Semi-Supervised and Unsupervised Sense Annotation via Translations

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, Lili Mou
Alberta Machine Intelligence Institute, Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,gkondrak,yixing1,amallik,lmou}@ualberta.ca

Abstract

Acquisition of multilingual training data continues to be a challenge in word sense disambiguation (WSD). To address this problem, unsupervised approaches have been proposed to automatically generate sense annotations for training supervised WSD systems. We present three new methods for creating sense-annotated corpora which leverage translations, parallel bitexts, lexical resources, as well as contextual and synset embeddings. Our semi-supervised method applies machine translation to transfer existing sense annotations to other languages. Our two unsupervised methods refine sense annotations produced by a knowledge-based WSD system via lexical translations in a parallel corpus. We obtain state-of-the-art results on standard WSD benchmarks.

1 Introduction

Word sense disambiguation, the task of identifying the meaning of a word in context, is one of the central problems in natural language understanding (Navigli, 2018). It is a well-studied benchmark for evaluating contextualized representations of words (Loureiro et al., 2021), and is better understood than tasks such as WiC (Pilehvar and Camacho-Collados, 2019). Modern WSD methods can be divided into supervised and knowledge-based approaches. The former depend on sense-annotated corpora, such as SemCor (Miller et al., 1994), while the latter rely instead on semantic knowledge bases such as WordNet (Miller, 1995).

While supervised WSD systems typically outperform knowledge-based systems (Scarlini et al., 2020b), their utility is limited by the availability of sufficiently large sense-annotated corpora for training. This includes systems based on contextualized embeddings (Bevilacqua and Navigli, 2020). In particular, there is a severe lack of high-quality sense-annotated corpora for languages other than English. This limitation has motivated the development of methods aimed at automatically disambiguating a large number of word tokens in a given unannotated corpus, ideally covering a wide range of word and sense types, while minimizing noise (Pasini and Navigli, 2017; Scarlini et al., 2019; Barba et al., 2020). The automatically tagged corpus can then be used to train a supervised WSD system, satisfying the dependency on training data without the need for manual annotation.

Following recent theoretical work Hauer and Kondrak (2020) on establishing the semantic equivalence of mutual translations, we introduce three translation-based methods for generating sense-tagged corpora. All three methods make use of lexical knowledge bases, and semantic information obtained from word-level translations. Semi-supervised LABELPROP creates a synthetic parallel corpus (*bitext*) by applying machine translation to a monolingual manually-annotated corpus, and projecting annotations to the target language. Similarly, unsupervised LABELGEN applies a knowledge-based WSD system to the English side of a bitext, and projects the resulting sense annotations across bitexts onto other languages. Finally, unsupervised LABELSYNC produces sense-annotated corpora in two languages at once by independently applying a knowledge-based WSD system to each side of a raw bitext, and then refining the initial annotations based on the confidence scores and multilingual information.

Our experiments on standard WSD test sets demonstrate that the new methods achieve state-of-the-art results in both semi-supervised and unsupervised sense annotation. We train two different reference supervised WSD systems on the generated data, and apply the resulting models to multilingual WSD benchmarks. Our results compare favourably to models trained on data produced by

Proceedings of Recent Advances in Natural Language Processing, pages 504–513
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_057

the previous state-of-the-art sense annotation methods. Indeed, some of the results obtained with our unsupervised methods rival those obtained by training on a manually sense-annotated corpus.

Our contributions are as follows: We present three novel, scalable methods that can generate annotated corpora for any language for which a suitable lexical knowledge base is available. We show that these methods achieve state-of-the-art results on multiple languages. We make our code and corpora available.[1]

2 Related Work

The sense tagging systems that we consider in this work, including our three novel methods, can be divided into four types according to two criteria (Figure 1). The first criterion is whether the method involves supervision in the form of a sense-annotated corpus. The second criterion is whether the method operates as a traditional self-contained WSD system, or instead assigns sense tags to a subset of the words in a corpus which can then be used to train a supervised WSD system. In this section, we discuss the most relevant examples of each of the four resulting types.

