Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia

Alessandro Raganato, Claudio Delli Bovi, and **Roberto Navigli**
Department of Computer Science
Sapienza University of Rome
{raganato,dellibovi,navigli}@di.uniroma1.it

Abstract

The hyperlink structure of Wikipedia constitutes a key resource for many Natural Language Processing tasks and applications, as it provides several million semantic annotations of entities in context. Yet only a small fraction of mentions across the entire Wikipedia corpus is linked. In this paper we present the automatic construction and evaluation of a Semantically Enriched Wikipedia (SEW) in which the overall number of linked mentions has been more than tripled solely by exploiting the structure of Wikipedia itself and the wide-coverage sense inventory of BabelNet. As a result we obtain a sense-annotated corpus with more than 200 million annotations of over 4 million different concepts and named entities. We then show that our corpus leads to competitive results on multiple tasks, such as Entity Linking and Word Similarity.

1 Introduction

One of the long-standing challenges of Artificial Intelligence is the automatic understanding of the meaning of text, i.e. Machine Reading [Etzioni *et al.*, 2006]. Over the last decade various lines of research have been geared towards achieving this goal, most notably Word Sense Disambiguation (WSD) [Navigli, 2009] and Entity Linking (EL) [Rao *et al.*, 2013]. In both tasks, semantically annotated corpora are indispensable in order to provide solid training and testing grounds for the development of disambiguation systems. Word-sense annotated corpora have been around for more than twenty years; however, even if named-entity annotated corpora have recently started to follow the same path, very few corpora to date comprise both kinds of annotations.

Indeed, encoding semantic information is a very demanding task, which can rarely be performed with high accuracy on a large scale. This is especially the case when such encoding requires both *lexicographic* (word senses) and *encyclopedic* knowledge (named entities) to be addressed [Schubert, 2006]. In this respect, semi-structured resources [Hovy *et al.*, 2013] stand as a convenient middle ground between high-quality, human-curated repositories and unstructured text; among others, Wikipedia constitutes an extraordinary source of semantic information for innumerable tasks in Natural Language Processing (NLP), from Named Entity Disambiguation [Cucerzan, 2007; Barrena *et al.*, 2015] to Semantic Similarity [Gabrilovich and Markovitch, 2007; Wu and Giles, 2015] and Information Extraction [Wu and Weld, 2010]. 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].

Unfortunately, only a fraction of linkable mentions in Wikipedia are in fact hyperlinked: out of over 580 million nouns across the whole corpus[1], those covered by hyperlinks (inter-page links) amount to just 116 million ($\sim$19%). Such link sparseness is partly due to Wikipedia style guidelines, which suggest linking each concept at most once within a page, and only when relevant and helpful in the context[2]. Being able to link appropriately every linkable Wikipedia mention would be a major step towards bridging this gap and turning Wikipedia into a full-fledged sense-annotated corpus. In the NLP community, the automatic identification and linking of referenced Wikipedia concepts and entities (*mentions*) across text is commonly referred to as *Wikification*. Resolving mention ambiguity, the key challenge of Wikification, has been addressed in various ways [Milne and Witten, 2008; Cheng and Roth, 2013]. Generally speaking, state-of-the-art WSD and EL systems with a Wikipedia-based sense inventory can be (and have been) used to this purpose [Scozzafava *et al.*, 2015]. However, although enriching Wikipedia can be seen as the special case of 'wikifying' Wikipedia articles, a system designed for general text does not take advantage of the existing Wikipedia structure at all.

In this paper our goal is to augment Wikipedia with as much semantic information as possible, by recovering potentially linkable mentions not covered by original hyperlinks. To achieve this, we rely only on the structure of Wikipedia itself, with no need for recourse to an off-the-shelf disambiguation system. We exploit direct connections among Wikipedia articles and categories in order to propagate hyperlink information across the corpus. We also leverage the wide-coverage semantic network of BabelNet [Navigli and Ponzetto, 2012] and its

