Improving Word Sense Disambiguation with Translations

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

Abstract

It has been conjectured that multilingual information can help monolingual word sense disambiguation (WSD). However, existing WSD systems rarely consider multilingual information, and no effective method has been proposed for improving WSD by generating translations. In this paper, we present a novel approach that improves the performance of a base WSD system using machine translation. Since our approach is language independent, we perform WSD experiments on several languages. The results demonstrate that our methods can consistently improve the performance of WSD systems, and obtain state-of-the-art results in both English and multilingual WSD. To facilitate the use of lexical translation information, we also propose BABALIGN, an precise bitext alignment algorithm which is guided by multilingual lexical correspondences from BabelNet.

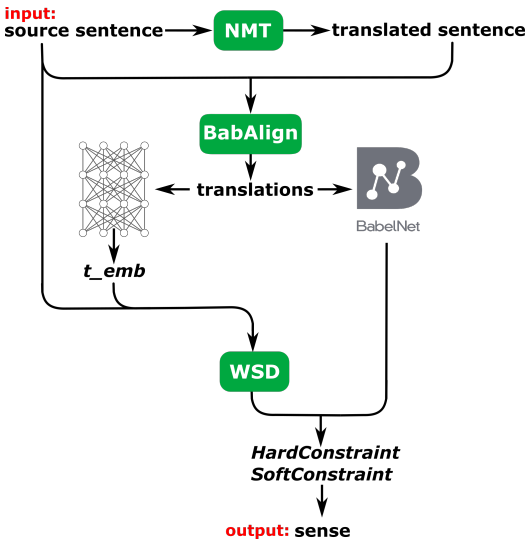

Figure 1: An overview of our approach to leverage translations to improve a base WSD system.

1 Introduction

Word sense disambiguation (WSD) is one of the core tasks in natural language processing. Given a predefined sense inventory, a WSD system aims to identify the correct sense of a content word in context. Although WSD is a monolingual task, it has been conjectured that multilingual information could help (Resnik and Yarowsky, 1999; Carpuat, 2009). Attempts have been made to leverage parallel corpora for sense tagging (Diab and Resnik, 2002), but no effective method for improving WSD with translations has been proposed to date.

Much of the history of WSD has been determined by the availability of manually created lexical resources in English, including SemCor (Miller et al., 1994) and WordNet (Miller, 1995). The situation changed with the introduction of BabelNet (Navigli and Ponzetto, 2012a), a massive multilingual semantic network, created by automatically integrating WordNet, Wikipedia, and other resources. In particular, BabelNet synsets contain translations in multiple languages for each individual word sense. Methods have been proposed to use multilingual information in BabelNet for WSD (Navigli and Ponzetto, 2012b; Apidianaki and Gong, 2015), but they do not directly exploit the mapping between senses and translations in multiple languages.

While there have been many attempts to apply WSD to machine translation (MT) (Liu et al., 2018; Pu et al., 2018), our goal instead is to harness advances in MT to improve WSD. Rather than develop a new WSD system, we propose a general method that can make existing and future systems more accurate by leveraging translations. We evaluate our methods with several supervised and knowledge-based WSD systems.

Our approach is based on the assumption of absolute synonymy between the senses of mutual translations in context (Hauer and Kondrak, 2020). The principal method SOFTCONSTRAINT refines sense

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4055–4065,
November 16–20, 2020.

predictions of a given base WSD system using sense-translation mappings from BabelNet. The approach is able to take advantage of translations in multiple languages, whether produced manually or by MT models. It is also able to leverage sense frequency information, which can be obtained in either a supervised or unsupervised manner. Another method that we test is *t_emb* which integrates translations as contextual word embeddings into a WSD system to bias its sense predictions. To obtain word-level translations from the translated contexts, we introduce BABALIGN, a precise alignment algorithm guided by BabelNet synsets. In Figure 1, we show the entire architecture of our model based on aforementioned components.

Our experimental results demonstrate that translations can significantly improve existing WSD systems. We perform several experiments on English and multilingual WSD with both manual and machine translations. In the English WSD experiments with manual translations and word-level alignments, we determine the potential of our methods in an ideal situation. In the experiments with machine translations, we validate that the methods are effective and robust by showing improvements over existing WSD systems. Finally, in the multilingual WSD experiments, we demonstrate the language independence of our methods.

The main contributions of this work are the following. (1) We propose the first effective method to improve WSD with automatically generated translations. (2) Our language-independent knowledge-based method achieves state-of-the-art results in both English all-words and multilingual WSD. (3) We introduce a bitext alignment algorithm that leverages information from BabelNet.

2 Related Work

The integration of multilingual information to improve English WSD has been considered in prior work. Through analyzing a multilingual dictionary, Resnik and Yarowsky (1999) observe that highly distinct senses can translate differently. Diab and Resnik (2002) propose a WSD system based on translation information extracted from a bitext, but it fails to outperform systems that rely on monolingual information only.

