Intelligenza Artificiale 7 (2013) 165–181
DOI 10.3233/IA-130057
IOS Press

BabelNet and Friends: A manifesto for multilingual semantic processing

Roberto Navigli*
Sapienza University of Rome, Italy

Abstract. Semantic processing is one of the most compelling and ambitious objectives in today's Natural Language Processing. Being able to process and understand text at the machine level can potentially enable powerful applications like semantically-aware statistical machine translation and semantic information retrieval, thereby having the potential to change the lives of everyday computer users.

In this paper I present a manifesto for the multilingual semantic processing of text. I illustrate the research vision that is pursued in my research group at the Sapienza University of Rome and describe the most recent results obtained. In the last part of the paper I outline a likely future for multilingual semantic processing focusing on the current directions and successes and highlighting on the major obstacles that make this task so hard.

Keywords: Multilingual semantic processing, Word sense disambiguation, Knowledge acquisition

1. Introduction

The lexical ambiguity of language is a crucial issue in the field of Natural Language Processing (NLP). For instance, given the following sentence:

[A] *Spring water* can be found at different *altitudes*,

intelligent systems would benefit from the ability to identify the correct meanings of *spring* (e.g., the geological vs. the season sense), *water* (e.g., the common vs. the chemical sense) and *altitude* (e.g., the geographical vs. the geometrical sense).

The task of computationally determining the meaning of a word in context is named Word Sense Disambiguation (WSD) [45, 46]. The most successful approaches to WSD are based on supervised machine learning. Among these, methods based on instance-based learning [15], Support Vector Machines [11, 27] and Latent Dirichlet Analysis integrated with Bayesian networks [10] have proven to provide state-of-the-art performance in many experimental settings. Given a word w, the typical process consists of learning a *word expert*, that is, a classifier that, given the context in which w occurs, assigns a sense label to it – e.g., *spring* in the example above is labeled with the geological sense. Typically, the sense inventory for a word is provided by a reference computational lexicon, that is, a highly structured lexical database (WordNet [22] being the most widespread). However, in order to attain state-of-the-art performance, supervised methods require large training sets tagged with word senses. It has been estimated that a high-accuracy supervised WSD system would probably need a text collection (i.e., a *corpus*) of about 3.2 million sense-annotated words [55]. At a throughput of one word per minute [18], this would require about 27 person-years of human annotation work. Even worse, all this effort would have to be repeated every time a new language became involved, or new sense inventories were adopted.

To overcome the demanding requirement for large amounts of hand-tagged data, unsupervised approaches

*Corresponding author: Roberto Navigli, Sapienza University of Rome, Italy. E-mail: navigli@di.uniromal.it.

(performing so-called Word Sense Induction) have been proposed in the literature (e.g., [16, 59, 68]). These methods typically rely on clustering techniques to induce groups of synonymous words based on their occurrences in similar contexts. While these methods do not require any labeled data, the sense distinctions modeled by the clusters are acquired dynamically and thus change not only when different algorithms are employed, but also when different parameter values are set within the same algorithm. As a result, performing comparisons is hard (although efforts in this direction do exist [3]). Moreover, lexical and semantic relations between the clusters (i.e., word senses) must also be automatically established in a later phase.

To deal with the above issues, recent research has leveraged the availability of wide-coverage lexical resources to develop knowledge-based algorithms. These approaches rely on an existing sense inventory (typically, WordNet), but do not require training. Instead, graphs are used to represent word senses (vertices) and their lexical and semantic connections (edges), as encoded by the reference knowledge resource. Next, graph-based algorithms are applied in order to perform WSD. These approaches have been shown to attain performance that is almost as good as supervised systems in domain-independent settings [52, 53], and even to surpass them on specific domains [2, 21, 47]. Given the above considerations, knowledge-based approaches can be considered the most promising in the short-medium term (cf. [45]).

Nonetheless, such algorithms are affected by the so-called *knowledge acquisition bottleneck* [24]. In fact, while WordNet encodes lexical and semantic relations of different kinds (e.g., hypernymy, such as *car* is-a *motor vehicle*; meronymy, such as *car* has-part *car door*, etc.), large amounts of non-taxonomic relations are needed to achieve high-performance WSD (e.g., *car* related-to *driver*). Recently, it has been shown that the richness of the knowledge resource greatly influences WSD performance [13, 49]. Figure 1 shows the trend of a simple graph-based algorithm on ambiguous words. The two lines graph the impact of two different resources: WordNet (dashed line), and EnWordNet (continuous line), a version of WordNet enriched with thousands of lexical semantic relations [44]. The graph shows that the richer the resource in terms of semantic connections (x axis), the higher the disambiguation performance (y axis), up to a very high performance in the range of 80–90% F1 measure [49]. Similar conclusions have also been drawn in different studies [13].

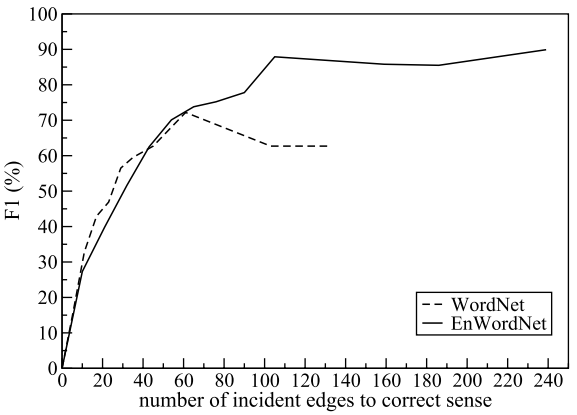


Fig. 1. WSD performance of vertex degree by number of incident edges for ambiguous words (picture from [49]).

Unfortunately, enriching lexical knowledge resources on a large scale and with high accuracy is a hard task. Current research on knowledge acquisition typically starts from existing lexical resources (such as WordNet) and applies some algorithm in order to collect new lexical and semantic information that is sense tagged and explicitly associated with concepts in the corresponding resource [56, 61, 62]. Other approaches to building an extended resource include heuristic methods based on the disambiguation of textual definitions [37]. As a result, new lexical and semantic relations between a concept and the disambiguated word in its definition can be established. More recently, a high-performance graph-based WSD algorithm, namely SSI, has been proposed for the disambiguation of WordNet textual definitions [53]. The output of the algorithm, manually validated, now consists of a 60,000-relation-edge knowledge base that enriches WordNet. A different approach is based on the automatic acquisition of relation triples [7, 20] or topic signatures [1], that is, words that cooccur frequently with a target word sense (e.g., *driver* and *fuel* frequently appear together with *car* in the sense of *automobile*). While cooccurrences are not explicitly disambiguated, more recent work has presented a novel graph-based algorithm, called SSI-Dijkstra, leading to the production of a large knowledge resource, named KnowNet [14].

The bulk of research on knowledge acquisition focuses on the enrichment of English resources, due to the availability of large-scale lexical resources such as WordNet. While lexical resources exist for other languages, they do not have enough coverage to enable accurate WSD. Recently, an unsupervised method for

the creation of a large-scale multilingual dictionary has been presented, called PanDictionary [35]. This work overcomes the limitation of working with a fixed number of languages, as it collects knowledge from hundreds of online dictionaries and Wiktionaries[1]. However, PanDictionary does not encode semantic relations between word senses, i.e., it does not fulfill the optimal criteria of a rich resource for WSD.

Another recent trend deals, instead, with harvesting semantic knowledge and structure from the so-called semi-structured resources, i.e., knowledge repositories which provide a "middle ground" between fully-structured resources like WordNet, and fully unstructured resources like raw text [30]. Wikipedia is a case in point of this research trend, being the largest and most popular collaborative and multilingual resource of world and linguistic knowledge. Resources drawing upon Wikipedia include YAGO [29], DBpedia [8], WikiNet [43], Freebase, etc. However, these resources mostly focus on encyclopedic aspects of knowledge, while neglecting lexicographic ones, i.e., the knowledge usually encoded within dictionaries.

An alternative to automatic methods is the manual development of multilingual wordnets. The first of this kind was EuroWordNet [69], a project funded by the EU FP4. EuroWordNet is a multilingual database building on top of WordNet and encoding synsets (i.e., concepts viewed as synonym sets) in 7 European languages, each aligned to the corresponding English synsets via an interlingua index. The coverage of the national wordnets ranges between around 6% (Estonian) and 38% (Dutch) of the English WordNet. A similar resource for East European languages has been developed, namely BalkaNet [67]. In Table 1 we report the statistics for concepts and relations in the two resources in comparison with WordNet. The table highlights the main problem of these wordnets for non-English languages, i.e., their limited coverage when compared to the original WordNet. The coverage problem has been partially tackled in the Multilingual Central Repository (MCR), an output of the EU FP5 Meaning Project [4]. MCR provides a resource containing a vast amount of semantic relations for many synsets. However, it relies on semi-automatic enrichment techniques, and focuses on a limited number of languages (i.e., Basque, Catalan, Italian, Spanish).

