Correcting Sense Annotations Using Wordnets and Translations

Arnob Mallik and **Grzegorz Kondrak**
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{amallik,gkondrak}@ualberta.ca

Abstract

Acquiring large amounts of high-quality annotated data is an open issue in word sense disambiguation. This problem has become more critical recently with the advent of supervised models based on neural networks, which require large amounts of annotated data. We propose two algorithms for making selective corrections on a sense-annotated parallel corpus, based on cross-lingual synset mappings. We show that, when applied to bilingual parallel corpora, these algorithms can rectify noisy sense annotations, and thereby produce multilingual sense-annotated data of improved quality.

1 Introduction

Word sense disambiguation (WSD) is the task of identifying the appropriate meaning of a word in context, from a predefined sense inventory, such as WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2012). It is one of the central problems in natural language understanding (Navigli, 2018). The primary approaches to tackle the WSD problem can be divided into supervised and knowledge-based methods. Supervised WSD systems have historically achieved the best overall results on standard WSD datasets (Raganato et al., 2017). However, these systems rely on large amounts of sense-annotated data for training, which is costly and difficult to produce. In particular, there is a severe lack of high-quality annotated data for languages other than English, which is known as the *knowledge acquisition bottleneck problem* (Pasini, 2020). To address this issue, various approaches have been proposed to automate the process of annotating texts in different languages at a large scale.

Some of the automated annotation approaches operate by leveraging translations from parallel corpora. The idea of using translations for WSD was considered by Resnik and Yarowsky (1997), based on the conjecture that different translations of an ambiguous source word in a target language could serve as sense-tagged training examples. This idea was put into practice by Ng et al. (2003), and then on a large scale by Chan and Ng (2005), as they implemented a semi-automatic approach of disambiguating English nouns using distinct Chinese translations, leveraged from an English-Chinese parallel corpora. Taghipour and Ng (2015) used a similar semi-automatic approach to create a WSD training set by leveraging the Chinese-English part of the MultiUN corpus (Eisele and Chen, 2010). Delli Bovi et al. (2017) removed the bottleneck of manual intervention, as they proposed a fully automated approach of producing multilingual sense-tagged corpora by jointly disambiguating multiple languages of a parallel corpus.

Our work is inspired by the central idea of the aforementioned research that translations may provide the necessary information to disambiguate an ambiguous word. However, we focus on leveraging translations to improve the quality of an already sense-tagged parallel corpus, rather than to annotate the corpus from scratch. We propose two algorithms for correcting sense annotations in a parallel corpus. The first algorithm attempts to rectify aligned senses that belong to different multi-synsets. The second algorithm considers all alignment links in a bitext to construct a one-to-one mapping between synsets in different languages. Both algorithms are based on the theory of synonymy and translational equivalence of Hauer and Kondrak (2020).

We empirically show that our algorithms achieve their goal of improving the quality of sense annotations in multiple languages. We extrinsically evaluate the proposed corrections by providing the corrected corpora as training data to a supervised WSD system. An intrinsic evaluation on a random sample of 200 corrected instances in English and Spanish confirms the improvement in the overall quality of the annotated corpora.

2 MultiWordNet (MWN) Algorithm

Algorithm 1 MWN

Input : set of aligned sense pairs (s, t)
$lex(s)$ - word of which s is a sense
$M(s)$ - multi-synset that contains sense s
$\mathcal{M}(w)$ - set of multi-synsets that contain word w

1: **for each** aligned sense pair (s, t) **do**
2: **if** $M(s) \neq M(t)$ **then**
3: $\mathcal{C} \leftarrow \mathcal{M}(lex(s)) \cap \mathcal{M}(lex(t))$
4: **if** $M(s) \in \mathcal{C}$ **and** $M(t) \notin \mathcal{C}$ **then**
5: CORRECT: $t \leftarrow (lex(t), M(s))$
6: **if** $M(s) \notin \mathcal{C}$ **and** $M(t) \in \mathcal{C}$ **then**
7: CORRECT: $s \leftarrow (lex(s), M(t))$

The MWN algorithm (Algorithm 1) is based on the simplifying assumption that the senses of aligned words are translationally equivalent (Hauer and Kondrak, 2020). The algorithm consults an existing multilingual wordnet (*multi-wordnet*) which is composed of multilingual synsets (*multi-synsets*) that include translationally-equivalent senses of words from both languages. Each polysemous word belongs to multiple multi-synsets. If the senses of the aligned words are found to belong to different multi-synsets, this is an indication of a possible annotation error that could be corrected.

