Multilingual Word Sense Disambiguation with Unified Sense Representation

Ying Su[1], Hongming Zhang[1,2], Yangqiu Song[1], Tong Zhang[1]
[1]HKUST, [2]Tencent AI lab, Seattle
ysuay@connect.ust.hk, {hzhangal,yqsong}@cse.ust.hk,
tongzhang@ust.hk

Abstract

As a key natural language processing (NLP) task, word sense disambiguation (WSD) evaluates how well NLP models can understand the lexical semantics of words under specific contexts. Benefited from the large-scale annotation, current WSD systems have achieved impressive performances in English by combining supervised learning with lexical knowledge. However, such success is hard to be replicated in other languages, where we only have limited annotations. In this paper, based on the multilingual lexicon BabelNet describing the same set of concepts across languages, we propose building knowledge and supervised-based Multilingual Word Sense Disambiguation (MWSD) systems. We build unified sense representations for multiple languages and address the annotation scarcity problem for MWSD by transferring annotations from rich-sourced languages to poorer ones. With the unified sense representations, annotations from multiple languages can be jointly trained to benefit the MWSD tasks. Evaluations of SemEval-13 and SemEval-15 datasets demonstrate the effectiveness of our methodology.

1 Introduction

As a critical natural language understanding task, word sense disambiguation (WSD) aims at classifying words into pre-defined senses. With such a disambiguation process, machines can understand the precise meanings of words. Previous researches have demonstrated that a sound WSD system could benefit many downstream NLP tasks, such as machine translation (Pu et al., 2018; Liu et al., 2018) and information extraction (Bovi et al., 2015).

Existing researches on word sense disambiguation mostly focus on English only. By leveraging lexical knowledge such as gloss (Iacobacci et al., 2016; Luo et al., 2018; Huang et al., 2019; Blevins and Zettlemoyer, 2020) or graph structure

---

WSD Example. Language: English
> *Context: Detailed studies of the **plan** were well underway*.
> plan#NOUN: plan%1:09:00:: **plan%1:09:01::**
>   plan%1:06:00::

MWSD Example. Language: French
> *Context: Le groupe des Nations_Unies a des projets de **plans** pour la réduction des émissions*.
> plan#NOUN: **bn:00062759n** bn:00062766n
>   bn:00005439n

Figure 1: Examples of WSD and Multilingual WSD (MWSD) task. The target words are indicated with the bold font in contexts. Candidate sense keys are listed below each context, and the one in blue is the correct sense.

(Banerjee et al., 2003; Kumar et al., 2019; Bevilacqua and Navigli, 2020) and supervised training over large-scale annotations, these models have achieved impressive performance on the standard English WSD task. However, though the English WSD task (Raganato et al., 2017) and multilingual WSD (MWSD) task (Navigli et al., 2013; Moro and Navigli, 2015) are of the same form as shown in Figure 1, this progress can not be easily applied across languages as the paucity of annotated training data and immense labor in handling diverse lexical knowledge of multiple languages separately.

BabelNet (Navigli and Ponzetto, 2012) is a multilingual semantic lexicon and contains a set of multilingually lexicalized concepts. Similar to WordNet (Miller, 1998), a Babel synset defines a concept shared by a group of words across languages with the same meaning. Based on the multilingual lexicon source BabelNet, we propose to build multilingual word sense disambiguation systems by inducing lexical knowledge and annotations from rich sourced language (e.g., English) to scarce sourced ones. First, as defined in BabelNet, words in each synset have the same sense, and the sense is described by lexical knowledge gloss despite the language forms. An example is

Figure 2: BabelNet contains inventories for multiple languages. Each word in a language has several senses, and different words across languages may share the same senses. For each sense across languages, glosses from various sources such as WordNet and Wikipedia are collected to describe its meaning.

shown in Figure 2. The knowledge can be injected into supervised MWSD systems. Second, the annotations acquired from rich sourced languages through machine translation and alignment tools, can be used as weak supervision. By utilizing the lexical knowledge and weak annotations, we can build a decent MWSD system for scarce sourced languages without further human effort.

