Semantic Specialization for Knowledge-based Word Sense Disambiguation

Sakae Mizuki and **Naoaki Okazaki**
Tokyo Institute of Technology
{sakae.mizuki@nlp., okazaki@}c.titech.ac.jp

Abstract

A promising approach for knowledge-based Word Sense Disambiguation (WSD) is to select the sense whose contextualized embeddings computed for its definition sentence are closest to those computed for a target word in a given sentence. This approach relies on the similarity of the *sense* and *context* embeddings computed by a pre-trained language model. We propose a semantic specialization for WSD where contextualized embeddings are adapted to the WSD task using solely lexical knowledge. The key idea is, for a given sense, to bring semantically related senses and contexts closer and send different/unrelated senses farther away. We realize this idea as the joint optimization of the Attract-Repel objective for sense pairs and the self-training objective for context-sense pairs while controlling deviations from the original embeddings. The proposed method outperformed previous studies that adapt contextualized embeddings. It achieved state-of-the-art performance on knowledge-based WSD when combined with the reranking heuristic that uses the sense inventory. We found that the similarity characteristics of specialized embeddings conform to the key idea. We also found that the (dis)similarity of embeddings between the related/different/unrelated senses correlates well with the performance of WSD.

1 Introduction

Word Sense Disambiguation (WSD) is the task of choosing the appropriate sense of a word from a given sense inventory using contextual information. WSD has proven its usefulness for Information Retrieval (Zhong and Ng, 2012) and Machine Translation (Campolungo et al., 2022). A series of extensive studies has led supervised WSD task performance to surpass the milestone of 80% accuracy (Bevilacqua and Navigli, 2020), which is the estimated human performance (Navigli, 2009).

In contrast, the goal of this study is *knowledge-based WSD*: a variant of WSD that does not rely on supervision data but only on lexical knowledge (e.g., word ontology). This task setting is practically appealing because it does not use a corpus with sense annotations (Bevilacqua et al., 2021), which is costly and labor-intensive to prepare.

A promising approach is based on similarity: to select the sense that is the nearest to a target word in the embedding space (Wang and Wang, 2020). Specifically, a pre-trained language model, typically BERT (Devlin et al., 2019), is used to compute *sense embeddings* for definition sentences. Similarly, a target word is encoded into a *context embedding* for a given sentence. Then, the model predicts the sense of the target word by finding the most similar sense embedding to the context.

The inherent challenge of the similarity-based approach is how we associate two different representations of word meanings, either by definition sentences or by words in context. Although the BERT embeddings capture the coarse-grained word meanings (Reif et al., 2019; Loureiro et al., 2021), there should be room for improvement. Notably, Wang and Wang (2020) proposed `SREF`, sense embedding adaptation by bringing semantically related senses closer. Extending their work, Wang et al. (2021b) proposed `COE`, context embedding enhancement heuristics during inference using the document-level global contexts of the given sentence, and reported the best performance. Despite being effective, `COE` cannot be applied to stand-alone texts, e.g., short messages on social media or search queries, limiting its applicability.

Our study aims to improve both accuracy and applicability to stand-alone texts. Specifically, we propose an adaptation method of the sense and context embeddings for the WSD task solely using lexical knowledge. Then, what are good embeddings for WSD? Our key idea is to 1) bring semantically related sense and context embeddings that convey the same meaning closer, and 2) send unrelated and/or different senses that share the same

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3457–3470

surface form farther away (Fig. 1-d). We formulate the idea as the Attract-Repel objective and self-training objective. The main novelty is the joint optimization to utilize their complementary nature: the former should improve the distinguishability between senses whereas the latter offers pseudo signals of context-sense associations, which has not been explored in previous methods.

The Attract-Repel objective, inspired by Vulic and Mrksic (2018), injects semantic relation knowledge into the similarity of sense pairs. Specifically, we make semantically related senses more similar while making different and unrelated senses more dissimilar (Fig. 1-a). While SREF performs Attract only, our method utilizes both Attract and Repel.

The self-training objective, inspired by the idea of retraining on the classifier's own predictions instead of annotated senses (Navigli, 2009), updates the similarity of context-sense pairs in a pseudo labeling manner (§ 6.1). Specifically, for each training step and given context, we bring the nearest neighbor sense among candidates closer (Fig. 1-b). We also impose distance constraints during adaptation to control the deviation from BERT embeddings (Fig. 1-c) because excessive deviation may cause an inaccurate nearest neighbor sense selection, which would cause a performance drop.

We call the overall proposed method SS-WSD, Semantic Specialization for WSD, following Vulic and Mrksic (2018). We evaluated SS-WSD using the standard evaluation protocol (Raganato et al., 2017) and confirmed that it outperforms the previous embeddings adaptation method. Furthermore, it achieved state-of-the-art (SoTA) performance when combined with the reranking heuristic that uses a sense inventory (Wang and Wang, 2021), and thus is applicable to stand-alone texts.

The contributions of our study are as follows:

- We proposed SS-WSD, an embedding adaptation method that achieves new SoTA in knowledge-based WSD, regardless of the availability of document-level global contexts.
- We found that the performance gain originates from the joint optimization of Attract-Repel and self-training objectives and the prevention of deviation from the original embeddings.
- Empirically, we found that the similarity of related/different/unrelated senses *relative to* the similarity of ground-truth context-sense pairs correlates well with the WSD performance.

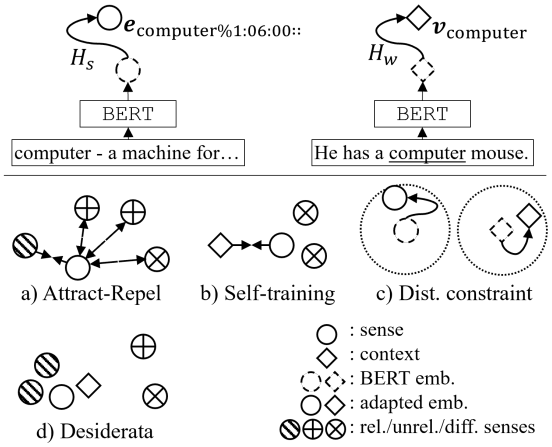

Figure 1: Schema of the proposed method. The BERT embeddings representing senses and contexts are adapted by transformation (top). Transformation functions are optimized using Attract-Repel and self-training objectives under distance constraints so that the adapted embeddings are effective for WSD (bottom).