Word sense induction (WSI) and cross-lingual WSD (CLWSD) are related tasks. WSI aims for automatically inducing word senses from corpora by clustering similar instances of words. Several prior works perform WSI based on bitexts to create bilingual sense inventory on word samples, where translations are treated as sense tags (Specia et al., 2007; Apidianaki, 2009). CLWSD is a task to predict a set of translations for a given ambiguous word in context. Attempts have been made to integrate translations as bag-of-words feature vectors to enhance CLWSD (Lefever et al., 2011). Since the goals of WSI and CLWSD differ from standard WSD with predefined senses, our approach is not directly comparable.

Navigli and Ponzetto (2012b) incorporate translations in BabelNet synsets as a feature in a graph-based WSD system. However, rather than apply translations of the focus word token as constraints, they simply consider all possible translations of the focus word type to enhance its sense distinctions.

Apidianaki and Gong (2015) directly apply sense-translation mappings in BabelNet as a hard constraint on sense predictions using translations from sense-annotated bitexts. Unlike our work, their approach is based on the BabelNet First Sense (BFS) baseline, rather than on an actual WSD system. Their results on English WSD fail to show improvement over the baseline, which may be due to the use of only a single target language, as well as word alignment errors.

3 Methods

We first formulate our WSD task. The input is a sentence, in which one word, e, is designated as the *focus word*. The set of possible senses of the focus word $S(e)$ comes from the sense inventory. We assume that a base WSD system assigns a probability or score to each sense, with the output being the sense with the maximum score. The objective is to determine which sense $s \in S(e)$ is the sense of e in this sentence.

We propose two methods, HARDCONSTRAINT and SOFTCONSTRAINT, which can be used to augment any base WSD system that meets the above specifications. Both methods leverage translations in order to constrain sense predictions made by a base WSD system. In addition, we introduce a method of leveraging contextual word embeddings to enhance the integration of translations in combination with those constraints. Finally, since our methods crucially depend upon identifying the translation of the focus word in the translated sentence, we also introduce a new knowledge-based word alignment algorithm.

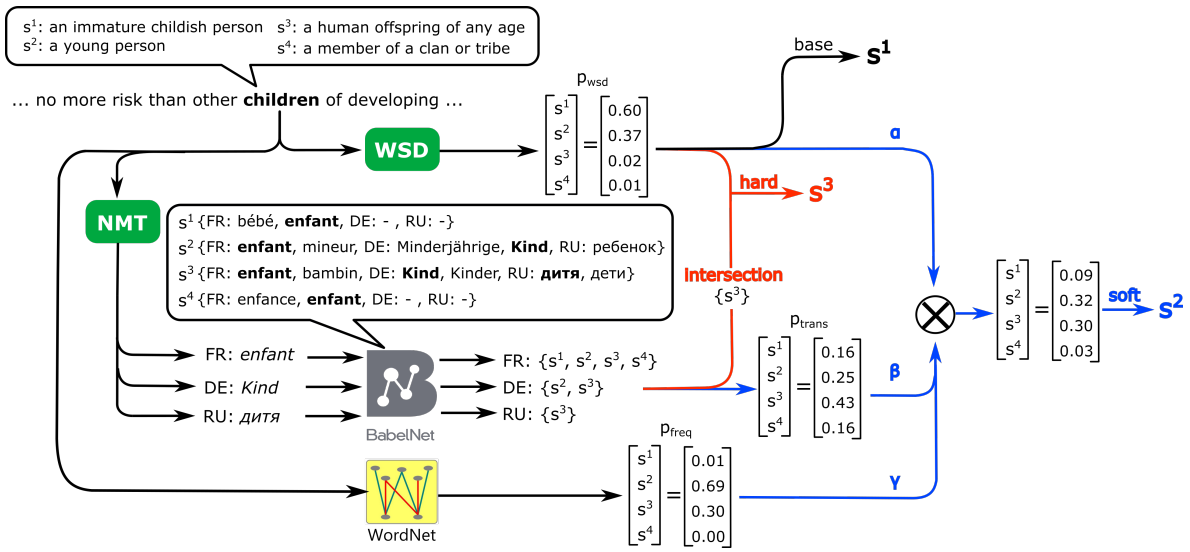

Figure 2: The application of HARDCONSTRAINT (red) and SOFTCONSTRAINT (blue) when disambiguating the word *children* in the given context (actual example from Senseval2 data where the correct sense is s^2).