To summarize, the contributions of this paper are two-fold: (1) We propose to build an MWSD system mBERT-UNI for multiple languages with transferred annotations from rich sourced languages and unified synsets with lexical knowledge, addressing the data paucity problem on the MWSD task; (2) Our system can be easily combined with other data generation efforts such as MuLaN (Barba et al., 2020), further boosting the system performance. Experiments results on benchmark SemEval-13 (Navigli et al., 2013) and SemEval-15 (Moro and Navigli, 2015) demonstrate the effectiveness of our methodology. Our code is open-resourced[1].

2 Related Work

This section introduces previous efforts on multilingual word sense disambiguation, which can be categorized into two streams: data-driven systems and knowledge-based systems.

2.1 Data-driven Systems

In the last decades, many efforts in the field of multilingual word sense disambiguation have been devoted to mitigating the knowledge acquisition bot-

tleneck problem (Gale et al., 1992; Pasini, 2020), which is hard to acquire sense-annotated corpora for multiple languages. To mitigate the paucity of annotations, many researchers have focused on automatically creating high-quality, sense-annotated training corpora (Pasini and Navigli, 2020). OM-SIT (Taghipour and Ng, 2015) proposed a semi-automatic approach to acquire one million training instances from MultiUN dataset (Eisele and Chen, 2010). OneSec (Scarlini et al., 2019) proposed to generate multilingual sense-annotated datasets on a large scale by mapping Wikipedia categories to word senses. MuLaN (Barba et al., 2020) utilized contextualized word embeddings to transfer sense annotations from labeled datasets SemCor (Miller et al., 1993) and WNG (Langone et al., 2004) to the unlabeled corpus from Wikipedia across languages. Hauer et al. (2021) proposed a label propagation approach for constructing multilingual sense-annotated corpora by machine translation. XL-WSD (Pasini et al., 2021) further enriches the annotations across 18 languages from six different linguistic families. Similar to Hauer et al. (2021), our automatic corpora generation method takes advantage of machine translation and alignment tools, while it is easy and feasible to use without additional resources. Moreover, we also induce lexical knowledge in building sense representations.

2.2 Knowledge-based Systems

Besides annotated corpora, lexical knowledge such as sense inventories is another key component in word sense disambiguation systems. Lexical knowledge sources such as WordNet and Babel-Net provide rich lexical knowledge, e.g., gloss or graph structure. Such knowledge has been exploited and shows decent performance in many supervised systems (Kumar et al., 2019; Loureiro and Jorge, 2019; Scarlini et al., 2020; Blevins and Zettlemoyer, 2020). Readers can refer to (Bevilacqua et al., 2021) for more details.

In this paper, we aim to induce lexical knowledge into MWSD systems. Based on the synsets and lexical knowledge in multilingual lexicon BabelNet, we propose to build unified sense representations that can be shared across languages. The sense representations can be incorporated into supervised systems to improve the performance of MWSD tasks.
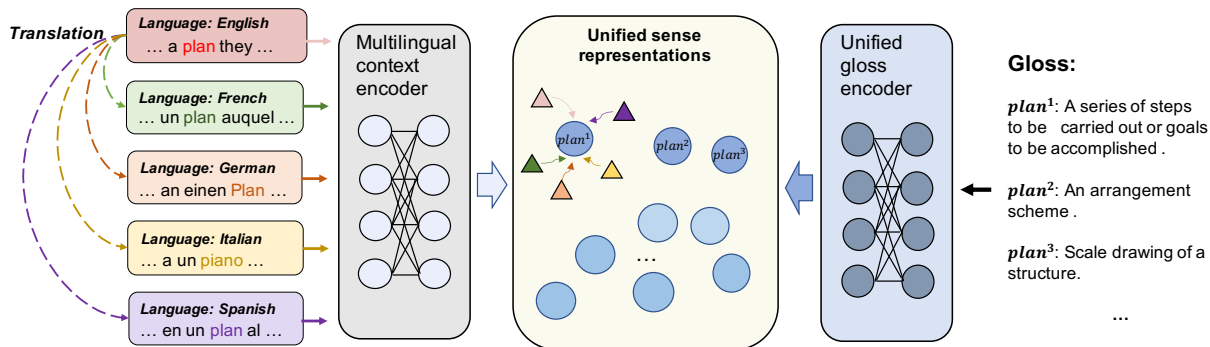
Figure 3: Architecture Overview of the mBERT-UNI model. We apply multilingual BERT for the encoders. The context encoder takes the multilingual context as input and generates representations for the target words. The gloss encoder takes the glosses as input and generates unified sense representations. The similarity scores between target word embedding and representation of candidate senses are calculated for each language. The sense with the highest score is the predicted sense by the model.