In general, while all these projects have developed meaningful open-domain resources, questions arise concerning their coverage, which strongly affects the quality of non-English disambiguation systems. Moreover, while WordNet is continuously updated, the same cannot be said for the above-mentioned resources.

Table 1
Statistics for EuroWordNet [69] & BalkaNet [60]. For each resource we provide coverage figures as the maximum and minimum number of items (i.e., lexical entries, synsets or semantic relations) across all languages in that resource

	Lexical entries	Synsets	Semantic relations
EuroWordNet	56,283–10,961	44,015–7,678	117,068–16,318
BalkaNet	7,891–24,118	25,453–4,557	n/a
WordNet 3.0	155,327	117,597	285,348

A solution to the low coverage issue is to perform WSD across languages, a task that has been shown to achieve state-of-the-art performance [54]. Cross-lingual WSD is the task of associating a word sense tag (i.e., a meaning) in a target language with a word in a source language. For example, consider again the English sentence:

[A] *Spring water* can be found at different *altitudes*.

Assuming English is the source and French the target language, our aim is to assign the geological French sense *source* to the English word *spring*. To perform cross-lingual WSD, different approaches are adopted. However, they all require either bilingual corpora for all language pairs of interest (e.g., [5, 17, 25, 28, 34]) or a multilingual knowledge resource [31, 32]. Unfortunately, the latter approach suffers from the coverage problems mentioned previously, whereas the former requires the availability of bilingual corpora for all pairs of languages of interest. Thus, if we aimed to cover n languages (e.g., the 23 official European languages), we would have to rely on the existence of (possibly aligned) bilingual corpora for $\binom{n}{2}$ language pairs (e.g., $\binom{23}{2} = 253$). While a large-scale parallel corpus exists for almost all European languages, i.e., JRC-Acquis [63], this corpus is aligned only at the sentence level, it is not sense tagged, and is heavily domain-dependent, since it contains EU documents of a predominantly legal nature.

Given the resource-related problems of cross-lingual disambiguation, we crucially note that no study to date has used – *jointly* and at the same time – the information *available* in all languages in order to perform Word Sense Disambiguation, a task we refer to as multilingual WSD. In general, multilingual WSD is the task of assigning a multilingual word sense to a word in context. For instance, given sentence (A) above, a central objective in multilingual WSD would be to tag the word

[1] http://www.wiktionary.org

spring with a multilingual sense, e.g., the set { *spring*, *source*, *Quelle*, ..., *sorgente* }. On the one hand, multilingual WSD implies working on many languages at the same time, thus posing additional challenges. However, on the other hand, the multilingual setting allows us to perform disambiguation for a specific language by leveraging all available resources, including, that is, those available in other languages, since the disambiguation is to be performed with multilingual sense tags. In practice, by working in a multilingual scenario, one can use the semantic relations available for one language for another language, e.g., even a language for which no relevant relations are available.

To summarize, multilingual WSD provides us with an experimental setting capable of overcoming the knowledge acquisition bottleneck. Given a multilingual resource whose lexical realizations in many languages are linked to a large-coverage knowledge base, it then becomes possible to develop a knowledge-based approach to multilingual WSD, in order to disambiguate for many languages – including those that are resource-poor. And it is the automatic creation of this multilingual lexical knowledge resource, i.e., BabelNet, that we are going to describe in the core part of this article.

2. Where to start from: WordNet and Wikipedia

Before delving into details on our vision of bringing together lexical knowledge in many languages within a unified resource, we describe two key complementary knowledge resources available online: WordNet and Wikipedia.

2.1. WordNet

WordNet [22, 39] is undoubtedly the most popular lexical knowledge resource in the area of NLP. It is a computational lexicon of English based on psycholinguistic principles. A concept in WordNet is represented as a synonym set (called *synset*), i.e., the set of words that share the same meaning. For instance, the concept of play as a dramatic work is expressed by the following synset[2]:

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

where the subscript and superscript of each word denote its part of speech (e.g., n stands for noun) and sense number, respectively. Words can be polysemous and consequently the same word, e.g., play, can appear in more than one synset. For example, WordNet represents the concept of dramatic play with the above synset and the concept of children's play activity with the following synset:

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

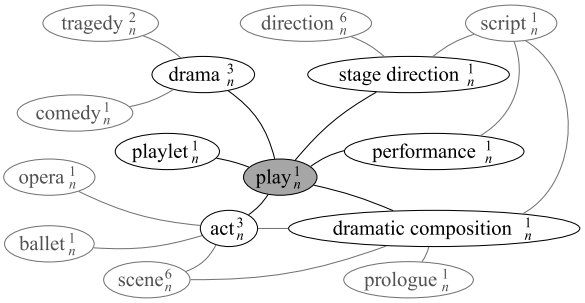
For each synset, WordNet provides a textual definition, or *gloss*. For example, the gloss of the first synset of play_n is: "a dramatic work intended for performance by actors on a stage". Synsets can contain small sentences which provide examples of their usage, e.g., "he wrote several plays but only one was produced on Broadway" for the dramatic work sense of play. Finally, WordNet provides *lexical* and *semantic relations* which relate synsets to each other. The inventory of semantic relations varies among parts of speech, including different kinds of relations between synsets. For instance, given two nominal synsets, typical semantic relations that can hold between them in WordNet include:

- *is-a* relations such as hypernymy (expressing concept generalization, e.g., play^1_n *is-a* dramatic composition^1_n) and hyponymy (expressing concept specialization): the *is-a* relation is by far the most common in WordNet. It structures the concepts expressed by synsets into a lexicalized taxonomy where each concept inherits information from its superordinate concepts.
- *instance-of* relations denoting set membership between a named entity and the class it belongs to (for instance, Shakespeare^1_n is an instance of dramatist^1_n).[3]
- *part-of* relations expressing the elements of a partition by means of meronymy (e.g., a stage direction^1_n is a meronym of play^1_n) and holonymy (e.g., a play^1_n is a holonym of stage direction^1_n).

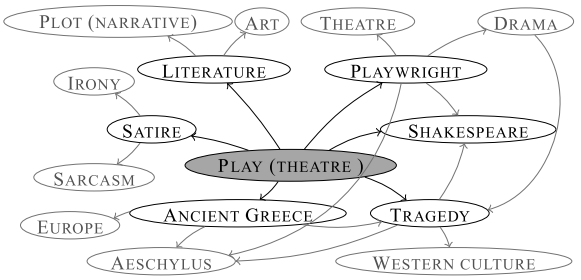
In addition to the standard WordNet relations, BabelNet also considers *gloss* relations. Given a synset S and its set of disambiguated gloss words $gloss(S) =$

[2] In the following we use WordNet version 3.0. Following [45] we denote with w^i_p the i-th sense of a word w with part of speech p (e.g., play^1_n denotes the first nominal sense of play). We use word senses to denote the corresponding synsets unambiguously (e.g., play^1_n for { play^1_n, drama^1_n, dramatic play^1_n }). Hereafter, we use *word sense* and *synset* interchangeably.

[3] This is a specific form of *is-a* introduced in WordNet 2.1 [40].



(a) Excerpt of the WordNet graph centered on the synset play^1_n.



(b) Excerpt of the Wikipedia graph centered on the Wikipage PLAY (THEATRE).

Fig. 2. Excerpts of the WordNet (a) and Wikipedia graphs (b). Both resources can be viewed as graphs by taking synsets (Wikipages, respectively) as nodes and lexical and semantic relations between synsets (hyperlinks between pages) as edges (picture from [51]).

$\{ \mathsf{s}_1, \ldots, \mathsf{s}_k \}$[4], we introduce a semantic gloss relation between S and each synset S_i which contains a sense $s_i \in gloss(S)$, $i = 1, \ldots, k$. For instance, the disambiguated gloss for play^1_n contains senses like actor^1_n and stage^3_n, so S – i.e., play^1_n – is related to both of the latter synsets via the gloss relation.

2.2. Wikipedia

Our second resource, Wikipedia, is a multilingual Web-based encyclopedia. It is a collaborative open source medium maintained by volunteers to provide a very large wide-coverage repository of encyclopedic information. Each article in Wikipedia is represented as a page (henceforth, Wikipage) and presents information about a specific concept (e.g., PLAY (THEATRE)) or named entity (e.g., WILLIAM SHAKESPEARE)[5]. The title of a Wikipage (e.g., PLAY (THEATRE)) is typically composed of the defined concept's lemma (e.g., play) plus an optional label in parentheses which specifies its meaning if the lemma is ambiguous (e.g., theatre vs. activity).