3.1 HARDCONSTRAINT

Our first method extends the idea of Apidianaki and Gong (2015) to constrain $S(e)$ based on sense-translation mappings in BabelNet. However, instead of relying on a single translation, we incorporate multiple languages by taking the intersection of the individual sets of senses; that is, we rule out senses if their corresponding BabelNet synsets do not contain translations from all target languages. This baseline method is simple but inflexible: the correct sense can be accidentally ruled out if the provided translation of the focus word is not found in the corresponding BabelNet synset.

Our implementation of HARDCONSTRAINT considers the intersection of the sets of synsets that contain translations from each language. Ideally, the intersection contains exactly one sense, which we take as the final prediction. (Such a case is illustrated in Figure 2.) Otherwise, if the intersection contains multiple senses, we choose the one with the highest score from the base WSD system. If the intersection happens to be empty, we back-off to the prediction of the base WSD system.

3.2 SOFTCONSTRAINT

HARDCONSTRAINT is effective at ruling out sense candidates, but is also sensitive to MT errors and BabelNet deficiencies. BabelNet contains translations for only 79% of the nominal senses in WordNet, and its multilingual lexicalizations have an average precision of only 72% (Navigli and Ponzetto, 2012a).

Our principal method, SOFTCONSTRAINT, is more robust in handling noisy MT translations and BabelNet gaps. It integrates information from three sources: the base WSD system, translations, and sense frequencies (Figure 2). From each of these sources, we derive a probability distribution over $S(e)$. We employ the product of experts (PoE) approach (Hinton, 2002) to combine the probabilities as follows:

$$\tilde{p}(s) = p_{wsd}(s)^{\alpha} \cdot p_{trans}(s)^{\beta} \cdot p_{freq}(s)^{\gamma}$$

The resulting score $\tilde{p}$ is an unnormalized measure of probability with tunable weights α, β, and γ. We tune those weights through grid-search. The sense that maximizes this measure is taken as the prediction. Below, we provide the details on each of the three distributions.

Probability p_{wsd} is obtained by simply normalizing the numerical scores from the base WSD system.

Probability p_{trans} is calculated on the basis of the set of translations for each focus word e in BabelNet. Given a source focus word e and a word f in another language, we obtain its sense coverage $c(e, f)$ representing the number of possible senses of e that are mapped to f, i.e., the number of BabelNet synsets containing both e and f. Based on the sense coverage, the word pair e and f is assigned a weight $w(e, f)$ that reflects its discrimination power:

$$w(e,f) = \begin{cases} \frac{1}{c(e,f)} & \text{if } c(e,f) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Now, we consider f to be a translation $t_L(e)$ for e in a target language $L \in \mathcal{L}$, where $\mathcal{L}$ stands for the set of target languages. The score of a candidate sense $s \in S(e)$ is then the sum of weights of the translations that are found in the corresponding BabelNet synset $BN(s)$:

$$score(s) = \sum_{L \in \mathcal{L}} w(e, t_L(e)) \cdot \mathbb{1}_{BN(s)}(t_L(e))$$

where $\mathbb{1}_{BN(s)}(t_L(e))$ is an indicator function that becomes 1 if $t_L(e) \in BN(s)$ and 0 otherwise. As with p_{wsd}, we normalize the scores into a proper probability distribution p_{trans} over the set of senses. To avoid zero values, we perform smoothing by adding a small positive value (a tunable parameter).

Probability p_{freq} represents the sense frequency information for a given lemma and part-of-speech (POS). This information is also used by most WSD systems. For English, we obtain sense frequencies from WordNet, which derives such information from SemCor, a sense-annotated corpus. To handle senses with zero frequency in SemCor, we also apply additive smoothing. To obtain p_{freq} for languages other than English, which lack large, high-quality sense annotated corpora, we use CluBERT (Pasini et al., 2020), the state-of-the-art system for unsupervised sense distribution learning, which applies a clustering algorithm to contextual embeddings from BERT (Devlin et al., 2019). Like our methods, CluBERT is language independent, has no additional training data requirements, and has been successfully integrated into WSD systems to improve their performance.

Figure 2 illustrates how SOFTCONSTRAINT combines the three probability distributions to correct an incorrect sense prediction produced by a base system.

3.3 Contextual Word Embeddings

Recent work has demonstrated the utility of contextual word embeddings for NLP tasks (Peters et al., 2018; Devlin et al., 2019). Accordingly, WSD systems such as SENSEMBERT (Scarlini et al., 2020) take a contextual embedding of the focus word as input, in order to leverage its dense encoding of relevant local information, which may be used to determine the correct sense.

In this section, we propose a method of adding translation information to the input of a WSD system by modifying the contextual embedding of the focus word to reflect its translation. We refer to this method as *t_emb*. Note that this method can be combined with either the HARDCONSTRAINT or SOFTCONSTRAINT methods. Unlike the constraint-based methods, which use translations of the focus word to post-process the output of a WSD system, *t_emb* provides the translation information in the form of an embedding directly as input to the WSD system. Thus, translation information is used as an additional feature to improve sense predictions of the base WSD system.