Supervised WSD systems rely on sense annotations to train disambiguation models, which are evaluated on benchmark datasets. Examples include GlossBERT (Huang et al., 2019), EWISE (Kumar et al., 2019), and EWISER (Bevilacqua and Navigli, 2020). Because they require labelled training data in the target language, such systems are generally impractical for languages other than English, nor are they directly comparable to our proposed methods.

Knowledge-Based WSD systems remain important due to the limited coverage of existing annotated corpora, as well as their English bias. These include graph-based systems such as UKB (Agirre et al., 2014), UKB enhanced with SyntagNet (Maru et al., 2019), and systems based on multilingual BERT (Devlin et al., 2019) and BabelNet (Navigli and Ponzetto, 2012), such as SensEmBERT (Scarlini et al., 2020a). We compare our unsupervised results to both UKB+SyntagNet and SensEmBERT.

Semi-Supervised Corpus Tagging systems depend on sense annotated corpora in one language to produce sense annotations in other languages. The current state-of-the-art method in this setting is MuLaN (Barba et al., 2020), which propagates sense

[1] https://www.cs.ualberta.ca/~kondrak

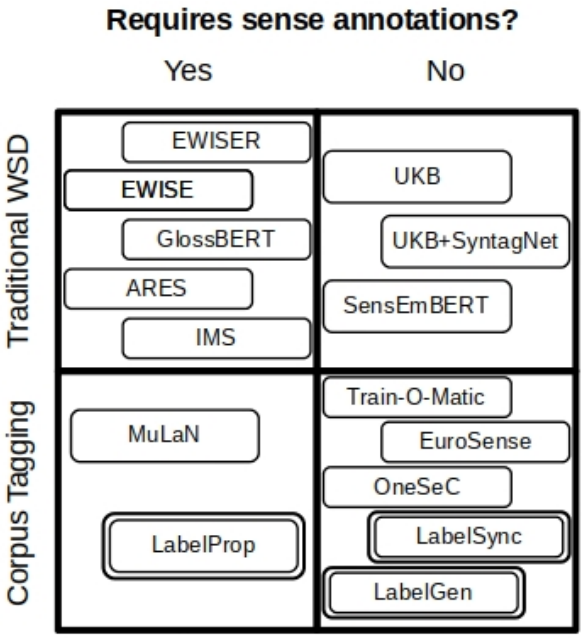

Figure 1: Typology of relevant sense tagging systems. Our own systems are shown in double-lined boxes.

annotations from SemCor and WordNet Gloss Corpus (WNG, Langone et al., 2004) to semantically similar contexts in Wikipedia corpora using contextual word representations from mBERT. Our LABELPROP method differs by leveraging machine translation to directly propagate sense annotations across word alignment links.

Unsupervised Corpus Tagging systems produce sense annotations "from scratch". Train-O-Matic (Pasini and Navigli, 2017) annotates Wikipedia in multiple languages by applying the Personalized PageRank (PPR) algorithm to BabelNet. OneSeC (Scarlini et al., 2019) combines Wikipedia categories and BabelNet synset representations to produce WSD training data, and outperforms Train-O-Matic. However, both Train-O-Matic and OneSeC annotate nominal instances only, and hence are not applicable to all-words WSD. On the other hand, EuroSense (Delli Bovi et al., 2017) jointly disambiguates content words of all parts of speech in a parallel corpus using a knowledge-based WSD system. Our LABELSYNC and LABELGEN methods differ in that they explicitly leverage lexical translation information obtained from a bitext.

Other work on using translations for WSD: Resnik and Yarowsky (1999) propose to distinguish senses only if a "minimum subset" of languages translate them differently. Apidianaki (2009) demonstrates how senses can be induced

by clustering lexical translations, and proposes an unsupervised WSD system based on such induced sense inventory and translation information. Lefever et al. (2011) frame WSD as translation selection, and propose a method based on multilingual feature vectors. Finally, Taghipour and Ng (2015) annotate English words with their Chinese translations using manually crafted sense-to-translation mappings. These methods are not comparable with our work as they do not link their sense annotations to the WordNet sense inventory, and therefore are not applicable to modern WSD datasets.

