of entities over relations is however very skewed, with 80.33% of the triples being instances of the generalizations' relationship. In contrast, ReVerb contains a highly sparse relation set (1,299,844 distinct relations) and more than 3 million distinct entities. Patty features the largest (and, together with WiSeNet, sparsest) set of triples, with 1,631,531 distinct relations and less than 10 triples per relation on average.

Figure 5.7. Precision (left) and coverage (right) of disambiguated seeds at different values of  $\text{con}_{\text{dis}}$  for (a) the whole set of triples in *Patty* and (b) the subset of ambiguous triples. Green circles represent the different values of  $\text{con}_{\text{dis}}$  considered.

### 5.2.2.1 Evaluating Knowledge Base Disambiguation

We tested KB-Unify's disambiguation pipeline experimentally in terms of both disambiguated seed quality and overall disambiguation performance. To this aim, we created a development set by extracting a subset of 6 million triples from the largest linked KB in our experimental setup, i.e. *Patty*. Triples in *Patty* are automatically linked to YAGO, which is in turn linked to WordNet and DBpedia. Since both resources are also linked by BabelNet, we mapped the original triples to the BabelNet sense inventory and used them to tune our disambiguation module. We also provide two baseline approaches: (1) direct disambiguation on individual triples with Babelify alone (without the seeds) and (2) direct disambiguation of the seeds only (without Babelify). We tuned our disambiguation algorithm by studying the quality of the disambiguated seeds extracted from the surface-text triples of *Patty*. Figure 5.7 shows precision and coverage for increasing values of the confidence threshold  $\text{con}_{\text{dis}}$ . We computed precision by checking each disambiguated seed against the corresponding linked triple in the development set, and coverage as the ratio of covered triples. We analyzed results for both the whole set of triples in *Patty* (Fig. 5.7a) and the subset of ambiguous triples (Fig. 5.7b), i.e. those triples whose subjects and objects have at least two candidate senses each in the BabelNet sense inventory. In both cases, precision of disambiguated seeds increases rapidly with  $\text{con}_{\text{dis}}$ , stabilizing above 90% with  $\text{con}_{\text{dis}} > 0.25$ . Coverage displays the opposite behavior, decreasing exponentially with more confident outcomes, from 6 million triples to less than a thousand (for seeds with confidence  $\text{con}_{\text{dis}} > 0.95$ ). As a result, we chose  $\text{con}_{\text{dis}} = 0.25$  as optimal threshold value for the subsequent experiments.

dis	SensEmbed			Baseline		
	0.5-0.7	0.7-0.9	0.9-1.0	0.5-0.7	0.7-0.9	0.9-1.0
Patty	.980	.980	1.000	.793	.780	1.000
WiSeNet	.958	.960	.973	.726	.786	.791
Nell	.955	.995	1.000	.800	.770	.885
ReVerb	.930	.940	.950	.775	.725	.920

Table 5.14. Disambiguation precision for all KB.

	spec = 0:8		spec = 0:5		spec = 0:3	
	all	only seeds	all	only seeds	all	only seeds
Patty	62.15	26.60	52.49	24.06	40.75	21.41
WiSeNet	60.00	37.46	54.44	22.26	53.58	16.62
Nell	76.97	62.98	50.95	20.71	44.70	4.36
ReVerb	41.20	38.57	25.14	23.70	13.37	12.75

Table 5.15. Coverage results (%) for all KBs.

**Manual Evaluation.** In addition, we manually evaluated the disambiguated seeds extracted from both linked KBs (Patty and WiSeNet) and unlinked KBs (Nell and ReVerb). For each KB, we extracted up to three random samples of 150 triples according to different levels of confidence  $dis$ : the first sample included extraction with  $0.5 \leq dis < 0.7$ , the second with  $0.7 \leq dis < 0.9$ , and the third with  $dis \geq 0.9$ . Each sample was evaluated by two human annotators: for each disambiguated triple  $\langle e_d; r; e_g \rangle$ , we presented the annotators with the surface-text arguments  $s_d; e_g$  and the relation string  $r$ , along with the two Babel synsets corresponding to the disambiguated arguments  $s_d; s_g$ , and we asked whether the association of each subject and object with the proposed Babel synset was correct. We then estimated precision as the average proportion of correctly disambiguated triples. For each sample we compared disambiguation precision using SensEmbed, as in Section 5.2.1.1, against the first baseline with Babelfy alone. Results, reported in Table 5.14, show that our approach consistently outperforms the baseline and provides high precision over all samples and KBs. We then evaluated the overall disambiguation output after specificity ranking (Section 5.2.1.2) and disambiguation with relation context using Babelfy (Section 5.2.1.3). We analyzed three configurations of the disambiguation pipeline, namely  $spec \in \{0:8; 0:5; 0:3\}$ . We ran the algorithm over both linked and unlinked KBs of our experimental setup, and computed the coverage for each KB as the overall ratio of disambiguated triples. Results are reported in Table 5.15 and compared to the coverage obtained from the disambiguated seeds only: context-aware disambiguation substantially increases coverage over all KBs. Table 5.15 also shows that a restrictive  $spec$  results in lower coverage values, due to the increased number of triples disambiguated without context.

**Automatic Evaluation.** We also evaluated the quality of disambiguation on a publicly available dataset (Dutta et al., 2014). This dataset provides a gold standard

	KB-Unify		Dutta et al.	Baseline
	all	only seeds	( $\alpha = 0.5$ )	
Precision	.852	.957	.931	.749
Recall	.875	.117	.799	.608
F-score	.864	.197	.857	.671

Table 5.16. Disambiguation results over the gold standard of Dutta et al. (2014).

of 1200 triples from *Nell* whose subjects and objects are manually assigned a proper DBpedia URI. We again used BabelNet’s inter-resource links to express DBpedia annotations with KB-Unify’s sense inventory and then checked, for each annotated triple in the dataset, the corresponding triple in the disambiguated version of *Nell* with  $\alpha_{dis} = 0.25$  and  $\alpha_{spec} = 0.8$ . We then repeated this process considering only the disambiguated seeds instead of the whole disambiguation pipeline (second baseline). In line with Dutta et al. (2014), we computed precision, recall and F-score for each setting. Results are reported in Table 5.16 and compared against those of Dutta et al. (2014) and against our first baseline with Babelify alone. KB-Unify achieves the best result, showing that a baseline based on straightforward disambiguation is negatively affected by the lack of context for each individual triple. In contrast, the baseline approach that relies only on the disambiguated seeds attains very high precision, but suffers from dramatically lower coverage.

### 5.2.2.2 Evaluating Specificity Ranking

We evaluated the specificity ranking (Section 5.2.1.2) generated by KB-Unify for all KBs of the experimental setup. First of all, we empirically validated our scoring function  $\text{Gen}(r)$  over each resource: for each relation we computed the average similarity among all its domain arguments  $s_D$  and among all its range arguments  $s_G$ .<sup>14</sup> We then plotted the averages of  $s_D$  and  $s_G$  against  $\text{Gen}(r)$  for each relation  $r$  (Figure 5.8). The overall trend shown by the four plots of Figure 5.8 suggests that, as observed in Section 5.2.1.2, the average similarity among domain and range arguments decreases for increasing values of  $\text{Gen}(r)$ , indicating that more general relations allow less semantically constrained subject-object types.

We then used human judgment to assess the quality of our specificity rankings. First, each ranking was split into four quartiles, and two human annotators were presented with a sample from the top quartile (i.e. a relation falling into the most general category) and a sample from the bottom quartile (i.e. a relation falling into the most specific category). We shuffled each relation pair, showed it to our human judges, and then asked which of the two relations they considered to be the more specific. Ranking precision was computed by considering those pairs where human choice agreed with the ranking. In addition, we also considered the agreement with a randomly shuffled version of each ranking, as a baseline comparison. Finally, we computed inter-annotator agreement on each ranking (except for *Nell*, due to the

<sup>14</sup>For both domain and range of  $r$ , we considered the disambiguated seed arguments from the previous step, and computed the cosine similarities of the corresponding sense embeddings pairwise; we then calculated the average of these similarities over the whole set.

Figure 5.8. Average argument similarity against Gen(r) for all the input KBs in the experimental setup.

	Nell	ReVerb	Patty	WiSeNet
Precision Gen(r)	.660	.715	.625	.750
Precision (random)	.504	.483	.525	.497
Cohen's kappa	-	.430	.620	.600

Table 5.17. Specificity ranking evaluation.

small sample size) with Cohen's kappa (Cohen, 1960). Results for each ranking and baseline are reported in Table 5.17, while some examples of general and specific relations for each KB are shown in Table 5.18. Disagreement between human choice and ranking is higher in Nell (where the set of relations is quite small compared to other KBs) and in Patty (due to a sparser set of relations, biased towards very specific patterns). Inter-annotator agreement is instead lower for ReVerb, where unconstrained Web harvesting often results in ambiguous relation strings.

### 5.2.2.3 Evaluating Relation Alignment

Due to the lack of available gold standards and test-beds, we evaluated the cross-resource relation alignment procedure of KB-Unify (Section 5.2.1.4) by exploiting human judgment once again. Given the results of Section 5.2.2.1, we considered the top 10k frequent relations for each KB and ran the algorithm over each possible pair of KBs with two different configurations:  $\text{align} = 0:7$  and  $\text{align} = 0:9$ . From each pair of KBs  $\{KB_i; KB_j\}$  we obtained a list of candidate alignments, i.e. pairs of relations  $\{r_i; r_j\}$  where  $r_i \in KB_i$  and  $r_j \in KB_j$ . From each list we then extracted a random sample of 150 candidate alignments. We showed each alignment  $\{r_i; r_j\}$  to

<sup>15</sup>In the case of relation synsets, such as Patty and WiSeNet, we selected up to three random relation phrases from each synset.

Nell	
High Gen(r)	agent created at location
Low Gen(r)	person in economic sector restaurant in city
ReVerb	
High Gen(r)	is for is in
Low Gen(r)	enter Taurus in carry oxygen to
Patty	
High Gen(r)	located in later served to
Low Gen(r)	starting pitcher who played league coach for
WiSeNet	
High Gen(r)	include is a type of
Low Gen(r)	lobe-finned fish lived during took part in the Eurovision contest

Table 5.18. Examples of general and specific relations for all KBs.

two human annotators, and asked whether  $r_i$  and  $r_j$  represented the same relation. The problem was presented in terms of paraphrasing: for each pair, we asked whether exchanging  $r_i$  and  $r_j$  within a sentence would have changed that sentence's meaning. In line with Section 5.2.2.2 we computed precision based on the agreement between human choice and automatic alignments. Results are reported in Table 5.19. Our alignment algorithm shows the highest precision in the pairings with  $\text{align} = 0:9$ . Alignment reliability decreases for lower  $\text{align}$ , as relation pairs where  $r_i$  is a generalization of  $r_j$  (or vice versa) tend to have similar centroids in  $V_S$ . The same holds for pairs where  $r_i$  is the negation of  $r_j$  (or vice versa). Even though one could certainly utilize measures based on relation string similarity (Dutta et al., 2015) to reduce wrong alignments in these cases, by relying on a purely semantic criterion we removed any prior assumption on the format of input KBs. Some

	Patty	-WiSeNet	Patty	-ReVerb	Nell	-ReVerb
$\text{align}$	0.7	0.9	0.7	0.9	0.7	0.9
Prec.	.68	.80	.58	.74	.61	.75
# Align.	128k	1.2k	47k	643	2.6k	88
	Patty	-Nell	WiSeNet	-Nell	WiSeNet	-ReVerb
$\text{align}$	0.7	0.9	0.7	0.9	0.7	0.9
Prec.	.66	1.00	.70	.84	.59	.87
# Align.	2.6k	57	381	34	9.9k	169

Table 5.19. Cross-resource alignment evaluation.

Patty-WiSeNet		align
portrayed	's character	0.84
debuted in	first appeared in	0.86
Patty-ReVerb		align
language in	is spoken in	0.81
mostly known for	plays the role of	0.70
Nell-ReVerb		align
bookwriter	is a novel by	0.88
personleadscity	is the mayor of	0.60
Nell-Patty		align
worksfor	was hired by	0.72
riveremptiesintoriver	tributary of	0.89
Nell-WiSeNet		align
animaleatfood	feeds on	0.72
teammhomestadium	play their home games at	0.88
ReVerb-WiSeNet		align
has a selection of	offers	0.82
had grown up in	was born and raised in	0.85

Table 5.20. Examples of cross-resource relation alignments and corresponding  $\text{align}$ .

examples of alignments are shown in Table 5.20.

To conclude, we report statistics regarding the unified KB produced from the initial set of resources in our experimental setup. We validated our thresholds for high-precision, and selected  $\text{dis} = 0.25$ ,  $\text{spec} = 0.8$  and  $\text{align} = 0.8$ . Our alignment algorithm produced 56,673 confident alignments, out of which 2,207 relation synsets were derived, with an average size of 16.82 individual relations per synset. As a result, we obtained a unified KB comprising 24,221,856 disambiguated triples defined over 1,952,716 distinct entities and 2,675,296 distinct relations.

**Final Remarks.** The rationale behind KB-Unify is that of bringing the semantic integration of lexical knowledge, pioneered by large-scale knowledge resources like BabelNet (Section 2.1.3) to the next level, by extending this approach and applying it to relational knowledge. To this aim, a fundamental first step is having an array of input KBs where relations arguments are disambiguated and linked to the same sense inventory, which requires us to design a disambiguation module that deals with OIE-derived knowledge extracted at the level of surface text. Despite the inherent difficulty of the disambiguation target (i.e. a set of possibly unrelated relation triples), we showed that devising a strategy to provide a richer and meaningful disambiguation context is key to obtain high quality disambiguation, a methodology that proved his effectiveness for both EuroSense (Section 4.2) and SenseDefs (Section 4.3). Indeed, KB-Unify achieves state-of-the-art disambiguation in our experimental setting (Section 5.2.2.1), and provides a general, resource-independent representation of semantic relations, suitable for any kind of KB. In this respect, the generality of

exibility of KB-Unify is another prominent feature of our approach, even if, as we show in Section 5.2.2.3, it might lead to suboptimal relation alignments when the support set of a relation is not large enough<sup>16</sup>. On the other hand, exploiting directly the semantic characterization of a relation phrase would lead, in our setting, to a loss of generality, because it assumes each and every relation to be somehow anchored to a textual representation. This is however not the case for many non-open IE systems, including *Nell*, where the relation inventory is hand-crafted by humans and not necessarily bound to surface text: since KB-Unify models the semantics of a relation using its arguments only, it is capable of handling this kind of relations seamlessly. Of course, there is still room for improvement in many (if not all) stages of the KB unification pipeline. For instance, one relevant aspect that we left mostly underinvestigated is the evaluation of the relation alignment step. Among other experiments, a thorough comparison between KB-Unify's purely distributional alignment module and, e.g. Universal Schema approaches (cf. Section 2.3.3) where relation alignments can be seen as implicitly learnt from the training data, is out of the scope of the present section<sup>17</sup>, but would constitute an important step forward for assessing the competitiveness of KB-Unify as a knowledge integration framework.

### 5.3 TaxoEmbed: Sense-Aware Hypernym Discovery

As we discussed in Section 5.2.2, one of the key design choices of KB-Unify is the generality of its unification approach, which makes no assumptions on the shape and features of a semantic relation inside a target KB, and hence can be seamlessly applied to 'closed' IE systems or, for that matter, manually curated knowledge resources (where semantic relations, pre-specified by human experts, are often not tied to any textual realization). In the present section we consider an IE scenario that lies at the opposite end of the spectrum compared to OIE: hypernym discovery which consists in the extraction of only one specific kind of semantic relation, i.e. the hypernymic ('is a') relation.<sup>18</sup>

**Why Hypernyms?** Hypernymy, i.e. the capability for generalization, lies at the core of human cognition. Unsurprisingly, identifying hypernymic relations has been pursued in NLP for approximately the last two decades, as successfully identifying this lexical relation contributes to improvements in Question Answering applications (Prager et al., 2000; Yahya et al., 2013) and Textual Entailment or Semantic Search systems (Hofmann et al., 2014; Roller and Erk, 2016). Moreover,

<sup>16</sup>Using only the semantic characterization of the arguments to ontologize relation is also a feature of *Patty*, and we showed that it leads to sparser results when compared with approaches that instead leverage directly the semantic characterization of a relation pattern (cf. Section 5.1).

<sup>17</sup>Beside its supervised nature, the Universal Schema paradigm focuses on modeling asymmetric implicature between relations rather than explicit (and symmetric) relation alignment, as in KB-Unify. This makes a direct and fair comparison difficult at this stage.