The algorithm operates on a sense-annotated parallel corpus (*bitext*). It performs annotation corrections on individual aligned word pairs (line 1) which are annotated with different multi-synsets (line 2). Each sense in a multi-wordnet is uniquely defined as a (word, synset) tuple. When applied to a sense, the *lex* and M operators return the first and second element of the tuple, respectively. We denote as $\mathcal{C}$ the set of all multi-synsets that contains both aligned words (Line 3).

The algorithm is designed to make selective corrections only in those alignment instances where there is little doubt about the appropriate correction. At most one of the two sense annotations in each instance can be corrected. A correction is made if and only if *exactly one* of the two aligned senses is found in $\mathcal{C}$ (lines 4-7). We do not attempt a correction if either both or none of the two senses are in $\mathcal{C}$. If both senses are outside of $\mathcal{C}$, we suspect multiple errors in bitext annotations and/or the multi-wordnet. On the other hand, if both senses are within of $\mathcal{C}$, it is not clear which of the two annotations may be incorrect.

3 Bipartite (BP) Algorithm

Algorithm 2 BP

Input : set of aligned sense pairs (s, t)
$lex(s)$ - word of which s is a sense
$S(s)$ - synset that contains sense s
$\mathcal{S}(w)$ - set of synsets that contain word w

1: G $\leftarrow \varnothing$
2: **for each** aligned sense pair (s, t) **do**
3: $weight(S(s), S(t))$++
4: $weight(S(s))$++
5: $weight(S(t))$++

6: $G' \leftarrow \varnothing$
7: **for each** edge $(S_1, S_2) \in G$ **do**
8: **if** $weight(S_1, S_2) \div weight(S_1) > \alpha$ **and**
9: $weight(S_1, S_2) \div weight(S_2) > \alpha$ **then**
10: $G' \leftarrow G' \cup (S_1, S_2)$

11: **for each** aligned sense pair (s, t) **do**
12: **if** $(S(s), S(t)) \notin G'$ **then**
13: **for each** $S' \in \mathcal{S}(lex(t))$ **do**
14: **if** $(S(s), S') \in G'$ **then**
15: CORRECT: $t \leftarrow (lex(t), S')$

The BP algorithm (Algorithm 2) is also based on the assumption that the aligned words should express exactly the same concept. However, it differs from the MWN algorithm in that it globally considers all the alignment links in a given bitext, and makes annotation corrections based on the most frequently observed links. Another difference is that BP only corrects the annotations in language L_2, based on the annotations in the *base language* L_1, which are assumed to be always correct. The algorithm is inspired by the *concept universality principle* of Hauer and Kondrak (2020) which states that each each monolingual synset corresponds to at most one synset in another language. No access to a multi-wordnet is assumed; instead the algorithm consults two language-specific wordnets, which are composed of monolingual synsets, rather than of multi-synsets.

The BP algorithm consists of three stages: (1) construct a bipartite graph G of synsets; (2) identify its subgraph G' of degree 1; and (3) correct sense annotations that are not found in subgraph G'. In fact, the first two stages constitute a stand-alone algorithm for creating a cross-lingual mapping be-

tween synsets. We describe the three stages in more detail below.

In the first stage (lines 1-5), we construct a *weighted undirected bipartite graph* $G = (V, E, weight)$ in which nodes represent monolingual synsets, and edges represent alignment links that are observed in the bitext. The weight of an edge is equal to the number of the observed alignment links in the bitext between the senses of the corresponding synsets. The weight of a node is simply the sum of the weights of all edges incident with the node, which is equal to the number of times the corresponding synset is used in aligned sense annotations in the bitext.