As before, our approach is to translate the context of the focus word, and use word alignment to identify the translation of the focus word. We compute a contextual embedding of this translation, just as we did for the focus word itself, and then concatenate the two embeddings. This produces a new embedding that can be provided to a base WSD system in place of the focus word embedding alone. However, since not all WSD systems use contextual embeddings, this method is less general, and we only apply it to some of our models and evaluation experiments.

3.4 Translation Alignment

The effectiveness of our approach for improving WSD depends on the correct identification of the word-level translations in each language. Even when the sentential context of the focus word is correctly rendered in another language, both HARDCONSTRAINT and SOFTCONSTRAINT rely on the proper alignment between the source focus word and its translation, which may be composed of multiple word tokens. Although attention weights in some NMT systems may be used to derive word alignment, such an approach is not necessarily more accurate than off-the-shelf alignment tools (Li et al., 2019). Therefore, our approach is to instead identify the word-level translations by performing a bitext-based alignment between the source focus words and their translations.

During development, we found that the accuracy of alignment tools such as FASTALIGN (Dyer et al., 2013) is limited by the size of the aligned bitext, as well as the lack of access to the translation information which is present in BabelNet. To mitigate these issues, we introduce a knowledge-based word alignment algorithm BABALIGN[1] that leverages translation information in BabelNet by post-processing the output of an off-the-shelf word aligner. BA-

[1]Implementation is available at: https://github.com/YixingLuan/BabAlign

BALIGN is shown to be more effective than existing word aligners in downstream tasks such as cross-lingual lexical entailment (Hauer et al., 2020). We first append our translated WSD data to a large lemmatized bitext. We further augment the bitext with the BabelNet translations for all WSD focus words. We then run the base aligner in both translation directions, and take the intersection of the two sets of alignment links. In its final stage, BABALIGN leverages the BabelNet translation pairs again, to post-process the generated alignment.

Algorithm 1 summarizes BABALIGN. The algorithm takes as input a source-language sentence and a target-language sentence, as well as the set of translations for each content word in the source sentence. As BABALIGN is an alignment post-processing algorithm, its input is the alignment of the two sentences from a base aligner.

If a source word w_s is aligned to a word w_t which is one of its translations, the alignment is considered correct. Since a possible translation may be composed of multiple words (e.g., French translation *salle d'audience* for *courtroom*), we attempt to expand a partial alignment by considering the adjacent word tokens. This is achieved by invoking *compound_search*, which takes the aligned token pair (w_s, w_t) and returns the longest sequence of target tokens c such that $bn(w_s)$ contains c, c contains w_t, and c does not contain any target tokens (except w_t) that are aligned by the base aligner. If no such compound is found, *compound_search* simply returns w_t, so no change in the alignment will be made.

On the other hand, if the source word w_s is aligned to a target word which is not among its translations, we invoke *bnlex_search*, which returns the longest sequence of target tokens l such that $bn(w_s)$ contains l, and l does not contain any tokens that are already aligned. Intuitively, this is an attempt to "repair" an incorrect alignment by searching for an unaligned target word which is known to be a translation of w_s. If such an l can be found (i.e. $l \neq None$), the alignment is modified so that w_s is aligned to l.

Algorithm 1 BABALIGN

Input:
list of all source tokens, $\sigma_s = (w_{s1}, \ldots, w_{sl})$
list of all target tokens, $\sigma_t = (w_{t1}, \ldots, w_{tm})$
BabelNet translations, $bn(w_s) = \{l_1, \ldots, l_n\}$

$A \leftarrow$ *BaseAligner*(σ_s, σ_t)
for each aligned word pair $(w_s, w_t) \in A$ **do**
  if $w_t \in bn(w_s)$ **then**
    $c \leftarrow$ *compound_search*(w_s, w_t)
    Modify A such that w_s aligns to c.
  else
    $l \leftarrow$ *bnlex_search*(w_s)
    if $l \neq None$ **then**
      Modify A such that w_s aligns to l.
return A

4 Word Alignment Evaluation

To show the effectiveness of BABALIGN, which combines an existing word aligner with translations from BabelNet, we evaluate the alignment performance using parallel datasets with gold alignment. We employ FASTALIGN as the base aligner. As the evaluation datasets, we use SemCor 3.0[2] and its translations, Multi SemCor (MSC) (Bentivogli and Pianta, 2005) and Japanese SemCor (JSC) (Bond et al., 2012), to evaluate English-Italian and English-Japanese alignment respectively. Both MSC and JSC contain manually annotated gold alignment for a subset of the sense-annotated content words in SemCor. We extract all English, Italian, and Japanese sentence triples where an English token has gold alignments in both the Italian and Japanese sides. We get 639 sentence triples with 2602 aligned tokens. We only evaluate the alignment performance for those 2602 sense-annotated tokens, and do not consider the alignment for other tokens, because our purpose here is to obtain proper translations for test words in the WSD setting.

