Sense-Annotated Corpora for Word Sense Disambiguation in Multiple Languages and Domains

Bianca Scarlini, Tommaso Pasini, Roberto Navigli
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{scarlini, pasini, navigli}@di.uniroma1.it

Abstract

The *knowledge acquisition bottleneck* problem dramatically hampers the creation of sense-annotated data for Word Sense Disambiguation (WSD). Sense-annotated data are scarce for English and almost absent for other languages. This limits the range of action of deep-learning approaches, which today are at the base of any NLP task and are hungry for data. We mitigate this issue and encourage further research in multilingual WSD by releasing to the NLP community five large datasets annotated with word senses in five different languages, namely, English, French, Italian, German and Spanish, and 5 distinct datasets in English, each for a different semantic domain. We show that supervised WSD models trained on our data attain higher performance than when trained on other automatically-created corpora. We release all our data containing more than 15 million annotated instances in 5 different languages at `http://trainomatic.org/onesec`.

Keywords: Word Sense Disambiguation, Multilinguality, Semantics

1. Introduction

Nowadays, deep neural models use data as their "fuel", and have proved to attain better performance the larger the training corpora they are provided with. Unfortunately, producing large manually-curated corpora is an expensive and time-consuming task. This is especially an issue for the field of Natural Language Processing (NLP), where data is usually tied to a specific language, thereby further increasing the time and cost of producing large-scale annotated corpora. An area of NLP that particularly suffers from such a problem is Word Sense Disambiguation (WSD), that is, the task of assigning a meaning to a word in a given context (Navigli, 2009). It is not by chance, in fact, that corpora annotated with senses are limited for English and almost absent for other languages. Indeed, while various sense-annotated corpora do exist for English, most of these are limited in terms of words and senses covered: MASC (Passonneau et al., 2012), a corpus providing 1M instances manually-annotated with senses, covers only 100 distinct words, OntoNotes (Hovy et al., 2006), instead, provides annotated examples for only 6500 concepts of less than 3500 words.

SemCor (Miller et al., 1993) is the largest manually-annotated English corpus currently available. It contains roughly 40K sentences and 200K content words annotated with their corresponding meanings. Despite being the standard training corpus for WSD, SemCor's main limitation resides in its low coverage of the English vocabulary in terms of both words and meanings. In fact, it provides annotations for only 22K unique lexemes, i.e., lemma and POS tag pairs, which cover not even 15% of the words comprised in WordNet (Miller et al., 1990), i.e., the largest and most-used electronic dictionary for English. Moreover, as we shift our focus towards lower-resourced languages, the need for semantically-annotated data becomes increasingly urgent. Indeed, SemCor, with some exceptions (Bentivogli and Pianta, 2005; Bond et al., 2012), lacks an adequate multilingual counterpart in most world languages and, hence, WSD models are limited when it comes to scaling over languages other than English. In this scenario, several automatic approaches for producing multilingual sense-annotated data (Pasini and Navigli, 2017; Pasini et al., 2018; Scarlini et al., 2019; Pasini and Navigli, 2020) have tried to mitigate the aforementioned shortcomings. In fact, being able to create silver annotated data on a large scale is crucial to freeing WSD models from dependence on exclusively those words and languages that are comprised within manually-curated resources.

In this paper we follow Scarlini et al. (2019) and apply their approach to all the nominal lexemes of 5 major European languages, i.e., English, Italian, Spanish, French and German, so as to ensure full coverage in terms of both words and senses. We release to the community training corpora in 5 different languages, including more than 15 million annotations. Our automatically-produced data lead a supervised model to state-of-the-art results in all multilingual WSD tasks, while remaining competitive with manually-curated resources on English. Furthermore, we exploited automatic approaches for inducing the distribution of word senses (Pasini and Navigli, 2018) to build 5 additional datasets for English, each peculiar to a different semantic domain. These domain-specific datasets proved to lead a supervised model to a significant gain in performance when it comes to in-domain WSD evaluation.