<sup>18</sup>In most of the literature on the subject, including standard evaluation benchmarks (Bordea et al., 2015, 2016), this task has actually been formulated as hypernym detection, i.e. the binary task consisting of, given a pair of words, deciding whether a hypernymic relation holds between them. The alleged simplification of this setting (Levy et al., 2015b; Camacho Collados, 2017) has led to reformulate the problem as hypernym discovery, i.e. given the search space of a domain's vocabulary, and given an input concept, discover its best (set of) candidate hypernyms.

hypernymic relations are the backbone of almost any ontology, semantic network and taxonomy, including all the structured knowledge resources we examined in Section 2.1, and represent a key concern also for general-purpose IE systems: for instance, *Nell* (Carlson et al., 2010) not only relies crucially on hand-crafted taxonomized concepts and their relations within its learning process, but also extracts and encodes a large amount of *is-a* relation triples among its content beliefs (cf. Section 5.2.2). Similarly, pattern-based OIE systems, including *Patty* (Section 3.2.1), *WiSeNet* (Section 3.2.2), and *DefIE* (Section 5.1), are capable of implicitly extracting hypernyms; however, as we discussed at the beginning of Section 5.2, finding a semantic relation that models hypernymy inside an OIE-derived KB necessarily requires manual inspection. In the same section, on the other hand, we also showed with *KB-Unify* that using semantic analysis explicitly provides a way to disambiguate, harmonize and unify OIE-derived knowledge, thereby making it better expendable in downstream applications.

**Hypernymy in the Vector Space.** Extracting hypernymic relations is the first and foremost step of taxonomy learning approaches. Apart from taxonomy learning, on which the scientific literature is broad and comprehensive (Wang et al., 2017), work stemming from distributional semantics has put forward a notion of linguistic regularities found in vector representations such as word embeddings (Mikolov et al., 2013c). In this area, supervised approaches, arguably the most popular nowadays, learn a feature vector between term-hypernym vector pairs and train classifiers to predict hypernymic relations (Carmona and Riedel, 2017). These pairs may be represented either as a concatenation of both vectors (Baroni et al., 2012), difference (Roller et al., 2014), dot-product (Mikolov et al., 2013b), or including additional linguistic information for LSTM-based learning (Shwartz et al., 2016). These approaches, however, tend to be less precise and seem to perform best in discovering broader semantic relations (Shwartz et al., 2016): a strategy to overcome this is proposed by Fu et al. (2014), where the fundamental idea is that of learning a hypernymic transformation matrix over a word embeddings space. Fu et al. (2014) show empirically that the hypernymic relation that holds for the pair *dragon*, *insect* differs from the one of, e.g., *carpenter*, *person*. Their system addresses this discrepancy via *k*-means clustering on the input space (tuned using a held-out development set), and then learns a piece-wise linear projection for each cluster.

**Motivation.** All the embedding approaches described above operate inherently at the surface level. This is partly due to the way evaluation is conducted, which is often limited to very specific domains with no integrative potential, such as taxonomies in food, science or equipment from Bordea et al. (2015), or restricted to lists of word pairs. Apart from the lexical ambiguity issues arising with IE systems in general, a specific drawback of surface-level taxonomy learning is that additional steps and error-prone procedures are required to identify semantic clusters (Fu et al., 2014). On the other hand, however, hypernym extraction at the sense level, to date, is performed almost exclusively by sense-aware OIE approaches (cf. Sections 3.2 and 5.1). In addition to not being usable explicitly without manual inspection, as we discussed earlier, hypernyms extracted with OIE techniques tend to be noisier, given

the high-coverage nature of these systems and their broader scope. Therefore, in line with all the approaches presented in this chapter, our strategy in the present section is that of reframing the supervised distributional approach of Fu et al. (2014) at the sense level: this allows us, on the one hand, to improve their domain adaptation procedure by leveraging the structured semantic knowledge in BabelNet (Section 2.1.3); on the other, it provides us with a flexible sense-level framework where we can rely on both manually-curated and OIE-derived hypernymic knowledge as training data. This approach, named TaxoEmbed (Espinosa Anke et al., 2016a),<sup>19</sup> is based on the sense embeddings of Iacobacci et al. (2015), and it is designed to discover hypernymic relations by exploiting linear transformations in the sense embedding space. Unlike previous approaches, TaxoEmbed leverages this intuition to learn a specific sense-aware transformation matrix for each domain of knowledge, using sense-level training data drawn from heterogeneous sources of hypernymic information. Being based on the sense inventory of BabelNet, TaxoEmbed performs jointly hypernym extraction and disambiguation, from which expanding existing ontologies becomes a trivial task. After explaining in detail the approach in Section 5.3.1, we carry out an extensive experimental evaluation in Section 5.3.2, showing that TaxoEmbed can effectively replicate the Wikidata is-a branch, and capture previously unseen relations in other reference taxonomies. Most notably, the best configuration of TaxoEmbed in our experiments considers two training sources: (1) Manually curated pairs from Wikidata (Vrandečić, 2012), and (2) hypernymy relations from KB-Unify (Section 5.2).

### 5.3.1 The TaxoEmbed pipeline

TaxoEmbed's approach can be described as a three-stage pipeline: in the first step, we take advantage of a clustering algorithm to associate each Babel synset in the training set with a domain cluster  $C$  (Section 5.3.1.1); then, we expand the training set by exploiting all the different English lexicalizations provided by BabelNet for each synset (Section 5.3.1.2); finally, we learn a cluster-wise linear projection matrix over all term-hypernym pairs in the expanded training set (Section 5.3.1.3). Throughout this process, we rely on SensEmbed as reference sense embedding space for TaxoEmbed, as we did for KB-Unify. As regards our initial training set, instead, we first leverage the portion of the hypernym branch of Wikidata (Vrandečić, 2012) included in BabelNet; as usual, in order to construct a training set  $W$  compliant with TaxoEmbed's sense inventory, we use BabelNet's inter-resource mapping to map each Wikidata item to the corresponding Babel synset. Besides  $W$ , we also construct a second training dataset, denoted as  $K$ , by leveraging OIE-derived knowledge from KB-Unify: specifically, we consider the unified KB generated in the experimental evaluation of Section 5.2.2, and identify the relation synset containing  $Nell$ 's is-a relation;<sup>20</sup> we then draw from the unified KB all the corresponding triples in which the arguments have a disambiguation confidence greater or equal than 0.9 (cf. Section 5.2.1.1). Initially,  $|W| = 5,301,867$  and  $|K| = 1,358,949$ .

<sup>19</sup> <http://wwwusers.di.uniroma1.it/~dellibovi/taxoembed>

<sup>20</sup> This relation is encoded in the KB of beliefs as 'generalizations' (cf. Section 5.2.2).

### 5.3.1.1 Domain Clustering

In contrast to Fu et al. (2014), where semantic clusters are induced via k-means, with  $k$  tuned on a development set, TaxoEmbed aim at learning a function sensitive to a predefined knowledge domain, under the assumption that vectors clustered with this criterion are likely to exhibit similar semantic properties. First, we allocate each Babel synset into its most representative domain, which is achieved by exploiting the set of thirty four domains available in the Wikipedia featured articles page<sup>21</sup>. We associate a given synset  $b$  with an appropriate domain using Nasari (Section 2.2.3.3): following Camacho Collados et al. (2016c), we build a lexical vector for each Wikipedia domain by concatenating all Wikipages representing a given domain  $d$  into a single text. Then, we calculate the similarities between the Nasari lexical vector corresponding to  $b$  and all the domain vectors, and select the domain  $\hat{d}$  with the highest similarity score:

$$\hat{d} = \operatorname{argmax}_{d \in D} \text{WO}(d; b) \quad (5.7)$$

where  $D$  is the set of all thirty-three domains,  $d$  is the vector of the domain  $d \in D$ ,  $b$  is the vector of the BabelNet synset  $b$ , and  $\text{WO}$  refers to the Weighted Overlap measure (Pilehvar et al., 2013) we used for comparison. In order to have a reliable set of domain labels, all the synsets with maximum similarity score below a specified threshold are not annotated with any domain. We fix the threshold to 0.35, which provides a fine balance between precision (estimated around 85%) and recall in our development set, and obtain almost 2 million synsets labeled with a domain.

### 5.3.1.2 Training Data Expansion

Prior to training our model, we benefit from the fact that a given Babel synset may be associated with a fixed number of lexicalizations (cf. Section 2.1.3). We take advantage of this synset representation to expand each term-hypernym synset pair. For each term-hypernym pair, defined at the level of Wikidata entities, the corresponding Babel synsets are used to retrieve all the associated English lexicalizations; in this way, each term-hypernym pair  $h; t$  in the training data is expanded into a set of  $\{L_t; j; L_h\}$  training pairs at the sense level<sup>22</sup> where  $L_t$  and  $L_h$  denote the set of lexicalizations available for  $t$  and  $h$ , respectively.

This expansion step yields the considerably larger sets  $\mathcal{W}$  and  $\mathcal{K}$ , where  $|\mathcal{W}| = 18,291,330$  and  $|\mathcal{K}| = 15,362,268$  (3 and 11 times bigger than their initial versions, respectively). These figures are higher than those reported in recent hypernym detection approaches, which exploited Chinese semantic thesauri along with manual validation of hypernym pairs (Fu et al., 2014) to obtain a total of 1,391 instances, or entity pairs from various knowledge resources (Shwartz et al., 2016), where the maximum reported split for training data (70%) amounted to 49,475 pairs.

<sup>21</sup> [https://en.wikipedia.org/wiki/Wikipedia:Featured\\_articles](https://en.wikipedia.org/wiki/Wikipedia:Featured_articles)

<sup>22</sup> In other words, each pair  $h; y$  drawn from  $h; t$  is such that  $x$  consists of the synset  $s_t$  associated with  $t$  paired with one of the lexicalizations in  $s_t$ , and  $y$  consists of the synset  $s_h$  associated with  $h$  paired with one of the lexicalizations in  $s_h$ .

### 5.3.1.3 Learning a Hypernym Detection Matrix

As discussed at the beginning of this section, the gist of TaxoEmbed lies in the capability of embedded vector space models to capture semantic relations (Mikolov et al., 2013b; Fu et al., 2014; Tan et al., 2015). However, instead of learning a global linear transformation function for a broad relation like hypernymy, learning a function sensitive to a given domain of knowledge has been proven more effective (Fu et al., 2014). Hence, TaxoEmbed follows an analogous strategy: given a specific domain  $d$  and the sense-level training set  $T$  (obtained from the previous step), we isolate the subset of  $T$  with pairs  $\{x_i, y_i\} \in C_d \times C_d$ , where  $C_d$  denotes the cluster of senses having the corresponding Babel synsets labeled with the domain  $d$ .

Then, for each domain-wise partition  $T^d$  of the expanded training set  $T$ , we construct a hyponym matrix  $X^d = [x_1 \dots x_n]$  (with all  $x_i \in C_d$ ) and a hypernym matrix  $Y^d = [y_1 \dots y_n]$  (with all  $y_i \in C_d$ ). Both  $X^d$  and  $Y^d$  comprise the SensEmbed vectors corresponding to the training pairs  $\{x_i, y_i\} \in C_d \times C_d; 0 \leq i < n$ .

Under the intuition that there exists a matrix  $A^d$  such that  $y^d = A^d x^d$ , we learn a transformation matrix for each domain cluster  $C_d$  by minimizing:

$$\min_{C_d} \sum_{i=1}^n \|x_i - A^d y_i\|^2 \quad (5.8)$$

Then, for any unseen Babel synsets labeled with the domain  $d$ , TaxoEmbed is able to compute a ranked list of the most probable hypernym vectors  $s$ , using cosine similarity as comparison measure:

$$\operatorname{argmax}_{s \in S} \frac{\langle x_j, s \rangle}{\|x_j\| \|s\|} \quad (5.9)$$

with  $S$  denoting the vector space of SensEmbed, and  $x_j \in L_s \times C_d$  representing the  $j$ -th lexicalization of  $s$ . At this point, TaxoEmbed associates with each Babel synset a ranked list of candidate hypernym vectors, each of which is associated with another Babel synset. This procedure allows us to cast the hypernym extraction task as a ranking problem, where TaxoEmbed is only provided with the hyponym term, and the most probable hypernym(s) must be discovered at testing time.

## 5.3.2 Experimental Evaluation

We evaluated experimentally the performance of TaxoEmbed by conducting several experiments, both automatic and manual. In Section 5.3.2.1 we assessed TaxoEmbed's ability to discover valid hypernyms for a given unseen term within a held-out evaluation dataset of 250 Wikidata term-hypernym pairs. In Section 5.3.2.2, instead, we evaluated the extent to which TaxoEmbed is able to correctly identify hypernyms outside of Wikidata. In both experiments, the evaluation benchmarks were defined at the sense level, i.e. composed of test pairs  $\{t, h\}$  where both  $t$  and  $h$  are Babel synsets.

### 5.3.2.1 Evaluating Hypernym Identification

**Experimental Setup.** For each domain, we retained 5k, 10k, 15k, 20k and 25k term-hypernym training pairs from  $W$  to generate five different configurations of

TaxoEmbed , and evaluated the resulting ve TaxoEmbed models on 250 test pairs for each of the 10 domains. We also experimented with two additional configurations of TaxoEmbed , which include 1k and 25k extra OIE-derived training pairs from  $K$  per domain. The resulting two models are denoted by  $25k+K_{1k}^d$  and  $25k+K_{25k}^d$ , respectively. Moreover, in order to validate the empirically grounded intuition of Fu et al. (2014), we introduced three non domain-sensitive configurations of TaxoEmbed : one configuration with 25k pairs from  $W$  and 50k additional pairs randomly sampled from  $K$  ( $25k+K_{50k}^r$ ), and two configurations with only random pairs coming from  $W$  ( $100K_{wd}^r$ ) and  $K$  ( $100k+K_{kbu}^r$ ).

**Baseline.** We included a distributional supervised baseline based on word analogies (Mikolov et al., 2013c), which works as follows: first, it calculates the difference vector of each training SensEmbed vector pair  $\langle x_i; y_i \rangle$  of a given domain  $d$ ; then, it averages all these difference vectors to obtain a global vector  $v_d$  for the domain  $d$ ; finally, given a test term  $t$  it calculates the vector closest to the sum of  $t$  (the corresponding term vector) and  $v_d$ :

$$\hat{h} = \operatorname{argmax}_{h \in S} \cos(v_d + t; h) \quad (5.10)$$

We trained this baseline using 25k domain-filtered pairs from  $W$ .

**Evaluation metrics.** We computed the following metrics for each domain and for all the configurations above: Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and R-Precision (R-P). These measures provide insights on different aspects of the outcome of the task, e.g. how often valid hypernyms were retrieved in the first positions of the rank (MRR), and if there were more than one valid hypernym, whether this set was correctly retrieved, (MAP and R-P).<sup>23</sup>

**Results and Discussion.** The outcome of all our experiments are summarized in Table 5.21. Results suggest that the performance of TaxoEmbed increases as training data expands, corroborating previous findings (Mikolov et al., 2013b). The improvement of TaxoEmbed over the baseline is consistent across most evaluation domain clusters and metrics, with domain-filtered data from  $K$  contributing positively in about two thirds of the evaluated configurations. As regards individual domains, the biology domain seems to be the easiest to model, likely due to the fact that fauna and flora are areas where hierarchical division of species is a field of study in itself, which traces back to Aristotelian times (Mayr, 1982), and therefore has been constantly refined over the years. This is the only domain in which training with no semantic awareness gives good results. We argue that this is due to the fact that a vast amount of synsets are allocated into the biology cluster (60% of the total, and up to 80% hypernyms). This produces the so-called lexical memorization phenomenon (Levy et al., 2015b), as the system memorizes prototypical biology-related hypernyms like taxon as valid hypernyms for many concepts. Another remarkable case involves the education and media domains, which experience the highest improvement with training data from  $K$  (5 and 6 MRR points, respectively).