The text in Wikipedia is partially structured, which makes it an important source of knowledge from which structured information can be harvested [30]. Apart from tables and infoboxes contained in Wikipages (infoboxes are a special kind of table which summarizes the most important attributes of the entity referred to by a page, such as the birth date and biographical details of a playwright like WILLIAM SHAKESPEARE), the pages are related by means of a number of relations, including:

- **Redirect pages:** These pages are used to forward to the Wikipage that contains the actual information about a certain concept. This is used to express alternative expressions for the same concept, thus modeling *synonymy*. For example, STAGEPLAY and THEATRICAL PLAY are both redirections to PLAY (THEATRE).
- **Disambiguation pages:** These pages contain links to a number of possible concepts which correspond to different meanings of a given expression. This models *homonymy* and *polysemy*, e.g., PLAY links to both pages PLAY (THEATRE) and PLAY (ACTIVITY).
- **Internal links:** Wikipages typically include hyperlinks to other Wikipages, which often refer to related concepts. For instance, PLAY (THEATRE) links to LITERATURE, PLAYWRIGHT, DIALOGUE, etc., PLAY (ACTIVITY) links to SOCIALIZATION, GAME, RECREATION, and so on.
- **Inter-language links:** Wikipages also provide links to their counterparts (i.e., corresponding concepts) contained within wikipedias written in other languages (e.g., the English Wikipage PLAY (THEATRE) links to the Italian DRAMMA and German BÜHNENWERK).
- **Categories:** Wikipages can be associated with one or more categories, i.e., category pages used to encode topics, e.g., PLAY (THEATRE) is categorized as THEATRE, DRAMA, LITERATURE, etc.

WordNet and Wikipedia can both be viewed as graphs: in the case of WordNet, nodes are synsets and

[4]Sense disambiguated glosses are available from the Princeton WordNet project at http://wordnet.princeton.edu/glosstag.shtml.

[5]Throughout this paper, unless otherwise stated, we use the general term *concept* to denote either a concept or a named entity.

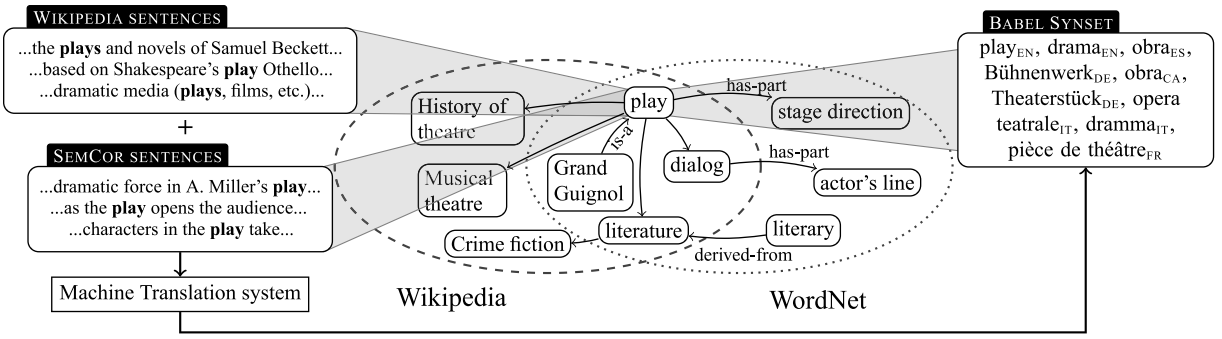


Fig. 3. An illustrative overview of BabelNet (we label nodes with English lexicalizations only): unlabled edges are obtained from links in the wikipage (e.g., PLAY (THEATRE) links to MUSICAL (THEATRE), whereas labled ones from WordNet (e.g., play^1_n has-part stage direction).

edges lexical and semantic relations between synsets, whereas, in the case of Wikipedia, nodes are Wikipages and edges the hyperlinks between them (i.e., the above-mentioned *internal* links). A small part of the WordNet and Wikipedia graphs centered on the synset play^1_n and Wikipage PLAY (THEATRE) is given in Fig. 2(a) and (b), respectively. The two graphs highlight the degree of complementarity of these two resources: while there are nodes in the two graphs which roughly correspond to the same concept (e.g., tragedy^2_n and TRAGEDY), each resource also contains specific knowledge which is missing in the other: this includes missing concepts (for instance, no Wikipage corresponding to direction^6_n), named entities (such as ANCIENT GREECE missing in WordNet), etc.

3. BabelNet

Given the above-mentioned highly complementary nature of WordNet, i.e., the largest machine-readable computational lexicon of English, and Wikipedia, i.e., the most popular multilingual encyclopedia, the next natural step was to integrate the two resources so as to create a large multilingual semantic network covering as many languages as possible. This resource, named BabelNet [51], and available online at `http://babelnet.org`, is therefore a large-scale "encyclopedic dictionary".

BabelNet encodes knowledge 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 (e.g., play *is-a* dramatic composition). Each edge is labeled with a *semantic relation* from R, e.g., {*is-a*, *part-of*, ..., ε}, where ε denotes an unspecified semantic relation. Importantly, each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., { play$_{EN}$, Theaterstück$_{DE}$, dramma$_{IT}$, obra$_{ES}$, ..., pièce de théâtre$_{FR}$ }. We call such multilingually lexicalized concepts *Babel synsets*. Concepts and relations in BabelNet are harvested from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia (introduced in Section 2). In order to construct the BabelNet graph, we extract at different stages: from WordNet, all available word senses (as *concepts*) and all the lexical and semantic pointers between synsets (as *relations*); from Wikipedia, all the Wikipages (i.e., Wikipages, as *concepts*) and semantically unspecified *relations* from their hyperlinks.

A graphical overview of BabelNet is given in Fig. 3. As can be seen, WordNet and Wikipedia overlap both in terms of concepts and relations: this overlap makes the merging between the two resources possible, enabling the creation of a *unified knowledge resource*. In order to enable multilinguality, we collect the lexical realizations of the available concepts in different languages. Finally, we connect the multilingual Babel synsets by establishing semantic relations between them. Thus, our methodology consists of three main steps:

1. We **integrate WordNet and Wikipedia** by automatically creating a mapping between WordNet senses and Wikipages (Section 3.1). This avoids duplicate concepts and allows their inventories of concepts to complement each other.
2. We **collect multilingual lexicalizations** of the newly-created concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (i.e., the *inter-language* links), as well as (b) a machine translation system

to translate occurrences of the concepts within sense-tagged corpora (Section 3.2).
3. We **create relations between Babel synsets** by harvesting all the relations in WordNet and in the wikipedias in the languages of interest (Section 3.3).

Throughout the remainder of this section, we will illustrate our approach by way of an example centered around the Wikipage PLAY (THEATRE) and the various WordNet senses of play.

3.1. Mapping Wikipedia to WordNet

During the first phase, we create links between Wikipages and WordNet senses. Given the full set of pages $Senses_{\text{Wiki}}$ and WordNet senses $Senses_{\text{WN}}$, we automatically obtain a mapping μ such that, for each Wikipage $w \in Senses_{\text{Wiki}}$, we have:

$$\mu(w) = \begin{cases} s \in Senses_{\text{WN}}(w) & \text{if a link can be established,} \\ \varepsilon & \text{otherwise,} \end{cases}$$

where $Senses_{\text{WN}}(w)$ is the set of senses of the *lemma* of w in WordNet. Given a Wikipage w, its corresponding lemma is given by either its title (tragedy for TRAGEDY) or the main token of a sense-labeled title (play for PLAY (THEATRE)). For instance, if we linked PLAY (THEATRE) to the corresponding WordNet sense play_n^1, we would have $\mu(\text{PLAY (THEATRE)}) = \text{play}_n^1$. Our approach is based on the following steps:

1. First, we apply a *mapping algorithm* (Section 3.1.1) that:
   (a) takes advantage of monosemous senses and redirections to establish immediate mappings;
   (b) given a Wikipage, determines the WordNet sense that maximizes the probability of the sense which is most suitable for that page.
2. In order to perform the mapping, we view the mapping process as a disambiguation problem, and create a *disambiguation context* for both WordNet senses and Wikipages (Section 3.1.2).
3. Finally, we provide two strategies to *estimate the conditional probability* of a WordNet sense given a Wikipage, both based on disambiguation contexts (Section 3.1.3). These strategies either:
   (a) make use of a simple bag-of-words (BoW) approach, or
   (b) exploit the graph structure of the target resource to perform the mapping.