3 Semi-Supervised LABELPROP

In this section, we introduce LABELPROP, a novel label propagation approach for constructing multilingual sense-annotated corpora. The idea is to translate a sense-annotated corpus in order to propagate the sense tags across the translations. No sense-annotated data is required in the target language. The method is composed of three steps: translation identification, knowledge-base filtering, and nearest neighbor filtering (Figure 2).

3.1 Translation Identification

Given a sense-annotated source corpus, we first translate the corpus into the target language using pre-trained neural machine translation models. Each sentence containing at least one source sense-annotated word is translated independently. If the translation of an annotated source word can be identified through word alignment, we annotate the translation with the same BabelNet synset as the aligned source word. This procedure is based on the assumption that lexical translations in context are semantically equivalent, and therefore very likely to express the same concept (Hauer and Kondrak, 2020).

For alignment, we use BABALIGN (Luan et al., 2020), a high-precision alignment tool which leverages translation information from BabelNet to improve on a base alignment system. In particular, BABALIGN augments the input corpus with lexical translation pairs to bias the aligner towards aligning words which are mutual translations. It also corrects alignments to maximize the number of aligned words that share BabelNet synsets. This emphasis on recovering word-level translation information makes BABALIGN particularly well-suited to our method.

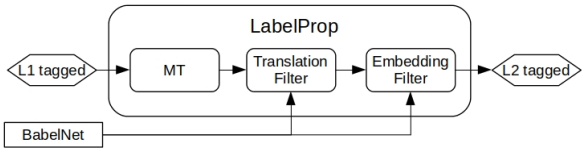

Figure 2: LABELPROP propagates senses from language L1 to language L2.

3.2 Knowledge-Based Filtering

The sense-projection procedure in the previous step may annotate a word with a BabelNet synset which does not actually contain that word. These invalid sense annotations may occur due to non-literal translation (i.e., the word and its translation do not express the same concept), errors in translation or alignment, or omissions in BabelNet. Since each sense of a word must correspond to a specific synset, such invalid annotations are discarded.

3.3 Nearest Neighbor Filtering

In order to further increase the precision, we apply a semi-supervised WSD method to each target translation that is sense annotated by the previous steps. For each word, we verify that the annotation propagated from the source-language corpus matches the annotation assigned by the WSD system; otherwise, we discard that sense annotation.

Our semi-supervised WSD method uses a one-nearest-neighbor approach with ARES multilingual synset embeddings (Scarlini et al., 2020b). We first obtain contextual word representations of each sense-annotated target translation by taking the sum of the last four layers of multilingual BERT (Devlin et al., 2019). Since ARES embeddings have twice the size of the original mBERT embeddings, we concatenate each obtained word representation with itself. We then compute the cosine similarity between the mBERT representation of the word, and the ARES representation of each synset containing the word. The synset that maximizes the similarity is taken as the output of this WSD system. To reiterate, we retain only the sense annotations from the previous step that agree with this WSD system.

4 Unsupervised Symmetric LABELSYNC

The LABELPROP method, presented in Section 3, is able to leverage existing sense annotated corpora, such as SemCor, to create comparable sense annotated corpora in other languages. However, the availability of sense-annotated corpora in other do-

mains and languages is very limited. On the other hand, large bitexts are relatively easy to obtain for many language pairs and domains.

To further reduce the dependency of WSD systems on *any* pre-existing annotated data, we introduce LABELSYNC, a method which annotates both sides of a given bitext. This method retains the idea of using word alignment to validate sense annotations, while eschewing the need for a sense annotated corpus. It is composed of three steps: monolingual word sense disambiguation, multilingual post-processing, and translation-based filtering (Figure 3).

4.1 Monolingual WSD

Our goal is to enrich both sides of the input bitext with sense tags. Since LABELSYNC does not assume access to any sense-annotated corpus, we employ a language-independent knowledge-based WSD system: a variant of UKB enhanced with SyntagNet (Maru et al., 2019). After each side of the bitext is annotated independently, we have two sense annotated corpora, one in each of the languages represented in the bitext.