<sup>23</sup>Bian et al. (2008) provide an in-depth analysis of all these metrics.

	art			biology			education			geography			health		
Train	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P
5k	0.12	0.12	0.12	0.63	0.63	0.59	0.00	0.00	0.00	0.08	0.07	0.07	0.08	0.08	0.07
15k	0.21	0.20	0.18	0.84	0.72	0.79	0.22	0.22	0.21	0.15	0.14	0.14	0.08	0.07	0.07
25k	0.29	0.27	0.26	0.84	0.83	0.81	0.33	0.32	0.30	0.23	0.22	0.21	0.09	0.09	0.08
25k+K <sub>1k</sub> <sup>d</sup>	0.29	0.28	0.26	0.84	0.80	0.79	0.32	0.29	0.27	0.22	0.22	0.21	0.09	0.09	0.08
25k+K <sub>25k</sub> <sup>d</sup>	0.26	0.24	0.22	0.70	0.63	0.56	0.38	0.36	0.33	0.15	0.13	0.12	0.11	0.11	0.10
25k+K <sub>50k</sub> <sup>r</sup>	0.28	0.26	0.24	0.82	0.77	0.72	0.36	0.33	0.30	0.17	0.16	0.16	0.12	0.11	0.10
100K <sub>wd</sub>	0.00	0.00	0.00	0.84	0.81	0.77	0.00	0.00	0.00	0.01	0.01	0.01	0.07	0.06	0.06
100K <sub>kb</sub>	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.12	0.12	0.11
Baseline	0.13	0.12	0.10	0.58	0.57	0.57	0.10	0.10	0.09	0.12	0.09	0.05	0.07	0.13	0.14
	media			music			physics			transport			warfare		
Train	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P	MRR	MAP	R-P
5k	0.28	0.28	0.27	0.10	0.10	0.09	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01
15k	0.14	0.13	0.12	0.08	0.07	0.07	0.36	0.35	0.34	0.25	0.23	0.21	0.01	0.01	0.01
25k	0.46	0.45	0.43	0.30	0.28	0.26	0.41	0.40	0.38	0.46	0.43	0.39	0.05	0.05	0.04
25k+K <sub>1k</sub> <sup>d</sup>	0.43	0.42	0.41	0.32	0.30	0.28	0.39	0.38	0.37	0.47	0.44	0.40	0.04	0.04	0.01
25k+K <sub>25k</sub> <sup>d</sup>	0.52	0.51	0.49	0.26	0.25	0.23	0.37	0.36	0.34	0.48	0.45	0.41	0.04	0.03	0.03
25k+K <sub>50k</sub> <sup>r</sup>	0.46	0.45	0.43	0.29	0.28	0.25	0.31	0.30	0.29	0.52	0.49	0.46	0.05	0.04	0.04
100K <sub>wd</sub>	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.01
100K <sub>kb</sub>	0.08	0.07	0.07	0.01	0.01	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.00	0.00
Baseline	0.57	0.43	0.52	0.03	0.03	0.03	0.05	0.04	0.04	0.29	0.25	0.21	0.04	0.04	0.04

Table 5.21. Overview of the performance of TaxoEmbed using different training data samples.

In fact, one of the main sources for in-domain relations in KB-Unify is Nell , which contains a large amount of relation triples between North American academic entities (professors, sports teams, alumni, donors; as well as media celebrities). Many of these entities are missing in Wikidata, and relations among them encoded in Nell are likely to be correct because in most cases these are unambiguous entities occurring in the same communicative contexts.

### 5.3.2.2 Evaluating Extra Coverage

**Experimental Setup.** For this experiment we used two configurations of TaxoEmbed: the first one includes 25k domain-wise training pairs from W (TaxE<sub>25k</sub>), and the second one includes also 1k pairs from K (TaxE<sub>25k+K<sup>d</sup></sub>). In order to evaluate these configurations on instances not included in Wikidata, we constructed a test set with 200 randomly extracted Babel synsets (20 per domain) for which no hypernym is available in Wikidata. Using this benchmark we compared TaxoEmbed against a number of taxonomy learning and IE systems, namely Yago (Hofmann et al., 2011a; Mahdisoltani et al., 2015), WiBi (Flati et al., 2016) and DefIE (Section 5.1). Then, three annotators assessed manually the validity of the hypernyms extracted by each system. Yago and WiBi can be viewed as upper bounds for TaxoEmbed, due to the nature of their hypernymic relations. In fact, both include a great number of manually-encoded taxonomies (e.g. exploiting WordNet and Wikipedia categories); Yago derives its taxonomic relations from an automatic mapping between WordNet and Wikipedia categories. WiBi, on the other hand, exploits a number of different Wikipedia-specific heuristics, Wikipedia categories, and the syntactic structure of Wikipedia-derived definitions (cf. Section 5.1). Finally, we included DefIE as comparison system by considering the knowledge base obtained for its experimental

	art			biology			education			geography			health		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TaxE <sub>25k</sub>	0.45	0.45	0.45	0.40	0.40	0.40	0.60	0.60	0.60	0.35	0.35	0.35	0.45	0.45	0.45
TaxE <sub>25k+K<sup>d</sup></sub>	0.50	0.50	0.50	0.40	0.40	0.40	0.55	0.55	0.55	0.35	0.35	0.35	0.45	0.45	0.45
DefIE	0.63	0.35	0.45	0.36	0.20	0.25	0.57	0.20	0.29	0.66	0.40	0.50	0.25	0.15	0.18
Yago	0.88	0.75	0.81	0.62	0.25	0.36	0.94	0.80	0.86	0.79	0.75	0.77	0.28	0.10	0.15
Wibi	0.70	0.70	0.70	0.58	0.50	0.54	0.94	0.80	0.86	0.75	0.75	0.75	0.66	0.50	0.57
	media			music			physics			transport			warfare		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
TaxE <sub>25k</sub>	0.10	0.10	0.10	0.45	0.45	0.45	0.15	0.15	0.15	0.35	0.35	0.35	0.25	0.25	0.25
TaxE <sub>25k+K<sup>d</sup></sub>	0.10	0.10	0.10	0.40	0.40	0.40	0.15	0.15	0.15	0.25	0.25	0.25	0.45	0.45	0.45
DefIE	0.81	0.45	0.58	0.71	0.50	0.58	0.42	0.15	0.22	0.54	0.30	0.38	0.60	0.30	0.40
Yago	0.76	0.65	0.70	0.84	0.55	0.67	0.80	0.40	0.53	0.93	0.70	0.80	0.81	0.65	0.72
Wibi	0.90	0.90	0.90	0.89	0.85	0.87	0.68	0.55	0.61	0.87	0.70	0.77	0.66	0.50	0.57

Table 5.22. Precision, recall and F-Measure outside Wikidata.

evaluation (Section 5.1.4) and identifying the hypernymic relations using a simple heuristic based on the relation pattern is  $a^{24}$

**Results and Discussion.** Table 5.22 shows the results of TaxoEmbed and all its comparison systems in detecting hypernyms outside Wikidata. As expected, Yago and WiBi achieve the best overall results. Nonetheless, TaxoEmbed, which relies solely on distributional information, performs competitively when compared to DefIE, improving recall over the latter in most domains, and even surpassing Yago in technical areas like biology or health. On the other hand, TaxoEmbed does not perform particularly well on media and physics. Overall, TaxoEmbed is able to discover novel hypernymic relations not captured by any other system (e.g. therapy for radiation treatment planning in the health domain, or decoration for molding in the art domain).

**Final Remarks.** TaxoEmbed, to best of our knowledge, is the first supervised hypernym discovery framework defined entirely at the sense level. As we validated experimentally throughout this section, this strategy allowed TaxoEmbed, on the one hand, to improve its domain adaptation procedure by exploiting the structured knowledge of BabelNet (Section 5.3.1.1) and, on the other, to expand effectively its training set by including heterogeneous OIE-derived knowledge from KB-Unify (Section 5.3.1.2). Moreover, even though all our experiments were carried out on test pairs also at the sense level, TaxoEmbed can be utilized to discover hypernyms at the word level by: (1) considering all the available senses of a given term inside BabelNet, or (2) exploiting the fact that SenseEmbed defines a shared vector space for word and senses (cf. Section 2.2.3.2) and considering the embedding associated with the term directly. In either case, one distinguishing feature of TaxoEmbed is its approach based on casting the hypernym extraction task as a ranking problem (cf. Section 5.3.1.3): this feature enabled a flexible evaluation framework, never applied,

<sup>24</sup>For all the reasons put forward in this section, this procedure is not exhaustive, and might miss some hypernymic relation instances extracted with different kinds of relation phrases: however, due to the nature of the definitional corpus targeted by DefIE, where sparsity and noise are limited, this heuristic arguably provides an accurate estimate.

to date, in the context of evaluating hypernym detection or taxonomy learning systems (Camacho Collados, 2017), and allowed us to combine highly demanding metrics for the quality of the candidate given at a certain rank, as well as other measures which consider the rank of the first valid retrieved candidate.

On the other hand, a current limitation of this framework, which was not addressed in the present section, is that its novel evaluation paradigm makes it difficult to carry out an extensive comparison between TaxoEmbed and most hypernym detection approaches published in the field. To this aim, a shared SemEval task specifically targeted to hypernym discovery has actually been organized, and it is ongoing at the time of writing.<sup>25</sup>

---

<sup>25</sup> <https://competitions.codalab.org/competitions/17119>



## Chapter 6

# Release

“ K +B Ç

[Continue to spur a running horse.]

Yamamoto Jin'emon

In this chapter we showcase all the resources and tools that have been released publicly in association with the contributions presented in Chapters 4 and 5. First of all, the backbone of every approach in terms of reference sense inventory and knowledge resource, i.e. BabelNet (Section 2.1.3) and its API, which we utilized to retrieve all the semantic information about synsets, their lexicalizations in the various languages, their connection inside the semantic network, as well as all the inter-resource mappings from and to WordNet, Wikipedia, Wikidata, DBpedia, Freebase, etc. The BabelNet data and API are freely available for research purposes and licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.<sup>1</sup> Of course, all the data drawn from external resources (including Wikipedia, Wikidata, WordNet, etc.) are released under the terms of the respective owners' licenses<sup>2</sup>. As regards the programmatic use of the BabelNet, we relied on the Java API for the majority of our implementations and experimental evaluations; however, the BabelNet API provides an HTTP RESTful service (accessible through an HTTP interface that returns JSON) along with the Java package, which makes it usable within any other programming language. In addition, starting from version 2.0 onward, BabelNet has been integrated in the so-called Linguistic Linked Open Data (LLOD) cloud, a part of the Linked Open Data cloud made up of interlinked linguistic resources (Chiarcos et al., 2011). This integration was achieved by encoding the knowledge in BabelNet using the Lemon RDF model (McCrae et al., 2011), and then providing a public SPARQL endpoint<sup>3</sup>. For any further detail about the use of

<sup>1</sup><http://creativecommons.org/licenses/by-nc-sa/3.0>

<sup>2</sup><http://babelnet.org/licenses>

<sup>3</sup><http://babelnet.org/sparql>

BabelNet, a comprehensive guide is available on the BabelNet website<sup>4</sup>.

All the tools based on BabelNet that we exploited throughout this thesis, including Babelfy (Section 2.2.2.3)<sup>5</sup>, Nasari (Section 2.2.3.3)<sup>6</sup>, and SensEmbed (Section 2.2.3.2)<sup>7</sup>, are all publicly available under the terms of the same license of BabelNet. In particular, Babelfy is equipped with an API along the lines of the BabelNet one, which includes a Java package and an HTTP RESTful service. The Babelfy API allows a programmatic use of Babelfy where the user can specify in detail the format of the input text, providing token-specific information such as part-of-speech tag, lemma, language, or even sense labels (which will be used by Babelfy as constraints when building the semantic graph). All the details for these use cases are reported in the online guide in the Babelfy website<sup>8</sup>.

Consistently with the above, most of the released material presented in this chapter complies with the same license, and is publicly available on dedicated websites. It is worth noting that all these contributions, together with the resource and tools we used, do not represent a collection of individual contributions per se, but rather a series of research efforts revolving around the common vision of the MultiJEDI<sup>9</sup> project, a 5-year ERC starting grant (2011-2016) with the objective of enabling multilingual text understanding. In fact, MultiJEDI led to the development of BabelNet in the first place, and defined the common thread that bundles together all the knowledge-based approaches that rely on it.

In the following sections we go over the released material in the same order the corresponding contributions have been treated across Chapters 4 and 5, i.e. Sew (Section 6.1), EuroSense (Section 6.2), SenseDefs (Section 6.3), and finally the released material associated with DefIE, KB-Unify and TaxoEmbed (Section 6.4). In each case, we give the details of the release and its format.

## 6.1 Sew

Sew (Raganato et al., 2016b)<sup>10</sup> is a sense-annotated corpus automatically built from Wikipedia, described in Section 4.1. In Sections 4.1.2 and 4.1.3 we considered a version of Sew obtained from an English Wikipedia dump of November 2014, which we subsequently used in the experimental evaluation. That specific version of Sew is the one we release publicly: it is available both as a complete (i.e. comprising all the sense annotations gathered by the hyperlink propagation pipeline before applying the conservative policy, including overlapping mentions) and as a conservative (i.e. the corpus after the final stage of the pipeline, henceafter applying the conservative policy). The former setting, with more than 44 million additional sense annotations, is suitable for high-coverage applications; the latter, designed to retain only the most confident propagations and no overlapping mention, is the one we used for both the intrinsic and extrinsic evaluations (cf. Section 4.1.3).

<sup>4</sup><http://babelnet.org/guide>

<sup>5</sup><http://babelfy.org>

<sup>6</sup><http://lcl.uniroma1.it/nasari>

<sup>7</sup><http://lcl.uniroma1.it/senseembed>

<sup>8</sup><http://babelfy.org/guide>

<sup>9</sup><http://multijedi.org>

<sup>10</sup><http://lcl.uniroma1.it/sew>

Figure 6.1. Excerpt from the XML sample of a Wikipage in Sew. The complete sample is available (in both XML and human-readable form) at: <http://lcl.uniroma1.it/sew/sample/sample.html>.

**Format.** Each Wikipage is stored in an individual XML file, named with the corresponding Wikipage title. The file contains a `wikiArticle` tag with the attributes `language` (ISO code of the language) and `title` (the actual title of the Wikipage). In turn, the `wikiArticle` tag contains two main tags:

- ^ `text` : the plain-text, one sentence per line, of the Wikipage as a whole, excluding infoboxes, image captions, references and categories<sup>11</sup>;
- ^ `annotations` : the complete list of sense annotations. Each sense annotation is encoded with an `annotation` tag having the following attributes:

`babelNetID` : the unique sense identifier as provided by BabelNet;

`mention` : the surface form of the mention as it appears in the plain-text of the Wikipage;

`anchorStart` : the token-based starting index (inclusive) of the sense annotation;

`anchorEnd` : the token-based ending index (exclusive) of the sense annotation;

`type` : the symbol associated with the propagation heuristic from which this sense annotation has been obtained (as reported in Table 4.1).

<sup>11</sup>This text has been escaped using XML entities. In order to retrieve the actual human-readable character, many functions can be used to unescape it (e.g., in Java, the class `StringEscapeUtils` provided by the `apache-commons-lang` API).

An excerpt of a sample XML file is given in Figure 6.1. All the sense annotations released with Sew are licensed under the same license of BabelNet, except for the original Wikipedia hyperlinks (marked with the type `HL`), which are compliant with Wikipedia and released under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC-BY-SA).<sup>12</sup>

In addition to the two versions of Sew, we also release the vector representations used in the experimental evaluations of Sections 4.1.3.3 and 4.1.4. As regards the sparse representations, both WBSew and SB-Sew (cf. Section 4.1.3.3) are available in two different versions, one where frequencies are estimated using raw counts (file ending with `rc.tsv`) and using lexical specificity (file ending with `ls.tsv`). Both versions consist in tab-separated files with a single vector for each line. The format is as follows:

```
ENTITY \t \t COMPONENT:VALUE \t ... \t COMPONENT:VALUE
```

where ENTITY is either the Babel synset (WB-Sew) or the Wikipage (SB-Sew) being represented by the vector, and the COMPONENT:VALUE pairs constitute the non-zero dimensions of the vector. Each COMPONENT is either a Wikipage (WB-Sew) or a BabelNet synset (SB-Sew).

As regards the dense representations, both Sew-Embed studied experimentally are available (cf. Section 4.1.4). Each version is encoded in a plain-text file with a single vector in each line, following the standard space-separated format of most released embedding representations:

```
SYNSET VALUE VALUE ... VALUEn
```

where SYNSET is the Babel synset being represented by the vector, and VALUE<sub>1</sub> : : VALUE<sub>n</sub> constitute the numerical components of the vector. Each vector has 400 dimensions (i.e.  $n = 400$ ) as both the external representations used are based on a 400-dimensional vector space (cf. Section 4.1.4). When a SYNSET is not covered by Sew-Embed is represented by an all-zero vector. Both sparse and dense representations are available under the same license of BabelNet (Creative Commons Attribution-Noncommercial-Share Alike 3.0 License).

## 6.2 EuroSense

EuroSense (Delli Bovi et al., 2017)<sup>13</sup> is a multilingual sense-annotated resource automatically built via the joint disambiguation of the Europarl parallel corpus in 21 languages, described in Section 4.2. We release two versions of EuroSense: a high-coverage version (i.e. the one obtained from the first stage of the EuroSense pipeline, described in Section 4.2.1), and a high-precision version (i.e. the one obtained at the end of the pipeline, described in Section 4.2.2). Similarly to Sew, the former version includes almost 93 million additional sense annotations (including overlapping mentions) and hence it is suitable for high-coverage applications, while the latter, refined with distributional semantic similarity, is oriented towards high-precision. The high-precision version of EuroSense is the one utilized in the extrinsic evaluation (Section 4.2.4.2).

<sup>12</sup><https://creativecommons.org/licenses/by-sa/3.0/>

<sup>13</sup><http://lcl.uniroma1.it/eurosense>

Figure 6.2. Excerpt from the XML sample of a sentence in the high-precision version of EuroSense .