[1] estimated from the Wikipedia dump of November 2014

[2] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style#Links

connections across Wikipedias in different languages, as well as across different lexicographic and encyclopedic resources. As a result, we obtain and make available to the community[3] a large sense-annotated corpus with more than 200 million annotations of over 4 million different concepts and named entities, covering almost 40% of the nouns in Wikipedia (compared to less than 20% covered by the original hyperlinks) and also including verbs, adjectives and adverbs. We evaluate annotation quality intrinsically (on a test set of hand-labeled hyperlinks) and extrinsically in two ways:

- using our sense annotations as a training set for EL with IMS [Zhong and Ng, 2010], an open-source supervised WSD system, and showing that it leads to performances in line with the state of the art on standard benchmarks;
- leveraging propagated hyperlinks to generate a simple, yet effective, *Wikipedia-based language-independent vector representation* that achieves competitive results on semantic similarity and sense clustering test sets.

Both experiments, in addition to confirming the quality of our annotations, also show that our corpus constitutes a key semantic resource, leading to important new performance baselines in Entity Linking and Semantic Similarity.

2 Related Work

Over the years, the WSD and EL communities have created a range of different sense-annotated datasets for a variety of evaluation tasks. A well-known example for WSD is the Senseval/SemEval competition series [Navigli *et al.*, 2013; Moro and Navigli, 2015], where manually annotated datasets are continuously released. Similarly, EL tasks are central in competitions such as TAC KBP, #Microposts and ERD.

The largest dataset manually annotated with word senses is SemCor [Miller *et al.*, 1993], a subset of the English Brown Corpus, with more than 200K content words tagged using the WordNet lexical database. Nevertheless, many instances of SemCor have very few annotations and only a small set of polysemous words is well covered. To bridge this gap, various automatic methods have been developed to generate training data on a larger scale, from unsupervised bootstrapping [Diab, 2004], to word alignments on parallel corpora [Zhong and Ng, 2009]. More recently Taghipour and Ng [2015] applied the latter approach to the MultiUN corpus and obtained one million training instances, which they released as the largest publicly available dataset for WSD. Being based on WordNet, however, their resource contains only lexicographic annotations.

As regards EL, Google has recently released two datasets containing web pages annotated with named entities: Wikilinks [Singh *et al.*, 2012], the result of a web crawl on roughly eleven million web pages that incorporate links to Wikipedia, and the Freebase annotation of the ClueWeb Corpora [Gabrilovich *et al.*, 2013], which comprises around 400 million web documents with six billion entity mentions automatically linked to Freebase. Despite their sizes, these corpora focus exclusively on named entities and ignore general concepts or non-nominal senses. They thereby constitute less rich and less structured resources compared to Wikipedia.

[3] http://lcl.uniroma1.it/sew

The recent development of joint approaches to WSD and EL, such as Babelfy [Moro *et al.*, 2014b], has enabled the automatic annotation of both word senses and named entities. Moro et al. [2014a] used Babelfy to annotate the MASC corpus, obtaining 286K annotations across 392 documents. However, this corpus is much smaller than the entire Wikipedia, and annotation quality is below 70%. Apart from exploiting off-the-shelf disambiguation systems, the specific task of detecting and annotating potentially linkable mentions in Wikipedia has been addressed in various ways, including gamification approaches [West *et al.*, 2015] and classifiers with Wikipedia-specific features [Noraset *et al.*, 2014]. Our approach is substantially different from both strategies. First, we do not rely on human intervention at all, nor do we utilize a trained and tuned learning system: our hyperlink propagation pipeline is fully automatic and based solely on the structure of Wikipedia. Second, we aim at covering as many mentions as possible, whereas Noraset et al. [2014] enforce a high-precision setting and West et al. [2015] focus only on hyperlinks that increase Wikipedia navigability.

Approaches similar to our hyperlink propagation methods proved to be effective in previous works for different purposes: a one-sense-per-page assumption is used in [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 semantic predicates [Flati and Navigli, 2013], and building taxonomies [Flati *et al.*, 2014]; finally, ingoing links to Wikipedia pages played a key role in the semantic representations of NASARI [Camacho-Collados *et al.*, 2015].

