View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Archivio della ricerca- Università di Roma La Sapienza

Huge Automatically Extracted Training Sets for Multilingual Word Sense Disambiguation

Tommaso Pasini, Francesco Elia, Roberto Navigli
Sapienza University of Rome
{pasini, elia, navigli}@di.uniroma1.it

Abstract

We release to the community six large-scale sense-annotated datasets in multiple language to pave the way for supervised multilingual Word Sense Disambiguation. Our datasets cover all the nouns in the English WordNet and their translations in other languages for a total of millions of sense-tagged sentences . Experiments prove that these corpora can be effectively used as training sets for supervised WSD systems, surpassing the state of the art for low-resourced languages and providing competitive results for English, where manually annotated training sets are accessible. The data is available at `trainomatic.org`.

Keywords: Multilingual Word Sense Disambiguation, Resource, Dataset

1. Introduction

Word Sense Disambiguation is a crucial task in Natural Language Processing as it can be beneficial to several downstream applications, i.e., natural language understanding, semantic parsing and question answering. Despite the task has been around for a long time, it is far from being solved as it presents several challenges that have not fully been addressed yet, starting from the theoretical difficulty of formally establishing what a "word sense" is and choosing a corresponding sense inventory to the more pragmatic problems of finding large-scale sense-annotated corpora to train supervised systems on. Although WordNet (Fellbaum, 1998) virtually solved the first problem at least for English, a wide range of other issues still remain open. In fact, since supervised WSD systems need to be trained on a word-by-word basis, creating effective datasets requires a huge effort, which is beyond reach even for resource-rich languages like English. Clearly, this issue is even more severe for systems that need both lexicographic and encyclopedic knowledge (Schubert, 2006) and/or need to work in a multilingual or domain-specific setting. Knowledge-based WSD, on the other hand, exploits the knowledge contained in resources like WordNet to build algorithms (e.g. densest subgraph (Moro et al., 2014) or personalized page rank (Agirre and Soroa, 2009)) that can choose the sense of a word in context, thus not requiring training data but usually adopting bag-of-words approaches that neglect the lexical and syntactic context of the word (information that is more easily exploited by supervised systems), which may be essential in some scenarios. Furthermore, performances of both types of systems are highly affected by distribution of word senses that are usually different for each domain of application (Pasini and Navigli, 2018).

In order to address these issues different solutions have been proposed in the past years, ranging from manually annotated resources that can be used to train WSD systems to automatic or semi-automatic approaches that aim at exploiting parallel corpora or partially annotated data in order to produce training corpora. One of the first attempt to produce a sense annotated corpus is SemCor (Miller et al., 1993), a collection of thousand sentences manually tagged with WordNet senses. While its quality is very high thanks to the effort of specialized annotators, it is far from covering the whole English vocabulary of words and senses. Moreover, such manual resources need extra effort to be maintained and updated to integrate new senses and words appearing in everyday language. Thus, in order to overcome these issues, semi-automatic or fully automatic approaches have been proposed over the past years.

Taghipour and Ng (2015) exploit a parallel corpus and the manual translations of senses to annotate the words in the corpus with senses. Similarly, but without the need for human intervention, Delli Bovi et al. (2017) and Camacho-Collados et al. (2016), rely on aligned sentences in order to create a richer context that can be beneficial to their disambiguation. Raganato et al. (2016), instead, designed a set of heuristics which exploit the human effort of the Wikipedia community in order to propagate and add sense annotations to the Wikipedia pages. Similarly Pasini and Navigli (2017) exploit a knowledge base in order to annotate sentences with sense tags and uses a measure of confidence in order to select the most correct annotated sentences. They show that, relying on a multilingual semantic network as the underlying knowledge base, they are able to create high-quality sense-tagged corpora for any languages supported by the semantic network.

Our work builds upon that of Pasini and Navigli (2017) in order to generate sense-tagged corpora for 5 major European languages (English, French, German, Spanish and Italian) and the most spoken language of Asia (Chinese) and paves the way for supervised Word Sense Disambiguation in multiple languages. Exploiting the knowledge contained in BabelNet (Navigli and Ponzetto, 2010; Navigli and Ponzetto, 2012) – a huge and multilingual semantic network containing both lexicographic and encyclopedic knowledge – and Wikipedia, we generated large corpora annotated with BabelNet senses for the 6 languages listed above.