2 Related Work

2.1 Knowledge-based WSD

Knowledge-based WSD is a variant of WSD that does not use a sense annotation corpora such as the SemCor (Miller et al., 1993) but uses lexical resources instead, typically WordNet. The majority vote based on sense frequencies, also known as the WordNet first sense heuristic (Jurafsky and Martin, 2009), is a simple but strong baseline method of this category. Sense definitions and usage examples are also used to measure the similarity of the target word in a sentence. The simplest method is based on word overlap (Lesk, 1986).

One recent direction is the use of BERT as a contextualized encoder. BERT embeddings showed empirical success on the supervised WSD task when used as features. Some analyses reported that BERT embeddings capture the coarse-grained word meanings (Reif et al., 2019; Loureiro et al., 2021). Wang and Wang (2020) proposed a similarity-based method in the embedding space. It chooses the sense which has the most similar embedding, formed from the concatenation of its lemma, definition, and usage examples, to the embedding of a target word. They also proposed the Semantic Relation Enhancement Framework (SREF_{emb}), which adapts sense embeddings by weighted averages over semantically related senses, e.g., hyponyms and derivations. SREF_{emb} is the most high-performing adaptation method so far. We report that our proposed method achieves better perfor-

mance.

2.2 Heuristics for Knowledge-based WSD

Another recent direction is the heuristics for choosing the most similar sense, which is further divided into those that use the sense inventory information and those that exploit the document-level global contexts of a given sentence. Wang and Wang (2020) proposed the former, the Try-again Mechanism (TaM). It reranks candidates by adding the similarity between the target word and the lexicographer class (supersense) that a candidate sense belongs to. Subsequent studies (Wang et al., 2021b; Wang and Wang, 2021) refined TaM using Coarse Sense Inventory (Lacerra et al., 2020). We examine the effectiveness of the proposed method combined with TaM because it can be applied to stand-alone texts.

Wang et al. (2021b) proposed contextual information enhancement (CIE), which enhances context embeddings by exploiting the document-level global contexts of a given sentence on evaluation. This idea originally stems from the one-sense-per-discourse hypothesis (Gale et al., 1992): that the sense of a word is highly consistent within a document.

2.3 Attract-Repel Framework

The Attract-Repel Framework is used to inject lexical knowledge into embeddings by encouraging similar instances to have closer embeddings while encouraging dissimilar instances to be farther away. Vulic and Mrksic (2018) and Mrkšić et al. (2017) reported that updating static word embeddings using lexical knowledge improves the performance of the word-level semantic relation classification task. Our study proposes its application to sense and context embeddings for the WSD task. We also reformulate the original loss function with the contrastive loss, inspired by its success in Computer Vision (Chen et al., 2020) and NLP (Gao et al., 2021; Wang et al., 2021a; Giorgi et al., 2021).

2.4 Supervised WSD

Supervised methods rely on corpora of sense-annotated contexts, such as SemCor, for training models. However, the coverage of words and senses is limited and biased towards more frequent senses (Pasini, 2020). Recent studies have addressed these limitations by incorporating lexical resources into the methods. Barba et al. (2021a) and its subsequent study (Barba et al., 2021b) reframed WSD as a span extraction task by appending definition sentences of candidate senses to the target context. They reached the SoTA performance among supervised methods.

Similarity-based approaches are also used with supervised methods. Supervised k-nearest neighbors (`Sup-kNN`) (Loureiro and Jorge, 2019) defines sense embeddings as the averaged context embeddings of annotated senses. The Bi-Encoder model (`BEM`) (Blevins and Zettlemoyer, 2020) jointly fine-tunes two BERT encoders for definition sentences and contexts, ensuring that context embeddings will be closer to the correct sense embeddings. The proposed method is similar in architectural design to `BEM`, but differs in that we do not fine-tune the BERT encoders. We will compare our results with `Sup-kNN` and `BEM` to assess the effect of using no sense annotation and of freezing BERT encoders on performance.

3 Semantic Specialization for WSD

3.1 Formalization of WSD

The proposed method adapts BERT embeddings by trainable transformation functions H_s and H_w:

$$\mathbf{v}_w = H_w(\hat{\mathbf{v}}_w), \tag{1}$$

$$\mathbf{e}_s = H_s(\hat{\mathbf{e}}_s), \tag{2}$$

where the inputs $\hat{\mathbf{v}}_w$ and $\hat{\mathbf{e}}_s$ are the context and sense embeddings computed by a BERT encoder and the outputs $\mathbf{v}_w$ and $\mathbf{e}_s$ are the specialized embeddings.

We train the transformation functions by minimizing the weighted sum of the Attract-Repel objective and the self-training objective on the specialized embeddings. Note that the BERT encoder is frozen (not fine-tuned). We integrate the constraints on the distance between the input and output into the architecture of transformation functions (§ 3.4).

To predict a sense for a given target word w, we look up the candidate senses $\mathcal{S}_w$ and compute their specialized sense embeddings using the learned function H_s. Similarly, we compute specialized context embeddings using H_w. Then, we select the nearest neighbor sense s^* using cosine similarity:

$$s^* = \underset{s' \in \mathcal{S}_w}{\arg\max}\, \rho_{w,s'}, \tag{3}$$

$$\rho_{w,s} = \cos(\mathbf{v}_w, \mathbf{e}_s) = \frac{\mathbf{v}_w \cdot \mathbf{e}_s}{\|\mathbf{v}_w\|\|\mathbf{e}_s\|}. \tag{4}$$

Element	Noun	Verb	Adj.	Adv.	All
# Lemmas	117,798	11,529	21,479	4,481	155,287
# Senses	146,320	25,047	30,002	5,580	206,949
Rel. senses	7.8	13.0	6.2	3.9	8.1
Diff. senses	0.8	4.1	1.2	0.7	1.3

Table 1: Summary statistics of lexical resources by part-of-speech tag. Values in the related and different senses rows indicate the average per sense.

3.2 Lexical Knowledge in WordNet

We use WordNet (Fellbaum, 1998) as a lexical resource and sense inventory. WordNet mainly consists of synsets, lemmas, and senses. A synset is a group of synonymous words that convey a specific meaning. A lemma presents a canonicalized form of a word and belongs to one or more synsets. A sense is the lemma disambiguated by a sense key, and belongs to a single synset. We use the sense key as the identifier of a sense.