In the second stage (lines 6-10), we construct a graph $G' = (V, E')$, which is a subgraph of G, such that every node has a degree of at most 1. The goal is to select the edges that represent the most frequent alignments. This is achieved by only retaining the edges with the relative weight above a threshold α (lines 8-9) in both directions. The threshold is constrained to be greater than 0.5, to guarantee that at most one edge per node is selected.

In the third stage (lines 11-15), annotation corrections are made based on the edges of the constructed bipartite graph G'. Unlike the MWN algorithm, the BP algorithm only corrects the annotations of words in language L_2. If an edge corresponding to a given alignment link is not found in G' (line 12), it attempts to correct the annotation in L_2 by following the edge in G' between the node $S(s)$, which represents the synset used to annotate the word in L_1, and the node S', which represents the synset in L_2 that expresses the same concept as $S(s)$.

4 Extrinsic WSD Evaluation

To extrinsically evaluate the algorithms, we apply them to *EuroSense* (Delli Bovi et al., 2017), an automatically constructed sense-annotated resource based on the *EuroParl* parallel corpus (Koehn, 2005). In EuroSense, words (which include non-compositional MWEs) are tagged with multilingual synsets from BabelNet 4.0 (Navigli and Ponzetto, 2012), and accompanied by their respective lemmatized forms.

We extract four sentence-aligned bitexts from EuroSense, by considering four different language pairs: English-Italian (EN-IT), English-German (EN-DE), English-French (EN-FR) and English-Spanish (EN-ES). We employ BABALIGN (Luan et al., 2020) to align the bitexts at the word level; the aligned word or phrase of each annotated token is taken as its translation.

Bitext	Sense Pairs	MWN	BP
EN → IT	4,713,589	541,326	82,685
EN → FR	5,219,146	664,253	106,023
EN → DE	3,083,325	179,400	59,446
EN → ES	5,015,140	518,488	92,634
IT → EN	4,713,589	235,087	89,798

Table 1: Number of sense corrections made by both algorithms.

The annotated translation pairs in EuroSense are filtered to remove non-existent senses, non-literal translations, and hypernym translations. A sense of a word is considered non-existent if it is not found in the respective BabelNet synset. If the aligned words have no synsets in common, they are treated as non-literal translations. Finally, we detect non-synonymous translations pairs by traversing hypernymy and hyponymy links in BabelNet (Hauer et al., 2020). In our development experiments, we found that approximately 3% of the pairs contain invalid senses, 13% are cases of non-literal translations. and 5% involve word entailment.

Following this filtering procedure, the remaining translation pairs are used as inputs to both algorithms to perform annotation corrections for each language separately. The BP threshold α is set to 0.8 on the basis of the development experiments. For IT, DE, FR and ES corrections, we use English as the base language. To perform EN corrections, we use Italian as the base language as it is reported to have good BabelNet coverage (Hauer et al., 2020). 75.3% of the English-Italian synset mappings returned by the BP algorithm match BabelNet concepts. Table 1 contains dataset and correction statistics for each of the five languages. The arrows in the leftmost column point from the base language to the corrected language.

We extrinsically evaluate the corrections by providing the corrected corpora as training data for a supervised WSD system, which is then evaluated on standard benchmark datasets. To this end, we employ IMS (Zhong and Ng, 2010), a supervised WSD system based on lexical features. To keep the corpus at a reasonable size, we consider a maximum of 10,000 randomly sampled training examples per sense. For English, in cases where

Train Set	Test Set							
	SemEval 2015			SemEval 2013				
	EN	IT	ES	EN	IT	FR	DE	ES
EuroSense	64.3	56.3	54.3	65.3	56.5	45.4	58.8	53.9
+ MWN	65.1	57.1	**55.3**	65.5	**58.3**	**48.0**	**60.0**	**56.7**
+ BP	64.5	**57.2**	**55.3**	65.4	56.7	45.9	59.1	54.1

Table 2: WSD F-score (%) of IMS trained on different corpora. A boldfaced result indicates a statistically significant improvement.