4.2 Multi-Lingual Post-Processing

Now that both sides of the bitext are annotated independently, we leverage the lexical translation information inherent in the bitext to increase the accuracy of the sense annotations. To improve the performance of our base WSD system, we employ the SOFTCONSTRAINT method of Luan et al. (2020). This method is applicable to any base WSD system which assigns a numerical score, such as a probability, to each sense of a disambiguated word. Most modern WSD systems, including UKB, satisfy this property.

The SOFTCONSTRAINT method depends on word-level translations of each annotated word, as well as translation information from BabelNet, which is based on the hypothesis that the translation of a word token provides semantic information about its sense (Hauer and Kondrak, 2020). In our case, translation information is readily available from the bitext. As with our LABELPROP method, we use BABALIGN (Luan et al., 2020) to word align the bitext. For each sense-annotated token, the aligned word or phrase is treated as its translation.

The SOFTCONSTRAINT method can also incorporate sense frequency information, to bias the annotations toward more probable senses. However, we exclude sense frequency information from this step, as it provided no discernible benefit in our development experiments.

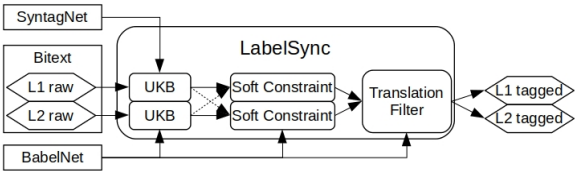

Figure 3: LABELSYNC assigns and refines sense annotations in two languages simultaneously.

4.3 Translation-Based Filtering

In the final step, we aim to further reduce the noise in our sense-annotated corpora by employing a BabelNet-based filtering method, similar to the one described in Section 3.2. As before, the key idea is to impose two constraints on our sense annotations: (1) a word should only be annotated with a synset that contains the word, and (2) aligned words should be annotated with the same synset. LABELPROP initially guarantees only the latter constraint, so it has to discard some annotations to ensure the former. In contrast, LABELSYNC initially guarantees the first constraint, as UKB can only annotate a word with a synset containing it. However, since each side of the corpus is annotated independently, the second constraint may not hold. The final step of LABELSYNC is aimed at resolving this problem by synchronizing the sense annotations across both sides of the bitext.

Unlike Delli Bovi et al. (2017), who leverage embeddings of concepts to filter questionable annotations, we adopt a binary alignment-based criteria using the assumption of semantic equivalence of lexical translations. We retain only those annotations that refer to the same multilingual synset as the sense annotations of their translations. We also retain annotations if the token cannot be aligned, or if its translation is not annotated.

5 Unsupervised Asymmetric LABELGEN

Unlike LABELSYNC, our second unsupervised method, LABELGEN, assumes that the source language is English, and treats the two sides of the bitext differently. The goal is to leverage available English resources to improve WSD performance on other languages, rather than to generate English sense annotations.

	LABELPROP				LABELSYNC			LABELGEN		
	Annotated Tokens	Annotated Word Types	Sense Types	Failed Alignments	Annotated Tokens	Annotated Word Types	Sense Types	Annotated Tokens	Annotated Word Types	Sense Types
EN	-	-	-	-	1,783,334	9,509	16,748	-	-	-
IT	399,569	25,361	29,290	30,763	2,083,741	10,910	22,211	1,372,876	8,355	16,046
ES	403,797	25,874	31,420	31,640	1,692,232	10,549	25,181	1,326,244	7,926	17,335
FR	407,590	25,193	32,129	32,181	1,458,588	7,776	11,529	1,433,647	7,712	17,980
DE	309,926	23,786	23,433	64,085	645,289	2,139	2,756	821,552	6,589	9,121

Table 1: Statistics of the sense-annotated corpora produced by each of our methods.

5.1 English WSD

Given a bitext, we first apply a knowledge-based WSD system to the English side *only*, as described in Section 4.1. The lexical information for other languages is retrieved from BabelNet multi-synsets which are aligned to WordNet 3.0 synsets. This automatic candidate retrieval process is noisy, because most BabelNet lexicalizations are automatically generated from various resources. In addition, while English WordNet contains the sense frequency estimates from the manually-annotated SemCor, such information is not readily available for other languages. Hence, WSD annotations are more accurate for English compared to other languages.