3 Building a Semantically Enriched Wikipedia

Our approach for building a Semantically Enriched Wikipedia (SEW) takes as input a Wikipedia dump and outputs a sense-annotated corpus, built upon the original Wikipedia text, where mentions are annotated according to the sense inventory of BabelNet [Navigli and Ponzetto, 2012]. BabelNet is a wide-coverage multilingual semantic network obtained from the automatic integration of various encyclopedias and dictionaries (WordNet and Wikipedia among others): being a superset of all these resources, BabelNet brings together lexicographic and encyclopedic knowledge and enables us to annotate both named entities and concepts (including verbs, adjectives and adverbs) using a common reference inventory. Furthermore, it provides convenient inter-resource mappings to directly convert from BabelNet annotations to, e.g., WordNet or Wikipedia annotations, and vice versa. Each item in the BabelNet inventory is represented as a *synset* (set of synonyms) and includes different surface forms (*lexicalizations*) of the same concept or entity.

Our pipeline applies some standard preprocessing in the first place, including tokenization, part-of-speech tagging and lemmatization. Disambiguation pages, '`List of`' articles and pages of common surnames are discarded, as they typically contain few lines of meaningful text and introduce noise into the propagation process. After preprocessing, we apply a

	Symbol	Heuristic 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 1: Summary of sense annotation types

cascade of *hyperlink propagation heuristics* to the corpus (Section 4). At each step a different heuristic is applied, enabling our algorithm to identify a list of synsets S^p to be propagated across a given Wikipedia page p; then, for each synset $s \in S^p$, occurrences of any lexicalization of s are detected and added as new annotations for p. All heuristics share a common assumption: given an ambiguous mention within a Wikipedia page, every occurrence of that mention refers to the same sense (*one sense per page*) and hence it is annotated with the same synset. Albeit simple, this assumption is surprisingly accurate[4] and increases coverage substantially.

As we apply a heuristic h to a given Wikipedia page p, we characterize h as being either *intra-page* (when it propagates synsets that occur as mentions within p itself) or *inter-page* (when it exploits the connections of p with other pages 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 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 and, if the mention is still ambiguous, we remove *all* its annotations. In other words, we do not attempt to annotate mentions that retain ambiguity even in the context of the same page (and connected pages). The set of annotation types is summarized in Table 1, while Section 4 describes each propagation heuristic in detail.

4 Propagation Heuristics

4.1 Intra-page Propagation Heuristics

Intra-page propagation heuristics collect a list of synsets S^p from the original hyperlinks across a Wikipedia page 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. Any mention discovered this way is then added to the list of sense annotations for p if part-of-speech tags are consistent. However, as potential mentions may contain punctuation or occur in some inflected form, propagation is performed as a two-pass procedure: a *surface mention propagation* (SP) over the original text of p before preprocessing, and a *lemmatized mention propagation* (LP) over tokenized and lemmatized text. Moreover, as people are not typically referred to by their full name inside the text of an article, we designed a specific heuristic to propagate *person mentions* (PP). 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 surrounding tokens against a precomputed set of first and last names, and added as annotation only if surrounding tokens do not match any person name. This allows us to avoid annotating false positives (e.g. siblings of s).

4.2 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 of Section 4.1. We exploited three inter-page heuristics:

Wikipedia Inlink Propagation (WIL) collects ingoing links to p inside Wikipedia (i.e. other Wikipedia pages where p is mentioned and hyperlinked) and adds the corresponding BabelNet synsets to S^p;

BabelNet Inlink Propagation (BIL), similarly to WIL, leverages ingoing links to the synset s_p that contains p in the BabelNet semantic network. These 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;

Category Propagation (CP) propagates hyperlinks across pages that belong to the same Wikipedia categories of p. Intuitively, pages belonging to the same categories tend to mention the same entities. Given a category c, we first harvest all hyperlinks appearing in all Wikipedia pages in c at least twice, and then we rank them by frequency count. In order to filter out categories that are too broad or uninformative (e.g. `Living people`) we associate with each category c a probability distribution over hyperlinks f^c, and compute the entropy $H(c)$ of such distribution as:

$$H(c) = -\sum_{h \in S^c} f^c(h)\, log_2\, f^c(h) \quad (1)$$

where h ranges over the set S^c of hyperlinks propagated through category c and $f^c(h)$ is computed as the normalized frequency count of h in S^c. Ranking categories by their entropy values allows us 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). Given a Wikipedia page p, we consider each category c_p of p where $H(c_p)$ is below a predefined threshold ρ_H[5], and add to S^p all the synsets that identify hyperlinks in S^{c_p}.