3 Approach

In this section, we first present the formal definition of the multilingual WSD task and used notations. After that, we present the details of the proposed system mBERT-UNI, a supervised framework incorporating lexical unified representation space for the MWSD task. From the overview in Figure 3, we can see that mBERT-UNI can be decomposed into four parts: (1) To address the data paucity issue, we first translate the annotated English corpus SemCor into other languages and use alignment tool to generate sense annotations; (2) A context encoder encodes the target words in multilingual context; (3) A gloss encoder encodes the glosses to produce unified sense representations; (4) The annotated corpus in several languages and unified sense representations are bound with a joint training setting.

3.1 Task Description and Notations

In the multilingual setting, the WSD task is to disambiguate the senses for a sequence of words $\{w_1, \cdots, w_m\}$ in a sentence S. The sentences come from various languages $L \in \{L_1, L_2, ..., L_n\}$. For each word w, the goal is to map it to a pre-defined sense $s \in \mathcal{S}_m$, where $\mathcal{S}_m = \{s_1, s_2, ..., s_k\}$ is the set of pre-defined candidate senses for w. The meaning of each sense is defined by the gloss. The candidate senses have a corresponding gloss set defined as $G = \{g_1, g_2, ..., g_k\}$. For the MWSD task, multiple languages have different inventories but share the same set of synsets and glosses as defined in BabelNet (Navigli and Ponzetto, 2012).

3.2 Multilingual Corpora Preparation

We use machine translation and alignment tools to acquire annotated training data for multiple languages. Machine translation has been developed for decades and has achieved remarkable progress (Wu et al., 2016; Tiedemann et al., 2020). Following (Luan et al., 2020), we use google translation to acquire parallel training corpora from English to other languages. Specifically, we translate SemCor (Miller et al., 1993) into the target languages with the Google translation tool[2] (Wu et al., 2016).

There are many research alignment methods to acquire aligned words across languages based on parallel corpora (Dyer et al., 2013; Östling and Tiedemann, 2016; Luan et al., 2020; Dou and Neubig, 2021). We use the early FastAlign tool[3] (Dyer et al., 2013) to align the words across languages for simplicity. Through the process, we propagate the annotations from English to multiple languages. The weak supervision signal in the transferred annotations can be further utilized in supervised systems.

An example of the context translation alignment and sense mapping between English and French is shown in Figure 4. Note that our method is language-independent and thus can be applied to many languages. For evaluating the MWSD system on SemEval-13 and SemEval-15, we apply the method to four languages, German (DE), French (FR), Spanish (ES), and Italian (IT).

3.3 Model Overview

The MWSD system mainly consists of an mBERT-UNI model, which is built upon the biencoder

---

[2]https://translate.google.com
[3]https://github.com/clab/fast_align

Context in English: Workers usually think more of a plan they contribute to .