5.2 Label Propagation

Having automatically sense-tagged the English side of the bitext, we propagate the labels to the non-English side using the procedure described in Section 3.1. In effect, we are applying the first part of LABELPROP, treating the English side as a sense-tagged corpus, and the other side as its translation. At the end of this process, both sides of the bitext are sense-annotated.

5.3 Re-Ranking and Filtering

We further refine the sense annotations on the non-English side of the bitext. We first apply the SOFT-CONSTRAINT method as described in Section 4.2, which re-ranks the possible senses for each annotated word using the assigned WSD scores. We then apply the filtering procedure from Section 3.2, which removes any sense annotations that do not exist in the BabelNet sense inventory.

6 Evaluation

Following prior work, we extrinsically evaluate our corpus construction approaches by providing the generated annotations as training data for supervised WSD systems (*reference systems*), which are then evaluated on standard multi-lingual WSD benchmarks. While our methods could also be applied to low-resource languages, the current lack of evaluation datasets precludes such experiments in this work.

6.1 Reference Supervised WSD Systems

We perform experiments with two reference supervised WSD systems: (1) IMS (Zhong and Ng, 2010) with the most-frequent-sense (MFS) backoff for English, and (2) *mBERT*, a transformer-based method, built on multilingual BERT (Devlin et al., 2019), as described by Barba et al. (2020). We use the default parameter settings and number of training epochs[2]. We train each model on each set of automatically produced sense annotations.

Following prior work, we use the SemEval-2007 dataset (Raganato et al., 2017) as our validation set for the English experiments. Because of the lack of standard validation sets for non-English languages, we use random samples of 1000 sentences from our training corpora. The hyperparameters of each system are held constant throughout all experiments.

6.2 Test Data

We test the reference WSD models on standard multilingual benchmark datasets: SemEval-2013 task 12 (Navigli et al., 2013), which contains data for Italian, Spanish, French, and German, and SemEval-2015 task 13 (Moro and Navigli, 2015), which covers Italian and Spanish. The SemEval-2013 datasets contain only nominal instances, while the SemEval-2015 datasets cover nouns, verbs, adjectives, and adverbs. We use the latest version of the datasets[3], which are annotated with synsets from BabelNet version 4.0.

For the experiments on English (Section 6.4.3), we use the standardized benchmarks of Raganato

[2]https://github.com/edobobo/transformers-wsd
[3]https://github.com/SapienzaNLP/mwsd-datasets

et al. (2017)[4], which comprise all-words test sets from five shared tasks: Senseval2 (Edmonds and Cotton, 2001), Senseval3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), and SemEval-2015 (Moro and Navigli, 2015). We also report the average results on the concatenation of all five test sets, which we refer to as ALL.

6.3 Semi-Supervised Approaches

This section is devoted to the empirical evaluation of the LABELPROP method from Section 3.

6.3.1 Experimental Setup

We apply LABELPROP to a sense-annotated English corpus comprised of SemCor (Miller et al., 1994) and the WordNet Gloss Corpus (WNG) (Langone et al., 2004). Following Luan et al. (2020), we translate each sentence of our English corpus with Google Translate independently into Italian, Spanish, French, and German. As described in Section 3.1, we induce word alignment by applying BABALIGN with FASTALIGN (Dyer et al., 2013) as its base aligner. Table 1 contains the statistics for the corpora created with our methods. "Sense types" indicates the number of distinct word senses in the corpus. "Failed alignments" refers to the number of English sense annotations that could not be propagated.

We compare LABELPROP to MuLaN (Barba et al., 2020), the current state-of-the-art system for semi-supervised corpus annotation, which also uses SemCor+WNG as its manually-annotated base English corpus. Specifically, we apply the same procedure to train the supervised reference system (IMS or mBERT) on the annotated data produced by each method. We train a single model for each system, using only the corpus produced by that system for each language, which limits the impact of language-specific issues. Nevertheless, due to both software and hardware variables and hyper-parameters, our MuLaN results differ from those reported in the original paper.