For SemCor, we continue to use the included tokenization, lemma, and POS information. For MSC and JSC, we do not use the tokenization, lemma, and POS information provided in the data to emulate the setting where we generate translations for monolingual WSD datasets. Instead, for MSC, JSC, and the additional bitexts, we employ morphological taggers to perform pre-processing: TreeTagger (Schmid, 1994) for Italian and MeCab (Kudo, 2005) for Japanese. The additional bitexts that we append to the data are from OpenSubtitles2018 (Lison and Tiedemann, 2016): English-Italian (37.8M sentences) and English-Japanese (2.2M sentences). We evaluate alignment performance in terms of

[2] We use SemCor 3.0 in the Natural Language Toolkit (NLTK) to keep the compatible file format with MSC and JSC.

Method	Data	En-It	En-Ja
FASTALIGN	test data only	80.4	36.0
	+OpenSub	93.3	75.6
	+OpenSub +pairs	93.6	81.9
BABALIGN	+OpenSub +pairs	**94.0**	**91.6**

Table 1: Alignment F-score on English-Italian and English-Japanese bitexts.

whether the lemma of the aligned translation corresponds to the lemma of the manually aligned translation in MSC or JSC.

Table 1 compares the alignment approaches. As expected, the concatenation of a large bitext to the test data (+OpenSub) dramatically reduces the number of errors. The addition of translation pairs from BabelNet (+pairs) yields further gains. BABALIGN itself improves the quality of the alignment on English-Japanese by nearly 10 points. The improvement on English-Italian is smaller, as the alignment between similar languages is easier, and the additional bitext is much larger. Japanese is particularly challenging, not only because it is typologically different, but also due to the frequency of multi-character compounds. The back-off strategy used by BABALIGN effectively leverages possible translations in BabelNet to recover tokenized compounds and missing alignment links. This mitigates the effect of alignment errors on our WSD results, which we describe in the next section.

5 WSD Evaluation

In this section, we first describe the WSD systems that we use in our experiments. We then show how our methods can improve existing WSD systems in the oracle setting for English all-words WSD. Finally, we report the results of the experiments on multilingual WSD and English all-words WSD with automatic translations.

5.1 WSD Systems

There are two main approaches to WSD: supervised and knowledge-based. Supervised systems are trained on sense-annotated corpora and generally outperform knowledge-based systems. On the other hand, knowledge-based systems usually apply graph-based algorithms to a semantic network and thus do not require any sense-annotated corpora. Since it is expensive to obtain manually sense-annotated corpora and such corpora exist mainly in English, it is often impractical to apply supervised systems to multilingual settings. Therefore, for multilingual WSD, knowledge-based approaches are typically employed.

Many effective WSD systems have been proposed; we include here only the systems that we use in our experiments. IMS (Zhong and Ng, 2010) is a canonical supervised WSD system, which uses support vector machines with various lexical features. LMMS (Loureiro and Jorge, 2019) leverages contextual word embeddings and surpasses the long-standing 70% F-score ceiling for supervised WSD. It learns supervised sense embeddings by applying BERT to SemCor, with additional semantic knowledge from WordNet. Among the knowledge-based systems, Babelfy (Moro et al., 2014) applies random walks with restarts to BabelNet to perform WSD and entity linking. Even though Babelfy is based on BabelNet, it does not make direct use of the translation information in BabelNet. Similarly, UKB (Agirre et al., 2014, 2018), which is based on personalized PageRank on WordNet, achieves state-of-the-art performance on English all-words WSD. Finally, utilizing contextual embeddings, SENSEMBERT (Scarlini et al., 2020) learns knowledge-based multilingual sense embeddings obtained by combining representations learned using BERT with knowledge obtained from BabelNet. This yields state-of-the-art results on English nouns WSD and multilingual WSD. We test these systems both without modification, and with the addition of our knowledge-based methods, to measure how much improvement can be obtained by leveraging translations.

5.2 Oracle WSD Experiments

Our first set of experiments aims at estimating the upper limits of our approach in an oracle setting of annotated and aligned bitexts with high-quality human translations.

Experimental Setup Our sense-annotated bitexts are MSC and JSC (Section 4), which contain manual translations of texts from SemCor. As in Section 4, we use 639 sentences with 2602 sense-annotated instances from MSC and JSC. We randomly sample 10% of the instances as the development set. We tune all parameters on the development set, and use the same hyperparameters throughout the experiment.