3.1.1. Mapping algorithm

We perform the following steps to link each Wikipage to a WordNet sense:

- Initially, our mapping μ is empty, i.e., each Wikipage w is linked to ε.
- For each Wikipage w whose lemma is monosemous both in Wikipedia and WordNet we map w to its only WordNet sense w_n^1.
- Finally, for each remaining Wikipage w for which no mapping was previously found (i.e., $\mu(w) = \varepsilon$), we do the following: for each Wikipage d which is a redirection to w, for which a mapping was previously found (i.e., $\mu(d) \neq \varepsilon$, that is, d is monosemous in both Wikipedia and WordNet) and such that it maps to a sense $\mu(d)$ in a synset S that also contains a sense of w, we map w to the corresponding sense in S. If a Wikipage w has not yet been linked, we assign the most likely sense to w based on the maximization of the conditional probabilities $p(s|w)$ over the senses $s \in Senses_{\text{WN}}(w)$ (no mapping is established if a tie occurs).

The algorithm returns the resulting mapping μ. At the core of our mapping algorithm lies the calculation of the conditional probability $p(s|w)$ of selecting the WordNet sense s given the Wikipage w. The sense s which maximizes this probability is determined as follows:

$$\mu(w) = \operatorname*{argmax}_{s \in Senses_{\text{WN}}(w)} p(s|w) = \operatorname*{argmax}_{s} \frac{p(s, w)}{p(w)} = \operatorname*{argmax}_{s} p(s, w). \quad (1)$$

The most appropriate sense s is obtained by maximizing the joint probability $p(s, w)$ of sense s and page w.

3.1.2. Disambiguation contexts

The joint probability of a WordNet sense and Wikipage is estimated by using the same technique as that adopted in Word Sense Disambiguation [45], i.e., we define a *disambiguation context* for each of the two concepts. Given a concept, i.e., a page or sense, this disambiguation context is a set of words obtained from the corresponding resource (i.e., Wikipedia or WordNet), whose senses are associated with the input concept through some semantic relation and which support a potential link in our mapping μ.

Disambiguation context of a Wikipage We use the following information as disambiguation context of a Wikipage w:

- **Sense labels:** e.g., given the page PLAY (THEATRE), the word theatre is added to the disambiguation context.
- **Links:** the lemmas of the titles of pages linked from the Wikipage w (i.e., outgoing links). For example, the links in the Wikipage PLAY (THEATRE) include literature, comedy, etc.
- **Redirections:** the lemmas of the titles of pages which redirect to w, e.g., PLAYLET redirects to PLAY (THEATRE), so playlet is included in the context.
- **Categories:** we include the syntactic heads of Wikipage categories as additional context. For example, the Wikipage PLAY (THEATRE) is categorized as PLAYS, DRAMA, THEATRE, etc.

The set of words obtained from all the sources above defines the disambiguation context $Ctx(w)$ of a Wikipage w. For example, $Ctx(\text{PLAY (THEATRE)}) = \{$ theatre, literature, comedy, ..., playlet, drama, ..., character $\}$.

Disambiguation context of a WordNet sense Given a WordNet sense s and its synset S, the following sources are used as disambiguation context:

- **Synonymy:** all synonyms of s in synset S. For example, given the synset of play_n^1, the context will include all its synonyms (that is, drama and dramatic play).
- **Hypernymy/Hyponymy:** all synonyms in the synsets H such that H is either a hypernym (i.e., a generalization) or a hyponym (i.e., a specialization) of S. For example, given play_n^1, we include its hypernym dramatic composition.
- **Gloss:** the set of lemmas of the content words occurring within the gloss of s. For instance, given $s = \text{play}_n^1$, defined as "a dramatic work intended for performance by actors on a stage", the disambiguation context of s will include the following lemmas: work, dramatic work, intend, performance, actor, stage.

We define the disambiguation context $Ctx(s)$ of a given a WordNet sense s as the set of words collected from some or all of the sources above. For example, $Ctx(\text{play}_n^1) = \{$drama, dramatic play, composition, work, intend, ..., actor, stage$\}$.

3.1.3. Probability estimation

Once the disambiguation contexts are determined, we can calculate the joint probability defined in Equation 1, i.e., the probability of a WordNet sense and Wikipage referring to the same concept. We calculate $p(s, w)$ as:

$$p(s, w) = \frac{score(s, w)}{\sum_{\substack{s' \in Senses_{\text{WN}}(w),\\ w' \in Senses_{\text{Wiki}}(w)}} score(s', w')}, \qquad (2)$$

We define two different ways of computing the $score(s, w)$ function:

- **Bag-of-words method:** computes $score(s, w) = |Ctx(s) \cap Ctx(w)| + 1$ (we add 1 as a smoothing factor). This is a simple method already proposed in [50], that determines the best sense s by computing the intersection of the disambiguation contexts of s and w, and thus it does not exploit the structural information available in WordNet or Wikipedia.
- **Graph-based method:** starts with the flat disambiguation context of the Wikipage $Ctx(w)$ and transforms it into the structured representation of a graph, which is then used to score the different senses of w in WordNet. A labeled directed graph $G = (V, E)$ is built following the same procedure outlined in [49] which connects possible senses of w's lemma with the senses of the words found in $Ctx(w)$. Specifically:
  1. We first define the set of nodes of G to be made up of all WordNet senses for the lemma of Wikipage w and for the words in $Ctx(w)$. Initially, the set of edges of G is empty, i.e., $E := \emptyset$.
  2. Next, we connect the nodes in V on the basis of the paths found between them in WordNet. Formally, for each vertex $v \in V$, we perform a depth-first search along the WordNet graph and every time we find a node $v' \in V$ $(v \neq v')$ along a simple path $v, v_1, \ldots, v_k, v'$ of maximal length L, we add all intermediate nodes and edges of such a path to G, i.e., $V := V \cup \{v_1, \ldots, v_k\}$, $E := E \cup \{(v, v_1), \ldots, (v_k, v')\}$.

  The result of this procedure is a subgraph of WordNet containing (1) the senses of the words in context, (2) all edges and intermediate senses found in WordNet along all paths of maximal length L that connect them. To compute $score(s, w)$ given a disambiguation graph G, we

define a scoring function $score(s, w)$ of the paths starting from s and ending in any of the senses of the context words $Ctx(w)$ by just summing up the values $e^{-(length(p)-1)}$ for each such path p between s and s' in WordNet, where $length(p)$ is the length of path p in terms of its number of edges.

3.2. Translating Babel synsets

We can now exploit our mapping between English Wikipages and WordNet senses to create our Babel synsets. Given a Wikipage w mapped to a sense s (i.e., $\mu(w) = s$), we create a Babel synset $S \cup W$, where S is the WordNet synset to which sense s belongs, and W includes: (i) w; (ii) the set of redirections to w; (iii) all the pages linked via its inter-language links; (iv) the redirections to the inter-language links present in the wikipedia of the target language. For example, having μ(PLAY (THEATRE)) = play^1_n, we create the following Babel synset: { play_{EN}, $\text{Bühnenwerk}_{\text{DE}}$, $\text{pièce de théâtre}_{\text{FR}}$, ..., $\text{opera teatrale}_{\text{IT}}$ }. The inclusion of redirections additionally enlarges the Babel synset with { $\text{Theaterstück}_{\text{DE}}$, $\text{texte dramatique}_{\text{FR}}$ }. However, a concept might be covered only in one of the two resources (either WordNet or Wikipedia), because no link could be established (e.g., with MUSICAL THEATRE or actor's line^1_n); alternatively, even if present in both resources, inter-language links might be missing for some language of interest (e.g., the Spanish and Catalan inter-language links for PLAY (THEATRE) are missing in Wikipedia).

To tackle the above issues and keep coverage high for all languages we translate the English senses in the Babel synset into missing languages. To do so, given a WordNet word sense in our Babel synset of focus (e.g., play^1_n) we collect its occurrences in SemCor [41], a corpus of more than 200,000 words annotated with WordNet senses. We do the same for Wikipages by collecting Wikipedia sentences with hyperlinks to the Wikipage of interest (e.g., PLAY (THEATRE)). By repeating this step for each English lexicalization in a Babel synset, we collect dozens of sentences for the synset (see left part of Fig. 3). Next, we apply a state-of-the-art Machine Translation system to translate these sentences. Given a specific term in the initial Babel synset, we collect the set of its translations. We then enrich the Babel synset with the most frequent translation in each language. For example, in order to collect missing translations for PLAY (THEATRE) and its corresponding WordNet sense play^1_n, we collect from Wikipedia occurrences of hyperlinks to the Wikipage and translate sentences such as the following:

(a) Best known for his **[[Play (theatre)|play]]** Ubu Roi, which is often cited as a forerunner to the surrealist theatre of the 1920s and 1930s, Jarry wrote in a variety of genres and styles.