Aligning target words:

Context in French: Les travailleurs pensent généralement davantage à un plan auquel ils contribuent .
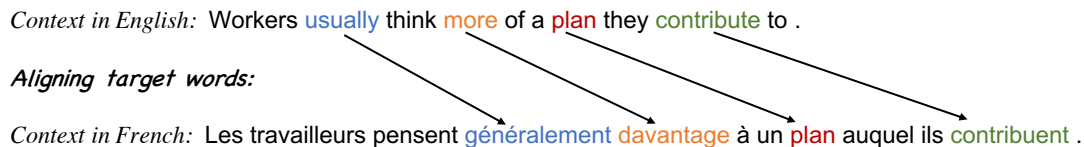
Figure 4: Example of training corpora preparation. The training corpora are first translated into multiple languages by the google translation tool. Then FastAlign tool is applied to find the alignment of labeled target words in English context to other languages. The matched aligned words in different languages are shown in the same color.

model for the English WSD task (Blevins and Zettlemoyer, 2020), where one encoder for encoding multilingual context and the other encoder for encoding unified gloss knowledge. The difference is that we apply the model to construct unified representations which can be used across languages. Both encoders apply the multilingual BERT (mBERT) transformer. mBERT (Kenton and Toutanova, 2019) is trained on 104 languages and is commonly used for cross-lingual semantic representation.

The model first extracts representations of context word and candidate sense representations. For a target word w in the language L, the context encoder generates its representation as e_w^L. Specifically, it is the average pooling of target word tokens in the output.

For sense representations, the gloss set of the corresponding target word is fed as input to the gloss encoder. Though different languages vary in the form of contexts, they share the synsets which can be described by gloss knowledge in a single language, e.g., English. The hidden state of [CLS] token in the gloss encoder output is the vector representation of the gloss. The representation of the candidate gloss set is $\{e_{g_1}, \cdots, e_{g_k}\}$.

The similarity scores between embedding of target word and embeddings of its candidate sense set are calculated by the dot product:

$$score^L(w, g_i) = e_w^L \cdot e_{g_i}, i \in [1, k].$$

The candidate sense with the highest score is the predicted sense produced by the system.

The system is trained with cross-entropy loss over the scores after a softmax activation under a supervised paradigm:

$$p_i = \frac{\exp(score^L(w, g_i))}{\sum_{i=1}^{k} \exp(score^L(w, g_i))},$$

$$loss^L(w) = - \sum_{i=1}^{k} [y_i \log(p_i)],$$

where y_i is 1 if the ith sense is the correct sense otherwise y_i is 0.

Joint Training Setting. As annotations are scarce for low-resourced languages, we further design the joint language setting to see if the unified sense representation can connect annotations across languages to benefit the MWSD task. Under the joint language training setting, the inputs to the context encoder can be from different sources and languages. In contrast, the gloss encoder still generates representation for the Babel synsets.

4 Experiment Setup

In this section, we introduce the experiment details including the evaluation dataset, the evaluation metric, training corpora, baseline methods, and implementation details.

4.1 Dataset and Evaluation Metric

Following the previous work (Barba et al., 2020), we evaluated the systems with the updated version of SemEval-13 and SemEval-15 [4](WordNet split). Specifically, SemEval-13 contains four low-resourced languages: Italian (IT), Spanish (ES), French (FR), and German (DE), and SemEval-15 contains Italian and Spanish. As no development dataset is provided, we randomly sample a small amount of the test instances as a development set for model selection. The instance number and the distribution of word sense number(#sense) on the test dataset are shown in Table 1. The word #sense distribution is calculated separately on word level and instance level. The test instance with the higher word sense is more difficult than those with lower word sense because there are more senses to be disambiguated. The F1 score(%) is used as the evaluation metric.

---

[4]https://github.com/SapienzaNLP/mwsd-datasets

Dataset	Inst num	Word avg	Inst avg
SemEval-13-IT	1,490	3.80	5.51
SemEval-13-ES	1,260	4.20	5.52
SemEval-13-FR	1,449	2.36	3.03
SemEval-13-DE	1,076	1.60	2.17
SemEval-15-IT	1,007	4.38	5.27
SemEval-15-ES	1,043	6.17	6.19

Table 1: Distribution of instance numbers, average word #sense on word level, and average word #sense on instance level for SemEval-13 and SemEval-15 test datasets.