We employ two knowledge-based WSD systems: Babelfy and UKB. Both systems have variants that take advantage of sense frequency information in

System	Translation	base	hard	soft
Babelfy	IT	50.7	60.3	58.6
	JA		65.8	65.8
	IT+JA		66.7	**68.6**
UKB	IT	58.0	64.1	64.2
	JA		72.0	72.1
	IT+JA		72.2	**73.3**
Babelfy + WN1st	IT	72.6	73.2	**73.6**
	JA		73.1	**73.6**
	IT+JA		73.4	**73.6**
UKB + dict_weight	IT	71.2	73.6	75.4
	JA		78.5	80.0
	IT+JA		77.8	**80.1**

Table 2: WSD F-score on the SemCor test set with Italian and Japanese translations.

WordNet. Babelfy backs off to the WordNet first sense (WN1st) using a fixed confidence threshold, which we set to 0.8 following Moro et al. (2014). UKB uses complete sense frequency distributions, which are referred to as the dictionary weight (*dict_weight*). We use the same parameter settings as Agirre et al. (2018). For fair comparison, when applying SOFTCONSTRAINT to a system variant without sense frequency information, we set γ to 0 to turn off the p_{freq} component.

Results The results in Table 2 demonstrate the efficacy of leveraging translations for WSD. The systems without sense frequency information are boosted by 15-18%, while the systems with full features get up to 9% absolute improvement. Also, SOFTCONSTRAINT consistently outperforms HARDCONSTRAINT. The modest improvement of 1% on Babelfy is due to the base system falling back on the WN1st sense in about 80% of test instances, precluding the use of translations.

Also, we observe that our approach is effective in combining translations from multiple languages. For instance, the F-score of 73.3% for plain UKB with SOFTCONSTRAINT (shown in Table 2) drops to 72.1% with only Japanese translations, to 64.2% with only Italian translations, and to 58.0% with no translations. These results also indicate that translations from a more distant language, i.e., Japanese, work better at discriminating senses.

5.3 Multilingual WSD Experiments

Since our methods are language-independent, we test them on standard multilingual WSD datasets.

Experimental Setup We perform our multilingual WSD evaluation on benchmark parallel datasets in English, Spanish, Italian, French, and German from SemEval-2013 task 12 (Navigli et al., 2013) and SemEval-2015 task 13 (Moro and Navigli, 2015).[3] The datasets contain manual reference translations, but are not word-aligned. We perform experiments in two settings, with either machine or human translations. To obtain automatic translations, we translate the test sets into English using Google Translate[4] because the pre-trained NMT models for test languages are not always available. For manual translations, we use the provided parallel datasets in all languages. For each individual language, we use BABALIGN to obtain translations of the focus word in other languages. We randomly sample 10% of test instances in each dataset to obtain development sets for parameter tuning.

We use two multilingual base WSD systems: IMS and SENSEMBERT. We train IMS on OneSeC (Scarlini et al., 2019), an automatically sense-annotated set of corpora in multiple languages.[5] For SENSEMBERT embeddings, when we integrate the translation embedding (*t_emb*), we concatenate the focus word embedding and its corresponding *t_emb*, as described in Section 3.3. To compute these contextual word embeddings for English translations, we use the 768-dimensional multilingual BERT cased pre-trained model (mBERT). Since both OneSeC and SENSEMBERT are limited to nouns, we follow Scarlini et al. (2019, 2020) in performing the evaluation on nominal instances only.

Since languages other than English lack large sense-annotated corpora, we employ two evaluation settings. In the default setting, sense frequency information is not used, with the parameter γ set to 0 in SOFTCONSTRAINT. In the other setting, we approximate sense distributions with CluBERT (Pasini et al., 2020).

Results In Tables 3 and 4, we report the WSD results on SemEval-2013 and SemEval-2015 datasets.[6] Surprisingly, the results with English-only Google Translate (GT) translations are only slightly lower on average than with manual translations from multiple languages. HARDCON-

[3] French and German are in SemEval-2013 only.

[4] https://translate.google.com/

[5] Iacobacci et al. (2016) propose an extended version of IMS that incorporates static English word embeddings; however, we are not aware of any IMS version with contextual word embeddings.

[6] Some combinations are omitted for clarity.

	Method	SE-13 DE	SE-13 ES	SE-13 FR	SE-13 IT	SE-15 ES	SE-15 IT
	base system	72.7	67.8	69.6	68.1	63.0	64.1
GT	soft ($\gamma = 0$)	73.7	71.4	73.3	74.9	65.0	70.8
GT	soft (CluBERT)	72.4	76.8	73.9	75.5	68.2	75.7
Manual	hard	72.0	71.2	74.3	73.4	65.5	70.0
Manual	soft ($\gamma = 0$)	73.5	75.0	**74.6**	**76.2**	65.5	71.1
Manual	soft (CluBERT)	**73.8**	**77.0**	74.5	74.9	**69.1**	**76.5**

Table 3: WSD F-score of IMS (OneSeC) with translations on the nominal instances of the SemEval-2013 and SemEval-2015 datasets, with Google Translate (English) and manual (all languages) translations.