Similarly, from SemCor we collect and translate, among others, the following sentence:

(b) The situation in which we find ourselves is brought out with dramatic force in Arthur Miller's $\textbf{play}^1_n$ The Crucible, which deals with the Salem witch trials.

As a result, we can augment the initial Babel synset with the following words: drame_{FR}, $\text{dramma}_{\text{IT}}$, obra_{CA}, obra_{ES}. Note that not only do we obtain translations for Catalan and Spanish which were initially unavailable, but we also obtain more lexicalizations for other languages, such as French and Italian.

3.3. Harvesting semantic relations

As the final step of our integration methodology, we establish semantic relations between our multilingual Babel synsets. We achieve this objective by, first, collecting the relations directly from WordNet and Wikipedia, and, second, weighting them using a relatedness measure based on the Dice coefficient. We first collect all lexical and semantic relations from WordNet (including the gloss relations introduced in Section 2.1). For example, given the Babel synset for play^1_n, we relate it to the Babel synsets of playlet^1_n, act^3_n, etc. (cf. Fig. 2(a)). We then add all relations from Wikipedia, by collecting all links occurring within each Wikipage and establishing an unspecified semantic relation ε between their corresponding Babel synsets (cf. the semantic relations for PLAY (THEATRE) in Fig. 2(b)).

We weight all the BabelNet edges so as to quantify the strength of association between Babel synsets. We use different strategies to take advantage of WordNet's and Wikipedia's respective distinctive properties – i.e., the availability of high-quality definitions from WordNet, and large amounts of hyperlinked text from Wikipedia – both based on the Dice coefficient. Given an existing, semantic relation between two WordNet synsets s and s', we calculate its corresponding weight using a method similar to the Extended Gloss Overlap measure for computing semantic relatedness [6]. We start by collecting (a) synonyms and (b) all gloss words from s and s', as well as their directly linked synsets,

into two bags of words S and S'. We remove stopwords and lemmatize the remaining words. We then compute the degree of association between the two synsets by computing the Dice coefficient as the number of words the two bags have in common normalized by the total number of words in the bags: $\frac{2\times|S\cap S'|}{|S|+|S'|}$.

In the case of edges corresponding to semantic relations between Wikipedia pages, instead, we calculate the degree of correlation between the two pages by using a co-occurrence based method which draws on large amounts of hyperlinked text. Given two Wikipages w and w', we compute the occurrence frequency of each individual page (f_w and $f_{w'}$) as the number of hyperlinks found in Wikipedia which link to it, and the co-occurrence frequency of w and w' ($f_{w,w'}$) as the number of times these links occur together within a sliding window of 40 words. The strength of association between w and w' is then obtained by calculating the Dice coefficient formula: $\frac{2\times f_{w,w'}}{f_w+f_{w'}}$. For instance, the Wikipages PLAY (THEATRE) and SATIRE occur as a link in Wikipedia 1,560 and 2,568 times, respectively, and co-occur 9 times within the same context. As a result, the Dice coefficient for these two pages is 0.0044.

3.4. BabelNet 2.0

The current brand-new version of the semantic network, i.e., BabelNet 2.0, covers 50 languages, and is available online, together with API for its programmatic use[6] BabelNet 2.0 integrates the following resources:

- WordNet [22], a popular computational lexicon of English (version 3.0),
- Open Multilingual WordNet [9], a collection of wordnets available in different languages (August 2013 dump),
- Wikipedia, the largest collaborative multilingual Web encyclopedia (October 2012 dumps),
- OmegaWiki[7], a large collaborative multilingual dictionary (01/09/2013 dump).

The number of lemmas for each language ranges between more than 8 million (English) and almost 100,000 (Latvian), with a dozen languages having more than 1 million lemmas. The number of polysemous terms ranges between almost 250,000 in English to only a few thousand for languages such as Galician, Latvian and Esperanto, with most languages having several tens of thousands of polysemous terms.

[6]http://babelnet.org

[7]http://omegawiki.org

BabelNet 2.0 contains about 9.3 million concepts, i.e., Babel synsets, and over 50 million word senses (regardless of their language). It also contains about 7.7 million images and almost 18 million textual definitions, i.e., glosses, for its Babel synsets. The synsets are linked to each other by a total of about 262 million semantic relations (mostly from Wikipedia). More statistics on the resource are available from http://babelnet.org/stats.jsp.

4. BabelNet's Friends

We now outline some of the applications that have been enabled in just a few months thanks to the availability of a large-scale multilingual semantic network such as BabelNet. Our perception is that this is just the tip of the iceberg, and that BabelNet could enable many more applications and uses in many areas, not only of NLP, but also related fields such as Information Retrieval. We will explore these opportunities in Section 5.

4.1. Multilingual joint WSD

As a first natural application of our multilingual semantic network we proposed a multilingual approach to WSD [52] which exploits three main factors:

1. the complementarity of the different translations of the most suitable senses of a target word in context;
2. the wide-coverage, multilingual lexical knowledge present in BabelNet;
3. the use of knowledge available in different languages to synergistically support disambiguation.

We called this approach *multilingual joint WSD*, since disambiguation is carried out by leveraging different languages *jointly and at the same time*. To this end, we first perform graph-based WSD using the target word in context as input, and then combine sense evidence from its translations using an ensemble method. The key idea of our joint approach is that sense evidence from translations in different languages provides complementary information for the senses of a target word in context. Therefore, more accurate sense predictions should be produced when such evidence is combined.

Given a word sequence $\sigma = (w_1, \ldots, w_n)$, and given a target word $w \in \sigma$, we disambiguate w as follows. We

start by collecting the knowledge needed for disambiguation. First, we collect the set S of Babel synsets corresponding to the different senses of the target word w. Next, we create the set T of multilingual lexicalizations of the target word w: to this end, we first include in T the word w itself, and then iterate through each synset $s \in S$ to collect the translations of each of its senses into the languages of focus. Finally, we create a disambiguation context *ctx* by taking the word sequence σ and removing w from it.

Next, we calculate a probability distribution over the different synsets S of w for each term $t_i \in T$. Each probability distribution quantifies the support for the different senses of the target word, determined using t_i and the context *ctx*: we save this information in a $|T| \times |S|$ matrix $LScore$, where each cell $lScore_{i,j}$ quantifies the support for synset $s_j \in S$, calculated using the term in $t_i \in T$. We determine the scores as follows:

- We select an element t_i from T at each step.
- Next, we create a multilingual context σ' by combining t_i with the words in *ctx*.
- We use σ' to build a graph $G_i = (V_i, E_i)$ by computing the paths in BabelNet which connect the synsets of t_i with those of the other words in σ', along the lines of [49]. Note that by selecting a different element from T at each step we create a new graph where different sets of Babel synsets get activated by the context words in *ctx*.
- Finally, we compute the support from term t_i for each synset $s_j \in S$ of the target word by applying a graph connectivity measure to G_i and store the result in $lScore_{i,j}$.

By repeating the process for each term in T we compute all values in the matrix $LScore$.

In the final phase we aggregate the scores associated with each term of T using an ensemble method M. For instance, M could simply consist of summing the scores associated with each sense over all distributions. As a result, the combined scoring distribution is returned. This sense distribution in turn can be used to select the best sense for the target word $w \in \sigma$.

Our experimental results on gold standard datasets show that, thanks to complementing wide-coverage multilingual lexical knowledge with robust graph-based algorithms and combination methods, we are able to achieve the state of the art in both monolingual all-words WSD and two different cross-lingual disambiguation tasks [52].

4.2. The SemEval-2013 multilingual word sense disambiguation task

The availability of BabelNet made it possible to organize a new task focused on multilingual WSD [48] as part of the SemEval-2013 semantic evaluation competition.[8]

The task required participating systems to annotate nouns in a test corpus with the most appropriate sense from the BabelNet sense inventory or, alternatively, from two main subsets of it, namely the WordNet or Wikipedia sense inventories. In contrast to previous all-words WSD tasks we did not focus on the other three open classes (i.e., verbs, adjectives and adverbs) since, when the task was organized, BabelNet 1.1.1 was used, and that version did not provide non-English coverage for them (the current version, i.e., 2.0, instead, does provide coverage).

The test set consisted of 13 articles obtained from the datasets available from the 2010, 2011 and 2012 editions of the workshop on Statistical Machine Translation (WSMT).[9] The articles cover different domains, ranging from sports to financial news.