Finally, in order to cover non-nominal content words, we apply a *Monosemous Content Word* (MP) heuristic to propagate verb, adjective and adverb senses that are monosemous according to our sense inventory.

[4] 98% of the Wikipedia pages support this assumption according to the estimate of Wu and Giles [2015]

[5] we used $\rho_H = 0.5$ in our experiments (Section 6)

	# Annotations	# Senses	# Documents	Ann. Type
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
SEW-WordNet	116 079 163	67 774	4 313 373	WordNet
SEW-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
MUN	1 357 922	31 956	62 815	WordNet
MASC	286 416	23 175	392	BabelNet

Table 2: Comparison of different sense-annotated corpora. Wikipedia (*first row*) refers to the Nov 2014 dump.

	Nouns	Verbs	Adjectives	Adverbs
SEW (all)	201 885 731	6 381 452	25 102 343	16 955 731
SEW (conservative)	162 674 740	5 987 696	20 923 743	16 889 181
MUN	687 871	412 482	251 362	6 207
MASC	131 688	82 489	30 015	23 685

Table 3: Sense annotations by part of speech

5 Statistics

We built SEW by applying the approach described in Sections 3 and 4 to the English Wikipedia dump of November 2014. We relied on BabelNet 3.0[6] as sense inventory, and exploited the Stanford CoreNLP pipeline[7] for preprocessing. Table 2 reports some general statistics: the original dump constitutes by itself a corpus of 4,313,373 Wikipedia pages with 71,457,658 sense annotations, covering 2,898,503 distinct synsets. 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 entities not covered by the original hyperlinks. 17.5% ambiguous annotations are removed by our conservative policy, but the overall synset coverage remains almost unchanged. Table 2 also includes statistics on SEW with only Wikipedia annotations (fifth row) and only WordNet annotations (fourth row).

The bottom rows of Table 2 report comparative statistics on other sense-annotated corpora mentioned in Section 2: Wikilinks [Singh *et al.*, 2012], FACC1 [Gabrilovich *et al.*, 2013], the sense-annotated MultiUN corpus [Taghipour and Ng, 2015] and the sense-annotated MASC corpus [Moro *et al.*, 2014a]. Compared to Wikilinks, which provides more than 40M annotations from over 10M web pages, 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 the sense-annotated MultiUN (MUN) corpus, the WordNet portion of SEW adds over 114M annotations and 35818 covered senses.

Table 3 shows sense annotations by part of speech before and after applying the conservative policy (Section 4). Most annotations are nouns (80.65%), followed by adjectives (10.03%), adverbs (6.77%) and verbs (2.55%). Proportions

[6]http://babelnet.org

[7]http://stanfordnlp.github.io/CoreNLP

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

Table 4: Sense annotations by annotation type

	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 5: Coverage of content words by part of speech

are somewhat skewed with respect to other corpora, such as MultiUN (50.65% of noun annotations) and the MASC corpus (45.97%), since we include non-noun annotations only when monosemous in our sense inventory.

Table 4 shows sense annotations by heuristic type for both intra-page heuristics (above) and inter-page heuristics (below). Each heuristic is identified by the corresponding names in Table 1. Apart from original hyperlinks (which provide 28.55% of the annotations) and monosemous mentions (19.27%), the Surface Mention Propagation (SP) and the BabelNet Inlink Propagation (BIL) heuristics provide 13.49% and 13.07% of annotations respectively, followed by the Category Propagation (CP) heuristic with 10.25%. As expected, annotations discarded after applying our conservative policy were mostly derived from inter-page heuristics (WIL, BIL, CP) which open up to a broader context with respect to intra-page ones (and are therefore prone to noisier propagations).

Finally, Table 5 reports the coverage at the word level with respect to the original Wikipedia. Out of 977,203,946 content words in total, our approach annotates with senses 38.75% of the nouns, 6.71% of the verbs, 27.87% of the adjectives, and 33.95% of the adverbs. In comparison, original 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 on all parts of speech, improving more than 16% with respect to the original Wikipedia (13.3%) and extending coverage to non-nominal content words (verbs, adverbs, adjectives).