The proposed method makes use of relational knowledge between senses for training the transformation functions. Specifically, for each sense s, we collect three sets of senses: *related* $\mathcal{S}_s^P$, *different* $\mathcal{S}_s^N$, and *unrelated* $\mathcal{S}_s^U$. The *related* set consists of sense keys of synonyms and semantically related senses (e.g., hyponyms) to the target sense. We followed the definition of related senses used in Wang and Wang (2020) (Appendix A). The *different* set consists of sense keys sharing the same lemma to the target sense excluding itself. In other words, the different senses correspond to the polysemy of the lemma of the target sense. The *unrelated* set presents sense keys that are randomly chosen from the sense inventory (see § 3.5.1 for details). Table 1 shows the statistics of lemmas and senses. See Table 6 (in Appendix A) for examples of the concepts explained in this subsection.

3.3 BERT Embeddings for Sense and Context

For obtaining BERT embeddings, we follow the standard practice of the previous studies (Wang et al., 2020; Bevilacqua and Navigli, 2020; Wang and Wang, 2020). Specifically, we use `bert-large-cased`[1] with special tokens `[CLS]` and `[SEP]`. For each subword, we compute a sum over outputs at the last four layers of Transformer blocks.

A context embedding is the average of BERT embeddings over constituent subwords. For the computation of sense embeddings, we follow the method that Wang and Wang (2020) used. See Appendix B for details.

3.4 Transformation Functions

The proposed method adapts embeddings by applying the trainable transformation, i.e., the specialization is learned by optimizing the transformation functions. This approach enables the adaptation of context embeddings on the fly during inference, which was not possible in the original approach that directly learns adapted embeddings (Vulic and Mrksic, 2018).

Let $\hat{\mathbf{v}}_w$ and $\hat{\mathbf{e}}_s$ be context and sense BERT embeddings. We transform them independently using residual mapping functions F_w and F_s, which are both two-layer feedforward networks, FFNN_w and FFNN_s. These networks are comprised of a linear layer with a ReLU activation, followed by a linear layer with a sigmoid activation.

$$\mathbf{v}_w = H_w(\hat{\mathbf{v}}_w) = \hat{\mathbf{v}}_w + \epsilon\|\hat{\mathbf{v}}_w\|F_w(\hat{\mathbf{v}}_w), \quad (5)$$

$$\mathbf{e}_s = H_s(\hat{\mathbf{e}}_s) = \hat{\mathbf{e}}_s + \epsilon\|\hat{\mathbf{e}}_s\|F_s(\hat{\mathbf{e}}_s), \quad (6)$$

$$F_w(\hat{\mathbf{v}}_w) = 2\sigma(\mathrm{FFNN}_w(\hat{\mathbf{v}}_w)) - 1, \quad (7)$$

$$F_s(\hat{\mathbf{e}}_s) = 2\sigma(\mathrm{FFNN}_s(\hat{\mathbf{e}}_s)) - 1, \quad (8)$$

where $\mathbf{v}_w$ and $\mathbf{e}_s$ are the specialized embeddings. ϵ is the hyperparameter that controls how far away the specialized embeddings can be. Specifically, the L2 distance relative to the original embedding $\|\mathbf{v}_w - \hat{\mathbf{v}}_w\|/\|\hat{\mathbf{v}}_w\|$ is bounded by $\epsilon\sqrt{N_d}$, where N_d is the dimension size of embeddings[2]. This is because the residual functions map the inputs to the space $[-1, +1]^{N_d}$.

3.5 Objectives

We jointly optimize the Attract-Repel objective for sense pairs and the self-training objective for context-sense pairs by minimizing the weighted sum of the loss functions,

$$L = L^{\mathrm{AR}} + \alpha L^{\mathrm{ST}}, \quad (9)$$

where α is the hyperparameter that determines the relative importance of the self-training objective.

The joint optimization is motivated by the complementary nature of these two objectives. The Attract-Repel objective should improve the separability of similar/different senses but does not contribute to determining which context and sense should be associated. In contrast, the self-training objective provides pseudo-supervision signals for

[1] We use `transformers` package (Wolf et al., 2020).

[2] $N_d = 1,024$ for `bert-large-cased`.

context-sense associations, although the informativeness is, when used alone, limited because it essentially reinforces the similarity to the initial nearest neighbor sense of the target context (§ 3.5.2).

3.5.1 Attract-Repel Objective

We formulate Attract-Repel objective loss L^{AR} using contrastive loss: we bring *related* senses closer while *different* and *unrelated* senses farther away[3] (§ 3.2). Specifically, for a given minibatch of senses $\mathcal{S}^B$ and a specific sense $s \in \mathcal{S}^B$, we define the subset excluding itself $\mathcal{S}^B \setminus \{s\}$ as the unrelated senses $\mathcal{S}_s^U$. Then, we randomly choose a sense s_p from the related senses $\mathcal{S}_s^P$. Similarly, we randomly choose up to five senses without replacement $\tilde{\mathcal{S}}_s^N$ from different senses $\mathcal{S}_s^N$. Finally, L^{AR} for the minibatch $\mathcal{S}^B$ is defined as follows:

$$L^{\mathrm{AR}} = -\sum_{s \in \mathcal{S}^B} \ln \frac{e^{\beta \rho_{s,s_p}}}{\sum_{s' \in \left(\{s_p\} \cup \mathcal{S}_s^U \cup \tilde{\mathcal{S}}_s^N\right)} e^{\beta \rho_{s,s'}}}, \quad (10)$$

$$\rho_{s,s'} = \cos(\mathbf{e}_s, \mathbf{e}_{s'}). \quad (11)$$

We set the scaling parameter β to 64, following the suggestions in metric learning studies (Deng et al., 2019; Wang et al., 2018).