The same article was available in 4 different languages (English, French, German and Spanish). In order to cover Italian, an Italian native speaker manually translated each article from English into Italian, with the support of an English mother tongue advisor. The overall number of content words annotated in each article ranges between about 1400 and above 1900 words, depending upon the language.

Interestingly, several different knowledge-based systems participated in the task, whereas no supervised system did, probably due to the lack of training data for non-English languages. State-of-the-art systems achieved results ranging between 61% and 71% F1 depending on the language. Several systems were able to outperform the competitive Most Frequent Sense baseline, except in the case of Wikipedia, but current performance leaves significant room for future improvement.

4.3. SPred: Harvesting semantic predicates

Another use of BabelNet is in the creation of a large repository of semantic predicates [23], i.e., predicates whose lexical arguments are replaced by their semantic classes. We start from lexical predicates, i.e., sequences

[8] http://www.cs.york.ac.uk/semeval-2013/task12/

[9] http://www.statmt.org/wmt12/

of the kind $w_1\ w_2\ \ldots\ w_i\ *\ w_{i+1}\ \ldots\ w_n$, where w_j are tokens ($j = 1, \ldots, n$), $*$ matches any sequence of one or more tokens, and $i \in \{0, \ldots, n\}$. We call the token sequence which matches $*$ the filling argument of the predicate. For example, *a * of milk* matches occurrences such as *a full bottle of milk*, *a glass of milk*, *a carton of milk*, etc. While in principle * could match any sequence of words, since we aim at generalizing nouns, in what follows we allow $*$ to match only noun phrases (e.g., *glass*, *hot cup*, *very big bottle*, etc.).

Our objective is to obtain semantic predicates from lexical ones. A semantic predicate is a sequence $w_1\ w_2\ \ldots\ w_i\ c\ w_{i+1}\ \ldots\ w_n$, where w_j are tokens ($j = 1, \ldots, n$), $c \in C$ is a semantic class selected from a fixed set C of classes, and $i \in \{0, \ldots, n\}$. As an example, consider the semantic predicate *cup of* BEVERAGE, where BEVERAGE is a semantic class representing beverages. This predicate matches phrases like *cup of coffee*, *cup of tea*, etc., but not *cup of sky*. Other examples include: MUSICAL INSTRUMENT *is played by*, *a* CONTAINER *of milk*, *break* AGREEMENT, etc.

Semantic predicates mix the lexical information of a given lexical predicate with the explicit semantic modeling of its argument. Importantly, the same lexical predicate can have different classes as its argument, like *cup of* FOOD vs. *cup of* BEVERAGE. Note, however, that different classes might convey different semantics for the same lexical predicate, such as *cup of* COUNTRY, referring to cup as a prize instead of cup as a container.

To harvest semantic predicates, for each lexical predicate of interest (e.g., *break* $*$):

1. We extract all its possible filling arguments from Wikipedia, e.g., *lease, contract, leg, arm*, etc.
2. We disambiguate as many filling arguments as possible using disambiguation heuristics based on Wikipedia and obtaining a set of corresponding Wikipedia pages for the filling arguments, e.g., *Lease*, *Contract*, etc. For instance, we propagate an existing annotation of a given lemma in the same Wikipedia page where an ambiguous predicate argument occurs to the argument itself.
3. We create the semantic predicates by generalizing the Wikipedia-linked filling argument to their most suitable semantic classes in WordNet, e.g., *break* AGREEMENT, *break* LIMB, etc. To do this, we leverage the Wikipedia-to-WordNet mapping available in BabelNet: we first link all our disambiguated arguments to WordNet and then leverage the WordNet taxonomy to populate a fixed set of semantic classes with the most suitable annotated arguments for a given predicate.

Finally, we can exploit the learned semantic predicates to assign the most suitable semantic class to new filling arguments for the given lexical predicate. We do this by using the following probability mixture:

$$P(c|\pi, a) = \alpha P_{distr}(c|\pi, a) + (1 - \alpha) P_{class}(c|\pi), \quad (3)$$

where $\alpha \in [0, 1]$ is an interpolation factor, $P_{distr}(c|\pi, a)$ is the distributional probability of a semantic class c of an argument a for a predicate π, and $P_{class}(c|\pi)$ is the conditional probability of a class c given π calculated on the basis of the number of Wikipedia sentences which contain an argument in that class.

Our experiments [23] show that we are able to create a large collection of semantic predicates from the Oxford Advanced Learner's Dictionary with high precision and recall, also in comparison with previous work on argument supertyping. Data can be found at: http://lcl.uniroma1.it/spred.

4.4. WiSeNet: A Wikipedia Semantic Network

Open Information Extraction (OIE) is a recent direction whose aim is to extract relations and instances together, typically on a large scale, in a single pass over the text and without human input [19, 70, 71]. As a result, recent efforts in this direction are able to produce millions of relation instances relating textual mentions of concepts and named entities by means of a textual relation phrase (e.g., *is a field of* in *(Natural language processing, is a field of, Computer science)*). However, these approaches pay little attention to making explicit the semantics of such extracted information. In other words, current OIE techniques do not provide a formal semantic representation for both the arguments and the labels of the harvested relations, which can denote different meanings due to the ambiguous nature of text. This is, instead, an important focus of knowledge acquisition techniques, which mine the semantic information available in semi-structured form in resources like WordNet [38], Wikipedia[10] and Freebase[11], among others, for extracting ontologies [58, 64] and semantic networks [43, 51, 66].

In our work [42] we addressed the limitations of OIE and knowledge acquisition by combining the best

[10] http://en.wikipedia.org/

[11] http://www.freebase.com/

of the two worlds. Key to our approach is the synergistic use of, first, OIE for the extraction of huge amounts of shallow knowledge from text and, second, knowledge acquisition for providing explicit semantics for this knowledge, i.e., ontologizing it. To this end we leverage Wikipedia as the primary source of semantic information. The final product of this process is WiSeNet: http://lcl.uniroma1.it/wisenet, a full-fledged Wikipedia-based Semantic Network with labeled, ontologized relations. Not only are the arguments of our relations connected to Wikipedia, and therefore to BabelNet, the relations themselves are also ontologized as sets of synonymous phrases.

Our approach consists of three steps: relation extraction, relation ontologization and relation disambiguation. During the first step we extract relational phrases from Wikipedia by exploiting deep syntactic analysis, e.g., we extract relational phrases such as *is a member of, is a part of, is a territory of*. In the second step we define a shortest path kernel similarity measure that integrates semantic and syntactic features to automatically build relation synsets, i.e., clusters of synonymous relational phrases with semantic type signatures for their domain and range. For instance, we cluster together the relational phrases *is a member of* and *is a part of* while, in another cluster, *is a part of* and *is a territory of*. Finally, we disambiguate the relational phrases extracted from Wikipedia using these relation synsets, obtaining a large set of automatically ontologized semantic relations, e.g., we recognize that the relational phrase *is a part of* is a synonym of *is a territory of* when we consider the sentence *Nunavut is a part of Canada*, while it is a synonym of *is a member of* for the sentence *Taproot Theatre Company is a part of Theatre Communications Group*. Our experimental results [42] show the high quality of the