4.2 Training Corpora

We utilize two types of automatically generated training datasets in our experiments, the dataset generated by our proposed translation-based method, and the dataset generated by a label propagation method MuLaN (Barba et al., 2020). Details of the dataset are shown in Table 2.

Translated Corpora: SemCor (Miller et al., 1993) is one of the largest annotated English WSD datasets, which contains 226,036 training instances covering 33,362 senses. We use SemCor 3.0 as the translation source. Due to differences in morphology between languages and inaccuracy brought by the alignment tool, a small amount of the annotated senses in English cannot be transferred to other languages. As a result, we get a comparable number of training instances.

MuLaN: MuLaN is one of the most representative works in automatically constructing the training corpus for the MWSD task. MuLaN has a broader coverage of sense keys as it utilizes the BabelNet inventory than SemCor which utilizes the WordNet inventory. Since we utilize the WordNet split of inventories and evaluation datasets, some of the words in the original MuLaN dataset are not in the used inventories. For fair comparison on mBERT-UNI model by inducing lexical gloss knowledge, we keep the instances with target words existing in the provided inventory, resulting in a filtered dataset MuLaN*.

4.3 Baselines

We compare the proposed mBERT-UNI model with the following baseline methods:

1. **BabelNet S1**: This baseline tags the target word with its most common sense. Following the ranking in BabelNet inventory, the top one

Language	EN	DE	FR	IT	ES
Translated	226k	169k	181k	181k	179k
MuLaN	–	245k	311k	416k	452k
MuLaN*	–	221k	270k	343k	394k

Table 2: Number of training instances for our translation based dataset, original MuLaN dataset, and filtered MuLaN* dataset.

ranked sense is the most common sense (MCS). The left senses are least common sense (LCS). We denote this frequency-based baseline as "BabelNet S1."

2. **mBERT-CLS**: The model is built on mBERT (Kenton and Toutanova, 2019). The pre-trained language model first extracts feature representation for target words in context sentences. On top of the frozen mBERT representation, a linear classifier is trained to classify the senses of target words. The model cannot be used to disambiguate unseen senses from the training dataset. Therefore, the model always predicts the most common sense for unseen senses as a back-off strategy.

4.4 Implementation Details

Both encoders in mBERT-CLS and mBERT-UNI models are initialized with a pre-trained Bert-base-multilingual-uncased model, which has 110M parameters. For both models, we use the Cross-Entropy loss as the training loss, and Adam (Kingma and Ba, 2015) as the optimization algorithm.

For mBERT-CLS, we fed the concatenation of the last four layers' output from mBERT encoder to a linear classifier. As discussed in (Blevins and Zettlemoyer, 2020), finetuning the mBERT-CLS does not improve the performance on the English WSD classification task. Therefore, we keep mBERT frozen and only train the linear classifier during training. The model is trained with a fixed learning rate $2 \cdot 10^{-5}$ for 50 epochs. The training batch size is 128.

For mBERT-UNI, the unified representation are generated from gloss knowledge in English, collecting from BabelNet and WordNet. For each sense key, BabelNet may have several gloss definitions and we select the source from WordNet for simplicity. The whole model is trained with the learning rate 10^{-5} for 20 epochs. We set the batch size at 40. The experiments are run on RTX 2080 and

Model	SemEval-13				SemEval-15	
	IT	ES	FR	DE	IT	ES
BabelNet S1	53.22	60.32	60.04	76.58	45.38	39.31
mBERT-CLS (Trans)	69.53	**70.32**	67.43	67.57	61.67	57.62
mBERT-UNI (SemCor)	65.70	67.14	**78.61**	79.74	67.23	64.91
mBERT-UNI (Trans)	**70.94**	69.68	77.29	**80.48**	**71.00**	**67.11**

Table 3: Results of mBERT-UNI on SemEval-13 and SemEval-15 test datasets. The training corpora generated from translation are briefly denoted as Trans.

the average running time for each experiment is 40 hours. For collecting glosses of word senses, we use BabelNet API [5].