2. Related Work

Tackling the Word Sense Disambiguation problem from a supervised point of view has been the focus of several recent works (Luo et al., 2018; Kumar et al., 2019; Bevilacqua and Navigli, 2019). While on the one hand these models proved to attain state-of-the-art results on the English WSD task, on the other hand they still depended heavily on sense-annotated training corpora, and hence were limited to a restricted set of words, senses and languages. In fact, the line of research focused on new methodologies for automatically creating high-quality sense-annotated datasets is long-standing (Taghipour and Ng, 2015; Raganato et al., 2016; Delli Bovi et al., 2017; Pasini and Navigli, 2017; Pasini et al., 2018; Scarlini et al., 2019; Pasini and Navigli, 2020). These latter five approaches enabled supervised systems to step outside the boundaries of standard English WSD tasks and allowed them to scale on languages where manually-curated resources are not available. Train-O-Matic (Pasini and Navigli, 2017) is a language-independent method for the creation of large-scale sense-annotated corpora. By exploiting the information enclosed within a semantic network it is able to provide high-quality annotations of raw sentences, leading a supervised model to competitive results on both English and multilingual WSD tasks. More recently, Scarlini et al. (2019) dropped the requirement to draw on the structure of a semantic network by exploiting a sparse vector representation of senses and Wikipedia's inner organization into pages and categories. Their automatically-created corpora proved to outperform all other automatic alternatives on English and achieve state-of-the-art results on all multilingual WSD datasets.

In this paper we complement the work by Scarlini et al. (2019, OneSeC) and apply their approach to all the words in the vocabulary of 5 different European languages. Moreover, we couple their approach with an automatic method for inducing word-sense distribution and create 5 different domain-specific datasets for English in order to enable domain-specific WSD on a large scale.

3. Preliminaries

OneSeC relies on three resources for producing sense-tagged corpora, namely, Wikipedia[1], the largest freely available electronic encyclopedia, BabelNet[2] (Navigli and Ponzetto, 2012), a multilingual semantic network, and NASARI[3] (Camacho-Collados et al., 2016), a sparse representation of BabelNet concepts.

Wikipedia Wikipedia is the largest encyclopedic corpus currently available. It is made up of approximately 300 separate editions, each comprising texts in a specific language. Wikipedia articles describe either abstract concepts or named entities, and are, in their turn, grouped together into Wikipedia categories. These categories collect a set of articles that share a common feature or characteristic, and organise the information in Wikipedia into macro areas. For example, the category Modernist writers contains, among others, the pages *Ernest Hemingway* and *F. Scott Fitzgerald*. For ease of reading, we define a *lemma occurring in a category* as the lemma that appears in a sentence of any page within that category and *the sentences of a category* as the sentences of all the pages belonging to the reference category.

BabelNet BabelNet is a multilingual semantic network built by merging together several heterogeneous resources, such as WordNet, Wikipedia, Wikidata etc. It is organised in the form of a graph, where nodes are concepts and edges are semantic relations between them. Each concept combines lexicalizations in different languages. For example, the *season* concept of spring#n[4] also includes the terms *printemps* (French), *Frühling* (German) and *primavera* (Italian).

NASARI NASARI vectors are sparse-vector representations of BabelNet concepts. Each dimension corresponds to a word that it is associated with a lexical specificity value (Lafon, 1980), which expresses how much the target word is relevant for the concept it represents. For example, the terms *calendar*, *year* and *season* are included in the NASARI vector corresponding to the *season* sense of spring#n. NASARI proved to be effective in different tasks, such as text classification (Sinoara et al., 2019) and Word Sense Disambiguation (Camacho-Collados et al., 2016). We note that the NASARI lexical vectors are available for nominal concepts only.

4. OneSeC

We now move on to describe OneSeC, which, by relying on the previously introduced resources, automatically produces sense-annotated data in multiple languages. It extends the "One Sense per Discourse" (Gale et al., 1992) assumption to "One Sense per Wikipedia Category", i.e., all the occurrences of a word in the same category share a common meaning, and leverages the information in a Wikipedia category to automatically tag the occurrences of the words therein.

Given a target lexeme l, OneSeC produces annotated examples for each of its senses by applying the following three steps.

Category Representation. First, we collect all the Wikipedia categories $C_1 \dots C_n$ such that l appears at least t times across the sentences of their pages. Then, for each category C of the lemma l, we create a Bag of Words representation B^l_C, i.e., a sparse vector in which dimensions are words scored by their frequency in the sentences of C where l appears.