3.5.2 Self-training Objective

We formulate the self-training objective loss L^{ST} so that we bring the contexts and nearest neighbor senses closer. In the self-training process, we label a word in context with the sense whose embedding is the closest to that of the word. Specifically, let $\mathcal{W}^B$ denote a minibatch of words. For a word $w \in \mathcal{W}^B$, we obtain a set of candidate senses[4] $\mathcal{S}_w$. Then, L^{ST} for the minibatch $\mathcal{W}^B$ is defined as,

$$L^{\mathrm{ST}} = \sum_{w \in \mathcal{W}^B} (1 - \max_{s \in \mathcal{S}_w} \rho_{w,s}), \quad (12)$$

$$\rho_{w,s} = \cos(\mathbf{v}_w, \mathbf{e}_s). \quad (13)$$

Note that the nearest neighbor sense for the same context changes during training as we update parameters of the transformation functions for embeddings. Our intention is to bootstrap the performance, which was impossible in the "static counterpart", e.g., pseudo-labeling with the WordNet first sense heuristic. That is also a motivation of introducing the distance constraint in Eq. 5 and 6: we were concerned about the performance drop when a large deviation occurs in the semantic specialization. We report empirical evidence that the constraint improves the performance (§ 6.3).

In principle, the training data can be any corpus annotated with lemmas and part-of-speech tags. Nevertheless, we used the SemCor (Miller et al., 1993) corpus with the sense annotations removed. This is because using these de-facto standard corpora contributes to better reproducibility and fairer comparisons.

3.6 Try-again Mechanism (TaM) Heuristic

We examine the effectiveness of the proposed method when combined with TaM. Specifically, we employ the variant (Wang and Wang, 2021)[5] that utilizes Coarse Sense Inventory (CSI) (Lacerra et al., 2020) because of its simplicity. In essence, TaM reranks candidate senses by updating similarities under the assumption that the context should be also similar to the coarse semantic category that the candidate sense belongs to. Let s_1 and s_2 be the top two nearest neighbors for the target word w and $\mathcal{S}_s^{\mathrm{CSI}}$ be the set of senses[6] belonging to the same CSI class as s belongs to. Then, we refine the similarity $\rho_{w,s}^+$ for each $s \in \{s_1, s_2\}$,

$$\rho_{w,s}^+ = \rho_{w,s} + \max_{s' \in \mathcal{S}_s^{\mathrm{CSI}}} \rho_{w,s'}. \quad (14)$$

Finally, we choose the sense from s_1 and s_2 with highest similarity using $\rho_{w,s}^+$, i.e., we use the refined similarity $\rho_{w,s}^+$ instead of $\rho_{w,s}$ (Eq. 3).

4 Experiment Settings

4.1 Training

We used WordNet senses for optimizing the Attract-Repel objective and the sense-annotated words in the SemCor corpus for the self-training objective. Note that we solely use lemmas and part-of-speech tags and disregard the sense annotations. The number of senses in WordNet is 206,949, and the number of words in the corpus is 226,036. We independently sampled minibatches N_B for each objective. For the Attract-Repel objective, we iterate

[3] In the contrastive learning literature, related, unrelated, and different senses correspond to the positives, weak negatives, and hard negative examples, respectively.

[4] Querying WordNet for a tuple of lemma and part-of-speech tag returns the candidate senses.

[5] We followed author's implementation: https://github.com/lwmlyy/SACE

[6] $\mathcal{S}_s^{\mathrm{CSI}}$ will be the empty set if s doesn't exist in the CSI because it does not cover all synsets.

over all sense keys in the WordNet with 15 epochs[7]. For hyperparameter optimization, we disabled TaM heuristics and used the evaluation set of SemEval-2007 as the development set, following the standard practice (Pasini et al., 2021). See Appendix C for details of the hyperparameter search. We set $N_B = 256$, $\alpha = 0.2$, and $\epsilon = 0.015$. We used the Adam optimizer with learning rate 0.001.

4.2 Evaluation

For evaluation, we used the WSD unified evaluation framework (Raganato et al., 2017)[8]. We used the nearest neighbor sense as the prediction (Eq. 3). For the evaluation metric, we adopt the micro-averaged F1 score[9] that is commonly used in the literature. Unless otherwise specified, we run the training process five times with different random seeds, and report the mean and standard deviations.

4.3 Baselines

We compare the proposed method in two experimental configurations: *Intrinsic* and *With Heuristics*. For the *Intrinsic* configuration, we compare it with the methods that do not use any heuristic. Specifically, we choose `PlainBERT` and $\texttt{SREF}_{\texttt{emb}}$ (Wang and Wang, 2020) as baselines. `PlainBERT` uses BERT embeddings $\hat{\mathbf{v}}_w$ and $\hat{\mathbf{e}}_s$ as is. $\texttt{SREF}_{\texttt{emb}}$[10] adapts sense embeddings so that it brings semantically related senses closer. For the *With Heuristics* configuration, we compare the proposed method with the methods that combine heuristics. Specifically, we choose $\texttt{SREF}_{\texttt{kb}}$ (Wang and Wang, 2020) and `COE` (Wang et al., 2021b) as baselines. $\texttt{SREF}_{\texttt{kb}}$ combines $\texttt{SREF}_{\texttt{emb}}$ with TaM. `COE` also utilizes $\texttt{SREF}_{\texttt{emb}}$, but it employs refined TaM and CIE. `COE` is the current SoTA method on knowledge-based WSD.

We also compare with supervised methods which employ the similarity-based approach to assess the effect of not using sense annotations and of freezing BERT encoders. Specifically, we compare with `Sup-kNN` (Loureiro and Jorge, 2019) and `BEM` (Blevins and Zettlemoyer, 2020) (§ 2.4), which both use SemCor as the trainset. `Sup-kNN` computes sense embeddings as the context embeddings averaged over the annotated senses. `BEM` fine-tunes BERT encoders so that context embeddings and correct sense embeddings are brought closer. We consider `BEM` as the de-facto upper bound of similarity-based approach, given its usage of a supervision signal to fine-tune the BERT encoders.

5 Experimental Results

Table 2 shows the WSD task performance. In both configuration, the proposed method $\texttt{SS-WSD}_{\texttt{emb}}$ outperformed all knowledge-based baselines.

In the *Intrinsic* configuration, $\texttt{SS-WSD}_{\texttt{emb}}$ outperformed $\texttt{SREF}_{\texttt{emb}}$ by 3.9pt, which is as much as a 9.3pt improvement over `PlainBERT`. Looking at the results for each part-of-speech, we observed the largest improvement over $\texttt{SREF}_{\texttt{emb}}$ for verbs (9.0pt). This result reflects the fact that verbs have the richer supervision signal for the Attract-Repel objective because of the largest number of related and different senses (Table 1) for verbs. This suggests that the richer semantic relation knowledge is, the higher performance the proposed method may achieve.