5 Result Analysis

In this section, we analyze the performances of our proposed mBERT-UNI model in two parts. We first introduce the effects of mBERT-UNI on MWSD task with our generated translated corpora. After that, we present further experiment results on the MWSD task under various settings.

5.1 Results of mBERT-UNI

We present the performance of mBERT-UNI and other baseline methods in Table 3. From the results, we can make the following observations:

1. Compared with BabelNet S1, knowledge and learning-based methods (i.e., mBERT-CLS and mBERT-UNI) can perform better in most languages. Such results show that even though we do not have any annotations for these languages, the corpus we translate from English can serve as a strong weak-supervision signal.

2. The only exception is German, in which Babel-Net S1 outperforms mBERT-CLS with translation. As shown in Table 1, this is potentially because words in German typically have much fewer candidate senses than in other languages. As a result, in most instances, simply predicting the most common sense will lead to the correct answer. In this case, the effect of learning is not as significant as in other languages. Even so, by carefully modeling the unified sense representations, the proposed model can still outperform the BabelNet S1 method by a 3.9 % F1 score.

3. Compared with the mBERT-CLS system, the proposed mBERT-UNI model outperforms on

five out of six datasets because of additional lexical knowledge from sense representations with the same translated corpora. Though mBERT-CLS has captured the transferred supervised signal from translated corpora, it is still not enough to disambiguate the senses well. By utilizing lexical knowledge from the unified sense definitions, mBERT-UNI can better disambiguate the word senses under a supervised setting.

4. The translated corpora benefit the MWSD system with external multilingual data. Compared to the mBERT-UNI system trained on original English SemCor and trained on the translated corpora, we can find that the system achieves performance gain on five out of six test datasets. This shows that though the machine translation and alignment tools may induce noise in the corpora preparation process, the resulting multilingual corpora still benefits the system on MWSD tasks. Future work may exploit in the direction of acquiring multilingual corpora of higher quality through automatic methods that can still benefit the system.

5.2 Further Analysis on mBERT-UNI

In this section, we conduct further analysis to show the effects of leveraging an additional corpus MuLaN (Barba et al., 2020) on mBERT-UNI, the effects of joint learning, and the performance on Least Common Sense (LCS). Details are as follows.

5.2.1 Effect of Adding MuLaN Corpora

To see if the knowledge brought by the unified sense representations can be helpful under a supervised paradigm with extra training corpora, we conduct experiments on MuLaN. The results are shown in Table 4.

By incorporating the unified sense representation, previous data generation methods such as Mu-LaN can further boost the performance of MWSD tasks. From the results, we can see even with the

Model	SemEval-13				SemEval-15	
	IT	ES	FR	DE	IT	ES
BabelNet S1	53.22	60.32	60.04	76.58	45.38	39.31
SensEmBERT (Scarlini et al., 2020)	69.80	73.40	77.80	79.20	-	-
OneSeC (Scarlini et al., 2019)	63.45	61.59	65.10	75.84	-	-
MuLaN (Barba et al., 2020)	**77.45**	77.70	80.12	82.09	70.31	68.73
mBERT-CLS (MuLaN*)	69.73	75.87	78.54	82.62	68.82	67.50
mBERT-UNI (MuLaN*)	75.64	**80.24**	81.64	83.27	72.99	**70.47**
mBERT-UNI (Trans+MuLaN*)	76.98	79.44	**82.68**	**83.83**	**74.58**	68.94

Table 4: Results of mBERT-UNI with extra data corpora MuLaN on SemEval-13 and SemEval-15 test dataset. mBERT-CLS (MuLaN*) is the performance on filtered dataset MuLaN*. mBERT-CLS (MuLaN*) is the performance of our implementation with filtered MuLaN as training data. MuLaN is the performance from original paper.

filtered training corpora of smaller size, mBERT-UNI still achieves performance gain over five out of six test datasets compared to MuLaN (Barba et al., 2020). Though the unified sense representations are built based on glosses from the English language only, it can still benefit multiple languages since words share a set of synsets. Future research may continue to find if enriching the sense representations with resources from different languages would still benefit the system.