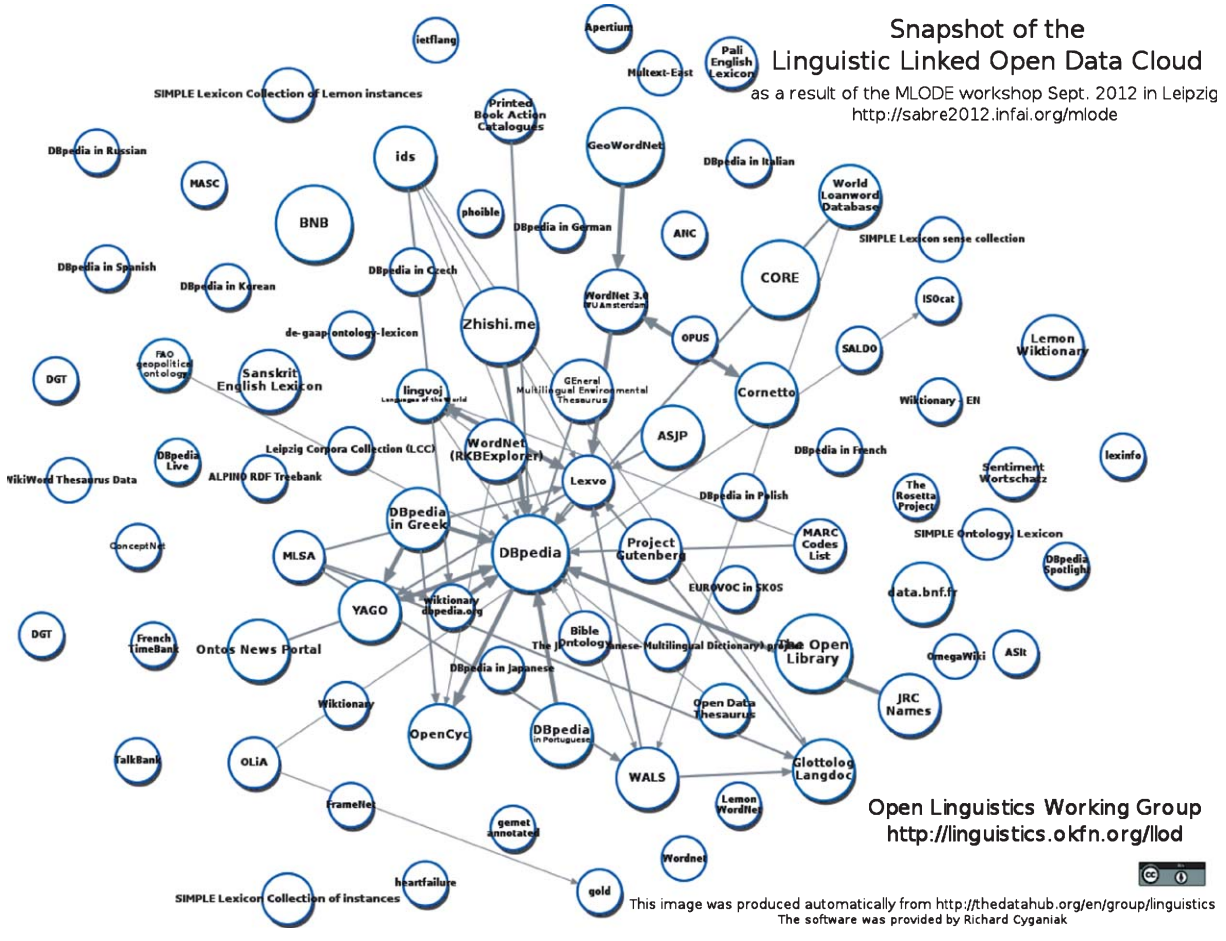


Fig. 4. Open Linguistics Working Group (2012), The Linguistic Linked Open Data cloud diagram (draft), version of September 2012, `http://linguistics.okfn.org/llod`.

acquired relation instances, synsets and disambiguated relation instances.

4.5. Connecting to the Linguistic Linked Open Data

Lexical semantic knowledge is an essential component not only for Natural Language Processing, it is also indispensable for the creation of the multilingual Semantic Web. Indeed, it is becoming increasingly critical that existing lexical resources be published as Linked Open Data (LOD), so as to foster integration, interoperability and reuse on the Semantic Web [26]. Thus, lexical resources provided in RDF format [33] can contribute to the creation of the so-called Linguistic Linked Open Data (LLOD, see Fig. 4), a vision fostered by the Open Linguistic Working Group (OWLG)[12], in which part of the Linked Open Data cloud is made up of interlinked linguistic resources [12]. The multilinguality aspect is key to this vision, in that it enables Natural Language Processing tasks which are not only cross-lingual, but also independent both of the language of the user input and of the linked data exploited to perform the task.

While the LOD is centered on DBpedia [8], the largest "hub" of Linked Data providing wide coverage of Named Entities, BabelNet focuses both on word senses and on Named Entities in many languages. Therefore, its aim is to provide full lexicographic and encyclopedic coverage. Compared to YAGO [65], BabelNet integrates WordNet and Wikipedia by means of a mapping strategy based on a disambiguation algorithm, and provides additional lexicalizations resulting from the application of MT and the integration of additional multilingual resources.

To integrate BabelNet into the LLOD, we encoded its content as LOD by using the Lemon RDF model [36]. Thanks to this model, we were able to put online a SPARQL endpoint and the Turtle RDF encodings of most of BabelNet 2.0 (see the BabelNet URL above). Our hope is that BabelNet will be used in the Semantic Web community as a "bridging tool" between real-world entities and lexical semantic knowledge.

5. Where to go from here

This is just the tip of the iceberg, and much still has to be done. Some promising directions, in which knowledge-based multilinguality can prove useful, are:

- **Semantic Textual Similarity**, the aim of which is to determine how similar two texts are at the semantic level, independently of how their content is expressed (i.e., which words are used). Recently, a unified approach based on WordNet has shown state-of-the-art performance when operating at multiple levels in a unified way [57]. We believe that the integration of multilingual lexical knowledge will provide a considerable contribution to this area.
- the **Semantic Web:** given the size of the Linked Open Data cloud, we believe that the recent availability of BabelNet as LOD might serve as a connecting resource, which could enable higher interoperability and alignment between heterogeneous data.
- **Multilingual disambiguation and entity linking**, which need very large amounts of knowledge lexicalized in many languages in order to perform a high-performance analysis of text not only at the semantic, but also at the syntactic, level. We also believe that domain-based WSD deserves more attention, seeing that many documents on the Web pertain to specific domains of interest [21]. Here, again, knowledge resources like BabelNet could be leveraged to carry out classification and disambiguation tasks.

6. Conclusions

In this paper I have presented a manifesto for research in multilingual semantic processing. After outlining the issues driving work in this research area, I illustrated the research vision that is pursued in my research group. This vision is largely based on the construction and enrichment of a multilingual "encyclopedic dictionary" and semantic network, i.e., BabelNet. BabelNet has already been demonstrated to enable a considerable number of usages and applications, described in Section 4. In the last part of the paper I outlined a likely future for multilingual semantic processing, focusing on current directions and successes.

Acknowledgements

The author gratefully acknowledges the support of the "MultiJEDI" ERC Starting Grant No. 259234. Thanks go to Francesco Maria Tucci and Jim McManus for making this crazy paper possible!

[12] http://linguistics.okfn.org

References

[1] E. Agirre, O. Ansa, D. Martinez and E. Hovy, Enriching WordNet concepts with topic signatures, In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pp. 23–28, Pittsburg, Penn., 2001.

[2] E. Agirre, O.L. de Lacalle and A. Soroa, Knowledge-based WSD on specific domains: Performing better than generic supervised WSD. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, Pasadena, Cal. 14–17 July 2009, pp. 1501–1506, 2009.

[3] E. Agirre and A. Soroa, Semeval-2007 task 2: Evaluating word sense induction and discrimination systems. In *Proc of SemEval-2007*, pp. 7–12, 2007.

[4] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini and P. Vossen, The MEANING multilingual central repository. In *Proceedings of the 2nd International Global WordNet Conference*, Brno, Czech Republic, 20–23 January 2004, pp. 80–210, 2004.

[5] C. Banea and R. Mihalcea, Word sense disambiguation with multilingual features, In *Proceedings of the Ninth International Conference on Computational Semantics*, pp. 25–34. Association for Computational Linguistics, 2011.

[6] S. Banerjee and T. Pedersen, Extended gloss overlap as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 9–15 August 2003, pp. 805–810, 2003.

[7] M. Banko, M.J. Cafarella, S. Soderland, M. Broadhead and O. Etzioni, Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6–12 January 2007, pp. 2670–2676, 2007.

[8] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann. Dbpedia - a crystallization point for the web of data, *Journal of Web Semantics* **7**(3):154–165, 2009.

[9] F. Bond and R. Foster, Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1352–1362, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10] J.F. Cai, W.S. Lee and Y.W. Teh, NUS-ML: Improving word sense disambiguation using topic features. In *Proc of SemEval-2007*, pp. 249–252, 2007.

[11] Y.S. Chan, H.T. Ng and Z. Zhong, NUS-ML: Exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, 23–24 June 2007, pp. 253–256, 2007.

[12] C. Chiarcos, S. Hellmann and S. Nordhoff, Towards a linguistic linked open data cloud: The Open Linguistics Working Group, *TAL* **52**(3):245–275, 2011.

[13] M. Cuadros and G. Rigau, Quality assessment of large scale knowledge resources. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 22–23 July 2006, pp. 534–541, 2006.

[14] M. Cuadros and G. Rigau, KnowNet: Building a large net of knowledge from the Web. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, U.K., 18–22 August 2008, pp. 161–168, 2008.

[15] B. Decadt, V. Hoste, W. Daelemans and A. van den Bosch, Gambl, genetic algorithm optimization of memory-based WSD. In *Proc of SENSEVAL-3*, pp. 108–112, 2004.

[16] A. Di Marco and R. Navigli, Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction, *Computational Linguistics* **39**(3):709–754, 2013.

[17] M. Diab, An unsupervised method for multilingual word sense tagging using parallel corpora: A preliminary investigation. In *Proceedings of the ACL-2000 workshop on Word senses and multilinguality*, pp. 1–9, 2000.