6 Experiments

We evaluated SEW by carrying out both an intrinsic (Section 6.1) and an extrinsic evaluation (Sections 6.2 and 6.3). In the former we compared our sense annotations against those discovered by 3W [Noraset *et al.*, 2014], a Wikipedia-specific system designed to add automatically high-precision hyperlinks to Wikipedia pages; in the latter we used SEW as a training set for Entity Linking (Section 6.2) and we exploited our propagated hyperlinks to develop Wikipedia-based language-independent vector representations for semantic similarity (Section 6.3). In both experiments of Sections 6.2 and 6.3 we compared against a baseline given by the original

	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 6: Results on the hand-labeled gold standard

	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**	-	-	0.821
DFKI	-	**0.889**	-	-
SUDOKU	-	0.870	-	-
Wikifier	-	-	**0.812**	0.724
M&W	-	-	0.685	**0.823**

Table 7: Results in terms of F-score on various WSD/EL datasets

Wikipedia.

6.1 Annotation Quality

We assessed the quality of our sense annotations on a hand-labeled evaluation set of 2,000 randomly selected Wikipedia pages, described in [Noraset *et al.*, 2014] and used for training, validating and testing 3W. We first ran our annotation pipeline (Sections 3-4) on it and then, following [Noraset *et al.*, 2014], we checked the 1530 solvable mentions against the gold standard by mapping our sense annotations from BabelNet synsets to Wikipedia pages. Results are reported in Table 6 and compared against 3W[8]: while obtaining a substantially higher recall, our approach 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 Wikipedia pages, 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 our approach annotates them with the corresponding correct WordNet adjectives (*American*^1_a, *German*^1_a). If we take these cases into account, our annotations achieve 96.5% precision and 64.4% F-score, showing that our propagation heuristics reach a precision level comparable to a trained and tuned high-precision linking system, while at the same time granting a much higher coverage, with an average of 31.3 new annotations per page (Section 5) against an estimate of 7 added by 3W [Noraset *et al.*, 2014].

We used the same gold standard to perform an ablation test on our propagation heuristics: for each heuristic h, we discarded annotations propagated by h and then repeated the experiment. Results (Table 6) 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).

6.2 Extrinsic Evaluation: Entity Linking

We evaluated SEW as a training set for EL using IMS [Zhong and Ng, 2010], a state-of-the-art supervised English all-words WSD system based on Support Vector Machines. We then tested IMS on four datasets: the English portion of the **SemEval-2013** task 12 dataset for multilingual WSD [Navigli *et al.*, 2013] and the English named entity portion of the **SemEval-2015** task 13 dataset for multilingual WSD and EL [Moro and Navigli, 2015], both with Wikipedia annotations; the **MSNBC** dataset [Cucerzan, 2007], with 756 mentions extracted from newswire text and linked to Wikipedia, and the test set of **AIDA-CoNLL** [Hoffart *et al.*, 2011]. Results are shown in Table 7 for all datasets in terms of F-score: IMS+SEW and IMS+HL represent IMS trained on SEW and IMS trained only on the original Wikipedia hyperlinks (HL), respectively. We include for each dataset a Most Frequent Sense (MFS) baseline provided by BabelNet, as well as results reported by other state-of-the-art EL systems in the literature: Babelfy [Moro *et al.*, 2014b] and the best performing system reported in [Navigli *et al.*, 2013] for SemEval-2013; the two best performing systems reported in [Moro and Navigli, 2015] for SemEval-2015; finally, Wikifier [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 trained on the original Wikipedia; this shows that our propagated hyperlinks lead to more accurate supervised models, adding semantic information that enables IMS to generalize better. 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 in line with more recent EL approaches, as well as systems specifically designed to exploit Wikipedia information. This suggests that, in general, our sense-annotated corpus has the potential to improve considerably the performance of Wikipedia-based EL systems.