In the *With Heuristics* configuration, $\texttt{SS-WSD}_{\texttt{kb}}$ outperformed `COE` by 0.8pt without using the CIE heuristic, which shows an advantage over the baselines regardless of whether the evaluation sentence is a stand-alone text or in a document. The improvement brought by TaM was 2.2pt. Although $\texttt{SS-WSD}_{\texttt{kb}}$ lagged behind `COE` on the SE07 (SemEval-2007) subset, we think this result is understandable because `COE` also used SE07 for hyperparameter optimization.

When compared to supervised methods, $\texttt{SS-WSD}_{\texttt{emb}}$ outperformed `Sup-kNN` by 1.4pt, while falling behind `BEM` by 4.1pt. The results indicate that the proposed method associates contexts with senses more precisely than the example-based sense embeddings computation using sense-annotated contexts. It also shows the effectiveness of the supervised fine-tuning of BERT encoders in `BEM`, as evidenced through their ablation study (Blevins and Zettlemoyer, 2020).

6 Analysis

6.1 Vanilla BERT Embeddings

The proposed method adapts the BERT embeddings (`PlainBERT`) by transformation. Therefore,

[7] In each epoch, we discarded the remaining examples in the self-training objective trainset once all sense keys have been traversed.

[8] Available at: `http://lcl.uniroma1.it/wsdeval/`

[9] Note that F1 score is equal to Precision and Recall (Pasini et al., 2021) because proposed method predicts a single sense.

[10] We applied their method to `PlainBERT`, consistent with the proposed method, to ensure a fair comparison of the effect of adaptation.

Method	TaM	CIE	By subset					By part-of-speech				All
			SE2	SE3	SE07	SE13	SE15	Noun	Verb	Adj.	Adv.	
Supervised												
Sup-kNN (Loureiro and Jorge, 2019)	×	×	76.3	73.2	66.2	71.7	74.1	—	—	—	—	73.5
BEM (Blevins and Zettlemoyer, 2020)	×	×	79.4	77.4	74.5	79.7	81.7	81.4	68.5	83.0	87.9	79.0
Knowledge-based, *Intrinsic* configuration												
PlainBERT	×	×	67.8	62.7	54.5	64.5	72.3	67.8	52.3	74.0	77.7	65.6
SREF$_{emb}$ (Wang and Wang, 2020)	×	×	70.3	68.0	60.4	74.2	77.4	76.3	53.5	75.2	76.3	71.0
SS-WSD$_{emb}$ (Ours)	×	×	**74.6*** (0.5)	**73.0*** (0.6)	**65.0*** (1.3)	**77.0*** (0.5)	**79.9*** (1.0)	**78.2*** (0.4)	**62.5*** (0.7)	**79.7*** (0.3)	**80.5*** (1.5)	**74.9*** (0.3)
Knowledge-based, *With Heuristics* configuration												
SREF$_{kb}$ (Wang and Wang, 2020)	✓	×	72.7	71.5	61.5	76.4	79.5	78.5	56.6	79.0	76.9	73.5
COE (Wang et al., 2021b)	✓	✓	76.0	74.2	**69.2**	**78.2**	80.9	**80.6**	61.4	80.5	81.8	76.3
SS-WSD$_{kb}$ (Ours)	✓	×	**77.7*** (0.5)	**75.9*** (0.6)	66.5 (1.0)	78.0 (0.5)	**81.6** (0.9)	79.3 (0.3)	**65.7*** (0.8)	**84.9*** (0.4)	**84.2*** (0.8)	**77.1*** (0.3)

Table 2: WSD performance by subset and part-of-speech tag. SS-WSD$_{emb,kb}$ are the proposed methods. Numbers in parentheses represent the standard deviation. Asterisks (*) indicate that the difference to the best baseline is statistically significant at $p < 0.05$ by the Student's t-test (two-tailed test). Checkmarks (✓) in the TaM and CIE columns represent the usage of those heuristics. We bolded the best result among knowledge-based methods in each configuration and underlined the objective for hyperparameter tuning. The scores of BEM, Sup-kNN, SREF$_{kb}$, and COE are taken from the original papers.

Method	WSD (All)
WN1stSense	65.2
PlainBERT	65.6

Table 3: F1 score of BERT embeddings (PlainBERT) and WordNet the first sense heuristic (WN1stSense).

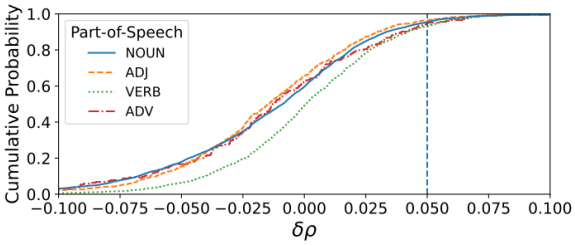

Figure 2: Cumulative distribution of the similarity margin between the incorrect sense and correct sense: $\delta\rho_w = \max_{s' \in \mathcal{S}_w \setminus \mathcal{S}_w^{gt}} \rho_{w,s'} - \max_{s' \in \mathcal{S}_w^{gt}} \rho_{w,s'}$, where $\mathcal{S}_w^{gt}$ is the set of ground-truth senses of the word w.

its performance is influenced by the ability of PlainBERT to disambiguate senses.

Table 3 shows the WSD task performance using PlainBERT. We also reported the WordNet first sense heuristic (WN1stSense) for reference. We observe that PlainBERT is comparable to WN1stSense, indicating that self-training is a more effective strategy than WN1stSense for obtaining pseudo sense labels.

Fig. 2 shows the distribution of the similarity margin (difference) between the nearest neighbor incorrect sense and ground-truth sense computed by PlainBERT. We used the evaluation set for this analysis. We found that the similarity margin is below 0.05 for approximately 90% of all instances. This indicates that a large deviation from PlainBERT is not necessary for replacing nearest neighbor senses with the ground-truth ones.

6.2 Effect of Objectives

Table 4 shows the performance comparison when we eliminate a specific component from the semantic specialization objectives (§ 3.5). We keep all hyperparameters unchanged.

When we exclude either the Attract-Repel objective or the self-training objective, we see the performance drop by 3.3pt and 4.4pt, respectively. This finding supports the claim that joint optimization is crucial for its complementary nature.