Experiments and statistics prove that these automatically created corpora are rich in terms of number of different lemmas annotated with a sense and number of sentences, and as such they can be a valuable resource for supervised WSD systems: in fact, systems trained on our datasets perform better or comparably to the state of the art across different languages. The added value is even more visible on low-resourced languages where such data is very scarce, if

at all available. We now give an overview of our corpus building procedure, including a brief description of Train-o-Matic; we then discuss features of the created datasets, our experimental setup for evaluation and its results.

2. Building the corpus

In order to build a sense annotated corpus for a given language L, our system takes as input a corpus of raw sentences C in the language L, a list of words W_L in the target language L and a semantic network G[1]. For each language L we apply (Pasini and Navigli, 2017)[Train-o-Matic] in order to annotate each target word $w \in W_L$ with a distribution over its senses.

For example given the ambiguous sentence *"A match is a lighter."* and the target word *"match"*, Train-o-Matic will output a sense distribution of the target word similar to the following:

$$[\text{match}_n^1 : 0.74, \text{match}_n^2 : 0.16, \text{match}_n^3 : 0.10]$$

where $word_{pos}^n$ follows the notation introduced in (Navigli, 2009) to indicate the *n-th* WordNet sense of $word$ with Part-of-Speech pos.

We chose Wikipedia in the language L as raw corpus C_L and BabelNet as the underlying semantic graph G because both are available for all the 6 languages of interest. BabelNet is also exploited in order to generate the lexicon W_L for each language L by collecting all the lexicaliztions of a synset in the graph in the given language L. Given the size of BabelNet we chose not to include all of its synsets, limiting our graph only to those that contain at least a sense from WordNet. We choose to keep all the BabelNet edges because they add many syntagmatic relations on top of the manually curated paradigmatic edges of WordNet.

To build each corpus we select all the sentences in each Wikipedia that contain at least one of the target words in W_L and then apply Train-o-Matic.

2.1. Train-o-Matic Overview

Train-o-Matic is a 3-step method to annotate a raw corpus of sentences.

1. Lexical Profiling Train-o-Matic exploits the semantic graph G in order to generate a lexical profile for each of the synsets in G. Such profile is computed by running the Personalized PageRank algorithm (Brin and Page, 1998) for each node in the graph. This means that, given the following formula:

$$v^{(t+1)} = (1-\alpha)v^{(0)} + \alpha M v^{(t)} \quad (1)$$

we set a 1 in the probability distribution v to the component that corresponds to the node for which we want to build the lexical profile. This procedure can also be interpreted as a random walk on the graph G where the walk is always restarted from the same initial node.

At the end of this step each synset s (i.e. node) in the graph has an associated vector in which each component represents another synset s' in the graph and the value of the component expresses the probability of reaching s' from s; this probability can be interpreted as a measure of relatedness between s and s'.

2. Sentence Scoring Once we have a distribution over the most related concepts for each synset in the graph, Train-o-Matic exploits them in order to annotate each target word in the raw corpus. For example, given the target word $w =$ "match", its set of senses retrieved from the semantic network $S_{\text{match}} = [match_n^1, match_n^2]$ and the sentence *"Messi didn't play the last match."* which contains the target word, the system creates a distribution over the senses in S_{match}.

This is done by approximating the probability of a sense given the target word and the sentence as follows:

$$P(s|\sigma, w) = \frac{P(\sigma|s,w)P(s|w)}{P(\sigma|w)} \quad (2)$$

$$\approx P(w_1|s,w)\dots P(w_n|s,w)P(s|w) \quad (3)$$

which assumes the independence of the words and removes the constant denominator. Each probability in (3) is computed exploiting the vectors previously computed. In fact, grounding the formula on our example, we have:

$$P(\text{match}_n^1|\text{Messi didn't play the last match, match}) = \quad (4)$$

$$P(\text{match}|\text{match}_n^1, \text{match})\times \quad (5)$$

$$P(\text{play}|\text{match}_n^1, \text{match})\times \quad (6)$$

$$P(\text{Messi}|\text{match}_n^1, \text{match}) \quad (7)$$

and each individual probability for the words w_i is computed by taking the value of the synset with the highest probability in the lexical profile of $match_n^1$ that contains the lemma w_i.