6.3 Extrinsic Evaluation: Semantic Similarity

Another interesting test bed for SEW is provided by vector representations for semantic similarity. In fact, several successful approaches to semantic similarity make explicit use of Wikipedia, from ESA [Gabrilovich and Markovitch, 2007] to NASARI [Camacho-Collados *et al.*, 2015]. Others, like SENSEMBED [Iacobacci *et al.*, 2015], report state-of-the-art results when trained on an automatically disambiguated version of Wikipedia. We argue that SEW constitutes a preferable starting point as compared to the original Wikipedia, both in terms of increased hyperlink connections (in the former case) and in terms of increased sense-annotated mentions (in the latter case). To test this experimentally, we designed two sense-based vector representations built upon our corpus:

- A *Wikipage-based representation* (WB-SEW) where we represented each sense s in our sense inventory as a vector v_s where dimensions are Wikipedia pages. We computed, for each page p, the corresponding component of v_s as

[8] using the recommended setting with threshold at 0.934

		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
	ρ	0.69	**0.70**	0.56	0.57	0.59	0.61	0.49	0.51
SimLex-666	r	**0.38**	**0.38**	0.26	0.34	0.32	0.32	0.28	0.31
	ρ	0.40	**0.41**	0.33	0.36	0.31	0.32	0.27	0.27

Table 8: Results on the word similarity task in terms of Pearson (r) and Spearman (ρ) correlation to human judgement

the frequency of s appearing as annotation in p;

- A *synset-based representation* (SB-SEW) where we represented each Wikipedia page p as a vector v_p where dimensions are BabelNet synsets. We computed, for each synset s, the corresponding component of v_p as the frequency of s appearing as annotation in p.

We estimated frequencies using both raw counts (RC) and lexical specificity (LS), as in [Camacho-Collados *et al.*, 2015]. Then we tested our vectors on the two largest standard benchmarks available for word similarity: the similarity portion of WordSim-353 (**WS-Sim**) and the noun portion of the SimLex-999 dataset (**SimLex-666**). In both cases we relied on *weighted overlap* [Pilehvar *et al.*, 2013] as similarity measure. Following other sense-based approaches [Pilehvar *et al.*, 2013; Camacho-Collados *et al.*, 2015] we adopted a conventional strategy for word similarity that selects, for each word pair, the closest pair of candidate senses. Table 8 reports our performance in comparison with baseline vectors (WB-HL and SB-HL) computed using only the original Wikipedia hyperlinks. Our vector representations improve consistently over the baseline in both datasets. On WS-Sim, in particular, we obtain higher correlation figures than approaches like ADW [Pilehvar *et al.*, 2013] ($r = 0.63$ and $\rho = 0.67$) and ESA ($r = 0.40$ and $\rho = 0.47$), achieving performances in line with the state of the art. Moreover, since our vector representations are defined with respect to a multilingual sense inventory, we also tested our best performing model (WB-SEW) on a multilingual benchmark given by the **RG-65** dataset and its translations (Table 9), consistently beating the baseline and showing a considerable improvement on French, German and Spanish over **Word2Vec**, both the original model[9] and the model retrofitted into WordNet [Faruqui *et al.*, 2015] (**retro**), and pre-trained embedding models in the individual languages from the Polyglot project[10] (**Polyglot**).

Finally, we tested our vector representations on the Wikipedia sense clustering task described in [Dandala *et al.*, 2013], evaluating on both benchmark datasets (**500-pair** and **SemEval**). For each sense pair we thus computed similarity as in the previous experiment, and then checked it against empirically validated clustering thresholds of $t = 0.1$ (WB-SEW) and $t = 0.5$ (SB-SEW). Results reported in Table 10 are consistent with the experiment on word similarity (Table 8) and show that our vector representations improve consistently over their baseline counterparts, with F-scores close to (or slightly above) the state of the art reported by NASARI (72% on 500-pair and 64.2% on SemEval).