Format. Both versions are stored in XML files with UTF-8 encoding. Each file contains a list of sentence tags, with an incremental id (starting from 0) as attribute. Each sentence contains a list of text tags, corresponding to the tokenized texts of the sentence in a given language (the ISO code of the language is encoded in the lang attribute), and an annotations tag, which includes all the sense annotations provided for that sentence. Each annotation includes a Babel synset identifier and has four (or six) attributes:

- lang: the language of the sense annotation (as ISO code);
- type (only in the high-precision version): whether the disambiguation was performed by `BABELFY` or `NASARJ`
- anchor: the exact surface-form match found in the sentence of the corresponding lang;
- lemma: the normalized form (lemma) of the annotation's anchor;
- coherenceScore: the coherence score associated with the annotation;
- nasariScore: the Nasari score associated with the annotation. This attribute is set to `--` when the annotation has type `BABELFY`.

An excerpt of a sample XML file is given in Figure 6.2. Both versions of EuroSense are available under the same license of BabelNet (Creative Commons Attribution-Noncommercial-Share Alike 3.0 License).

### 6.3 SenseDefs

SenseDefs (Camacho Collados et al., 2016a)<sup>14</sup> is a multilingual large-scale corpus of automatically disambiguated definitions coming from BabelNet, described in Section 4.3. Since SenseDefs was obtained by adapting the same disambiguation pipeline used for EuroSense (Section 4.2), the released resource is also available in two versions, complete and high-precision, obtained after the first and after the final stage of the pipeline, respectively.

<sup>14</sup><http://lcl.uniroma1.it/disambiguated-glosses>

Figure 6.3. Excerpts from the English XML sample of SenseDefs showing two definitions for Palaeochiropteryx and Abraham Lincoln drawn from, respectively, Wikipedia and WordNet. The former is taken from the complete version of SenseDefs, while the latter from the high-precision version.

**Format.** The release of SenseDefs is split according to the specific source of the definitions (WordNet, Open Multilingual WordNet, Wiktionary, Wikipedia, Wikipedia disambiguation pages, OmegaWiki, and WikiData). For each subset of the corpus, definitions are further divided by language, with each language being encoded in an individual XML file named with the corresponding ISO code. Each file contains a list of definition tags, with the respective id (e.g. page titles in Wikipedia, or osets in WordNet) as attribute. Then, each definition contains the plain-text of the original definition (as available in the given resource), as well as the set of sense annotations. A sense annotation comprises a Babel synset identifiers enclosed in an annotation tag with the following attributes:

- source: whether the annotation was disambiguated by BABELFY, the `MCS` back-off strategy (only in the complete version), or `NASARI` (only in the high-precision version);<sup>15</sup>
- anchor: the exact surface-form match found in the text of the definitions;
- bfScore: the internal confidence score used by BabelFY to enable/disable the back-off strategy on the annotation;
- coherenceScore: the coherence score associated with the annotation;
- nasariScore: the Nasari score associated with the annotation (only in the high-precision version).

An excerpt of a sample XML file is given in Figure 6.3. Both versions of SenseDefs are available under the same license of BabelNet (Creative Commons Attribution-Noncommercial-Share Alike 3.0 License).

<sup>15</sup>When the annotation has source `MCS` both the bfScore and the coherenceScore attributes are set to `--`.

## 6.4 OIE-derived Resources

In Chapter 5 we studied three diverse but effective approaches to the extraction of relational knowledge at the sense level. Even though none of these approaches was directly designed to generate a full-edged knowledge resource, but rather they were focused on either defining a prototypical sense-aware extraction pipeline (Section 5.1), or on establishing a robust and general sense-aware framework (Sections 5.2 and 5.3), we still utilized them to produce relational knowledge tailored to the experimental evaluation of their performances. Hence, along the lines of the resources presented in the previous sections, we also released most of this evaluation material for the use and scrutiny of the research community. These experimental data can be used as a comparison when developing alternative or more sophisticated systems for the tasks we address, as relational knowledge for a variety of downstream NLP systems to build upon, or as a way to replicate our results.

### DefIE

DefIE (Delli Bovi et al., 2015b)<sup>16</sup> is a full-edged sense-aware OIE pipeline designed for definitional knowledge, described in Section 5.1. For the purpose of its broad experimental evaluation in Section 5.1.4, we considered the output of DefIE over an input corpus composed of the whole set of textual definitions in BabelNet 2.5, which we used for all our experiments. We release the following:

- ^ The complete set of semantic relations extracted from the input corpus with more than 10 relation instances, which is available as plain-text file with the following format:

```
RELATION ID \t RELATION PATH \t RELATION STRING
```

where **RELATION ID** is a unique identifier for the relation, **RELATION PATH** is the corresponding path in the syntactic-semantic graph and **RELATION STRING** is the lemmatized relation pattern;

- ^ Contains the complete set of extracted relation triples for the relations above, as plain-text file. The format is the following:

```
SUBJECT \t RELATION ID \t OBJECT
```

where **RELATION ID** is the relation identifier and **SUBJECT**(**OBJECT**) refer to the Babel synset identifier of the subject (object) of the triple;

- ^ The subject and object semantic type distributions for each semantic relation (cf. Section 5.1.2), as plain-text file formatted as follows:

```
RELATION ID_X \t DOMAIN CLASS_ID \t PROBABILITY \t ...
RELATION ID_Y \t RANGE CLASS_ID \t PROBABILITY \t ...
```

<sup>16</sup><http://lcl.uniroma1.it/defie>

where `RELATION ID` is the relation identifier, `DOMAIN CLASS ID` (`RANGE CLASS ID`) are Babel synset identifiers associated to domain (range) semantic classes and `PROBABILITY` is the corresponding probability value. For each distribution, semantic classes are sorted by decreasing probability value;

- ^ The relation taxonomy derived from the extracted set of semantic relations (cf. Section 5.1.3), as plain-text file. Each line of the file encodes a single edge of the taxonomy graph in the following format:

```
HYPONYM RELATION ID \t HYPERNYM RELATION ID
```

where `HYPONYM RELATION ID` and `HYPERNYM RELATION ID` denote the relation identifiers of the hyponym and hypernym relation, respectively.

### KB-Unify

KB-Unify (Delli Bovi et al., 2015a)<sup>17</sup> is a sense-aware framework for integrating the output of different OIE systems into a single unified and fully disambiguated knowledge repository, described in Section 5.2. We carried out its experimental evaluation in Section 5.2.2, after running its unification pipeline of a set on four individual OIE-derived KBs (cf. Section 5.2.2). We release the following:

- ^ A disambiguated version of the two unlinked KBs in the experimental setup i.e. `Nell` and `ReVerb`, obtained as output of the disambiguation module (cf. Section 5.2.2.1). They are both available as plain-text files with the following format:

```
ARGUMENT1 ARGUMENT2 \t RELATION 1 \t ... \t RELATION N
```

where `ARGUMENT1` and `ARGUMENT2` are either Babel synset identifiers (if the corresponding triple was disambiguated) or the original argument strings (otherwise). `RELATION 1` to `RELATION N` denote all the original relation strings in which the two arguments occurred in the original KBs;

- ^ All the cross-resource relation alignments obtained as output of the alignment procedure (Section 5.2.1.4), one plain-text file per KB pair. Each file encodes an alignment on each line as tab-separated string containing the two relation identifiers and the corresponding alignment confidence. Each individual KB is also included separately in the package;
- ^ The unified KB obtained as a final result of KB-Unify's unification pipeline. We include several versions of the unification output constructed with different thresholds for the alignment confidence, as well as reduced versions of these outputs obtained by considering only the top 10k semantic relations from each individual KB. Each version of the unified KB comprises two files, one with the complete set of relation synsets (including singletons) and another one with the complete set of relation triples. The former encodes one relation synset per line, using a tab-separated string where an incremental relation identifier is associated with the following string:

<sup>17</sup><http://lcl.uniroma1.it/kb-unify>

$$\{ [KB1] :R1 \quad \backslash t \quad [KB2] :R2 \quad \dots \quad \backslash t \quad [KBN] :RN \}$$

where  $KB1 : : KBN$  denote the individual source KBs of relations  $R1 : : RN$  respectively. The latter file, instead, contains all the relation triples, one per line, encoded as tab-separated strings in the same way as in the OIE release.

Finally, we release the evaluation data used in all the experiments of Section 5.2.2 for replication purposes, including the random samples, the corresponding gold standards, and the guidelines provided to the annotators for each task.

### TaxoEmbed

TaxoEmbed (Espinosa Anke et al., 2016a)<sup>18</sup> is a supervised distributional framework for domain-specific hypernym discovery at the sense level, described in Section 5.3. Its performance was evaluated experimentally in Section 5.3.2, where we trained a variety of TaxoEmbed models on a large heterogeneous training set drawn from both Wikidata and KB-Unify (cf. Section 5.3.1). In order to enable the research community to replicate our results, we release the complete training dataset of TaxoEmbed, available as a package comprising two files (one for the Wikidata pairs, and another one for the OIE-derived pairs), together with the domain labels obtained using Nasari (cf. Section 5.3.1.1) and the SensEmbed vector space. All the data are expressed with respect to the BabelNet sense inventory (i.e. with Babel synset identifiers). Finally, we also release the Python implementation of TaxoEmbed used in our experiments, available from an open-source BitBucket repository<sup>19</sup>.

<sup>18</sup><http://wwwusers.di.uniroma1.it/~dellibovi/taxoembed>

<sup>19</sup><https://bitbucket.org/luisespinosa/taxoembed>



## Chapter 7

# Conclusion

We can only see a short distance ahead  
but we can see plenty there that needs to be done.  
Alan M. Turing

In this thesis we looked closely at the intersection between two prominent areas of Natural Language Processing: Information Extraction, i.e. the automatic extraction and formalization of machine-readable knowledge from natural language text (cf. Section 2.3), and Lexical Semantics, the field of study concerned with establishing and modeling the meaning of lexical items in a computational way. We saw that one crucial issue that bundles them together is lexical ambiguity. Broadly speaking, dealing with ambiguity is indeed one of the long-standing challenges in NLP, as various types of ambiguity (lexical, structural, pragmatic) can arise at many different levels within the process of understanding natural language utterances; when it comes to identify, extract and encode effectively factual content, which is the focus of IE, ambiguity at the lexical level is a particularly striking problem. Let us consider once again the second example of Section 3.2:

Washington is the capital of the United States

In this case both the subject argument and the relation phrase are ambiguous, and resolving these ambiguities is crucial for encoding this piece of factual knowledge correctly. Among the various techniques for automatically linking and disambiguating lexical items (Sections 2.2.1 and 2.2.2), we saw that a promising strategy of dealing with this problem on a large scale consists in leveraging knowledge resources (Section 2.1): in fact, efforts in creating, developing, managing, integrating and interconnecting structured knowledge using a variety of lexico-semantic resources (lexicons, dictionaries, encyclopedias, databases, knowledge graphs) are widespread in the research community (Gurevych et al., 2016).

The key role of knowledge resources in NLP is what enabled us to draw an important connection between the challenge of lexical ambiguity and the knowledge acquisition bottleneck phenomenon, which we discussed in Chapter 1. In fact, on the one hand, we showed that extracting information from open text (Information Extraction) is one of those tasks where facing lexical ambiguity is of the utmost importance; on the other, a key step towards developing large-scale high-quality disambiguation systems consists in populating and enriching knowledge resources (i.e. overcoming the knowledge acquisition bottleneck), especially with the kinds of syntagmatic relations that are usually encoded implicitly in open and unstructured text. In the example above, having a lexicalized semantic network at our disposal, where the concept of capital as official seat of a country's government is connected (i.e. has a semantic relation) with the entity Washington as the U.S. capital, would be decisive to resolve all lexical ambiguities in that relation instance.

In light of the above, our main objective in this thesis (Section 1.1) was that of addressing both problems in a synergistic way, by developing a principled approach to open-text knowledge extraction based on explicit semantic analysis at the sense level. To this aim, we operatively considered a two-fold objective: (1) developing reliable methods to harvest sense-level information on a large scale, and (2) introducing sense-aware techniques into the well-established OIE paradigm for extracting relational knowledge. In tackling both tasks we adopted a knowledge-based strategy and leveraged a wide-coverage, multilingual knowledge base and semantic network, i.e. BabelNet (Section 2.1.3), as a fundamental backbone. In fact, resources like BabelNet, where lexicographic and encyclopedic knowledge is seamlessly integrated, represent a first important step in the direction of large-scale sense-level approaches designed to scale up in terms of scope and languages (Delli Bovi and Navigli, 2017). Throughout Chapters 4 and 5 we relied on BabelNet not only as a wide-coverage sense inventory for the disambiguation tools and sense-aware methods we used or developed, but we also took advantage of the scaffolding of structured lexico-semantic information it provides in a variety of ways. From this perspective, the contributions we put forward in this thesis build upon BabelNet, in the attempt to:

1. Overcome the knowledge acquisition bottleneck with respect to sense-level information, by constructing and delivering to the research community a series of large-scale corpora of various kinds (encyclopedic text, parallel text, definitional text), all equipped with sense annotations from the BabelNet sense inventory. In fact, even though WordNet and Wikipedia have been the de facto standards in terms of reference sense inventories for WSD and EL, respectively, we saw in Section 3.1 that BabelNet takes the best of both worlds, enabling sense-annotated corpora suitable for both WSD and EL, and based on a unified sense inventory that extends to all the languages covered by Wikipedia;
2. Extending the key idea behind BabelNet (i.e. integrating and unifying complementary information from many individual resources using semantic analysis) from lexical knowledge to relational knowledge, both because the semantic network of BabelNet fails to cover explicitly a great deal of relational knowledge (cf. Section 2.1.3), and because most repositories of semantic relations to date, especially when derived from IE or OIE approaches, are designed as stand-alone contributions with their own structures and type inventories.

## 7.1 Wrapping Up

As discussed in Chapter 1, the main focus of this thesis is on OIE, inherently unsupervised, as strategy to extract relational knowledge. Thus, in contrast to other paradigms geared towards the same goal (e.g. Knowledge Base Completion), our starting point was solely open and unstructured text in natural language. This is why our first and foremost task, addressed in Chapter 4, was that of developing robust, flexible and reliable methods to automatically obtain sense-level information on a large scale. Given our choice of BabelNet as underlying knowledge resource, we saw in Section 3.1 that the size and scope of the sense inventory do not allow any degree of human intervention: on the other hand, while off-the-shelf disambiguation systems have proven to be a viable way of harvesting sense annotations, there is still large room for improvement in fully automatic pipeline based on them, as we discussed at the beginning of Chapter 4. Throughout that chapter, which tackles the first objective of Section 1.1, we showed how exploiting at best the shape and features of the target corpus is key to achieve our goal. We considered three different disambiguation scenarios, where we adopted a similar methodological approach: first, we investigated a disambiguation pipeline suitable for the target text; then, we applied it to produce and release to the community a full-fledged sense-annotated resource; finally, we carried out an extensive evaluation, both intrinsic and extrinsic, to assess the sense annotation quality of such resource.

We started in Section 4.1 with a semi-structured resource (i.e. Wikipedia) as disambiguation target, and developed Sew (Raganato et al., 2016b), a Wikipedia-based sense-annotated corpus which, to date, constitutes the largest BabelNet-annotated resource available. With the broad and comprehensive experimental evaluation of Section 4.1.3, which also includes a dedicated study on vector representations (Delli Bovi and Raganato, 2017), we demonstrated that, in the special case of Wikipedia, a large amount of high-quality sense annotations can be obtained automatically without employing off-the-shelf disambiguation systems at all. Furthermore, our extrinsic experiments showed that having this unprecedented number of sense annotations can greatly boost simple vanilla approaches, enabling them to perform on par with more sophisticated state-of-the-art systems in their respective tasks, thereby setting new performance baselines in the field.

In Sections 4.2 and 4.3, instead, we shifted our focus to a parallel corpus, where we brought together equivalent translations of the same English sentences, and to a definitional corpus, where we gathered all the textual definitions associated with a given denotandum from different resources and languages. In both cases, we could not rely on semi-structured knowledge already embedded in the corpus (as in Section 4.1), and we designed a two-stage pipeline using two external tools: Babelfy (Section 2.2.2.3), a state-of-the-art graph-based WSD/EL system, and Nasari (Section 2.2.3.3), a vector representation for all the nominal concepts and entities in BabelNet. The gist of this disambiguation pipeline was exploiting at best an enriched multilingual context to harvest as many sense annotations as possible with Babelfy, which implicitly enforced cross-language semantic coherence, and then refining the disambiguation output using distributional semantic similarity to correct the structural bias of Babelfy. By applying this pipeline on the two corpora mentioned above, we obtained two resources EuroSense (Delli Bovi et al.,

2017) and SenseDefs (Camacho Collados et al., 2016a), which constitute the largest sense-annotated parallel corpus and the largest sense-annotated definitional corpus, respectively. Also, compared to sense-annotated corpora obtained using only Babelify (and without taking into account the features of the target text), our disambiguation pipeline improved considerably the estimated accuracy of sense annotations, as shown in Table 4.20. An additional advantage of this pipeline, compared to the Wikipedia-specific methods used to construct Sew, is that using a two-stage process enabled us to release two versions of the corresponding resource, each suitable to certain sets of applications, and to associate one or more confidence scores to each sense annotation (see Sections 6.2 and 6.3), which can be used to further tune the resource for a specific task, application, or use.