3. Sentence Ranking The last step aims at sorting and removing the sentences which are less likely to be correctly tagged. The sentences are in fact ranked by a confidence score which is computed by considering the difference between the most likely and second most likely senses of the target word. For example, referring to the previous example sentence, if match_n^1 received a probability of .7 and match_n^2 one of .3 then the sentence score will be .4. For each sense of a given word w, the candidate sentences are sorted using the confidence score. In order to select how many sentences to include in total, we set a parameter K that represents how many sentences must be included for the first sense of the given target word (i.e., the most common sense), with subsequent senses (according to the BabelNet ordering) for the same word receiving a decreasing number of examples computed according to a Zipf's distribution.

The following formula better explains the computation of the number of sentences assigned to each sense in a given ordering o.

$$\text{examples}_s = K \times \frac{1}{index(o,s)^z}$$

where $index(o, s)$ is a function that returns the position of a synset s in the ordering o. So, for example, if K is set to

[1]We consider a WordNet-like structure of the semantic network, where the nodes are synsets (concepts) which contain a set of lemmas that can express that concept.

	Total	English	French	German	Italian	Spanish	Chinese
Number of Annotations	17,987,488	12,722,530	1,597,230	1,213,634	1,037,253	935,713	481,128
Distinct lemmas covered	146,068	51,395	25,689	22,300	19,192	14,596	12,896
Distinct senses covered	63,613	56,229	33,843	23,526	22,587	21,388	12,485
Average # of sentences per sense	75.5	226.3	47.2	51.6	45.9	43.7	38.5
Average confidence score	56.74	71.64	22.07	89.19	19.40	50.41	87.75
Average Polisemy	1.71	1.56	1.78	1.66	1.80	1.74	1.76

Table 1: Statistics for each corpus in each language.

Corpus	Sentences	Annotations	Unique Words
SemCor	37,176	226,036	22,436
SemCor+OMSTI	850,974	1,137,170	22,437
Train-o-Matic	12,722,530	12,722,530	51,395

Table 2: Statistics of SemCor, OMSTI and Train-o-Matic about the number of sentences, annotations and unique words.

100, the first sense of the target word will receive K examples, the second one $\frac{K}{2^z}$ and so on; z is another parameter of the system.

3. Statistics

In this section we report some features of the corpora produced by Train-o-Matic, in order to give a complete overview of the data.

In Table 1 we show the number of annotations for each language as well as the number of distinct words and senses that have at least one example in our corpora and the number of sentences for each sense on average.

Train-o-Matic was able to generate around $18M$ annotated sentences for roughly $146K$ distinct lemmas and $63K$ distinct senses across languages. These corpora proved also to be of high quality, taking supervised system on par with or beyond state of the art results (Section 4.1.). The number of annotations is bigger for English and comparable across other languages: this is both because, for English, we set the value of the parameter K (see Section 2.1.) to 500 instead of 100, and because BabelNet, on average, contains more English senses compared to other languages.

As can be seen, each language has an average of 75 different sentences for each sense in the corpus, with English having the highest number of sentences per sense. Note that the total number of distinct senses covered is not equal to the sum of distinct senses for each sense due to the fact that we use a language-independent sense inventory (i.e. BabelNet) similarly to Otegi et al. (2016) and Delli Bovi et al. (2017). Thus many senses are shared across languages. The average confidence score measures how confident the system was on average when annotating the given language, meaning that the resulting data is most likely better: this score depends on both the average ambiguity of each lemma and on the quality of the relations in BabelNet. As expected, the system confidence score is highest in languages that have the lowest polisemy, i.e. English and German, which have the lowest average number of senses for nouns. As regards the average number of sentences for each sense, it directly depends on the parameter K and z that we set experimentally (see Section 2.1.). All corpora but English proved to lead supervised system to better performance when K was set to 100 and z between 2.0 and 3.0, thus we preferred to keep a lower number of more accurate sentences (50 for each sense). The English corpus, instead, was generated with K equal to 500 and z equal to 2.0 and thus it has a higher average number of sentences for each sense.