Sense Assignment. We assign a sense to each lexeme-category pair (l, C) by leveraging the NASARI lexical vectors (see Section 3.). That is, as first step we compute the similarity between B^l_C and each NASARI vector v_s associated with a sense s of l by means of the Weighted Overlap (Pilehvar et al., 2013)[WO] measure[5]. The WO of two vectors v_1 and v_2 is computed as follows:

$$WO(v_1, v_2) = ln(|I|+1) \left(\sum_{w \in I} \frac{1}{r^{v_1}_w + r^{v_2}_w} \right) \left(\sum_{i=1}^{|I|} \frac{1}{2i} \right)^{-1}$$

where $r^{v_i}_w$ is the rank of the word w in the vector v_i and I is the set of intersecting dimensions of v_1 and v_2.

Then, we assign to (l, C) the sense that maximizes the WO similarity with B^l_C. Formally, being $v_{s_1} \dots v_{s_n}$ the NASARI vectors of the senses of l and

[1] https://www.wikipedia.org/

[2] https://babelnet.org/

[3] http://nlp.uniroma1.it/nasari/

[4] We use the notation lemma#pos.

[5] We preferred the Weighted Overlap over the most common Cosine Similarity as it has been proved to be more effective in capturing the distance between sparse vectors (Pilehvar et al., 2013).

Corpus	Sentences	Annotations	Distinct Nouns	Distinct Senses
SemCor	18,531	87,002	11,391	15,906
SemCor+OMSTI	455,762	558,383	11,396	16,194
Train-O-Matic	12,722,530	12,722,530	51,395	56,229
OneSeC	8,813,642	8,813,642	28,383	39,848

Table 1: Statistics on the English corpora restricted to the nominal words of SemCor, SemCor+OMSTI, Train-O-Matic and OneSeC in terms of number of sentences, annotations, unique nouns and unique nominal senses.

$WO(B^l_C, v_{s_1}) \ldots WO(B^l_C, v_{s_n})$ the list of WO similarities, we compute the sense s to be associated with (l, C) as follows:

$$sense(l, C) = \arg\max_{s_i}(WO(B^l_C, s_i))$$

At the end of this step, each pair (l, C) is mapped to the potentially most suitable sense s of l.

Sentence Sampling. Given the mapping between senses and categories we computed in the previous step, we now select the set of training examples for each sense s_i of the target lexeme l. To this end, we first set the number of sentences to draw for s_i by applying a Zipfian distribution as follows:

$$\mathcal{K}_{s_i} = \mathcal{K}(i^z)^{-1} \quad (1)$$

where i is the rank of s_i according to BabelNet's ordering of senses, $\mathcal{K}$ is a parameter of the system and determines the number of examples for the first sense of l and z is the value of the exponent that controls how fast the distribution decreases. Then, for each sense s_i of l we take all the categories in which l appears that are associated with s_i and sample a number of sentences for each category that is proportional to its weighted overlap similarity with respect to s_i. Formally, given the list of categories $C_1 \ldots C_m$ associated with s_i and ordered according to their weighted overlap score, we define the number of sentences to draw from C_j as follows:

$$\mathcal{K}^{C_j}_{s_i} = \mathcal{K}_{s_i} \frac{1}{j} \left(\sum_{j'=1}^{m} \frac{1}{j'} \right)^{-1}$$

Finally, for each category C_j we perform a weighted sampling in which the probability of a sentence being selected is determined by its perplexity[6], i.e., the lower the perplexity the higher the probability of a sentence being sampled. At the end of this procedure every occurrence of l in the selected sentences is tagged with the sense s_i.

5. Creating the Corpora

We apply OneSeC's procedure to all the nominal words in the vocabulary of five major European languages, i.e., English, Italian, Spanish, French and German. More in detail, as regards English, we consider the whole set of nouns of WordNet. As for the other languages, we take into account the set of nouns in the WordNet part of BabelNet, i.e., we take all the synsets in BabelNet that contain a WordNet sense and collect all the nouns in the target language therein.