Chapter 5, instead, is devoted to the second objective of Section 1.1: once equipped with reliable automatic methods to harvest sense-level information from open text, we were able to reframe the OIE paradigm at the sense level, studying where and how sense-aware methods can effectively enhance the process of extracting relational knowledge. Even though there have been previous attempts to inject semantic features into OIE systems, including approaches dealing explicitly with phenomena like lexical ambiguity and synonymy (Section 3.2), we showed that they still have a number of practical limitations, mostly connected with the need to cope with data sparsity and noisy extractions, which prevent them to enforce a deeper semantic analysis. We addressed this issue in Section 5.1, where we designed a full-edged OIE pipeline targeted at the denser, virtually noise-free setting of definitional text, and we integrated a fully sense-aware approach into the extraction process. The resulting quasi-OIE system, DefIE (Delli Bovi et al., 2015b), exploited a comprehensive semantic analysis to extract unambiguous relation triples, with 'semantic' relation patterns that could be effectively arranged in a relation taxonomy without devising complex alignments or subsumption strategies (cf. Section 3.2.1). With the broad experimental evaluation of DefIE (Section 5.1.4) we showed that a fully sense-aware OIE pipeline on a considerably smaller corpus results in comparable (or greater) performances than standard OIE pipelines on massive, even Web-scale, noisy corpora, in addition to all the advantages of anchoring the extracted knowledge to the semantic network of BabelNet. The results obtained with DefIE suggest that, when extracting factual information from open text, it might be convenient to analyze the target corpus at hand, and perhaps try to isolate knowledge-rich pieces of text (e.g. definitions), instead of blindly process massive amounts of noisy data and then devise sophisticated strategies to refine incorrect extractions.

Another issue of current OIE systems, pointed out at the beginning of Chapter 5, is the fact that they tend to produce isolated repositories of relational knowledge, typically featuring their own internal structure and type inventories (cf. Section 2.3.1), and with very few attempts of integration or interoperability among them. This issue becomes even more critical for OIE systems, where the relation inventory is not specified in advance and there is no way of establishing, without manual inspection, whether two systems have extracted the same piece of information, or whether they have discovered the same semantic relation. Not even sense-level approaches, such as PATTY (Section 3.2.1), WiSeNet (Section 3.2.2), or DefIE (Section 5.1), deal with the problem. This situation motivated the development of KB-Unify (Delli Bovi et al., 2015a), a sense-aware framework for integrating

the outputs of different OIE systems into a single, unified and fully disambiguated knowledge repository: in Section 5.2, where we described the unification pipeline of KB-Unify and its experimental evaluation, we demonstrated that semantic analysis at the sense level can be used to interconnect relational knowledge, even when derived from very heterogeneous sources, such as the human-crafted semantic relations of Nell and the relation synsets generated by Patty and WiSeNet. Since the unification procedure operated at the sense level, we devised an ad-hoc disambiguation strategy for a collection of relation triples (Section 5.2.1.3) where, similarly to the disambiguation approaches of Chapter 4.1, providing a rich and meaningful context for disambiguation was key to obtain high-quality results.

Finally, in Section 5.3 we shifted the focus from OIE, based on the unconstrained extraction of an unspecified number of semantic relations, to the opposite end of the IE spectrum, i.e. hypernym discovery, which is concerned with extracting only one specific kind of semantic relation: hypernymy. Apart from the prominent role of hypernymic information in the field, we were motivated by the fact that the IE and OIE systems treated in the previous sections were also capable of extracting valuable hypernymic information. Thanks to KB-Unify, then, this information could be 'semanticized' and integrated with hypernymic knowledge drawn from different sources (e.g. Wikidata) and used to train a supervised hypernym discovery model at the sense level. This was indeed the gist of TaxoEmbed (Espinosa Anke et al., 2016a), a sense-level framework for supervised hypernym discovery that relied on the vector space of SensEmbed (Iacobacci et al., 2015) to learn a domain-specific linear transformation from hyponyms to hypernyms. By redefining hypernym discovery at the sense level, TaxoEmbed was able to leverage an heterogeneous training set with both human-curated hypernymic knowledge from Wikidata and OIE-derived hypernymic knowledge from KB-Unify, thereby achieving its best performance in our experimental evaluation (Section 5.3.2).

## 7.2 Future Work and Perspectives

Notwithstanding the key contributions presented in this thesis, showcased in Section 1.3 and then further discussed in Section 7.1, all the proposed approaches do not represent conclusive solutions to the tasks they address, but rather constitute a leap forward that opens up avenues for future work. In fact, in some cases they contributed to reshape the landscape of their field of study by putting forward resources that were not available before, while at the same time setting new performance baselines on standard evaluation benchmarks thanks to these resources (cf. Chapter 4). In other cases they opened a new perspective on their fields (cf. Chapter 5), shedding a light on some aspects that were previously overlooked or neglected, and proposing new ways and new methodologies to confront with the problem at hand.

First of all, a number of short- and medium-term improvements can be envisaged for each contribution presented in this thesis, some of which are currently under investigation. As regards Sew (Section 4.1), for instance, there are additional ways of exploiting Wikipedia-derived knowledge to propagate sense annotations: in particular, an aspect that was not completely captured by the propagation pipeline described in Section 4.1.1 is multilinguality. The pipeline was developed and applied

only on the English Wikipedia, but it does not actually carry language-specific features<sup>1</sup> and could be extended to other languages. Furthermore, an array of propagation heuristics can be developed on the basis of the fact that the various monolingual Wikipedias can be seen altogether as a massive comparable corpus, where two Wikispaces referring to the same entity not only come from the same topic, but they actually describe the same subject. Speaking of multilinguality, the disambiguation pipeline used in Sections 4.2 and 4.3 (which, instead, relies heavily on multilinguality) can also be improved in this respect: when constructing a multilingual context for Babelfy, for both EuroSense and SenseDefs, all the languages were treated equally, neglecting the fact that Babelfy, as any other off-the-shelf system, does not perform equally well on all languages. As noted in Section 2.2.1, the performance of a knowledge-based system depends strictly on the quality of the underlying resource, and hence ultimately on the structured knowledge available for a given language in that resource. This means that, while the performance of less-resourced languages might be improved by propagating disambiguation decisions from more content languages, the vice versa could also happen, especially when a large amount of languages is considered as the same time, as in SenseDefs.<sup>2</sup> Therefore, in order to effectively create and maintain a positive synergistic effect, the mutual interaction among languages at disambiguation time should be further studied and controlled. Finally, a shortcoming of all the sense-annotated resources in Chapter 4 is that verbal senses are very rarely captured. Sew manages to annotate only monosemous verbs from WordNet with a specific heuristic, as all the Wikipedia-based propagation hyperlinks are only associated with nominal senses (in fact, the Wikipedia sense inventory itself is strictly nominal). On the other hand, the disambiguation pipeline of EuroSense and SenseDefs utilizes Babelfy to disambiguate verbs in the first stage, but in the refinement step only nominal senses are considered. Disambiguating verbal senses with high accuracy is a well-known problem in WSD (Raganato et al., 2017a): verbs represent the word class with the highest average polysemy, and where many sense-level distinctions are extremely fine-grained. Nevertheless, they often constitute a crucial component to model, bearing most of the semantic content of a sentence.

Similarly, the approaches presented in Chapter 5 are not flawless and improvable in many ways. The competitive performances of DefIE (Section 5.1) result from the effective interplay between the characteristics of the target text and the less-constrained extraction procedure, designed to trust the underlying data. As shown in Section 5.1.4.5, applying the same quasi-OIE pipeline to non-definitional text causes extraction accuracy to drop substantially. To broaden the scope of DefIE there are two viable options: enforcing syntactic and semantic constraints to cope with noisy extractions, as in Patty and WiSeNet (cf. Section 3.2), or designing a general-purpose pipeline where DefIE is coupled with a definition extraction module,

<sup>1</sup>Apart from the preprocessing phase (tokenization, part-of-speech tagging and lemmatization) which, however, is available for many languages nowadays, and the number of languages covered will likely be increasingly large in the future, thanks to projects like The Universal Dependencies (Nivre et al., 2016).

<sup>2</sup>In fact, Table 4.17 showed that, while our disambiguation pipeline consistently achieved the highest F-scores, in some cases the Babelfy baseline reported a higher precision.

<sup>3</sup>In fact, Nasari relies on Wikipedia to construct its vector representations (cf. Section 2.2.3.3).

designed to spot and isolate definitional knowledge across the target text (Navigli and Velardi, 2010; Benedictis et al., 2013; Espinosa Anke and Saggion, 2014; Espinosa Anke et al., 2015; Dalvi et al., 2015). Also, the quality of DefIE's extractions depend crucially on the quality of disambiguation: as we discussed above, disambiguation systems to date tend to struggle with verbal senses which, instead, are usually the head component in OIE-derived relation phrases (cf. Section 2.3.2). Finally, as regards KB-Unify and TaxoEmbed (Sections 5.2 and 5.3), the unification pipeline proposed in Section 5.2.1, albeit effective, represents only a first attempt towards this largely unexplored task. First of all, the disambiguation and alignment stages, which KB-Unify performs as two successive steps, represent two tightly interconnected processes: in some cases, cross-resource alignment could still be carried out at the word level (e.g. when the subject and object arguments are less ambiguous); at the same time, some cross-resource alignments could also be exploited at disambiguation time, to further enrich the context of an ambiguous relation triple with knowledge from aligned resources. In other words, a potentially more effective solution would be that of performing disambiguation and alignment jointly. In addition, the alignment procedure also suffers from a structural disadvantage: as we showed in *Patty* (Section 3.2.1), modeling the semantics of a relation using only the semantics of its arguments can lead to false positives, i.e. very different relations that are defined on very similar argument sets (e.g. *is the mother of* vs. *is the father of*, or *is a city in* vs. *is the capital city of*). On the other hand, only considering features of the relation phrase is also suboptimal, since very similar relation phrases might identify very different semantic relations (e.g. *played a* vs. *played with*, *played in*, *played the*). The specific case of hypernymic relations is even more critical, since they can be identified with a plethora of different patterns and, at the same time, the generality of these relations makes it unpractical to study the shape of their arguments sets in the vector space, especially when such diverse sets are replaced by a single centroid, as KB-Unify does: this shortcoming, in turn, reflects on TaxoEmbed, where many valuable OIE-derived training pairs, not aligned with *Nell* (i.e. a relation in the first place (cf. Section 5.3.1), might not have been captured.

**Long-Term Perspectives.** From a broader standpoint, the advancements described in this thesis, along with many other research efforts that are pushing forward the computational study of Lexical Semantics, pave the way for a greater, more general reshaping and adaptation of the field's landscape. When the deep learning tsunami hit the shores of NLP, the general feeling was that in a few years' time the next big step of deep neural networks would have been full natural language understanding (Manning, 2015): however, the impressive results achieved by these architectures in other areas of AI, such as Computer Vision, seemed difficult to replicate in the domain of NLP. Among many other important reasons, this difficulty is also connected with the lack of large-scale labeled datasets for many NLP tasks; in this respect, Lexical Semantics is one of the most problematic areas, also due to phenomena like the knowledge acquisition bottleneck. In fact, for the very same reasons, not even wide-coverage multilingual semantic resources, including BabelNet, are enough to solve Lexical Semantics, as discussed in Chapter 1. With this thesis, we point at two promising avenues, strongly connected but also complementary to a

certain extent, that we believe should be further explored in future work:

1. Establishing a unified, consistent and effective framework to develop data-driven sense-level models on a large scale : this goal requires, first of all, to have reliable and scalable automatic methods to construct high-quality sense-annotated corpora. Starting from those developed in Chapter 4, with their strengths and weaknesses, but also including other recent approaches geared towards the same task (Pasini and Navigli, 2017), a general-purpose disambiguation pipeline would be a major next step, coupled with an adequate evaluation framework that provides a level playing field for both training and testing purposes (Usbeck et al., 2015; Raganato et al., 2017a). Then, of course, we would need to develop scalable supervised models. For instance, in the domain of WSD (Section 2.2.1), this would translate into challenging the well-established word expert paradigm (which requires casting a new classification problem for each and every target word type) and at the same time going beyond language-specific models. Efforts in this direction are under way (Raganato et al., 2017b), but investigating whether and how these data-driven models can scale to extremely large training datasets and vocabularies, or whether structured semantic knowledge from resources like BabelNet can be effectively exploited for this purpose, still constitutes an important open problem;
2. Developing a principled approach to extract, ontologize, align and unify relational knowledge : this goal consists, on the one hand, in designing large-scale sense-level IE/OIE approaches to extract and ontologize relational knowledge from open text, a research thread, started with Patty and WiSeNet (Section 3.2) and brought forward by DefIE (Section 5.1), which also relies on the availability of sense-annotated data (see the point above). On the other hand, semantic analysis is also needed to integrate, harmonize and unify effectively IE/OIE-derived knowledge, as we showed with KB-Unify (cf. Section 5.2), where we moved our first steps towards the construction of a BabelNet-like repository for relational knowledge. However, while we restricted ourselves mainly to OIE-derived resources in the experimental evaluation, the unification framework we designed is capable of handling any kind of resource. On a wider perspective, the ultimate task we are envisioning with such a framework is that of unifying closed and open IE, an ambitious research objective for which integration approaches such as Universal Schemas (Section 2.3.3) have already started paving the way. In this respect, the long-term goal is that of developing models that are able to autonomously spot and extract a semantic relation across natural language text, and then decide whether such relation should be classified as one of the relations already encoded in a given type inventory, or rather considered as a novel, unseen relation to be modeled separately.

As a final remark, a key aspect of both the perspectives treated above is multilinguality . In fact, whenever we refer to 'large-scale' approaches, we implicitly consider the capability of including multiple languages into the picture, a requirement that is becoming increasingly important nowadays. Focusing on frameworks that are as language-independent as possible has been one of the major concern throughout this

thesis, reflected in the central role of BabelNet in all the approaches we studied. The sense-annotated resources developed in Chapter 4 are either already multilingual (Sections 4.2 and 4.3), or easily extendable to multiple languages (Section 4.1); similarly, the sense-aware approaches of Chapter 5, being defined at the level of language-independent Babel synsets, are also open to a multilingual extension that would not require, e.g., off-the-shelf Machine Translation systems (Faruqui and Kumar, 2015). As a matter of fact, KB-Unify and TaxoEmbed do not rely on any language-specific module, and are already capable of working with relation triples or term-hypernym pairs regardless of the specific language they are encoded in. By coupling these frameworks with multilingual corpora, such as EuroSense or SenseDefs, we are indeed laying the foundations for the development of wide-coverage systems targeted at the multilingual extraction of relational knowledge at the sense level. These systems should be able to build upon multilingual lexico-semantic resources, harvesting relation instances that are lexically unambiguous and seamlessly ontologized, and at the same time building, brick by brick, a multilingual repository of relational knowledge on top of lexical resources, with sets of multilingual paraphrases and synonymous patterns representing the same underlying semantic relation. Similarly to resources like BabelNet, these language-independent relations could then be ontologized into a lexicalized semantic network and connected with semantic links representing, e.g., hypernymy (is a university and is an educational institution), subsumption (knows and has a romantic relationship with), inference (is awarded with and is nominated for), or even statistical co-occurrence (works in and lives in). With such a semantic resource as a backbone for data-driven models, the deep learning tsunami could hit once again the shores of Natural Language Understanding, but this time with its full force.



# Bibliography

- Eugene Agichtein and Luis Gravano. Snowball: Extracting Relations from Large Plain-text Collections. In Proc. of ACM-DL , pages 85 94, 2000.
- Eneko Agirre and David Martínez. Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web. In Proc. of the COLING Workshop on Semantic Annotation and Intelligent Content, pages 11 19, 2000.
- Eneko Agirre and David Martínez. Unsupervised WSD Based on Automatically Retrieved Examples: The Importance of Bias. InProc. of EMNLP , pages 25 33, 2004.
- Eneko Agirre and Aitor Soroa. Personalizing PageRank for Word Sense Disambiguation. In Proc. of EACL , pages 33 41, 2009.
- Eneko Agirre, Olatz Ansa, Eduard Hovy, and David Martínez. Enriching Very Large Ontologies using the WWW. In Proc. of OL, pages 25 30, 2000.
- Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, Karmele Mendizabal, Eli Pociello, and Mikel Quintian. Improving the basque wordnet by corpus annotation. In Proc. of GWC, pages 287 289, 2005.
- Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. Knowledge-Based WSD on Specific Domains: performing Better than Generic Supervised WSD. InProc. of IJCAI , pages 1501 1506, 2009.
- Eneko Agirre, Xabier Arregi, and Arantxa Otegi. Document Expansion Based on WordNet for Robust IR. In Proc. of COLING , pages 9 17, 2010.
- Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. Random Walks for Knowledge-Based Word Sense Disambiguation. Computational Linguistics, 40(1):57 84, 2014.
- Alan Akbik and Alexander Löser. KrakeN: N-ary Facts in Open Information Extraction. In Proc. of AKBC-WEKEX , pages 52 56, 2012.
- Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. Polyglot: Distributed Word Representations for Multilingual NLP. In Proc. of CoNLL , pages 183 192, 2013.
- Nikolaos Aletras and Mark Stevenson. A Hybrid Distributional and Knowledge-based Model of Lexical Semantics. InProc. of SEM, pages 20 29, 2015.

- Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In Proc. of ACL, pages 344 354, 2015.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A Latent Variable Model Approach to PMI-based Word Embeddings. TACL, 4: 385 399, 2016.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning Principled Bilingual Mappings of Word Embeddings while Preserving Monolingual Invariance. In Proc. of EMNLP, pages 2289 2294, 2016.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proc. of LREC, pages 2200 2204, 2010.
- Timothy Baldwin, Nam Kim Su, Francis Bond, Sanae Fujita, David Martinez, and Takaaki Tanaka. MRD-based Word Sense Disambiguation: Further Extending Lesk. In Proc. of IJCNLP, pages 775 780, 2008.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Gritt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for Sembanking. In Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178 186, 2013.
- Satanjeev Banerjee and Ted Pedersen. Extended Gloss Overlap as a Measure of Semantic Relatedness. In Proc. of IJCAI, pages 805 810, 2003.
- Michele Banko and Oren Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. In Proc. of ACL-HLT, pages 28 36, 2008.
- Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the Web. In Proc. of IJCAI, pages 2670 2676, 2007.
- Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chungchieh Shan. Entailment Above the Word Level in Distributional Semantics. In Proc. of EACL, pages 23 32, 2012.
- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In Proc. of COLING, pages 1591 1600, 2014.
- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. UNIBA: Combining Distributional Semantic Models and Sense Distribution for Multilingual All-Words Sense Disambiguation and Entity Linking. In Proc. of SemEval-2015, pages 360 364, 2015.
- Flavio De Benedictis, Stefano Faralli, and Roberto Navigli. GlossBoot: Bootstrapping Multilingual Domain Glossaries from the Web. In Proc. of ACL, pages 528 538, 2013.

- Luisa Bentivogli and Emanuele Pianta. Exploiting Parallel Texts in the Creation of Multilingual Semantically Annotated Resources: The MultiSemCor Corpus. *Natural Language Engineering* 11(3):247-261, 2005.
- Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. Revising the Wordnet Domains Hierarchy: Semantics, Coverage and Balancing. *IrProc. of MLR*, pages 101-108, 2004.
- Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. *InProc. of WWW*, pages 467-476, 2008.
- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. *In Proc. of SIGMOD*, pages 1247-1250, 2008.
- Giulia Bonansinga and Francis Bond. Multilingual Sense Intersection in a Parallel Corpus with Diverse Language Families. *InProc. of GWC*, pages 44-49, 2016.
- Francis Bond and Ryan Foster. Linking and Extending an Open Multilingual Wordnet. *In Proc. of ACL*, pages 1352-1362, 2013.
- Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. Semeval-2015 task 17: Taxonomy Extraction Evaluation (TExEval). *In Proc. of SemEval* pages 902-910, 2015.
- Georgeta Bordea, Els Lefever, and Paul Buitelaar. Semeval-2016 task 13: Taxonomy Extraction Evaluation (TExEval-2). *In Proc. of SemEval* pages 1081-1091, 2016.
- Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning Structured Embeddings of Knowledge Bases. *IrProc. of AAAI*, pages 301-306, 2011.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. *In Proc. of NIPS*, pages 2787-2795, 2013.
- Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. *InProc. of EMNLP*, pages 615-620, 2014.
- Ronald J. Brachman. What 'is-a' is and isn't: An Analysis of Taxonomic Links in Semantic Networks. *IEEE Computer*, 16(10):30-36, 1983.
- Samuel Brody and Mirella Lapata. Bayesian Word Sense Induction. *InProc. of EACL*, pages 103-111, 2009.
- Martin Brümmer, Milan Dojchinovski, and Sebastian Hellmann. DBpedia Abstracts: A Large-Scale, Open, Multilingual NLP Training Corpus. *In Proc. of LREC*, pages 3339-3343, 2016.
- Bruce G. Buchanan and David C. Wilkins, editors. *Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

- Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, 2006.
- Razvan C. Bunescu and Raymond J. Mooney. Learning to Extract Relations from the Web using Minimal Supervision. In *Proc. of ACL*, pages 576–583, 2007.
- Razvan C. Bunescu and Marius Pa<sup>3</sup>ca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proc. of EACL*, pages 9–16, 2006.
- Hiram Calvo and Alexander Gelbukh. Is the Most Frequent Sense of a Word Better Connected in a Semantic Network? In *Proc. of ICIC*, pages 491–499, 2015.
- José Camacho Collados. Why we have switched from building full-edged taxonomies to simply detecting hypernymy relations. *arXiv preprint arXiv:1703.04178*, 2017.
- José Camacho Collados, Mohammad Taher Pilehvar, and Roberto Navigli. A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets. In *Proc. of ACL*, pages 1–7, 2015a.
- José Camacho Collados, Mohammad Taher Pilehvar, and Roberto Navigli. A Unified Multilingual Semantic Representation of Concepts. In *Proc. of ACL*, pages 741–751, 2015b.
- José Camacho Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. A Large-Scale Multilingual Disambiguation of Glosses. In *Proc. of LREC*, pages 1701–1708, 2016a.
- José Camacho Collados, Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Semantic Representations of Word Senses and Concepts. *ACL Tutorial*, 2016b.
- José Camacho Collados, Mohammad Taher Pilehvar, and Roberto Navigli. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016c.
- José Camacho Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proc. of SemEval*, pages 15–26, 2017.
- Andrew Carlson and Charles Schafer. Bootstrapping Information Extraction from Semi-structured Web Pages. In *Proc. of ECML-PKDD*, pages 195–210, 2008.
- Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In *Proc. of AACL*, pages 1306–1313, 2010.
- V. Ivan Sanchez Carmona and Sebastian Riedel. How Well Can We Predict Hypernyms from Word Embeddings? A Dataset-Centric Analysis. In *Proc. of EACL*, pages 401–407, 2017.
- Marine Carpuat and Dekai Wu. Word Sense Disambiguation vs. Statistical Machine Translation. In *Proc. of ACL*, pages 387–394, 2005.

- Yee Seng Chan and Hwee Tou Ng. Scaling up Word Sense Disambiguation via Parallel Texts. In Proc. of AAAI , pages 1037 1042, 2005.
- Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. In Proc. of ACL , pages 33 40, 2007.
- Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. Improving Distributed Representation of Word Sense via WordNet Gloss Composition and Context Clustering. In Proc. of ACL , pages 15 20, 2015.
- Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A Unified Model for Word Sense Representation and Disambiguation. In Proc. of EMNLP , pages 1025 1035, 2014.
- Xiao Cheng and Dan Roth. Relational Inference for Wikification. In Proc. of EMNLP , pages 1787 1796, 2013.
- Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Towards a Linguistic Linked Open Data cloud : The Open Linguistics Working Group. TAL , 52 (3):245 275, 2011.
- Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. Semantic Role Labeling for Open Information Extraction. In Proc. of FAM-LbR , pages 52 60, 2010.
- Stephen Clark and James R. Curran. Wide-coverage Efficient Statistical Parsing with CCG and Log-Linear Models. Computational Linguistics, 33(4):493 552, 2007.
- Stephen Clark and David Weir. Class-Based Probability Estimation Using a Semantic Hierarchy. Computational Linguistics, 28(2):187 206, 2002.
- Jacob Cohen. A Coefficient of Agreement of Nominal Scales. Educational and Psychological Measurement 20(1):37 46, 1960.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12:2493 2537, 2011.
- Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A Framework for Benchmarking Entity-annotation Systems. In Proc. of WWW , pages 249 260, 2013.
- Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. Trans-gram, Fast Cross-lingual Word-embeddings. In Proc. of EMNLP , pages 1109 1113, 2015.
- D. Alan Cruse. Lexical Semantics Cambridge University Press, 1986.
- Montse Cuadros and German Rigau. Quality Assessment of Large Scale Knowledge Resources. In Proc. of EMNLP , pages 534 541, 2006.
- Silviu Cucerzan. Large-scale Named Entity Disambiguation based on Wikipedia data. In Proc. of EMNLP-CoNLL , pages 708 716, 2007.

- Bhavana Dalvi, Einat Minkov, Partha P. Talukdar, and William W. Cohen. Automatic Gloss Finding for a Knowledge Base using Ontological Constraints. In Proc. of WSDM, pages 369–378, 2015.
- Bharath Dandala, Chris Hokamp, Rada Mihalcea, and Razvan C. Bunescu. Sense Clustering Using Wikipedia. In Proc. of RANLP, pages 164–171, 2013.
- Jeroen de Knij, Kevin Meijer, Flavius Frasincar, and Frederik Hogenboom. Word Sense Disambiguation for Automatic Taxonomy Construction from Text-Based Web Corpora. In Proc. of WISE, pages 241–248, 2011.
- Oier Lopez de Lacalle and Eneko Agirre. A Methodology for Word Sense Disambiguation at 90% based on large-scale Crowd Sourcing. In Proc. of SEM, pages 61–70, 2015.
- Marie-Catherine de Marne e, Bill MacCartney, and Christopher D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In Proc. of LREC, pages 449–454, 2006.
- Gerard de Melo and Gerhard Weikum. MENTA: Inducing Multilingual Taxonomies from Wikipedia. In Proc. of CIKM, pages 1099–1108, 2010.
- Luciano Del Corro and Rainer Gemulla. ClausIE: Clause-based Open Information Extraction. In Proc. of WWW, pages 355–366, 2013.
- Claudio Delli Bovi and Roberto Navigli. Multilingual semantic dictionaries for natural language processing: The case of BabelNet Encyclopedia with Semantic Computing and Robotic Intelligence 1(1):1630015, 2017.
- Claudio Delli Bovi and Alessandro Raganato. Sew-Embed at SemEval-2017 Task 2: Language-Independent Concept Representations from a Semantically Enriched Wikipedia. In Proc. of SemEval, pages 252–257, 2017.
- Claudio Delli Bovi, Luis Espinosa Anke, and Roberto Navigli. Knowledge Base Unification via Sense Embeddings and Disambiguation. In Proc. of EMNLP, pages 726–736, 2015a.
- Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis. Transactions of the Association for Computational Linguistics, 3:529–543, 2015b.
- Claudio Delli Bovi, José Camacho Collados, Alessandro Raganato, and Roberto Navigli. EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text. In Proc. of ACL, pages 594–600, 2017.
- Mona Diab and Philip Resnik. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In Proc. of ACL, pages 255–262, 2002.
- Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge Vault: a Web-Scale Approach to Probabilistic Knowledge Fusion. In Proc. of KDD, pages 601–610, 2014.

- Greg Durrett and Dan Klein. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Transactions of the Association for Computational Linguistics*, 2: 477–490, 2014.
- Arnab Dutta, Christian Meilicke, and Simone Paolo Ponzetto. A Probabilistic Approach for Integrating Heterogeneous Knowledge Sources. *Proc. of ESWC*, pages 286–301, 2014.
- Arnab Dutta, Christian Meilicke, and Heiner Stuckenschmidt. Enriching Structured Knowledge with Open Information. In *Proc. of WWW*, pages 267–277, 2015.
- Philip Edmonds and Scott Cotton. Senseval-2: Overview. In *Proc. of SENSEVAL*, pages 1–5, 2001.
- Philip Edmonds and Adam Kilgarri. Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems. *Natural Language Engineering* 8(4):279–291, 2002.
- Andreas Eisele and Yu Chen. MultiUN: A Multilingual Corpus from United Nation Documents. In *Proc. of LREC*, pages 2868–2872, 2010.
- Nicolai Erbs, Torsten Zesch, and Iryna Gurevych. Link discovery: A comprehensive analysis. In *Proc. of ICSC*, pages 83–86, 2011.
- Katrin Erk. A Simple, Similarity-based Model for Selectional Preferences. In *Proc. of ACL*, pages 216–223, 2007.
- Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. Named Entity Disambiguation for Noisy Text. In *Proc. of CoNLL*, pages 58–68, 2017.
- Luis Espinosa Anke and Horacio Saggion. Applying Dependency Relations to Definition Extraction. In *Proc. of NLDB*, pages 63–74, 2014.
- Luis Espinosa Anke, Horacio Saggion, and Claudio Delli Bovi. Definition extraction using sense-based embeddings. *Proc. of IWES*, pages 10–15, 2015.
- Luis Espinosa Anke, José Camacho Collados, Claudio Delli Bovi, and Horacio Saggion. Supervised Distributional Hypernym Discovery via Domain Adaptation. In *Proc. of EMNLP*, pages 424–435, 2016a.
- Luis Espinosa Anke, Horacio Saggion, and Francesco Ronzano. TALN at SemEval-2016 Task 14: Semantic Taxonomy Enrichment Via Sense-Based Embeddings. In *Proc. of SemEval*, pages 1332–1336, 2016b.
- Luis Espinosa Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. ExTaSem! Extending, Taxonomizing and Semantifying Domain Terminologies. In *Proc. of AAAI*, pages 2594–2600, 2016c.
- Allyson Ettinger, Philip Resnik, and Marine Carpuat. Retrieving Sense-Specific Word Vectors Using Parallel Text. In *Proc. of NAACL-HLT*, pages 1378–1383, 2016.

- Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised Named-entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91-134, 2005.
- Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. Open Information Extraction: The Second Generation. In *Proc. of IJCAI*, pages 3-10, 2011.
- Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying Relations for Open Information Extraction. In *Proc. of EMNLP*, pages 1535-1545, 2011.
- Tobias Falke, Gabriel Stanovsky, Iryna Gurevych, and Ido Dagan. Porting an Open Information Extraction System from English to German. In *Proc. of EMNLP*, pages 892-898, 2016.
- Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, and Edward Y. Chang. Distant Supervision for Relation Extraction with Matrix Completion. In *Proc. of ACL*, pages 839-849, 2014.
- Stefano Faralli and Roberto Navigli. A New Minimally-Supervised Framework for Domain Word Sense Disambiguation. In *Proc. of EMNLP-CoNLL*, pages 1411-1422, 2012.
- Manaal Faruqui and Shankar Kumar. Multilingual Open Relation Extraction using Cross-lingual Projection. In *Proc. of NAACL-HLT*, pages 1351-1356, 2015.
- Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrotting Word Vectors to Semantic Lexicons. In *Proc. of NAACL*, pages 945-955, 2015.
- Christiane Fellbaum, editor. *WordNet: an electronic lexical database* MIT Press, Cambridge, MA, USA, 1998.
- Erwin Fernandez-Ordonez, Rada Mihalcea, and Samer Hassan. Unsupervised Word Sense Disambiguation with Multilingual Representations. In *Proc. of LREC*, pages 847-851, 2012.
- Paolo Ferragina and Ugo Scaiella. Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *IEEE Software*, 29(1):70-75, 2012.
- Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20:116-131, 2002.
- John R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis*, 1952-59:1-32, 1957.
- Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. In *Proc. of ACL*, pages 945-955, 2014.

- Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. MultiWiBi: The multilingual Wikipedia bitaxonomy project. *Artificial Intelligence*, 241:66–102, 2016.
- Robert W. Floyd. Algorithm 97: Shortest Path. *Communications of the ACM*, 5(6): 345–345, 1962.
- W. Nelson Francis and Henry Kucera. *Brown Corpus Manual*. Technical report, Department of Linguistics, Brown University, 1979.
- Marc Franco-Salvador, Paolo Rosso, and Manuel Montes y Gómez. A Systematic Study of Knowledge Graph Analysis for Cross-Language Plagiarism Detection. *Information Processing & Management*, 52(4):550–570, 2016.
- Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning Semantic Hierarchies via Word Embeddings. In *Proc. of ACL*, pages 1199–1209, 2014.
- Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proc. of IJCAI*, pages 1606–1611, 2007.
- Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0), 2013. URL <http://lemurproject.org/clueweb12>.
- William A. Gale, Kenneth W. Church, and David Yarowsky. A method for disambiguating word senses in a corpus. *Computers and the Humanities* 26:415–439, 1992a.
- William A. Gale, Kenneth W. Church, and David Yarowsky. Estimating Upper and Lower Bounds on the Performance of Word-sense Disambiguation Programs. In *Proc. of ACL*, pages 249–256, 1992b.
- William A. Gale, Kenneth W. Church, and David Yarowsky. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proc. of TMI*, pages 101–112, 1992c.
- Pablo Gamallo and Marcos Garcia. Multilingual Open Information Extraction. In *Proc. of EPIA*, pages 711–722, 2015.
- Matt Gardner and Tom Mitchell. Efficient and Expressive Knowledge Base Completion using Subgraph Feature Extraction. In *Proc. of EMNLP*, pages 1488–1498, 2015.
- Matt Gardner, Partha Pratim Talukdar, Bryan Kisiel, and Tom Mitchell. Improving Learning and Inference in a Large Knowledge-Base using Latent Syntactic Cues. In *Proc. of EMNLP*, pages 833–838, 2013.