When reporting the results achieved with mBERT, we also include the results of two recent WSD systems: ARES, using the reported results from Scarlini et al. (2020b), and 0-Shot WSD, with results replicated using the code provided by Barba et al. (2020). Since they are not designed to create annotated training data for other WSD systems, they are not directly comparable to LABELPROP.

[4]http://nlp.uniroma1.it/wsdeval

	SemEval-2013				SemEval-2015		
Model	IT	ES	FR	DE	IT	ES	AVG
MCS	44.2	37.1	53.2	**70.2**	44.6	39.6	48.2
MuLaN	65.6	65.6	68.1	69.7	63.7	59.9	65.4
LABELPROP	71.4	71.0	65.1	62.8	67.1	64.0	66.9
-NN	70.5	70.1	62.7	63.5	66.3	61.8	65.8
-NN -KB	66.7	68.2	60.2	56.7	61.2	60.2	62.2

Table 2: WSD F-score obtained with IMS trained on the corpora generated by MuLaN and LABELPROP.

	SemEval-2013				SemEval-2015		
Model	IT	ES	FR	DE	IT	ES	AVG
ARES	77.0	75.3	81.2	79.6	71.4	70.1	75.7
0-shot$_{SC+WNG}$	78.3	77.6	80.8	78.3	70.5	68.6	75.7
MuLaN	76.8	78.4	80.4	78.8	68.7	67.8	75.2
LABELPROP	78.4	78.5	80.4	77.8	72.7	68.9	76.1

Table 3: WSD F-score obtained with mBERT trained on the corpora generated by MuLaN and LABELPROP. Results of two semi-supervised WSD systems are included for reference.

6.3.2 Results

Table 2 presents the multilingual WSD results obtained by IMS on standard test sets. While IMS is no longer a state-of-the-art system, it is still commonly used as a benchmark for evaluating automatically generated corpora (Scarlini et al., 2019). The results demonstrate the relative quality of the generated corpora: LABELPROP is better than MuLaN on Italian and Spanish, as well as on average. The difference in performance is found to be statistically significant across all six datasets ($p < 0.05$ with McNemar's test). Neither of the approaches outperforms the most common sense (MCS) baseline on German, which we discuss further in Section 6.3.3. The ablation results in the last two rows show that both the nearest neighbour WSD filter (NN) and the translation-based filter (KB) improve the quality of the annotations.

Table 3 presents the corresponding results using the more recent mBERT as the reference system. Our results are slightly better on average than those of 0-shot and ARES. However, only the MuLaN results are directly comparable to our LABELPROP results, as both systems produce training data for a supervised WSD reference system. LABELPROP matches or outperforms MuLaN on every dataset except German SemEval-2013, and achieves better results on average compared to the results we replicated. The difference in F-score between LABELPROP and our replicated MuLaN experiment is significant for the SemEval 2015 Italian dataset ($p < 0.05$ with McNemar's test).

6.3.3 Error Analysis

Error analysis suggests two reasons for the relatively low results on the German data. First, English multi-word compounds often correspond to single words in German, which makes it difficult to properly propagate English sense annotations. For example, the two words in *giveaway program*, which is a translation of *Werbeprogramm*, are separately annotated with different senses. The second issue is the quality of the BabelNet translation coverage. We observe that among 69,402 BabelNet synsets, that correspond to word senses appearing in SemCor+WNG, only 40,490 synsets contain at least one German translation, compared to over 50,000 synsets in each of the other three languages.

6.4 Unsupervised Approaches

In this section, we evaluate our unsupervised methods, LABELSYNC and LABELGEN, against comparable systems.

6.4.1 Experimental Setup

We adopt UKB (Agirre et al., 2014) as the base knowledge-based WSD system used in the first step of both LABELSYNC and LABELGEN to perform the initial tagging of a bitext. (This base WSD system is not to be confused with the reference supervised WSD system that is only used for the purpose of corpus evaluation.) Following Maru et al. (2019), we use WordNet as a lexical knowledge base, enriching it with information from WNG, and syntagmatic information from SyntagNet. BabelNet is the source of multilingual lexicalization information. When applying UKB, the PPR_{w2w} variant of the personalized PageRank algorithm is run separately for each word, while concentrating the initial probability mass in the senses of the context words rather than the focus word.