Moreover, we slightly modify OneSeC to create domain-specific datasets for all English nominal lexemes by exploiting DaD (Pasini et al., 2018), i.e., a knowledge-based method for inducing in-domain sense distributions. DaD computes the distribution over BabelNet synsets by first approximating the probability of a BabelDomain (Camacho-Collados and Navigli, 2017) being a topic covered in the corpus of raw sentences, and then leverages the connections in BabelNet to propagate the domains' probabilities over all BabelNet synsets. Hence, in order to select the number of sentences to assign to each sense of a lemma l (see Equation 1), for each domain d we rank the senses of l according to DaD's sense distributions associated with d, instead of following BabelNet's ordering of senses. We create sense-annotated corpora according to the domain-aware sense distributions for 5 different domains, i.e., Biology, Health Care, Politics, Social Issues and Maths & PC.

We note that for building OneSeC's corpora we have to set 3 different system parameters: i) t, i.e., the minimum number of occurrences of a lemma in a Wikipedia category, ii) $\mathcal{K}$, i.e., the maximum number of sentences to be assigned with the first sense of a lemma, iii) z, i.e., the exponent that controls the Zipf's distribution (see Section 4.). For both English and multilingual datasets, we set $t = 20$. In the case that we find no categories for a lemma with $t = 20$, we relax the constraint and set $t = 10$. As for the other parameters, we follow Scarlini et al. (2019) and set $\mathcal{K} = 700$ and $z = 2.1$ for English and $\mathcal{K} = 200$ and $z = 2.0$ for the other languages.

6. Statistics

In this Section we show the statistics of our corpora in order to give a general overview of them. In Table 2 we report the characteristics of each corpus divided per language. As can be seen, the English corpus is the richest. The average polysemy of OneSeC on English, in fact, is higher than the one of WordNet[7], meaning that the lemmas covered by OneSeC are among those with the highest polysemy and most of the lemmas that are not covered are, instead, monosemous. The average number of sentences per sense is higher for English than for the other languages inasmuch as it directly depends on the parameter K (Section 5.) The French corpus is the second richest one in terms of number of annotations and lemmas and senses covered. In to-

[6] The perplexity is computed by means of the Neural Language Model presented in Howard and Ruder (2018).

[7] 1.24 according to the statistics at https://wordnet.princeton.edu/documentation/wnstats7wn.

Language	Annotations	Distinct Nouns	Distinct Senses	Avg Sentences Per Sense	Avg Polysemy
English	8,813,642	28,383	39,848	221.18	1.40
Italian	1,384,771	12,155	10,795	128.27	1.14
Spanish	1,259,779	12,698	12,429	101.35	1.19
French	2,141,391	15,976	15,454	138.56	1.10
German	1,724,003	15,646	14,171	121.65	1.04
Aggregated	15,323,586	84,734	40,043	142.20	1.17

Table 2: Statistics breakdown for each language of the OneSeC's corpora.

Domain	Annotations	Distinct Nouns	Distinct Senses	Avg Sentences Per Sense	Avg Polisemy
Biology	8,807,313	28,383	39,754	220.86	1.40
Health Care	8,803,621	28,383	39,714	219.86	1.40
Politics	8,811,269	28,383	39,761	220.75	1.40
Social Issue	8,817,313	28,383	39,793	220.55	1.40
Math	8,801,351	28,383	39,706	217.13	1.40
Aggregated	44,040,867	28,383	39,813	219.83	1.40

Table 3: Statistics breakdown for each domain-specific dataset on English.

Dataset	OneSeC	TOM	OMSTI	SemCor	MFS
Senseval-2	73.2	70.5	74.1	**76.8**	72.1
Senseval-3	68.2	67.4	67.2	**73.8**	72.0
SemEval-07	63.5	59.8	62.3	**67.3**	65.4
SemEval-13	**66.5**	65.5	62.8	65.5	63.0
SemEval-15	**70.8**	68.6	63.1	66.1	66.3
ALL	69.0	67.3†	66.4†	**70.4**	67.6

Table 4: Results of IMS trained on different corpora on the English WSD tasks. † marks statistical significance between OneSeC and its competitors.

tal, our corpora all together cover 40,043 distinct meanings with more than 15M annotations.