	Method	SE-13 DE	SE-13 ES	SE-13 FR	SE-13 IT	SE-15 ES	SE-15 IT
	base system	76.7	74.7	77.6	70.7	64.4	68.7
GT	soft ($\gamma = 0$)	77.7	80.8	79.4	76.8	65.0	74.1
GT	soft (CluBERT)	78.1	80.4	80.7	78.9	65.7	**78.7**
GT	soft (CluBERT+t_emb)	78.2	80.8	80.9	79.4	65.9	**78.7**
Manual	hard	77.1	80.1	79.3	76.6	63.5	72.8
Manual	soft ($\gamma = 0$)	76.8	**81.9**	80.8	78.3	64.6	73.6
Manual	soft (CluBERT)	76.8	79.2	**81.5**	**79.8**	66.4	**78.7**
Manual	soft (CluBERT+t_emb)	**79.6**	81.4	**81.5**	78.9	**66.6**	**78.7**

Table 4: WSD F-score of SENSEMBERT with translations on the nominal instances of the SemEval-2013 and SemEval-2015 datasets, with Google Translate (English) and manual (all languages) translations.

STRAINT performs well in this set of experiments, as nouns are very well covered by BabelNet.[7] SOFTCONSTRAINT achieves an average improvement of several F1 points on both systems, even without sense frequency information. The best results are obtained using sense frequency estimates from CluBERT, especially when they can be combined with mBERT-based contextual translation embeddings (*t_emb*), neither of which requires manually sense-annotated corpora. We interpret these results as the new state of the art in multilingual WSD based on the consistent improvement over SENSEMBERT.

To evaluate the potential of using translations from a replicable NMT model, we obtain English translations for test words in the SemEval-2013 German dataset with a pre-trained transformer model (Ng et al., 2019) available in the fairseq toolkit (Ott et al., 2019). In this setting, we use only English translations for both constraints and *t_emb*. The results on both WSD systems with the pre-trained model are almost the same as with GT, and slightly better than with English-only manual translations. According to our preliminary analysis, machine translations may sometimes work better because they tend to be more literal, and easier to align with the source focus words. This suggests that our methods can effectively leverage translations from different kinds of sources.

5.4 English WSD Experiments with NMT

In the final set of experiments, we evaluate our methods on standard monolingual benchmark datasets using NMT translations from multiple languages.

Experimental Setup We evaluate on five English all-words datasets: Senseval2, Senseval3, SemEval-2007, SemEval-2013, and SemEval-2015 from the unified framework made available by Raganato et al. (2017). Since these datasets are not accompanied by translations, we automatically obtain the translations from NMT models. We tune parameters on Senseval2, and apply the same parameter settings in all datasets.

We test our methods with four base WSD systems: Babelfy and UKB (knowledge-based), and IMS and LMMS (supervised), trained on SemCor 3.0 provided in Raganato et al. (2017). Our replication experiments match the reported results for these systems ($\pm$0.2% on average). For translations, we employ pre-trained transformer models from the fairseq toolkit: English-French and English-German models from Ott et al. (2018), and an English-Russian model from Ng et al. (2019). We choose French, German, and Russian as target languages due to the availability of pre-trained models. Note that unlike multilingual WSD experiments (Section 5.3), we do not use Google Translate in the following experiments. We compare plain Babelfy and UKB to SOFTCONSTRAINT without p_{freq}. For other systems, we derive p_{freq} from sense frequency information from WordNet 3.0.[8]

Results Table 5 shows the results on the standard English all-words WSD datasets. While HARDCONSTRAINT is not sufficiently robust to improve complex WSD systems with automatically generated translations, SOFTCONSTRAINT shows statistically significant improvements over the original performance for all base systems.

[7] Over 99% of the words in BabelNet are nouns (Navigli and Ponzetto, 2012a). On average, 92% of the SemEval translations are in the BabelNet synsets of the correct senses.

[8] Due to the complexity of transforming mBERT representations into different dimensionalities and vector spaces, translation embeddings are not used in these experiments.