Table 2, instead, shows the comparison, in terms of number of sentences, annotations and unique words covered, between our automatically generated English corpus and two other corpora:

- SemCor (Miller et al., 1993), a corpus containing about 226,000 tokens annotated manually with WordNet senses.
- One Million Sense-Tagged Instances (Taghipour and Ng, 2015)[SemCor+OMSTI], a sense-annotated dataset obtained via a semi-automatic approach based on the disambiguation of a parallel corpus, i.e., the United Nations Parallel Corpus, performed by exploiting manually translated word senses. It also contains SemCor.

In terms of number of annotated sentences and number of annotations, our corpus is significantly bigger than SemCor and SemCor+OMSTI (by a factor of 200 and 10 respectively). More importantly, however, it covers double the amount of nouns that are covered by these two corpora, allowing supervised systems to have higher recall and to rely less on the Most Frequent Sense heuristic.

4. Experimental Setup

In order to evaluate the quality of the corpora we tested the performance of IMS, a state-of-the-art WSD system, when trained on our datasets.

English setup: For English, we compare the performance of IMS when trained on Train-o-Matic to that obtained against training with:

Test Set	Language	Train-o-Matic			Best System
		Precision	Recall	F1	F1
SemEval 2013	German	0.66	0.61	**0.63**	0.62
	French	0.61	0.60	**0.61**	**0.61**
	Spanish	0.68	0.66	0.67	**0.71**
	Italian	0.71	0.66	**0.68**	0.66
SemEval 2015	Spanish	61.3	54.8	**57.9**	56.3
	Italian	65.1	55.6	**59.9**	56.6

Table 3: Precision, Recall and F1 of IMS trained on Train-o-Matic, against the best performing system on SemEval-13 and SemEval-15.

Dataset	Train-o-Matic	OMSTI	SemCor	MFS
Senseval-2	70.5	74.1	**76.8**	72.1
Senseval-3	67.4	67.2	**73.8**	72.0
SemEval-07	59.8	62.3	**67.3**	65.4
SemEval-13	**65.5**	62.8	**65.5**	63.0
SemEval-15	**68.6**	63.1	66.1	66.3
ALL	67.3	66.4	**70.4**	67.6

Table 4: F1 of IMS trained on Train-o-Matic, OMSTI and SemCor, and MFS for the Senseval-2, Senseval-3, SemEval-07, SemEval-13 and SemEval-15 datasets.

The evaluation has been performed using the unified evaluation framework for Word Sense Disambiguation made available by Raganato et al. (2017), thus considering the following WSD shared tasks: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Navigli et al., 2007), SemEval-2013 (Navigli et al., 2013) and SemEval-2015 (Moro and Navigli, 2015). We set the two Train-o-Matic parameters K to 500 and z to 2.0 experimentally, testing the models learned by IMS on a small in-house development set[2] and choosing the one with the highest performance.

Multilingual setup: For the other languages we tuned the two paramenter K and z in the same way we did for English. The corpora proved to be more effective with K set to 100, for all the languages, and z ranging in $[2.0, 3.0]$. To prove that the generated data in the other languages are also high quality we also report the performance of IMS when trained on Train-o-Matic corpora for Italian and Spanish on the Multilingual WSD task of SemEval-2015 (Moro and Navigli, 2015), and for German, French, Spanish and Italian on the Multilingual WSD task of SemEval-2013 (Navigli et al., 2013) which focuses on nouns only. Given that no supervised system have been submitted to this task[3] we compare against the best performing knowledge-based systems of the two SemEvals.

4.1. Results

English results: As can be seen in Table 4 IMS trained on our corpus is always comparable, if not better (from 2 to 3 points), than OMSTI[4]. SemCor, instead, provides better training data for 3 out of 5 datasets, while the performance of IMS is comparable on the SemEval-2013 and SemEval-2015. This shows that our automatically generated data can lead to better performance than semi-automatic datasets and, in some situations, even surpass that of manually annotated ones. More interestingly, the ability to automatically generate high-quality sense-annotated data enables the creation of domain-specific datasets that could be used to train WSD systems on particular domains of interest. Given that such a system would most likely outperform a system trained on non-specialized data (e.g. because the latter may have learned a Most Frequent Sense bias that is not accurate for the domain at hand), this is often a need for companies which need to specialize their software on a specific use case (see (Pasini and Navigli, 2017) for experiments on domain specific tasks).