As can be seen in Table 1, when compared to other manually- and semi-automatically annotated corpora, i.e., SemCor (Miller et al., 1993) and SemCor+OMSTI (Taghipour and Ng, 2015), OneSeC covers more than double their lemmas and senses with one order of magnitude more annotations. Compared to SemCor+OMSTI, in fact, OneSeC covers almost three times more nouns and two times more senses. Furthermore, OMSTI covers roughly the same number of nouns as SemCor with a slightly higher number of meanings. Train-O-Matic, instead, is the training corpus with the largest coverage in terms of both lemmas and senses. One reason why OneSeC falls behind Train-O-Matic in terms of coverage is that it filters out all the categories that contain less than 10 sentences in which the target word occurs (see Section 4. and 5.). Therefore, it may happen that a word-sense pair has no category associated with it and hence no annotations can be provided. Even though OneSeC covers less nouns, it is more accurate, as shown in our experiments. OneSeC, in fact, leads a supervised WSD model to attain higher results in comparison to the same WSD model trained on Train-O-Matic data.

In Table 3 we break down the statistics for each corpus we created for 5 distinct semantic domains on English. As can be seen, OneSeC can find as many examples in each domain as in the general-domain corpus (Table 2), meaning that it can also cover senses that are, instead, peculiar to a specific domain.

7. Experimental Setup

For assessing the quality of OneSeC annotations we trained a supervised WSD model on our automatically-generated data and tested its performance on 5 standard WSD benchmarks for English and for another four languages, i.e., Italian, Spanish, French and German.

Reference WSD models As for English, we used It Makes Sense (Zhong and Ng, 2010, IMS), a support vector machine-system which builds a single model for each target word in the training set. For the other languages, instead, we employed the BiLSTM-based model proposed by Raganato et al. (2017b).

Test Data We tested on all the nominal instances comprised in the 5 standard English datasets included in the framework of Raganato et al. (2017a), namely, Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-07 (Pradhan et al., 2007), SemEval-13 (Navigli et al., 2013) and SemEval-15 (Moro and Navigli, 2015). Furthermore, we report the results on the ALL dataset, i.e., the concatenation of all the aforementioned test sets. For evaluating the corpora in the other four languages, instead, we tested the reference WSD model (BiLSTM) on the multilingual datasets of SemEval-2013 task 12 (Navigli et al., 2013) (Italian, Spanish, French and German) and SemEval-2015 task 13 (Moro and Navigli, 2015) (Italian and Spanish).

Domain-Specific Evaluation To evaluate the domain-specific corpora produced by OneSeC we generated a specific training corpus (OneSeC$_{dom}$) for each of the following five domains: Biology, Health Care, Politics, Social Issues and Maths & PC, as explained at the end of Section 5. Then, we trained IMS on each training set separately, and tested it on the documents of SemEval-13 and SemEval-15 corresponding to each of the aforementioned domains.

Dataset	Domain	Size	Backoff	OneSeC$_{dom}$			OneSeC			TOM			MFS
				P	R	F1	P	R	F1	P	R	F1	
SemEval-13	Biology	135	MFS	80.7	80.7	**80.7**	67.4	67.4	67.4	63.0	63.0	63.0	64.4
			-	79.4	74.1	**76.6**	65.1	60.7	62.8	59.0	53.3	56.0	
	Health Care	138	MFS	67.4	67.4	**67.4**	65.9	65.9	65.9	65.2	65.2	65.2	56.5
			-	64	50.8	**60.8**	62.4	56.5	59.3	61.3	55.1	58.0	
	Politics	279	MFS	73.5	73.5	**73.5**	68.8	68.8	68.8	65.2	65.2	65.2	67.7
			-	72.0	68.1	**70.0**	67.0	63.4	65.2	62.5	54.8	58.4	
	Social Issues	349	MFS	73.6	73.6	**73.6**	66.5	66.5	66.5	68.5	68.5	68.5	67.6
			-	70.7	62.2	**66.2**	62.5	55.0	58.5	63.1	53.0	57.6	
SemEval-15	Maths & Pc	97	MFS	63.0	63.0	**63.0**	60.0	60.0	60.0	50.0	50.0	50.0	40.9
			-	62.9	61.0	**61.9**	59.8	58.0	58.9	50.0	47.0	48.5	

Table 5: Domain-specific evaluation on SemEval-2013 and SemEval-2015 of IMS trained on OneSeC$_{dom}$, OneSeC and TOM.