[9] we report results of pre-trained vectors over the Google News corpus (EN) and 1 billion tokens from Wikipedia (DE and FR)

[10] https://sites.google.com/site/rmyeid/projects/polyglot

		WB-SEW		WB-HL		Word2Vec		Polyglot
		RC	LS	RC	LS	original	retro	
EN	r	0.673	**0.674**	0.619	0.614	-	-	0.51
	ρ	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
	ρ	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
	ρ	0.689	**0.695**	0.637	0.615	0.53	0.60	0.15
ES	r	**0.811**	0.804	0.757	0.740	-	-	0.51
	ρ	**0.815**	0.812	0.764	0.759	-	-	0.56

Table 9: Pearson (r) and Spearman (ρ) correlation results for multilingual semantic similarity on the RG-65 dataset

	WB-SEW		SB-SEW		WB-HL		SB-HL	
	RC	LS	RC	LS	RC	LS	RC	LS
500-pair	0.668	0.668	**0.707**	0.674	0.671	0.654	0.233	0.186
SemEval	0.630	0.642	0.630	**0.645**	0.562	0.558	0.294	0.239

Table 10: F-score results on Wikipedia sense clustering

7 Conclusion and Future Work

We have presented the automatic construction and evaluation of SEW, a Semantically Enriched Wikipedia, where the overall number of linked mentions has been more than tripled by exploiting at best the hyperlink structure of Wikipedia and the wide-coverage sense inventory of BabelNet. Based on a cascade of hyperlink propagation modules (which need no training, validation or tuning) our approach generated a large sense-annotated corpus with over 200M annotations and 4M different concepts and entities, providing a coverage of almost 30% over all content words across Wikipedia (including verb, adjective and adverb senses). To the best of our knowledge, SEW is the largest available resource that comprises word senses and named entity mentions together, annotated using the same sense inventory. This makes it a suitable dataset for tasks such as Entity Linking and Word Similarity, that usually require dedicated training sets. We assessed the quality of our annotations with both intrinsic and extrinsic evaluations (Section 6) and we showed that our corpus sets important performance baselines for multiple tasks and datasets (even across languages, as shown in Section 6.3). SEW stands, therefore, as a key semantic resource not only in terms of size (i.e. amount of sense annotations and coverage), but also in terms of scope (i.e. lexicographic and encyclopedic knowledge from a multilingual wide-coverage inventory); we demonstrated its potential for markedly improving on the plethora of Wikipedia-based NLP systems currently being developed by the research community, thanks to its greatly increased number of hyperlinks, which, in turn, result in many more sense-annotated mentions across each Wikipedia page.

As future work we plan to further refine the quality of our sense annotations by imposing semantic coherence at the paragraph level, especially in larger and structured Wikipedia pages where the one-sense-per-page assumption is more likely to fail; at the same time, we are devising new strategies to increase coverage even further, perhaps exploiting a Wikification/EL system trained on our propagated hyperlinks. Fu-

ture perspectives include the extension of our approach to Wikipedias in other languages, moving towards the construction of a large, multilingual sense-annotated corpus.

References

[Barrena *et al.*, 2015] A. Barrena, A. Soroa, and E. Agirre. Combining mention context and hyperlinks from Wikipedia for named entity disambiguation. *SEM, 2015.

[Camacho-Collados *et al.*, 2015] J. Camacho-Collados, M. Pilehvar, and R. Navigli. NASARI: a novel approach to a semantically-aware representation of items. NAACL, 2015.

[Cheng and Roth, 2013] X. Cheng and D. Roth. Relational inference for Wikification. EMNLP, 2013.

[Cucerzan, 2007] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. EMNLP, 2007.

[Dandala *et al.*, 2013] B. Dandala, C. Hokamp, R. Mihalcea, and R. Bunescu. Sense clustering using Wikipedia. RANLP, 2013.

[Diab, 2004] M. Diab. Relieving the data acquisition bottleneck in word sense disambiguation. ACL, 2004.

[Etzioni *et al.*, 2006] O. Etzioni, M. Banko, and M. Cafarella. Machine Reading. AAAI, 2006.

[Faruqui *et al.*, 2015] M. Faruqui, J. Dodge, S. Jauhar, C. Dyer, E. Hovy, and N. Smith. Retrofitting word vectors to semantic lexicons. NAACL, 2015.

[Flati and Navigli, 2013] T. Flati and R. Navigli. SPred: Large-scale Harvesting of Semantic Predicates. ACL, 2013.

[Flati *et al.*, 2014] T. Flati, D. Vannella, T. Pasini, and R. Navigli. Two is bigger (and better) than one: the Wikipedia bitaxonomy project. ACL, 2014.