Lang.	Algorithm	original correct	algorithm correct	neither correct
English	MWN	6	**18**	26
	BP	12	18	20
Spanish	MWN	11	**33**	6
	BP	17	20	13

Table 3: Intrinsic evaluation results. A boldfaced result indicates a statistically significant improvement.

the system fails to make a prediction, we back off to the most frequent sense. For all languages, any monosemous words are automatically tagged with their single possible sense.

Table 2 presents the WSD results of IMS models trained on the corrected corpora, along with the results of models trained on the original EuroSense corpus. The evaluation is performed on benchmark multilingual datasets from SemEval-2013 task 12 (Navigli et al., 2013) and SemEval-2015 task 13 (Moro and Navigli, 2015). The results show that IMS achieves better results when trained on the corrected corpora. The MWN improvements are statistically significant ($p < 0.05$ using McNemar's test) over the results obtained by the original corpus for all languages except English. The BP improvements are smaller but consistent. This verifies the utility of the annotation corrections made by two algorithms when the information is transferred from English to less-resourced languages.

5 Intrinsic Evaluation

To intrinsically evaluate the quality of the sense annotation corrections made by the algorithms, a random sample of 200 English and Spanish instances were annotated manually. For each instance, an annotator was shown the corresponding sentence from EuroSense, and asked to decide whether the focus word is used in the original or the corrected sense (or neither). The senses were defined using BabelNet glosses and synonyms. and provided in a random order.

The results in Table 3 indicate that both algorithms improve the quality of the annotations in both languages. The improvements are statistically significant for the MWN algorithm ($p < 0.05$ with McNemar's test).

The wrong corrections may be grouped into three types:

Incomplete multi-synsets Many BabelNet synsets do not contain all possible lexicalizations of the concept that it represents. For example, the synset *bn:00109131a*, which is glossed in English as "related to the future", contains the Spanish adjective *futuro* but not its English translation *future*. Such omissions, which are frequent in BabelNet because of its semi-automatic construction method, prevent the MWN algorithm from making a correction.

Noise in the bitext The English-German bitext slice of EuroSense contains a total of 19,230 distinct English synsets, among which only 10,661 (55%) have matching German synsets in the dataset. This implies that nearly half of concepts represented in English are not expressed by German words, which makes it impossible to match concepts across languages. The issue may be related to the high frequency of nominal compound words in German, which are often translated as multi-word expressions in English (e.g., *Versicherungskaufmann* "insurance salesman").

Excessive granularity of senses Some instances involved a choice between fine-grained senses. For example, in the Spanish phrase *"la conclusión real de este fin de semana"* ("the actual conclusion of this weekend") the annotator found it difficult to decide whether the Spanish noun *conclusión* is used in the sense of "the temporal end; the concluding time" or "a concluding action."

6 Conclusion

Our extrinsic and intrinsic evaluation results constitute a strong proof-of-concept that translations and wordnets can be leveraged to make effective annotation corrections in a sense-annotated bitext. Manual analysis indicates that most of the invalid corrections can be traced to errors and omissions in existing lexical resources. In the future, we plan to investigate the use of machine translation instead of bitexts for the purpose of automatically annotating raw monolingual text corpora.

Acknowledgments

We thank Eduardo Montemayor Castillo and Dawn McKnight for help with manual annotation, We thank Bradley Hauer for comments on the final version of the paper.

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Yee Seng Chan and Hwee Tou Ng. 2005. Scaling up word sense disambiguation via parallel texts. In *AAAI*, volume 5, pages 1037–1042.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 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: Short Papers)*, pages 594–600.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In *LREC*.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 263–269, Barcelona (online).

Bradley Hauer and Grzegorz Kondrak. 2020. Synonymy = translational equivalence. *arXiv preprint arXiv:2004.13886*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 288–297.

Roberto Navigli. 2018. Natural language understanding: Instructions for (present and future) use. In *IJCAI*, pages 5697–5702.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, pages 222–231.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462.

Tommaso Pasini. 2020. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4936–4942.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Tagging Text with Lexical Semantics: Why, What, and How?*, pages 79–86.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344.

Zhi Zhong and Hwee Tou Ng. 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.