Lang	OneSeC			TOM - Bi-LSTM			Best
	P	R	F1	P	R	F1	F1
IT	72.3	64.5	**68.2**	65.4	60.6	62.9	68.0 ◇
ES	76.0	68.3	**72.0**	71.7	66.8	69.2	71.0 ∗
FR	79.2	70.9	**74.8**	71.0	64.2	67.4	61.0 ◇∗
DE	83.0	68.5	**75.1**	77.5	64.1	70.2	63.0 ◇

Table 6: Comparison of Bi-LSTM trained on OneSeC and TOM with the best system (Best) on SemEval-2013. ◇ Train-O-Matic, ∗ UMCC-DLSI.

Lang	OneSeC			TOM - Bi-LSTM			Best
	P	R	F1	P	R	F1	F1
IT	65.0	60.2	**62.5**	61.6	58.3	59.9	59.9 ◇
ES	67.8	58.4	**62.8**	62.5	56.1	59.2	57.9 ◇

Table 7: Comparison of Bi-LSTM trained on OneSeC and TOM with the best system (Best) on SemEval-2015. ◇ Train-O-Matic.

Competitors We compared the results attained by the reference WSD models trained on OneSeC and other three training corpora for English:

- **Train-O-Matic** (TOM), a knowledge-based approach for producing training corpora for WSD in English and all the languages supported by BabelNet.
- **OMSTI**, a semi-automatic approach for generating sense-annotated data that leverages parallel corpora and manual annotations.
- **SemCor**, a manually-annotated corpus which is the de-facto standard training set for English Word Sense Disambiguation models.

As for the multilingual setting, we compared OneSeC with the best performing model in each of the tested datasets and with the reference multilingual WSD model (Bi-LSTM) when trained on Train-O-Matic corpora.

8. Results

As a first result in Table 4 we report the performance of IMS on the 5 general-domain English WSD benchmarks. As can be seen, OneSeC corpora lead IMS to attain the highest results across the board when compared to its automatic and semi-automatic competitors, i.e., Train-O-Matic and OMSTI. OneSeC, in fact, ranks second only in comparison to SemCor, which, however, is a manually-curated resource. It performs only 1.4 F1 points lower than SemCor on ALL, a loss that is due to the automatic nature of OneSeC and to the unavoidable noise that can be found in silver data. Nevertheless, we note that OneSeC outperforms SemCor in 2 out of 5 of the tested datasets, i.e., SemEval-13 and SemEval-15, with an increment of 1 and 4.7 F1 points, respectively.

In Table 5 we report the results on the domains of SemEval-13 and SemEval-15 attained by IMS trained on the 5 OneSeC$_{dom}$ corpora and compare them with the results achieved when training IMS on the general-domain training corpora of OneSeC and Train-O-Matic (TOM). OneSeC$_{dom}$ leads IMS to attain the best results across each domain, with the highest boost of more than 13 points compared to its general version on the Biology domain, and of more than 17 points compared to Train-O-Matic. This shows the effectiveness of OneSeC in providing examples for senses that are specific to a given domain.

As regards the multilingual evaluation, in Tables 6 and 7 we report the results attained by the BiLSTM-based model on each specific language, separately. To set a level playing field with Train-O-Matic, which reported the results attained by IMS on the multilingual evaluation, we trained the same BiLSTM-based model on their data as well. OneSeC attains, also in this case, the highest results across the board, repeatedly beating both Train-O-Matic and the best performing system (UMCC-DLSI (Gutiérrez et al., 2010)) on each language by several points.

Overall, our corpora proved to be of high quality in different settings and languages. In fact, OneSeC places itself as the best alternative when it comes to train a model on a specific domain or on lower-resourced languages, while remaining competitive with manually-curated resources on English.