	System	Method	SE-2	SE-3	SE-07	SE-13	SE-15	ALL
Knowledge-based	WN1st sense baseline	-	66.8	66.2	55.2	63.0	67.8	65.2
	Babelfy	base system	50.2	46.4	38.9	55.6	54.3	50.3
		hard	53.0*	49.2*	41.7*	55.6	55.9*	52.3*
		soft ($\gamma = 0$)	**57.7***	**54.3***	**47.0***	**60.1***	**61.8***	**57.3***
	UKB	base system	64.2	54.8	40.0	**64.5**	64.5	60.4
		hard	65.3*	57.4*	44.0*	62.6	66.2*	61.5*
		soft ($\gamma = 0$)	**67.6***	**58.8***	**48.6***	**64.5**	**71.1***	**64.0***
	Babelfy + WN1st	base system	66.6	65.5	53.0	63.0	**68.5**	64.9
		hard	66.7	65.5	53.4	62.7	**68.5**	64.9
		soft	**67.4***	**65.9**	**54.3***	**63.4**	68.3	**65.4***
	UKB + dict_weight	base system	68.8	66.1	53.0	68.8	70.3	67.3
		hard	68.5	65.5	53.6	64.5	69.7	66.1
		soft	**71.3***	**66.8**	**54.1**	**69.0**	**74.2***	**68.9***
Supervised	IMS	base system	71.3	**69.1**	**61.5**	65.1	68.3	68.3
		hard	71.0	68.2	60.7	62.0	67.6	67.1
		soft	**72.3**	68.7	59.8	**65.8**	**71.7***	**69.0***
	LMMS	base system	76.3	75.4	67.9	75.0	76.9	75.3
		hard	75.9	74.1	66.2	70.9	75.7	73.6
		soft	**77.2**	**77.1***	**69.2**	**76.1**	**77.2**	**76.4***

Table 5: English all-words WSD F-score on standard evaluation datasets with translations from 3 languages (French, German, and Russian). The results show statistically significant improvement over the base system are marked with * (McNemar's Test, $p < 0.05$).

method	trans	SE-2	SE-3	SE-07	SE-13	SE-15	ALL
base	-	68.8	66.1	53.0	68.8	70.3	67.3
soft	FR	70.0	67.9	54.5	67.6	70.6	68.0
	DE	70.2	66.4	**55.4**	67.5	71.3	67.8
	RU	69.6	66.6	53.4	68.7	71.7	67.9
	FR+DE+RU	**71.3**	**66.8**	54.1	**69.0**	**74.2**	**68.9**

Table 6: WSD F-score of UKB + dict_weight with translations from a single language only.

For example, UKB with *dict_weight* correctly predicts the sense of "earth" in *"the* **world***'s most influential countries."* However, English *world* and its three translations, *monde*, *Welt*, and *mir*, are only found in the BabelNet synset glossed as "populace", while the Russian translation *mir* happens to be missing from the BabelNet synset glossed as "earth", perhaps because there is no Russian link to the English Wikipedia page for *World*. Hence, while HARDCONSTRAINT miscorrects the UKB prediction to the sense of "populace", SOFTCONSTRAINT keeps it unchanged by leveraging sense frequencies and the base system scores.

In Table 6, we show additional results on UKB with *dict_weight* when using only a single language to derive translations. All three languages show similar improvements, and we can obtain better improvements by combining multiple languages.

In summary, these results again demonstrate that our method can effectively integrate information from the WSD system itself, translations, and sense frequency even with noisy translations generated by NMT models. While translations are shown to help even strong supervised WSD systems, the improvements are particularly impressive on knowledge-based systems. The SOFTCONSTRAINT result on UKB with *dict_weight* sets a new state of the art for knowledge-based systems.

6 Conclusion

We proposed a novel approach that improves WSD by leveraging translations from multiple languages, which incorporates a knowledge-based bitext alignment. We tested our methods with several base WSD systems. We demonstrated experimentally that SOFTCONSTRAINT can consistently improve WSD performance even when no manual translations are available, leading to state-of-the-art results on multilingual and English all-words WSD. We make the source code available at https://github.com/YixingLuan/translations4wsd.

Acknowledgments

We thank Tommaso Pasini for the assistance with the multilingual WSD datasets. This research has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Alberta Machine Intelligence Institute (Amii), Amii Fellow Program, CIFAR AI Chair Program, and AltaML.

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.

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 29–33.

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 Association for Computational Linguistics (EACL-09)*, pages 77–85.

Marianna Apidianaki and Li Gong. 2015. LIMSI: Translations as source of indirect supervision for multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 298–302.

Carmen Banea and Rada Mihalcea. 2011. Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, pages 25–34.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27.

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.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 255–262.

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.

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 14th International Workshop on Semantic Evaluation*.

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

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*.

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.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303, Florence, Italy. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691.

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 *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, pages 240–243.

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 (SemEval 2015)*, pages 288–297.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012b. Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1399–1410.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Tommaso Pasini, Federico Scozzafava, and Bianca Scarlini. 2020. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4008–4018, Online. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

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. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the Thirty-Fourth Conference on Artificial Intelligence*, pages 8758–8765. Association for the Advancement of Artificial Intelligence.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New methods in Language Processing*, pages 44–49.

Lucia Specia, Maria GV Nunes, and Mark Stevenson. 2007. Exploiting parallel texts to produce a multilingual sense tagged corpus for word sense disambiguation. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 292:277.

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.