Both of our unsupervised methods operate on an unannotated bitext. To keep the corpus size manageable, we randomly sample 200k sentences with English, French, German, Italian, and Spanish translations from EuroSense (Delli Bovi et al., 2017) discarding its existing sense annotations. This produces four bitexts with English as one of the languages, which we align at the word level using BABALIGN. The SOFTCONSTRAINT method employed by LABELSYNC to refine the initial sense annotations leverages the lexical translations. Table 1 presents the statistics of the produced corpora.

	SE-2013				SE-2015		
Model	IT	ES	FR	DE	IT	ES	AVG
MCS	44.2	37.1	53.2	70.2	44.6	39.6	48.2
UKB+SyntagNet	72.1	74.1	70.3	76.4	69.0	63.4	70.9
SENSEMBERT	69.8	73.4	77.8	79.2	-	-	-
OneSeC	63.5	61.6	65.1	75.8	-	-	-
LABELSYNC	75.7	78.2	72.4	75.3	70.8	66.3	73.1
LABELGEN	77.8	80.5	80.7	75.4	68.7	66.1	74.9

Table 4: WSD F-score obtained with mBERT trained on the corpora generated by LABELSYNC and LABELGEN.

The direct competitor of LABELSYNC and LABELGEN is OneSeC (Scarlini et al., 2019), an unsupervised system which produces sense-annotated data by leveraging the semantic information within Wikipedia categories. Since OneSeC can only tag nouns, any model trained on a corpus it produces will likewise only be able to disambiguate nouns. Therefore, we do not apply models trained on OneSeC to the SemEval-2015 datasets, which include verb, adjective, and adverb instances. For our multilingual experiments, we also compare to two knowledge-based WSD systems described in Section 2: UKB with SyntagNet (Maru et al., 2019), and SENSEMBERT (Scarlini et al., 2020a).

6.4.2 Multilingual Results

Table 4 presents the multilingual WSD results when using mBERT as the reference WSD system. With the consistent exception of German, the results of mBERT trained on the annotations produced by LABELSYNC are substantially better than those trained on the corpus generated by OneSeC, which is the previous state-of-the-art for unsupervised corpora tagging. Unlike OneSeC, our unsupervised methods can annotate tokens representing all parts of speech, and can therefore be applied to the SemEval 2015 datasets. LABELSYNC also outperforms both knowledge-based WSD systems, UKB+SyntagNet and SENSEMBERT, and the most common sense (MCS) baseline. LABELGEN further improves on LABELSYNC by 1.8% on average. This makes it our best performing system, which sets a new state-of-the-art on the SemEval-2013 Italian, Spanish, and French datasets.

6.4.3 English Results

In this section, we evaluate LABELSYNC on English WSD. We do not test LABELGEN on English, as it was specifically designed to tag non-English corpora. Furthermore, because OneSeC annotates only nominal instances, we conduct separate all-

	SE2	SE3	S07	S13	S15	ALL
MFS	66.8	66.2	55.2	63.0	67.8	65.2
SemCor	71.3	69.1	61.5	65.5	68.3	68.3
EuroSense + SemCor	-	-	-	66.4	69.5	-
LABELSYNC	69.4	64.5	57.4	71.7	72.9	68.4

Table 5: WSD F-score on all instances obtained with IMS trained on the corpora generated by LABELSYNC.

	SE2	SE3	S07	S13	S15	ALL
MFS	66.8	66.2	55.2	63.0	67.8	65.2
SemCor	74.8	73.1	64.2	69.9	74.7	72.6
LABELSYNC	69.6	65.9	55.2	71.4	75.1	68.9

Table 6: WSD F-score on all instances obtained with mBERT trained on the corpora generated by LABELSYNC.

words and nouns-only experiments.