Moreover, the unified sense representations encoded with the gloss knowledge from BabelNet, are of high quality. SensEmBERT (Scarlini et al., 2020) produced BERT-based sense embeddings by exploiting mostly the semantic relations in BabelNet and Wikipedia for multiple languages separately. Compared with SensEmBERT, our unified sense representation can be simply acquired from the single lexical knowledge source WordNet and even achieves higher performance on the MWSD task.

The mBERT-UNI also supports merging multiple sources of data generation efforts. Combining MuLaN* with our generated dataset, mBERT-UNI can boost the performance on four out of six test datasets. The only exception is Spanish (ES). As shown in Table 1, the test instances in ES are more challenging than in other languages, and thus they are potentially more vulnerable to the noise in the automatically generated training corpora.

5.2.2 Effect of Joint Training

In this section, we conduct experiments to study the effect of the proposed joint learning setting. We are interested in two questions: (1) Whether the joint learning setting can help models solve the MWSD problem or not? (2) Whether the joint learning setting will have a negative effect on English WSD
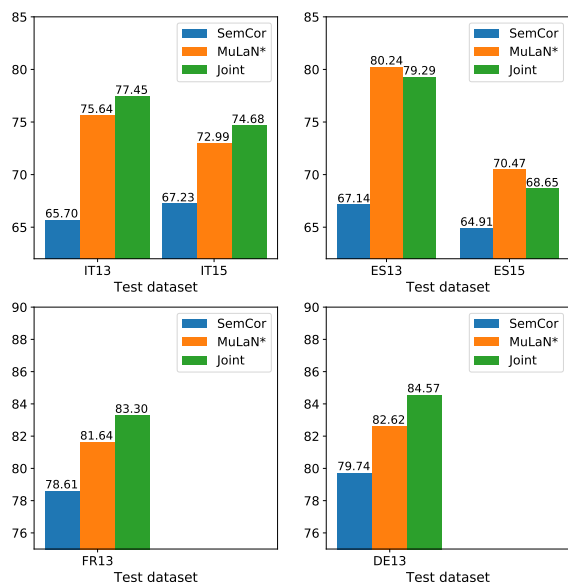

Figure 5: Results of joint learning on the MWSD task. For joint training setting of each language, training data contains SemCor (English) and MuLaN* with the corresponding language part, e.g., SemCor and MuLaN* (Italian) for Italian (*IT13* and *IT15*).

or not. To answer the question, we conduct experiments on training with monolingual datasets and multilingual datasets.

To answer the first question, we present the performances on the MWSD task in Figure 5. We can see that joint learning can achieve better performance on four out of the six datasets and comparable performance on the other two (ES13, ES15). For each language, we combine the MuLaN* with English SemCor as a new training dataset. This result shows that with the unified sense representation, jointly training instances from different languages can improve the annotation usage efficiency across languages. For language ES, the higher

Model	SemEval-13				SemEval-15	
	IT	ES	FR	DE	IT	ES
mBERT-CLS (Trans)	60.02	63.11	57.16	45.58	49.26	53.21
mBERT-UNI (Trans)	62.59	61.92	67.38	59.83	63.90	64.15
mBERT-CLS (MuLaN*)	61.45	68.74	68.36	64.38	59.02	63.61
mBERT-UNI (MuLaN*)	68.24	75.81	72.77	66.38	66.83	63.47

Table 5: Results of Least Common Senses (LCS) on SemEval-13 and SemEval-15 test dataset.