[18] P. Edmonds, *Designing a task for SENSEVAL-2*, Technical report, University of Brighton, U.K., 2000.

[19] O. Etzioni, M. Banko, S. Soderland and D.S. Weld, Open information extraction from the web, *Commun ACM* **51**(12):68–74, 2008.

[20] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld and A. Yates, Web-scale information extraction in KnowItAll (Preliminary results). In *Proc of WWW-04*, pp. 100–110, 2004.

[21] S. Faralli and R. Navigli, A new minimally-supervised framework for domain Word Sense Disambiguation. In *Proc. of the 2012 Conference on Empirical Methods in Natural Language Processing*, Jeju, Korea, July 12–14, pp. 1411–1422, 2012.

[22] C. Fellbaum, editor. *WordNet: An Electronic Database*, MIT Press, Cambridge, MA 1998.

[23] T. Flati and R. Navigli, SPred: Large-scale Harvesting of Semantic Predicates. In *Proceedings of the 51st AnnualMeeting of the Association for Computational Linguistics*, pp. 1222–1232, Sofia, Bulgaria 2013.

[24] W.A. Gale, K. Church and D. Yarowsky, A method for disambiguating word senses in a corpus, *Computers and the Humanities* **26**:415–439, 1992.

[25] W.A. Gale, K. Church and D. Yarowsky, Using bilingual materials to develop Word Sense Disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal*, Canada, 25-27 June 1992, pp. 101–112, 1992.

[26] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar and J. McCrae, Challenges for the multilingual web of data, *J Web Sem* **11**:63–71, 2012.

[27] C. Grozea, Finding optimal parameter settings for high performance Word Sense Disambiguation. In *Proc Of SENSEVAL-3*, pp. 125–128, 2004.

[28] W. Guo and M.T. Diab, Combining orthogonal monolingual and multilingual sources of evidence for all words WSD. In *Proc of ACL*, pp. 1542–1551, 2010.

[29] J. Hoffart, F.M. Suchanek, K. Berberich and G. Weikum, Yago2: A spatially and temporally enhanced knowledge base from wikipedia, *Artificial Intelligence* **194**:28–61, 2013.

[30] E.H. Hovy, R. Navigli and S.P. Ponzetto, Collaboratively built semi-structured content and artificial intelligence: The story so far, *Artificial Intelligence* **194**:2–27, 2013.

[31] N. Ide, Cross-lingual sense determination: Can it work? *Computers and the Humanities* **34**:223–234, 2000.

[32] R. Ion and D. Tufiş, Multilingual Word Sense Disambiguation using aligned wordnets, *Romanian Journal on Science and Technology of Information, Special Issue on Balka-Net* **7**(1-2):183–200, 2004.

[33] O. Lassila and R.R. Swick, Resource description framework (RDF) model and syntax specification. In *Technical report, World Wide Web Consortium*, 1999.

[34] E. Lefever, V. Hoste and M. De Cock, Parasense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 317–322, 2011.

[35] Mausam, S. Soderland, O. Etzioni, D. Weld, M. Skinner and J. Bilmes, Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Singapore, 2–7 July 2009, pp. 262–270, 2009.

[36] J. McCrae, D. Spohr and P. Cimiano, Linking lexical resources and ontologies on the Semantic Web with Lemon. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC)*, pp. 245–259, Heraklion, Crete, Greece 2011.

[37] R. Mihalcea and D. Moldovan, eXtended WordNet: Progress report. In *Proceedings of the NAACL-01 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Penn., June 2001, pp. 95–100, 2001.

[38] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller, Introduction to WordNet: An On-Line Lexical Database, *International Journal of Lexicography* **3**(4):235–244, 1990.

[39] G.A. Miller, R.T. Beckwith, C.D. Fellbaum, D. Gross and K. Miller, WordNet: An online lexical database, *International Journal of Lexicography* **3**(4):235–244, 1990.

[40] G.A. Miller and F. Hristea, WordNet nouns: Classes and instances, *Computational Linguistics* **32**(1):1–3, 2006.

[41] G.A. Miller, C. Leacock, R. Tengi and R. Bunker, A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pp. 303–308, Plainsboro, N.J., 1993.

[42] A. Moro, R. Navigli, Integrating syntactic and semantic analysis into the open information extraction paradigm. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 2148–2154, Beijing, China, 2013.

[43] V. Nastase and M. Strube, Transforming wikipedia into a large scale multilingual concept network, *Artificial Intelligence* **194**:62–85, 2013.

[44] R. Navigli, Semi-automatic extension of large-scale linguistic knowledge bases. In *Proceedings of the 18th Internationa Florida AI Research Symposium Conference*, Clearwater Beach, Flo., 15–17 May 2005, pp. 548–553, 2005.

[45] R. Navigli, Word Sense Disambiguation: A survey, *ACM Computing Surveys* **41**(2):1–69, 2009.

[46] R. Navigli, A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pp. 115–129, 2012.

[47] R. Navigli, S. Faralli, A. Soroa, O. Lopez de Lacalle and E. Agirre, Two birds with one stone: Learning semantic models for Text Categorization and Word Sense Disambiguation. In *Proceedings of the Twentieth ACMConference on Information and Knowledge Management*, Glasgow, Scotland, U.K. 24–28 October 2011, pp. 2317–2320, 2011.

[48] R. Navigli, D. Jurgens and D. Vannella, Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pp. 222–231, Atlanta, USA 2013.

[49] R. Navigli and M. 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.

[50] R. Navigli and S.P. Ponzetto, BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11–16 July 2010, pp. 216–225, 2010.

[51] R. Navigli and S.P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* **193**:217–250, 2012.

[52] R. Navigli and S.P. Ponzetto, Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1399–1410, 2012.

[53] R. Navigli and P. 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.

[54] H.T. Ng, B. Wang and Y.S. Chan, Exploiting parallel texts for Word Sense Disambiguation: An empirical study. In *Proc of ACL-03*, pp. 455–462, 2003.

[55] H.T. Ng, Getting serious about Word Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* pp. 1–7, Washington D.C, USA, 1997.

[56] M. Pennacchiotti and P. Pantel, Ontologizing semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney*, Australia, 17–21 July 2006, pp. 793–800, 2006.

[57] M.T. Pilehvar, D. Jurgens and R. Navigli, Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1341–1351, Sofia, Bulgaria 2013.

[58] H. Poon and P. Domingos, Unsupervised ontology induction from text. In *Proc of ACL*, pp. 296–305, 2010.

[59] H. Schütze, Automatic word sense discrimination, *Computational Linguistics* **24**(1):97–124, 1998.

[60] P. Smrz, Quality control for WordNet development. In *Proc of GWC-04*, pp. 206–212, 2004.

[61] R. Snow, D. Jurafsky and A. Ng, Semantic taxonomy induction from heterogeneous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17–21 July 2006, pp. 801–808, 2006.

[62] R. Snow, D. Jurafsky and A.Y. Ng, Learning syntactic patterns for automatic hypernym discovery. In Lawrence K. Saul, Yair Weiss and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pp. 1297–1304, Cambridge, Mass., 2005. MIT Press.

[63] R. Steinberger, B. Pouliquen, A. Widiger, C. Ig-nat, T. Erjavec, D. Tufiş and D. Varga, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc of LREC '06*, 2006.

[64] F.M. Suchanek, G. Kasneci and G. Weikum, Yago - A Core of Semantic Knowledge. In *Proc of WWW*, pp. 697–706, 2007.

[65] F.M. Suchanek, G. Kasneci and G. Weikum, Yago: A large ontology from Wikipedia and WordNet, *Journal of Web Semantics* **6**(3):203–217, 2008.

[66] S. Szumlanski and F. Gomez, Automatically acquiring a semantic network of related concepts. In *Proc of CIKM*, pp. 19–28, 2010.

[67] D. Tufiš, D. Cristea and S. Stamou, BalkaNet: Aims, methods, results and perspectives. a general overview, *Romanian Jour-*

nal on Science and Technology of Information **7**(1-2):9–43, 2004.

[68] J. Véronis, Hyperlex: lexical cartography for information retrieval, *Computer Speech and Language* **18**(3):223–252, 2004.

[69] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht, The Netherlands, 1998.

[70] F. Wu and D.S. Weld, Open Information Extraction Using Wikipedia. In *Proc of ACL*, pp. 118–127, 2010.

[71] J. Zhu, Z. Nie, X. Liu, B. Zhang and J. Wen, StatSnowball: A Statistical Approach to Extracting Entity Relationships. *Proc of WWW*, pp. 101–110, 2009.