- Kiril Gashteovski, Rainer Gemulla, and Luciano Del Corro. MinIE: Minimizing Facts in Open Information Extraction. In Proc. of EMNLP , pages 2620 2630, 2017.
- Jim Giles. Internet encyclopaedias go head to headNature, 438:900 901, 2005.
- Hila Gonen and Yoav Goldberg. Semi Supervised Preposition-Sense Disambiguation using Multilingual Data. In Proc. of COLING , pages 2718 2729, 2016.
- Aitor González, German Rigau, and Mauro Castillo. A Graph-Based Method to Improve WordNet Domains. In Proc. of CICLING , pages 17 28, 2012.
- Stephan Gouws, Yoshua Bengio, and Greg Corrado. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. In Proc. of ICML , pages 748 756, 2015.
- Ralph Grishman. Information Extraction: Techniques and Challenges. In Proc. of SCIE, pages 10 27, 1997.
- Adam Grycner and Gerhard Weikum. HARPY: Hypernyms and Alignment of Relational Paraphrases. In Proc. of COLING , pages 2195 2204, 2014.
- Adam Grycner, Gerhard Weikum, Jay Pujara, James Foulds, and Lise Getoor. RELLY: Inferring Hypernym Relationships Between Relational Phrases. In Proc. of EMNLP , pages 971 981, 2015.
- Iryna Gurevych. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In Proc. of IJCNLP , pages 767 778, 2005.
- Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. Uby: A Large-scale Uni ed Lexical-semantic Resource Based on LMF. In Proc. of EACL , pages 580 590, 2012.
- Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek. Linked lexical knowledge bases: Foundations and applicationsSynthesis Lectures on Human Language Technologies9(3):1 146, 2016.
- Younggyun Hahm, Jungyeul Park, Kyungtae Lim, Youngsik Kim, Dosam Hwang, and Key-Sun Choi. Named Entity Corpus Construction using Wikipedia and DBpedia Ontology. In Proc. of LREC , pages 2565 2569, 2014.
- Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke Zettlemoyer. Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves. In Proc. of EMNLP , pages 289 299, 2013.
- Alon Halevy, Peter Norvig, and Fernando Pereira. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24(2):8 12, 2009.
- Birgit Hamp and Helmut Feldweg. Germanet a lexical-semantic net for german. In Proc. of ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications pages 9 15, 1997.

- Xianpei Han and Le Sun. An Entity-topic Model for Entity Linking. In Proc. of EMNLP-CoNLL , pages 105 115, 2012.
- Patrick Hanks. Do word meanings exist? Computers and the Humanities 34(1 2): 205 215, 2000.
- Zellig Harris. Distributional structure. Word, 10:146 162, 1954.
- Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. Learning Entity Representation for Entity Disambiguation. In Proc. of ACL , pages 30 34, 2013a.
- Zhengyan He, Shujie Liu, Yang Song, Mu Li, Ming Zhou, and Houfeng Wang. Efficient Collective Entity Linking with Stacking. In Proc. of EMNLP , pages 426 435, 2013b.
- Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. WebCage: a Web-Harvested Corpus Annotated with GermaNet Senses. In Proc. of EACL , pages 387 396, 2012a.
- Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. An Automatic Method for Creating a Sense-Annotated Corpus Harvested from the Web. International Journal of Computational Linguistics and Applications, 3(2):35 50, 2012b.
- Karl Moritz Hermann and Phil Blunsom. Multilingual Distributed Representations without Word Alignment. In Proc. of ICLR , pages 1 9, 2014.
- José María Gómez Hidalgo, Manuel de Buenaga Rodríguez, and José Carlos Cortizo Pérez. The role of word sense disambiguation in automated text categorization. In Proc. of NLDB , pages 298 309, 2005.
- Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. ArXiv:1408.3456, 2014.
- Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. Learning to Understand Phrases by Embedding the Dictionary. Transactions of the Association for Computational Linguistics, 4:17 30, 2016.
- Johannes Hortsch, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages. In Proc. of WWW , pages 229 232, 2011a.
- Johannes Hortsch, Mohammed Amir Yosef, Ilaria Bordino, Hagen Fürstenauf, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust Disambiguation of Named Entities in Text. In Proc. of EMNLP , pages 782 792, 2011b.
- Johannes Hortsch, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. KORE: Keyphrase Overlap Relatedness for Entity Disambiguation. In Proc. of CIKM , pages 545 554, 2012.

- Johannes Hoart, Dragan Milchevski, and Gerhard Weikum. STICS: Searching with Strings, Things, and Cats. In Proc. of SIGIR, pages 1247–1248, 2014.
- Raphael Hofmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations. In Proc. of HLT, pages 541–550, 2011.
- Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% Solution. In Proc. of NAACL, pages 57–60, 2006.
- Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence, 194:2–27, 2013.
- Linmei Hu, Xuzhong Wang, Mengdi Zhang, Juanzi Li, Xiaoli Li, Chao Shao, Jie Tang, and Yongbin Liu. Learning Topic Hierarchies for Wikipedia Categories. In Proc. of ACL, pages 346–351, 2015.
- Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving Word Representations via Global Context and Multiple Word Prototypes. In Proc. of ACL, pages 873–882, 2012.
- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In Proc. of ACL, pages 95–105, 2015.
- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Embeddings for Word Sense Disambiguation: An Evaluation Study. In Proc. of ACL, pages 897–907, 2016.
- Nancy Ide. MultiMASC: An Open Linguistic Infrastructure for Language Research. In Proc. of the 5th Workshop on Building and Using Comparable Corpora, pages 42–48, 2012.
- Nancy Ide, Tomaž Erjavec, and Dan Tu. Automatic Sense Tagging using Parallel Corpora. In Proc. of NLPRS, pages 83–89, 2001.
- Nancy Ide, Tomaž Erjavec, and Dan Tu. Sense Discrimination with Parallel Corpora. In Proc. of the ACL Workshop on WSD, pages 61–66, 2002.
- Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. Development of the Japanese wordnet. In Proc. of LREC, pages 2420–2420, 2008.
- Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models. In Proc. of NAACL, pages 683–693, 2015.
- Colette Joubarne and Diana Inkpen. Comparison of Semantic Similarity for Different Languages using the Google N-Gram Corpus and Second-Order Co-Occurrence Measures. In Advances in Artificial Intelligence, pages 216–221, 2011.

- Mikael Kågebäck and Hans Salomonsson. Word Sense Disambiguation using a Bidirectional LSTM. In *Proceedings of CogALex* pages 51–56, 2016.
- Muhammad Faheem Khan, Aurangzeb Khan, and Khairullah Khan. Efficient Word Sense Disambiguation Technique for Sentence Level Sentiment Classification of Online Reviews. *Science International (Lahore)*, 25:937–943, 2013.
- Tushar Khot, Ashish Sabharwal, and Peter Clark. Answering Complex Questions using Open Information Extraction. In *Proc. of ACL*, pages 311–316, 2017.
- Adam Kilgarri. What is Word Sense Disambiguation Good For? In *Proc. of NLPRS*, pages 209–214, 1997.
- Adam Kilgarri. English Lexical Sample Task Description. In *Proc. of Senseval* pages 17–20, 2001.
- Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT summit*, volume 5, pages 79–86, 2005.
- Stanley Kok and Pedro Domingos. Extracting Semantic Networks from Text Via Relational Clustering. In *Proc. of ECML-PKDD*, pages 624–639, 2008.
- Zornitsa Kozareva and Eduard Hovy. Learning Arguments and Supertypes of Semantic Relations using Recursive Patterns. In *Proc. of ACL*, pages 1482–1491, 2010.
- Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective Annotation of Wikipedia Entities in Web Text. In *Proc. of KDD*, pages 457–466, 2009.
- Pierre Lafon. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1(1):127–165, 1980.
- Ni Lao, Tom Mitchell, and William W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proc. of EMNLP*, pages 529–539, 2011.
- Ni Lao, Amarnag Subramanya, Fernando Pereira, and William W. Cohen. Reading the Web with Learned Syntactic-Semantic Inference Rules. In *Proc. of EMNLP-CoNLL*, pages 1017–1026, 2012.
- Claudia Leacock, George A. Miller, and Martin Chodorow. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–165, 1998.
- Els Lefever and Véronique Hoste. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proc. of SemEval* pages 15–20, 2010.
- Els Lefever and Véronique Hoste. SemEval-2013 task 10: Cross-lingual Word Sense Disambiguation. In *Proc. of SemEval* pages 158–166, 2013.
- Els Lefever, Véronique Hoste, and Martine De Cock. ParaSense or How to Use Parallel Corpora for Word Sense Disambiguation. In *Proc. of ACL-HLT*, pages 317–322, 2011.

- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* pages 1-29, 2014.
- Michael Lesk. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proc. of SIGDOC*, pages 24-26, 1986.
- Omer Levy, Yoav Goldberg, and Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *TACL*, 3:211-225, 2015a.
- Omer Levy, Steven Remus, Chris Biemann, and Ido Dagan. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proc. of NAACL*, pages 970-976, 2015b.
- Jiwei Li and Dan Jurafsky. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *Proc. of EMNLP*, pages 1722-1732, 2015.
- Thomas Lin, Mausam, and Oren Etzioni. No Noun Phrase Left Behind: Detecting and Typing Unlinkable Entities. In *Proc. of EMNLP-CoNLL*, pages 893-903, 2012.
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural Relation Extraction with Selective Attention over Instances. In *Proc. of ACL*, pages 2124-2133, 2016.
- Xiao Ling and Daniel S. Weld. Fine-grained Entity Recognition. In *Proc. of AAAI*, pages 94-100, 2012.
- Xiao Ling, Sameer Singh, and Daniel S. Weld. Design Challenges for Entity Linking. *Transactions of the Association for Computational Linguistics*, 3:315-328, 2015.
- Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proc. of LREC*, pages 923-929, 2016.
- Kenneth C. Litkowski. Senseval-3 Task: Word Sense Disambiguation of WordNet Glosses. In *Proc. of Senseval-3* pages 13-16, 2004.
- Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *Proc. of CIDR*, 2015.
- Massimiliano Mancini, José Camacho Collados, Ignacio Iacobacci, and Roberto Navigli. Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. In *Proc. of CoNLL*, pages 100-111, 2017.
- Jean M. Mandler. Preverbal representation and language. In P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett, editors, *Language, speech, and communication. Language and space* chapter 9, pages 365-384. MIT Press, Cambridge, MA, USA, 1996.

- Christopher D. Manning. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4):701-707, 2015.
- Antonio Di Marco and Roberto Navigli. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, 39(3):709-754, 2013.
- Tamara Martín-Wanton, Aurora Pons-Porrata, Andrés Montoyo-Guijarro, and Alexandra Balahur. Opinion Polarity Detection - Using Word Sense Disambiguation to Determine the Polarity of Opinions. In *Proc. of ICAART*, pages 483-486, 2010.
- Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. Open Language Learning for Information Extraction. In *Proc. of EMNLP-CoNLL*, pages 523-534, 2012.
- Ernst Mayr. *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press, 1982.
- Diana McCarthy, Marianna Apidianaki, and Katrin Erk. Word Sense Clustering and Clusterability. *Computational Linguistics*, 42(2):245-275, 2016.
- John McCrae, Dennis Spohr, and Philipp Cimiano. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In *Proc. of ESWC*, pages 245-259, 2011.
- Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie M. Strassel. An Evaluation of Technologies for Knowledge Base Population. In *Proc. of LREC*, pages 369-372, 2009.
- Douglas L. Medin, Robert L. Goldstone, and Dedre Gentner. Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science* 1(1):54-69, 1990.
- Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proc. of CoNLL*, pages 51-61, 2016.
- Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBPedia Spotlight: Shedding Light on the Web of Documents. In *Proc. of I-Semantics*, pages 1-8, 2011.
- Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. Effectiveness and Efficiency of Open Relation Extraction. In *Proc. of EMNLP*, pages 447-457, 2013.
- Rada Mihalcea. Bootstrapping Large Sense Tagged Corpora. In *Proc. of LREC*, pages 1407-1411, 2002.
- Rada Mihalcea and Andras Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proc. of CIKM*, pages 233-242, 2007.

- Rada Mihalcea and Dan Moldovan. eXtended WordNet: Progress report. In Proc. of the NAACL Workshop on WordNet and Other Lexical Resources, pages 95–100, 2001.
- Rada Mihalcea and Dan I. Moldovan. An Automatic Method for Generating Sense Tagged Corpora. In Proc. of AAAI/IAAI, pages 461–466, 1999.
- Rada Mihalcea, Timothy Chklovski, and Adam Kilgarri. The Senseval-3 English Lexical Sample Task. In Proc. of Senseval, pages 1–4, 2004.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In CLR Workshop, 2013a.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Advances in NIPS, pages 3111–3119, 2013b.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proc. of NAACL-HLT, pages 746–751, 2013c.
- George A. Miller. On knowing a word. *Annual Review of Psychology* 50(1):1–19, 1999.
- George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1):1–28, 1991.
- George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. WordNet: an online lexical database. *International Journal of Lexicography*, 3(4): 235–244, 1990.
- George A. Miller, Claudia Leacock, Randee Tenji, and Ross T. Bunker. A semantic concordance. In Proc. of HLT, pages 303–308, 1993.
- David Milne and Ian H. Witten. Learning to link with Wikipedia. In Proc. of CIKM, pages 509–518, 2008.
- David Milne and Ian H. Witten. An Open-Source Toolkit for Mining Wikipedia. *Artificial Intelligence*, 194:222–239, 2013.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant Supervision for Relation Extraction Without Labeled Data. In Proc. of ACL, pages 1003–1011, 2009.
- Tom M. Mitchell. Reading the Web: A Breakthrough Goal for AI. *AI Magazine*, 2005.
- Makoto Miwa and Mohit Bansal. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In Proc. of ACL, pages 1105–1116, 2016.
- Dan Moldovan and Adrian Novischi. Word Sense Disambiguation of WordNet Glosses. *Computer Speech & Language* 18(3):301–317, 2004.

- Raymond J. Mooney and Razvan C. Bunescu. Subsequence Kernels for Relation Extraction. In Y. Weiss, P. B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems* 18, pages 171–178. MIT Press, 2006.
- Andrea Moro and Roberto Navigli. WiSeNet: Building a Wikipedia-based Semantic Network with Ontologized Relations. In *Proc. of CIKM*, pages 1672–1676, 2012.
- Andrea Moro and Roberto Navigli. Integrating Syntactic and Semantic Analysis into the Open Information Extraction Paradigm. In *Proc. of IJCAI*, pages 2148–2154, 2013.
- Andrea Moro and Roberto Navigli. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval*, pages 288–297, 2015.
- Andrea Moro, Hong Li, Sebastian Krause, Feiyu Xu, Roberto Navigli, and Hans Uszkoreit. Semantic Rule Filtering for Web-Scale Relation Extraction. In *Proc. of ISWC*, pages 347–362, 2013.
- Andrea Moro, Roberto Navigli, Francesco Maria Tucci, and Rebecca J. Passonneau. Annotating the MASC Corpus with BabelNet. In *Proc. of LREC*, pages 4214–4219, 2014a.
- Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL*, 2:231–244, 2014b.
- Dana Movshovitz-Attias and William W. Cohen. KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts. In *Proc. of ACL*, pages 1449–1459, 2015.
- David Nadeau and Satoshi Sekine. A Survey of Named Entity Recognition and Classification. *Linguistic Investigations* 30(1):3–26, 2007.
- Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *Proc. of EMNLP-CoNLL*, pages 1135–1145, 2012.
- Vivi Nastase and Michael Strube. Decoding Wikipedia Categories for Knowledge Acquisition. In *Proc. of AACL*, pages 1219–1224, 2008.
- Vivi Nastase and Michael Strube. Transforming Wikipedia Into a Large Scale Multilingual Concept Network. *Artificial Intelligence*, 194:62–85, 2013.
- Roberto Navigli. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proc. of ACL*, pages 105–112, 2006.
- Roberto Navigli. Word Sense Disambiguation: A survey. *ACM Computing Surveys* 41(2):1–69, 2009.
- Roberto Navigli. A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In *Proc. of SOFSEM*, pages 115–129, 2012.