Dataset	SemCor	+IT	+ES	+FR	+DE
ALL	75.70	75.94	75.82	75.76	75.31
Δ	-	+0.24	+0.12	+0.06	-0.39

Table 6: Results of joint learning on the ALL test dataset of English WSD task. *+IT* means that training dataset contains the SemCor (English) and MuLaN* (Italian).

word #sense of the test instances may account for the performance drop under a joint training setting than using MuLaN* alone.

To answer the second question, we report the performance of the mBERT-UNI model trained with English SemCor as well as another joint trained setting on the all-words English WSD datasets proposed by (Raganato et al., 2017). The test dataset "ALL" covers all five datasets, including semeval 2007 (Pradhan et al., 2007), senseval-2 (Palmer et al., 2001), senseval-3 (Snyder and Palmer, 2004), semeval2012 (Navigli et al., 2013), and semeval2015 (Moro and Navigli, 2015). We show the results in Table 6. We can see that the overall performance in English is comparable in different settings. Since the MuLaN dataset is specially designed for other languages and the propagated annotations mainly come from SemCor, the joint training does not benefit the English WSD task much. However, joint training enables a single mBERT-UNI to generate unified sense representations, which can be used in disambiguating word senses in multiple languages. In future work, the unified sense representation may be applied in cross-lingual representation learning.

5.2.3 LCS Analysis

In this section, we analyze the influence of unified sense representation on the performance of the least common senses (LCS). We split the test instances into two parts, one part with annotation of BabelNet S1 and one part with annotations except BabelNet S1 as least common senses. Compared with most common senses, less common ones are more difficult to disambiguate for MWSD systems

because of fewer training instances on average.

We show the performances on the two groups of different systems are shown in Table 5. Comparing mBERT-CLS and mBERT-UNI, adding the sense representations can help improve models' performance significantly on the least common senses. The improvement is consistent on different training corpora for both the translated corpora and MuLaN. It can be concluded that unified sense representation with lexical knowledge improves the ability of deep models to disambiguate the least common senses. This is because mBERT-UNI can still generate and learn unique sense representations for the least common senses even with no or very few training instances. However, while the systems achieve decent performance on the overall performance, disambiguating the lease common senses is still a challenging problem.

6 Conclusion and Future Work

In this paper, to build feasible knowledge and supervised based systems for multilingual word sense disambiguation, we propose to construct unified sense representation by utilizing Babel synsets, and transferred annotations from rich source languages by machine translation and alignment tools. With the unified representations, previous data generation efforts can be combined and further boost the performance. Moreover, annotations from different languages can be jointly trained and benefit the multilingual word sense disambiguation task. Experiments on standard evaluation multilingual word sense disambiguation benchmarks demonstrate the effectiveness of the proposed method.

Future work can be extended on how to induce more lexical knowledge from various languages to improve the representation learning. Moreover, based on the fact that multiple languages share a set of concepts described by Babel synsets, the generated representations may benefit cross lingual representation for other natural language understanding tasks.

References

Satanjeev Banerjee, Ted Pedersen, et al. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Ijcai*, volume 3, pages 805–810. Citeseer.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. Mulan: Multilingual label propagation for word sense disambiguation. In *Proc. of IJCAI*, 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.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *ACL*.

Claudio Delli Bovi, Luis Espinosa Anke, and Roberto Navigli. 2015. Knowledge base unification via sense embeddings and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 726–736.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

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.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *LREC*.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5):415–439.

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. 2021. Semi-supervised and unsupervised sense annotation via translations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 504–513.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing 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 3509–3514.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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 Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 63–69.

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 Alipio 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.

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.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2473–2482.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

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.

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. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217 – 250.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24.

Tommaso Pasini. 2020. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *IJCAI*, pages 4936–4942.

Tommaso Pasini and Roberto Navigli. 2020. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103215.

Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

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 (SemEval-2007)*, pages 87–92.

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.

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 AAAI Conference on Artificial Intelligence*, volume 34, pages 8758–8765.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of SENSEVAL-3, the 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.

Jörg Tiedemann, Santhosh Thottingal, et al. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.