9. Conclusion

In this paper we propose a total of 10 different datasets automatically annotated with sense labels for 5 different languages and 5 distinct semantic domains. Our experiments proved that our datasets are of high quality and can be used as training data by supervised models to perform Word Sense Disambiguation in each of the 5 different languages. Furthermore, when it comes to in-domain WSD, our domain-specific datasets proved to be effective in all the tested scenarios, aiding WSD approaches to perform better in each specific domain.

As future work, we plan to refine our approach by exploiting other knowledge resources, such as SensEmBERT (Scarlini et al., 2020), VerbAtlas (Di Fabio et al., 2019) or SyntagNet (Maru et al., 2019).

We release all the data at `http://trainomatic.org/onesec` comprising more than 15 million annotations across 5 different languages. We plan to include more languages and domains in the future.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

10. Bibliographical References

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Bevilacqua, M. and Navigli, R. (2019). Quasi Bidirectional Encoder Representations from Transformers for Word Sense Disambiguation. In *Proceedings of RANLP*, pages 122–131.

Bond, F., Baldwin, T., Fothergill, R., and Uchimoto, K. (2012). Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th global WordNet conference (GWC 2012)*, pages 56–63. Citeseer.

Camacho-Collados, J. and Navigli, R. (2017). BabelDomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 223–228.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2016). Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Delli Bovi, C., Camacho-Collados, J., Raganato, A., and Navigli, R. (2017). Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 594–600.

Di Fabio, A., Conia, S., and Navigli, R. (2019). VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling. In *Proceedings of EMNLP-IJCNLP*, pages 627–637.

Edmonds, P. and Cotton, S. (2001). Senseval-2: overview. In *Proceedings of SENSEVAL*, pages 1–5. Association for Computational Linguistics.

Gale, W. A., Church, K., and Yarowsky, D. (1992). One sense per discourse. In *Proceedings of DARPA Speech and natural language Workshop*, pages 233–237.

Gutiérrez, Y., Fernàndez, A., Montoyo, A., and Vázquez, S. (2010). Umcc-dlsi: Integrative resource for disambiguation task. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 427–432. Association for Computational Linguistics.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). Ontonotes: The 90% solution. In *Companion Volume to the Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 57–60.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339.

Kumar, S., Jat, S., Saxena, K., and Talukdar, P. (2019). Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of ACL*, pages 5670–5681.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots.*, 1(1):127–165.

Luo, F., Liu, T., He, Z., Xia, Q., Sui, Z., and Chang, B. (2018). Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of EMNLP*, pages 1402–1411, Brussels, Belgium.

Maru, M., Scozzafava, F., Martelli, F., and Navigli, R. (2019). SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In *Proceedings of EMNLP-IJCNLP*, pages 3525–3531.

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

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

Moro, A. and Navigli, R. (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of SemEval-2015*, pages 288–297.

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., Jurgens, D., and Vannella, D. (2013). Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of Semeval 2013*, volume 2, pages 222–231.

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

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.

Pasini, T. and Navigli, R. (2018). Two knowledge-based methods for high-performance sense distribution learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5374–5381.

Pasini, T. and Navigli, R. (2020). Train-O-Matic: Supervised Word Sense Disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103–215.

Pasini, T., Elia, F. M., and Navigli, R. (2018). Huge Automatically Extracted Training-Sets for Multilingual Word Sense Disambiguation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation.*, pages 1694 – 1698.

Passonneau, R. J., Baker, C., Fellbaum, C., and Ide, N. (2012). The masc word sense sentence corpus. In *Proceedings of LREC*, pages 3026–3030.

Pilehvar, M. T., Jurgens, D., and Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of ACL*, pages 1341–1351.

Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M. (2007). Semeval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of Semeval-2007*, pages 87–92.

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.

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

Raganato, A., Delli Bovi, C., and Navigli, R. (2017b). Neural sequence learning models for word sense disambiguation. In *Proceedings of EMNLP*, pages 1156–1167.

Scarlini, B., Pasini, T., and Navigli, R. (2019). Just “OneSeC” for Producing Multilingual Sense-Annotated Data. In *Proceedings of ACL*, volume 1, pages 699–709.

Scarlini, B., Pasini, T., and Navigli, R. (2020). SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the Thirty-Fourth Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., and Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163:955–971.

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

Taghipour, K. and Ng, H. T. (2015). One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 338–344.

Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83.