When we remove either the unrelated senses or different senses from the Attract-Repel objective, we also see the performance drop by 5.0pt and 1.4pt, respectively. This result supports the idea that bringing semantically unrelated and different senses farther away contributes to performance. We also find that unrelated senses are more effective than different senses. A possible cause is the number of examples: while the number of unrelated

Ablation	WSD (All)	Δ[pt]
SS-WSD$_{\texttt{emb}}$	74.9	—
-Attract-Repel *objective*	71.6	-3.3
-Self-training *objective*	70.5	-4.4
-Unrelated senses $\mathcal{S}^U$ *repelling*	69.9	-5.0
-Different senses $\mathcal{S}^N$ *repelling*	73.5	-1.4
-Context *adaptation*	71.7	-3.2

Table 4: Ablation study of training objective. *Objective* rows represent the corresponding objective is excluded. *Repelling* rows represent the corresponding sense pairs are removed from the Attract-Repel objective (Eq. 10). *Adaptation* rows represent the usage of identity transformation. All differences are statistically significant at $p < 0.05$ by Welch's *t*-test (two-tailed test).

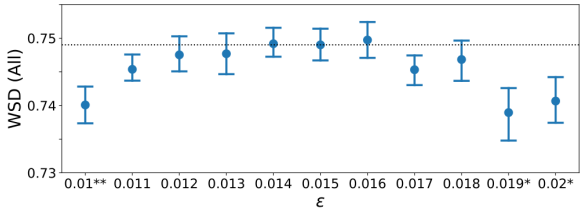

Figure 3: Ablation study of hyperparameter ϵ (§ 3.4). Dot and error bar represent the mean and standard deviation, respectively. Horizontal line represents the default setting ($\epsilon = 0.015$) performance. Asterisks indicate that the difference with respect to the default setting is statistically significant at $p < 0.05$ (*) and $p < 0.005$ (**) by Welch's *t*-test (two-tailed test).

senses is always[11] 255, the number of different senses is, on average, just 1.3 (see Table 1)[12].

Disabling the adaptation of context embeddings (by using identity transformation) caused a performance drop of 3.2pt, indicating that adapting both sense and context embeddings is necessary.

6.3 Effect of Distance Constraint

Fig. 3 shows the performance comparison when we change ϵ, the hyperparameter that bounds how farther away the specialized embeddings can be, in the interval [0.01,0.02] with a step size of 0.001. We found that performance follows an inverted U-shaped curve along ϵ, indicating that a sweet spot exists. Briefly, it shows that a severe constraint (small ϵ) results in an insufficient update for replacing nearest neighbors with ground-truth senses. In contrast, a looser constraint (large ϵ) results in a substantial deviation, eventually making the self-training less effective in the training process. The latter fact supports the claim that controlling the deviation from the original embeddings is necessary.

[11]Minibatch size (=256) minus one yields 255.

[12]In fact, only 38% of all senses have different senses.

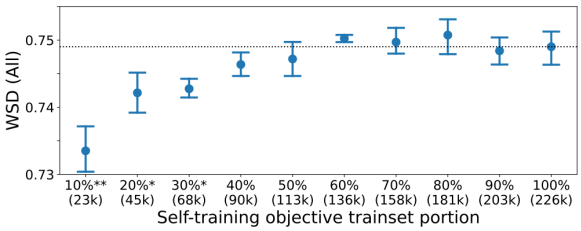

Figure 4: Impact of varying the self-training dataset size from 10% (23k examples) to 100% (224k). The dot and error bar indicates the mean and standard deviation, respectively. The horizontal line represents the performance when utilizing the 100% examples. Asterisks denote that the deviation from the 100% is statistically significant at $p < 0.05$ (*) and $p < 0.005$ (**) by Welch's *t*-test (two-tailed test).

7 Effect of Self-training Dataset Size

Fig. 4 illustrates the impact of varying the number of examples used for the self-training objective on the WSD task performance. It should be noted that 100% in the figure corresponds to using all examples in the SemCor corpus. We found that performance improves as the number of examples increases and reaches a saturation point at 60%, corresponding to 136k examples. While the coverage of words and senses appearing in the contexts also matters, it indicates that the benefits of self-training do not necessarily increase with the scaling to millions of examples.

7.1 Similarity Characteristics

We quantitatively investigate how well the proposed method achieved the key idea (Fig. 1-d): bringing related senses and contexts closer while unrelated and different senses farther away. Specifically, in Table 5, we reported averages of similarity values between related senses $\rho_{\mathcal{S}^P}$, unrelated senses $\rho_{\mathcal{S}^U}$, and different senses $\rho_{\mathcal{S}^N}$, along with averages of similarity values between ground-truth context-sense pairs[13] $\rho_{\mathcal{W}^{\text{gt}}}$. See Appendix D for formal definitions. We found that the proposed method SS-WSD$_{\texttt{emb}}$ brought context-sense pairs closer than PlainBERT ($\rho_{\mathcal{W}^{\text{gt}}}$: 0.64 → 0.77). In contrast, it pushed the unrelated and different senses away: $\rho_{\mathcal{S}^U}$:0.77 → 0.64 and $\rho_{\mathcal{S}^N}$:0.87 → 0.78. These results demonstrate that joint optimization of the Attract-Repel and self-training objectives realized the key idea successfully.

Can we expect better performance if we realize the key idea more precisely? We investigated the

[13]We used sense-annotated words in the evaluation dataset.