[Gabrilovich and Markovitch, 2007] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. IJCAI, 2007.

[Gabrilovich *et al.*, 2013] E. Gabrilovich, M. Ringgaard, and A. Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0), 2013.

[Hoffart *et al.*, 2011] J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. EMNLP, 2011.

[Hovy *et al.*, 2013] E. Hovy, R. Navigli, and S. Ponzetto. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *AIJ*, 194:2–27, 2013.

[Hu *et al.*, 2015] L. Hu, X. Wang, M. Zhang, J. Li, X. Li, C. Shao, J. Tang, and Y. Liu. Learning topic hierarchies for Wikipedia categories. ACL, 2015.

[Iacobacci *et al.*, 2015] I. Iacobacci, M. Pilehvar, and R. Navigli. SensEmbed: Learning sense embeddings for word and relational similarity. ACL, 2015.

[Miller *et al.*, 1993] G.A. Miller, C. Leacock, R. Tengi, and R. Bunker. A semantic concordance. HLT, 1993.

[Milne and Witten, 2008] D. Milne and I. Witten. Learning to link with Wikipedia. CIKM, 2008.

[Moro and Navigli, 2015] A. Moro and R. Navigli. SemEval-2015 task 13: multilingual all-words sense disambiguation and entity linking. SemEval, 2015.

[Moro *et al.*, 2014a] A. Moro, R. Navigli, F. Tucci, and R. Passonneau. Annotating the MASC Corpus with BabelNet. LREC, 2014.

[Moro *et al.*, 2014b] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *TACL*, 2:231–244, 2014.

[Nastase and Strube, 2008] V. Nastase and M. Strube. Decoding Wikipedia categories for knowledge acquisition. AAAI, 2008.

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

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

[Navigli *et al.*, 2013] R. Navigli, D. Jurgens, and D. Vannella. SemEval-2013 task 12: multilingual word sense disambiguation. SemEval, 2013.

[Navigli, 2009] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10, 2009.

[Noraset *et al.*, 2014] T. Noraset, C. Bhagavatula, and D. Downey. Adding high-precision links to Wikipedia. EMNLP, 2014.

[Pilehvar *et al.*, 2013] M. Pilehvar, D. Jurgens, and R. Navigli. Align, Disambiguate and Walk: A unified approach for measuring semantic similarity. ACL, 2013.

[Ponzetto and Strube, 2011] S. Ponzetto and M. Strube. Taxonomy induction based on a collaboratively built knowledge repository. *AIJ*, 175(9-10):1737–1756, 2011.

[Rao *et al.*, 2013] D. Rao, P. McNamee, and M. Dredze. Entity Linking: Finding extracted entities in a knowledge base. *Multi-Source, Multilingual Information Extraction and Summarization*, 11:93–115, 2013.

[Schubert, 2006] L. Schubert. Turing's dream and the knowledge challenge. AAAI, 2006.

[Scozzafava *et al.*, 2015] F. Scozzafava, A. Raganato, A. Moro, and R. Navigli. Automatic identification and disambiguation of concepts and named entities in the multilingual Wikipedia. AI*IA, 2015.

[Singh *et al.*, 2012] S. Singh, A. Subramanya, F. Pereira, and A. 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.

[Taghipour and Ng, 2015] K. Taghipour and H. Ng. One million sense-tagged instances for word sense disambiguation and induction. CoNLL, 2015.

[West *et al.*, 2015] R. West, A. Paranjape, and J. Leskovec. Mining missing hyperlinks from human navigation traces: A case study of Wikipedia. WWW, 2015.

[Wu and Giles, 2015] Z. Wu and C. Lee Giles. Sense-aware semantic analysis: A multi-prototype word representation model using Wikipedia. AAAI, 2015.

[Wu and Weld, 2010] F. Wu and D. Weld. Open information extraction using Wikipedia. ACL, 2010.

[Zhong and Ng, 2009] Z. Zhong and H. Ng. Word sense disambiguation for all words without hard labor. IJCAI, 2009.

[Zhong and Ng, 2010] Z. Zhong and H. Ng. It makes sense: a wide-coverage word sense disambiguation system for free text. ACL, 2010.