Multilingual results: Looking now at results in Table 3 it is clear that the best improvement in performance, compared to the current state of the art, is obtained on low-resourced languages, which was our main objective. We note that IMS, when trained on Train-o-Matic corpora, is able to score from 1 to 3 points more than the best system of each language and each of the two SemEval (i.e. SemEval 2013 and SemEval 2015) but Spanish in SemEval 2013.
This comes as expected as supervised systems perform better than knowledge-based ones (Raganato et al., 2017) when enough training data is available. Still, it is not the purpose of this paper to show that these datasets provide the best possible training sets in all scenarios, but rather that they can be very valuable in low-resourced languages, for which training supervised systems would be otherwise impossible.

5. Conclusion

We release to the community 6 sense-annotated corpora for the 5 major European languages (English, French, Spanish, German and Italian) plus Chinese, each containing on average more than 1 million sentences from Wikipedia articles and automatically annotated using Train-o-Matic.
Our experiments proved that these corpora provide effective training ground for supervised WSD system, especially in a multilingual setting where sense annotated data is scarce, if at all available. As a matter of fact, the performance of supervised systems trained on this data is better or comparable to those trained on semi-automatically and, in some cases, manually-curated data. Given the lack of

[2]The development set contains roughly 50 items per language.
[3]Note that no supervised system have ever been submitted for a multilingual WSD task.
[4]We recall that OMSTI has been built using a semi-automatic approach and contains SemCor

such data for languages other than English, most WSD systems that target these languages usually adopt a knowledge-based approach, thus neglecting syntactic and contextual information that may be essential in some scenarios. This point is confirmed by the fact that we are able to outperform such systems by using these corpora as training set. All these points show that our corpora are able to address the need for sense-annotate data in low-resources languages. The data is available at `trainomatic.org`.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487.

6. Bibliographical References

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009*, pages 33–41.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Camacho-Collados, J., Bovi, C. D., Raganato, A., and Navigli, R. (2016). A Large-Scale Multilingual Disambiguation of Glosses. In *Proceedings of LREC*, pages 1701–1708, Portoroz, Slovenia.

Delli Bovi, C., Camacho-Collados, J., Raganato, A., and Navigli, R. (2017). EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proc.of ACL*, volume 2, pages 594–600.

Edmonds, P. and Cotton, S. (2001). Senseval-2: overview. In *Proc. of SensEval 2*, pages 1–5. ACL.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Moro, A. and Navigli, R. (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proc. of SemEval-2015*.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transaction of ACL (TACL)*, 2:231–244.

Navigli, R. and Ponzetto, S. P. (2010). 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*, pages 216–225.

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

Navigli, R., Litkowski, K. C., and Hargraves, O. (2007). Semeval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007*, pages 30–35.

Navigli, R., Jurgens, D., and Vannella, D. (2013). Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM '13)*, volume 2, pages 222–231.

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

Otegi, A., Aranberri, N., Branco, A., Hajic, J., Neale, S., Osenova, P., Pereira, R., Popel, M., Silva, J., Simov, K., et al. (2016). Qtleap wsd/ned corpora: Semantic annotation of parallel corpora in six languages. In *Proceedings of the 10th Language Resources and Evaluation Conference, LREC*, pages 3023–3030.

Pasini, T. and Navigli, R. (2017). Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88. Association for Computational Linguistics.

Pasini, T. and Navigli, R. (2018). Two knowledge-based methods for high-performance sense distribution learning. In *AAAI-2018*.

Raganato, A., Delli Bovi, C., and Navigli, R. (2016). Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In *Proceedings of IJCAI*, pages 2894–2900, New York City, NY, USA, July.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proc. of EACL*, pages 99–110, Valencia, Spain.

Schubert, L. K. (2006). Turing's dream and the knowledge challenge. In *Proc. of AAAI-06*, pages 1534–1538.

Snyder, B. and Palmer, M. (2004). The english all-words task. In *Proc. of Senseval-3*, pages 41–43, Barcelona, Spain.

Taghipour, Kaveh and Ng, Hwee Tou. (2015). *One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction*. Association for Computational Linguistics.

7. Language Resource References

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