Table 5 presents the English WSD results on all test instances with IMS as the reference system. The corpus annotated by LABELSYNC is a subset of the corpus annotated by EuroSense. The results of LABELSYNC on S13 and S15 are much better than the results using EuroSense augmented with SemCor as reported by Delli Bovi et al. (2017), which we attribute to the explicit use of translation information. On the concatenation of all five test sets, the unsupervised IMS+LABELSYNC results rival the supervised results of IMS trained on SemCor, a manually sense-annotated corpus.

In Table 6, we see mBERT performing much better than IMS when trained on SemCor. Remarkably, the corpus generated in an unsupervised manner by LABELSYNC yields results on S13 and S15 that surpass those obtained by training mBERT directly on SemCor. These results are impressive because LABELSYNC makes no use of manual sense annotation. We speculate that this may be due to the difference in domain between SemCor and the corpus annotated by LABELSYNC.

Tables 7 and 8 present the results of our final set of experiments, in which the reference systems are tested on English nominal instances only. Here, we can compare LABELSYNC directly to its competitor, OneSeC. Both of our reference WSD systems, IMS and mBERT, clearly perform better across all datasets when trained on the corpus produced by our LABELSYNC method, compared to training on the corpus produced by OneSeC. These results establish LABELSYNC as the new state of the art for unsupervised English corpus sense tagging, and a step towards overcoming the knowledge acquisition bottleneck in WSD.

	SE2	SE3	S07	S13	S15	ALL
MFS	72.1	72.0	65.4	63.0	66.3	67.6
SemCor	76.8	73.8	67.3	65.5	66.1	70.4
OneSeC	73.2	68.2	63.5	66.5	70.8	69.0
LABELSYNC	76.1	70.0	68.6	71.7	72.1	72.3

Table 7: English WSD F-score on nominal instances obtained with IMS as the reference WSD system,

	SE2	SE3	S07	S13	S15	ALL
MFS	72.1	72.0	65.4	63.0	66.3	67.6
SemCor	79.7	75.4	67.9	69.9	75.0	74.0
OneSeC	74.2	67.1	62.9	68.8	74.2	70.2
LABELSYNC	76.8	70.8	66.0	71.4	75.7	73.0

Table 8: English WSD F-score on nominal instances obtained with mBERT as the reference WSD system.

7 Conclusion

We have introduced new methods to address the knowledge acquisition bottleneck in word sense disambiguation in both the semi-supervised and unsupervised settings. The methods leverage recent advances in machine translation, alignment, and contextual embeddings. Extrinsic experiments with a variety of WSD systems demonstrate that the quality of the corpora created by our methods is substantially higher compared to those produced by prior work. Our methods for automatic sense tagging can produce annotated corpora for many languages, and approach the quality of manual annotation in some cases. We make our corpora available for further research.

One advantage of our unsupervised methods is that they can be applied to annotate any bitext involving any languages. We posit that our results could be further improved by annotating corpora with broader domain coverage, or by matching the domain of the source corpus to the domain of the data to be disambiguated. We leave this as a direction for future work.

Acknowledgments

The authors of this paper are listed in alphabetical order. Yixing Luan and Arnob Mallik conducted the experiments with the semi-supervised and unsupervised methods, respectively. Bradley Hauer and Grzegorz Kondrak prepared 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

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 77–85.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLaN: Multilingual label propagation for word sense disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5.

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

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3507–3512.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

Helen Langone, Benjamin R Haskell, and George A Miller. 2004. Annotating WordNet. In *Proceedings of the Workshop On Frontiers In Corpus Annotation*, pages 63–69.

Els Lefever, Véronique Hoste, and Martine De Cock. 2011. ParaSense or how to use parallel corpora for word sense disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, pages 1–55.

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.

Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. 2019. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3534–3540.

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

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of a Workshop on Human Language Technology*, pages 240–243. Association for Computational Linguistics.

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.

Tommaso Pasini and Roberto Navigli. 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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, pages 1267–1273.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 87–92.

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. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "OneSeC" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the Thirty-Fourth Conference on Artificial Intelligence*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

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.