Models	$\rho_{\mathcal{S}^P}$	$\rho_{\mathcal{S}^U}$	$\rho_{\mathcal{S}^N}$	$\rho_{\mathcal{W}^{\text{gt}}}$	$\Delta\rho_{\mathcal{S}^P}\uparrow$	$\Delta\rho_{\mathcal{S}^U}\downarrow$	$\Delta\rho_{\mathcal{S}^N}\downarrow$	$\overline{\Delta\rho}\uparrow$	WSD (All)
PlainBERT	0.91	0.77	0.87	0.64	0.27	0.12	0.23	-0.030	65.6
SS-WSD$_{\text{emb}}$	0.88	0.64	0.78	0.77	0.11	-0.13	0.01	0.078	74.9
-Attract-Repel	0.92	0.79	0.90	0.81	0.11	-0.02	0.08	0.014	71.6
-Self-training	0.88	0.64	0.78	0.61	0.27	0.02	0.17	0.027	70.5
-Unrelated senses	0.90	0.73	0.79	0.73	0.17	0.00	0.06	0.033	69.9
-Different senses	0.87	0.61	0.79	0.77	0.09	-0.17	0.02	0.081	73.5
-Context adaptation	0.88	0.64	0.78	0.63	0.25	0.01	0.15	0.032	71.7

Table 5: Similarity characteristics of sense pairs and context-sense pairs. $\rho_{\mathcal{S}^P}$, $\rho_{\mathcal{S}^U}$, and $\rho_{\mathcal{S}^N}$ are the similarity to related, unrelated, and different senses (Eq. 15). $\rho_{\mathcal{W}^{\text{gt}}}$ is the similarity of the context and its ground-truth senses (Eq. 16). $\Delta\rho_*$ is the difference to $\rho_{\mathcal{W}^{\text{gt}}}$ (Eq. 17). $\overline{\Delta\rho} = \frac{1}{3}(\Delta\rho_{\mathcal{S}^P} - \Delta\rho_{\mathcal{S}^U} - \Delta\rho_{\mathcal{S}^N})$. Uparrow↑ (downarrow↓) represents the positive (negative) direction is favorable. WSD (All) are replicated from Tables 2 and 4 for reference.

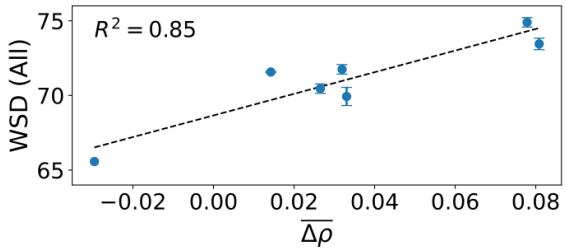

Figure 5: The relationship between the similarity characteristic metric $\overline{\Delta\rho}$ and WSD performance in Table 5.

relationship between these similarity metrics and WSD task performance. Specifically, we subtract $\rho_{\mathcal{W}^{\text{gt}}}$ from each metric in order to capture the closeness of senses *relative to* the correct context-sense pairs, defining $\Delta\rho_*$ as $\rho_* - \rho_{\mathcal{W}^{\text{gt}}}$. For example, $\Delta\rho_{\mathcal{S}^N} = \rho_{\mathcal{S}^N} - \rho_{\mathcal{W}^{\text{gt}}}$ should be a negative value because the average similarity among different senses $\rho_{\mathcal{S}^N}$ should be smaller than that among correct context-sense pairs $\rho_{\mathcal{W}^{\text{gt}}}$. Therefore, we compute the value $\overline{\Delta\rho} = \frac{1}{3}(\Delta\rho_{\mathcal{S}^P} - \Delta\rho_{\mathcal{S}^U} - \Delta\rho_{\mathcal{S}^N})$ to estimate the WSD performance.

Fig. 5 shows that $\overline{\Delta\rho}$ correlates well with WSD task performance ($R^2 = 0.85$). It suggests that if we achieve the key idea more precisely, we may improve the WSD performance. For instance, using a richer lexical relation knowledge, exploitation of the monosemous words, and self-training with confidence thresholding may be promising. We leave it for future work.

8 Conclusion

In this paper, we proposed SS-WSD: Semantic Specialization for WSD[14]. The proposed method learns how to adapt BERT embeddings by transformation and uses the semantic relation knowledge as a supervision signal. The key idea is the desired characteristics of similarities: bringing related senses and the contexts closer while unrelated senses and different senses farther away. We realized it as the joint optimization of the Attract-Repel and self-training objectives while preventing large deviations from original embeddings. Experiments showed that the proposed method outperformed the previous embedding adaptation method. When combined with the reranking heuristic that can be applied to stand-alone texts, it established a new SoTA performance on knowledge-based WSD. The proposed method performs well regardless of the availability of global contexts beyond the target sentence during inference, which the previous study did not achieve. Several analyses showed the effectiveness of the objectives and constraints introduced for specialization. We also found that the closeness of semantically related/different/unrelated senses relative to the closeness of correct context-sense pairs positively correlates with the WSD task performance.

9 Future Work

Given that the proposed method only necessitates lexical resources, it has the potential to effectively address the knowledge acquisition bottleneck problem (Pasini, 2020). Thus, we are interested in applying the proposed method to multilingual WSD using multilingual language models as contextualized encoders. One approach is the zero-shot cross-lingual transfer, which involves learning embeddings adaptation using only English lexical resources. Another option is the joint training of all target languages using multilingual lexical resources such as BabelNet (Navigli et al., 2021). We are also interested in integrating the proposed method into supervised WSD and applying the transfer learning of the specialized embeddings to other NLP tasks.

[14]The source code is available at: https://github.com/s-mizuki-nlp/semantic_specialization_for_wsd

10 Limitations

One limitation of this work is that it is specific to BERT. Although this is in line with the standard practice in previous studies, experimenting with other pre-trained language models is preferred to assess the utility of the proposed method, or to improve the performance further. Another limitation is that it is evaluated on a single dataset and task. While we also followed the de-facto standard protocol, evaluating on rare senses (Maru et al., 2022) or Word-in-Context task (Pilehvar and Camacho-Collados, 2019; Martelli et al., 2021) will bring us more comprehensive insights on the effectiveness and applicability.

11 Ethics Statement

This work does not involve the presentation of a new dataset, nor the utilization of demographic or identity characteristics in formation. In this work, we propose a method for adapting contextualized embeddings for WSD using lexical resources. The proposed method is not limited to a specific resource, we used WordNet as the source of semantic relation knowledge and sense inventory. Therefore, adapted embeddings and sense disambiguation behavior may reflect the incomplete lexical diversity of WordNet in culture, language (Liu et al., 2021), and gender (Hicks et al., 2016).

12 Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19H01118. We thank Marco Cognetta for his valuable input and for reviewing the manuscript.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021a. ESC: redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672. Association for Computational Linguistics.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021b. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

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, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4330–4338.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1597–1607.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

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.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

William A. Gale, Kenneth Ward Church, and David Yarowsky. 1992. One sense per discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (Volume 1: Long Papers)*, pages 879–895.