- Roberto Navigli and Mirella Lapata. An experimental study on graph connectivity for unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(4):678-692, 2010.
- Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *AIJ*, 193:217-250, 2012.
- Roberto Navigli and Paola Velardi. Structural Semantic Interconnections: a knowledge-based approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(7):1075-1088, 2005.
- Roberto Navigli and Paola Velardi. Learning Word-Class Lattices for Definition and Hypernym Extraction. In *Proc. of ACL*, pages 1318-1327, 2010.
- Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. SemEval-2007 Task 07: Coarse-grained English All-words Task. In *Proc. of SemEval*, pages 30-35, 2007.
- Roberto Navigli, David Jurgens, and Daniele Vannella. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proc. of SemEval*, volume 2, pages 222-231, 2013.
- Steven Neale, Luis Gomes, Eneko Agirre, Oier Lopez de Lacalle, and António Branco. Word Sense-Aware Machine Translation: Including Senses as Contextual Features for Improved Translation Models. In *Proc. of LREC*, pages 2777-2783, 2016.
- Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *Proc. of EMNLP*, pages 1059-1069, 2014.
- Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. Compositional Vector Space Models for Knowledge Base Completion. In *Proc. of ACL*, pages 156-166, 2015.
- Hwee Tou Ng and Hian Beng Lee. Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-based Approach. In *Proc. of ACL*, pages 40-47, 1996.
- Hwee Tou Ng, Bin Wang, and Yee Seng Chan. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proc. of ACL*, pages 455-462, 2003.
- Dat Nguyen, Martin Theobald, and Gerhard Weikum. J-NERD: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features. *Transactions of the Association for Computational Linguistics*, 4:215-229, 2016.
- Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing YAGO: Scalable Machine Learning for Linked Data. In *Proc. of WWW*, pages 271-280, 2012.

- Joakim Nivre, Marie-Catherine de Marne e, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A Multilingual Treebank Collection. In Proc. of LREC, pages 215 229, 2016.
- Thanapon Noraset, Chandra Bhagavatula, and Doug Downey. Adding High-Precision Links to Wikipedia. In Proc. of EMNLP, pages 651 656, 2014.
- Adrian Novischi. Accurate Semantic Annotations via Pattern Matching. In Proc. of FLAIRS, pages 375 379, 2002.
- Arantxa Otegi, Nora Aranberri, Antonio Branco, Jan Hajic, Steven Neale, Petya Osenova, Rita Pereira, Martin Popel, Joao Silva, Kiril Simov, and Eneko Agirre. QLeap WSD/NED Corpora: Semantic Annotation of Parallel Corpora in Six Languages. In Proc. of LREC, pages 3023 3030, 2016.
- Rasmus Berg Palm, Dirk Hovy, Florian Laws, and Ole Winther. End-to-End Information Extraction without Token-Level Supervision. In Proc. of the Workshop on Speech-Centric NLP, pages 48 52, 2017.
- Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. Making ne-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* 13(2):137 163, 2007.
- Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation. In Proc. of EACL, pages 86 98, 2017.
- Simone Papandrea, Alessandro Raganato, and Claudio Delli Bovi. SupWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation. In Proc. of EMNLP: System Demonstrations, pages 103 108, 2017.
- Tommaso Pasini and Roberto Navigli. Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data. In Proc. of EMNLP: System Demonstrations, pages 78 88, 2017.
- Rebecca J. Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. The MASC Word Sense Sentence Corpus. In Proc. of LREC, pages 3025 3030, 2012.
- Marco Pennacchiotti and Patrick Pantel. Ontologizing Semantic Relations. In Proc. of ACL, pages 793 800, 2006.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. In Proc. of EMNLP, pages 1532 1543, 2014.
- Tommaso Petrolito and Francis Bond. A Survey of WordNet Annotated Corpora. In Proc. of GWC, pages 236 245, 2014.
- Francesco Piccinno and Paolo Ferragina. From TagME to WAT: a New Entity Annotator. In Proc. of ERD, pages 55 62, 2014.

- Mohammad Taher Pilehvar and Nigel Collier. De-Contextualized Semantic Representations. In Proc. of EMNLP , pages 1680-1690, 2016.
- Mohammad Taher Pilehvar and Roberto Navigli. A Large-scale Pseudoword-based Evaluation Framework for State-of-the-art Word Sense Disambiguation. Computational Linguistics, 40(4):837-881, 2014.
- Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In Proc. of ACL , pages 1341-1351, 2013.
- Mohammad Taher Pilehvar, José Camacho Collados, Roberto Navigli, and Nigel Collier. Towards a Seamless Integration of Word Senses into Downstream NLP Applications. In Proc. of ACL , pages 1857-1869, 2017.
- Simone Paolo Ponzetto and Michael Strube. Taxonomy Induction Based on a Collaboratively Built Knowledge Repository. Artificial Intelligence , 175(9-10): 1737-1756, 2011.
- Marten Postma, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. Addressing the MFS Bias in WSD systems. In Proc. of LREC , pages 1695-1700, 2016.
- Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In Proc. of SemEval-2007, pages 87-92, 2007.
- John Prager, Eric Brown, Anni Coden, and Dragomir Radev. Question-answering by Predictive Annotation. In Proc. of SIGIR , pages 184-191, 2000.
- Xiao Pu, Nikolaos Pappas, and Andrei Popescu-Belis. Sense-Aware Statistical Machine Translation using Adaptive Context-Dependent Clustering. In Proc. of WMT , volume 1, pages 1-10, 2017.
- James Pustejovsky. The Generative Lexicon. Computational Linguistics, 17(4), 1991.
- Alessandro Raganato, José Camacho Collados, Antonio Raganato, and Yunseo Jung. Semantic Indexing of Multilingual Corpora and its Application on the History Domain. In Proc. of LT4DH , pages 140-147, 2016a.
- Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In Proc. of IJCAI , pages 2894-2900, 2016b.
- Alessandro Raganato, José Camacho-Collados, and Roberto Navigli. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In Proc. of EACL , pages 99-110, 2017a.
- Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. Neural Sequence Learning Models for Word Sense Disambiguation. In Proc. of EMNLP , pages 1167-1178, 2017b.

- Delip Rao, Paul McNamee, and Mark Dredze. Entity Linking: Finding Extracted Entities in a Knowledge Base. *Multi-Source, Multilingual Information Extraction and Summarization*, 11:93-115, 2013.
- Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proc. of ACL-HLT*, pages 1375-1384, 2011.
- Deepak Ravichandran and Eduard Hovy. Learning Surface Text Patterns for a Question Answering System. In *Proc. of ACL*, pages 41-47, 2002.
- Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proc. of IJCAI*, pages 448-453, 1995.
- Philip Resnik. Selectional Constraints: An Information-Theoretic Model and its Computational Realization. *Cognition*, 61(1-2):127-159, 1996.
- Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. MindNet: Acquiring and Structuring Semantic Information from Text. In *Proc. of ACL*, pages 1098-1102, 1998.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling Relations and Their Mentions Without Labeled Text. In *Proc. of ECML-PKDD*, pages 148-163, 2010.
- Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. Relation Extraction with Matrix Factorization and Universal Schemas. In *Proc. of NAACL-HLT*, pages 74-84, 2013.
- Alan Ritter, Mausam, and Oren Etzioni. A Latent Dirichlet Allocation Method for Selectional Preferences. In *Proc. of ACL*, pages 424-434, 2010.
- Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. Injecting Logical Background Knowledge into Embeddings for Relation Extraction. In *Proc. of NAACL-HLT*, pages 1119-1129, 2015.
- Stephen Roller and Katrin Erk. Relations such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment. In *Proc. of EMNLP*, pages 2163-2163, 2016.
- Stephen Roller, Katrin Erk, and Gemma Boleda. Inclusive Yet Selective: Supervised Distributional Hypernymy Detection. In *Proc. of COLING*, pages 1025-1036, 2014.
- Sascha Rothe and Hinrich Schütze. AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proc. of ACL*, pages 1793-1803, 2015.
- Benjamin Rozenfeld and Ronen Feldman. Self-supervised Relation Extraction from the Web. *Knowledge and Information Systems* 17(1):17-33, 2008.
- Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627-633, 1965.

- Gerard Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613-620, 1975.
- Celina Santamaría, Julio Gonzalo, and Felisa Verdejo. Automatic Association of Web Directories with Word Senses. *Computational Linguistics*, 29(3):485-502, 2003.
- Helmut Schmid. Improvements In Part-of-Speech Tagging With an Application To German. In *Proc. of the ACL SIGDAT-Workshop*, pages 47-50, 1995.
- Hinrich Schütze. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97-123, 1998.
- Federico Scozzafava, Alessandro Raganato, Andrea Moro, and Roberto Navigli. Automatic identification and disambiguation of concepts and named entities in the multilingual Wikipedia. In *Proc. of AI\*IA*, pages 357-366, 2015.
- Hui Shen, Razvan Bunescu, and Rada Mihalcea. Coarse to Fine Grained Sense Disambiguation in Wikipedia. In *Proc. of SEM*, pages 22-31, 2013.
- Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, and Sebastian Riedel. Neural Architectures for Fine-grained Entity Type Classification. In *Proc. of EACL*, pages 1271-1280, 2017.
- Vered Shwartz, Yoav Goldberg, and Ido Dagan. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In *Proc. of ACL*, pages 2389-2398, 2016.
- Avirup Sil and Alexander Yates. Re-ranking for joint named-entity recognition and linking. In *Proc. of CIKM*, pages 2369-2374, 2013.
- Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. Wikilinks: A large-scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia. Technical Report UM-CS-2012-015, University of Massachusetts, Amherst, 2012.
- Benjamin Snyder and Martha Palmer. The english all-words task. In *Proc. of Senseval-3* pages 41-43, 2004.
- Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Proc. of NIPS*, pages 926-934, 2013.
- Gabriel Stanovsky and Ido Dagan. Creating a Large Benchmark for Open Information Extraction. In *Proc. of EMNLP*, pages 2300-2305, 2016.
- Gabriel Stanovsky, Ido Dagan, and Mausam. Open IE as an Intermediate Structure for Semantic Tasks. In *Proc. of ACL*, pages 303-308, 2015.
- Mark Steedman. *The Syntactic Process* MIT Press, Cambridge, MA, USA, 2000.

- Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybylski, and Signe Gilbro. An Overview of the European Union's Highly Multilingual Parallel Corpora. *Language Resources and Evaluation*, 48(4):679-707, 2014.
- Katy Steinmetz. Redefining the modern dictionary. *TIME*, pages 20-21, May 23rd 2016.
- Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. SOFIE: A Self-organizing Framework for Information Extraction. In *Proc. of WWW*, pages 631-640, 2009.
- Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. Multi-instance Multi-label Learning for Relation Extraction. In *Proc. of EMNLP-CoNLL*, pages 455-465, 2012.
- Simon 'uster, Ivan Titov, and Gertjan van Noord. Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders. In *Proc. of NAACL-HLT*, pages 1346-1356, 2016.
- Kaveh Taghipour and Hwee Tou Ng. Semi-Supervised Word Sense Disambiguation using Word Embeddings in General and Specific Domains. In *Proc. of NAACL-HLT*, pages 314-323, 2015a.
- Kaveh Taghipour and Hwee Tou Ng. One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. In *Proc. of CoNLL*, pages 338-344, 2015b.
- Luchen Tan, Haotian Zhang, Charles L.A. Clarke, and Mark D. Smucker. Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings. In *Proc. of ACL*, pages 657-661, 2015.
- Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A Probabilistic Model for Learning Multi-Prototype Word Embeddings. In *Proc. of COLING*, pages 151-160, 2014.
- Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proc. of LREC*, pages 2214-2218, 2012.
- Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random Walk with Restart: Fast Solutions and Applications. *Knowledge and Information Systems* 14(3): 327-346, 2008.
- Antonio Toral, Stefania Brancale, Monica Monachini, and Claudia Soria. Rejuvenating the Italian WordNet: Upgrading, Standardising, Extending. In *Proc. of GWC*, 2010.
- Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing Text for Joint Embedding of Text and Knowledge Bases. In *Proc. of EMNLP*, pages 1499-1509, 2015.
- Rocco Tripodi and Marcello Pelillo. A Game-Theoretic Approach to Word Sense Disambiguation. *Computational Linguistics*, 43(1):31-70, 2017.

- Peter D. Turney and Patrick Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.
- Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke, Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald Sack, René Speck, Raphaël Troncy, Jörg Waitelonis, and Lars Wesemann. GERBIL: General Entity Annotator Benchmarking Framework. In *Proc. of WWW*, pages 1133–1143, 2015.
- Paola Velardi, Stefano Faralli, and Roberto Navigli. OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3):665–707, 2013.
- Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. Multilingual Relation Extraction using Compositional Universal Schema. In *Proc. of NAACL-HLT*, pages 886–896, 2016.
- Denny Vrandečić. Wikidata: A New Platform for Collaborative Data Collection. In *Proc. of WWW*, pages 1063–1064, 2012.
- Ivan Vulić and Anna Korhonen. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proc. of ACL*, pages 247–257, 2016.
- Yogarshi Vyas and Marine Carpuat. Sparse Bilingual Word Representations for Cross-lingual Lexical Entailment. In *Proc. of NAACL-HLT*, pages 1187–1197, 2016.
- Chengyu Wang, Xiaofeng He, and Aoying Zhou. A Short Survey on Taxonomy Learning from Text Corpora: Issues, Resources and Recent Advances. In *Proc. of EMNLP*, pages 1201–1214, 2017.
- William Yang Wang and William W. Cohen. Joint Information Extraction and Reasoning: A Scalable Statistical Relational Learning Approach. In *Proc. of ACL*, pages 355–364, 2015.
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge Graph and Text Jointly Embedding. In *Proc. of EMNLP*, pages 1591–1601, 2014.
- David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. Structured Training for Neural Network Transition-Based Parsing. In *Proc. of ACL*, pages 323–333, 2015.
- Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. In *Proc. of ACL*, pages 596–605, 2015.
- Daniel S. Weld, Raphael Hofmann, and Fei Wu. Using Wikipedia to Bootstrap Open Information Extraction. *ACM SIGMOD Record*, 37(4):62–68, 2009.

- Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In Proc. of WWW , pages 515 526, 2014.
- Robert West, Ashwin Paranjape, and Jure Leskovec. Mining Missing Hyperlinks from Human Navigation Traces: A Case Study of Wikipedia. In Proc. of WWW , pages 1242 1252, 2015.
- Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction. In Proc. of EMNLP , pages 1366 1371, 2013.
- Anna Wierzbicka, editor. Semantics: Primes and Universals Oxford University Press, Oxford, UK, 1996.
- Dekai Wu and Pascale Fung. Can Semantic Role Labeling Improve SMT? In Proc. of EAMT , pages 218 225, 2009.
- Fei Wu and Daniel S. Weld. Open information extraction using wikipedia. In Proc. of ACL , pages 118 127, 2010.
- Zhaohui Wu and C. Lee Giles. Sense-aware semantic analysis: A multi-prototype word representation model using Wikipedia. In Proc. of AAAI , pages 118 127, 2015.
- Yadollah Yaghoobzadeh and Hinrich Schütze. Intrinsic Subspace Evaluation of Word Embedding Representations. In Proc. of ACL , pages 236 246, 2016.
- Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, and Gerhard Weikum. Robust Question Answering over the Web of Linked Data. In Proc. of CIKM , pages 1107 1116, 2013.
- Mohamed Yahya, Steven Euijong Whang, Rahul Gupta, and Alon Halevy. ReNoun: Fact Extraction for Nominal Attributes. In Proc. of EMNLP , pages 325 335, 2014.
- Xuchen Yao, Benjamin Van Durme, and Chris Callison-Burch. Expectations of Word Sense in Parallel Corpora. In Proc. of NAACL-HLT , pages 621 625, 2012.
- Xuchen Yao, Jonathan Berant, and Benjamin Van Durme. Freebase QA: Information Extraction or Semantic Parsing? In Proc. of the ACL Workshop on Semantic Parsing, pages 82 86, 2014.
- David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proc. of ACL , pages 189 196, 1995.
- Alexander Yates and Oren Etzioni. Unsupervised Methods for Determining Object and Relation Synonyms on the Web. Journal of Artificial Intelligence Research , 34(1):255 296, 2009.
- Dayu Yuan, Ryan Doherty, Julian Richardson, Colin Evans, and Eric Altendorf. Semi-supervised Word Sense Disambiguation with Neural Models. In Proc. of COLING , pages 1374 1385, 2016.

- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Proc. of EMNLP , pages 1753 1762, 2015.
- Rong Zhang and Alexander I. Rudnicky. A Large Scale Clustering Scheme for Kernel K-Means. In Proc. of ICPR , pages 289 292, 2002.
- Sheng Zhang, Kevin Duh, and Benjamin Van Durme. MT/IE: Cross-lingual Open Information Extraction with Neural Sequence-to-Sequence Models. In Proc. of EACL , pages 64 70, 2017.
- Shubin Zhao and Ralph Grishman. Extracting Relations with Integrated Information using Kernel Methods. In Proc. of ACL , pages 419 426, 2005.
- Zhi Zhong and Hwee Tou Ng. Word Sense Disambiguation for All Words Without Hard Labor. In Proc. of IJCAI , pages 1616 1621, 2009.
- Zhi Zhong and Hwee Tou Ng. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In Proc. of ACL: System Demonstrations, pages 78 83, 2010.
- Zhi Zhong and Hwee Tou Ng. Word Sense Disambiguation Improves Information Retrieval. In Proc. of ACL , pages 273 282, 2012.
- Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. StatSnowball: A Statistical Approach to Extracting Entity Relationships. In Proc of. WWW , pages 101 110, 2009.
- Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. Bilingual Word Embeddings for Phrase-Based Machine Translation. In Proc. of EMNLP , pages 1393 1398, 2013.