Amanda Hicks, Michael W. Rutherford, Christiane Fellbaum, and Jiang Bian. 2016. An analysis of wordnet's coverage of gender identity using twitter and the national transgender discrimination survey. In *Proceedings of the 8th Global WordNet Conference*, pages 123–130.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*, chapter 20. Prentice-Hall, Inc.

Caterina Lacerra, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. CSI: A coarse sense inventory for 85% word sense disambiguation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, pages 8123–8130.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986*, pages 24–26.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

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 Conference of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5682–5691.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and José Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Comput. Linguistics*, 47(2):387–443.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mcl-wic). In *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021*, pages 24–36.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of word sense disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, USA, March 21-24, 1993*.

Nikola Mrkšić, Ivan Vulić, Diarmuid O Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10:1–10:69.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of babelnet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4559–4567. ijcai.org.

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

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. XL-WSD: an extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021*, pages 13648–13656.

Mohammad Taher Pilehvar and José Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273.

Alessandro Raganato, José 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.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 8592–8600.

Ivan Vulic and Nikola Mrksic. 2018. Specialising word vectors for lexical entailment. 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 1134–1145.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021a. CLINE: contrastive learning with semantic negative examples for natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (Volume 1: Long Papers)*, pages 2332–2342.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6229–6240.

Ming Wang and Yinglin Wang. 2021. Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5218–5229.

Ming Wang, Jianzhang Zhang, and Yinglin Wang. 2021b. Enhancing the context representation in similarity-based word sense disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8965–8973.

Yinglin Wang, Ming Wang, and Hamido Fujita. 2020. Word sense disambiguation: A comprehensive knowledge exploitation framework. *Knowledge Based System*, 190:105030.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 2012. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–282.

A Lexical Resources

Table 6 shows an example of lexical resources for a sense key `computer%1:06:00::`. Note that unrelated senses are randomly chosen in practice.

For *related* senses lookup, we followed Wang and Wang (2020)'s paper and implementation[15]. Briefly, for a given sense key, we collect the synsets that encompass either itself or the sense keys connected by `derivationally_related_forms` relation. Then, for each collected synset, we extend the synsets via semantic relations shown in Table 7. Finally, we collect the sense keys that belong to either one of the synsets in the extended set of synsets, together with those connected to a given sense key by semantic relations shown in Table 7. We used the `nltk.corpus.wordnet` package for implementation.

B BERT Embeddings for Sense

For the computation of sense embeddings, we followed Wang and Wang (2020)'s method. Specifically, for a given sense key, we generate a sentence by filling in the following template using the lemma, synset lemmas, definition, and examples:

```
[lemma] - [syn. lemma 1], ...,
[syn. lemma n] - [definition]
[example 1] ... [example m],
```

where `n` and `m` represent the number of synonym lemmas and the number of examples. Then we take the average over all subwords in a sentence. For example, applying the template to the sense `computer%1:06:00::` will produce the following sentence.

> computer - computer, computing device, data processor, ... - a machine for performing calculations automatically

We solely use the examples available in WordNet Gloss Corpus and do not use the augmented examples that Wang and Wang (2020) collected.

C Hyperparameter Search

For the hyperparameter search, we first jointly optimized on the number of minibatches N_B, relative importance between objectives α (Eq. 9), and constraint on the distance from BERT embeddings ϵ (Eq. 3.5). We used `TPESampler` in the `optuna` package (Akiba et al., 2019) for optimization. We run hyperparameter search over $N_B \in \{64, 128, 256, 512, 1024\}$, $\alpha \in [0.1, 10]$, and $\epsilon \in [0.001, 0.1]$. The number of search trials is 210. Then, we ran a grid search on ϵ over the interval in [0.01,0.02] using a step size of 0.001. During hyperparameter search, we observed that 1) large minibatch size of 256 or above doesn't produce any statistically significant difference and 2) α is much less sensitive compared to ϵ.

[15] https://github.com/lwmlyy/SREF

D Analysis of Similarity Characteristics

We quantify the similarity characteristic as the macro average of similarity between senses and the similarity of ground-truth context-sense pairs. Specifically, for a given sense s, we calculate the average similarity to its related senses $\mathcal{S}_s^P$, unrelated senses $\mathcal{S}_s^U$, and different senses $\mathcal{S}_s^N$. Following Attract-Repel objective (§ 3.5.1), we define the minibatch excluding itself as the unrelated senses: $\mathcal{S}_s^U = \mathcal{S}^B \setminus \{s\}$. Then, we take the average over all senses $\mathcal{S}$, yielding the similarity among related senses $\rho_{\mathcal{S}^P}$, unrelated senses $\rho_{\mathcal{S}^U}$, and different senses $\rho_{\mathcal{S}^N}$ as follows:

$$\begin{aligned}
\rho_{\mathcal{S}^P} &= \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{1}{|\mathcal{S}_s^P|}\sum_{s'\in\mathcal{S}_s^P}\rho_{s,s'},\\
\rho_{\mathcal{S}^U} &= \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{1}{|\mathcal{S}_s^U|}\sum_{s'\in\mathcal{S}_s^U}\rho_{s,s'},\\
\rho_{\mathcal{S}^N} &= \frac{1}{|\mathcal{S}^N|}\sum_{s\in\mathcal{S}^N}\frac{1}{|\mathcal{S}_s^N|}\sum_{s'\in\mathcal{S}_s^N}\rho_{s,s'},
\end{aligned} \tag{15}$$

where $\mathcal{S}^N = \{s; |\mathcal{S}_s^N| > 0\}$.

For the similarity of ground-truth context-sense pairs $\rho_{\mathcal{W}^{\mathrm{gt}}}$, we use the pairs of the word and annotated senses in the evaluation dataset (§ 4.2). For a given word w, we calculate the average similarity to its ground-truth senses $\mathcal{S}_w^{\mathrm{gt}}$. Then, we take the average over all words $\mathcal{W}$ as follows:

$$\rho_{\mathcal{W}^{\mathrm{gt}}} = \frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}}\frac{1}{|\mathcal{S}_w^{\mathrm{gt}}|}\sum_{s\in\mathcal{S}_w^{\mathrm{gt}}}\rho_{w,s}. \tag{16}$$

Finally, we define $\Delta\rho_*$ as the difference to $\rho_{\mathcal{W}^{\mathrm{gt}}}$ for each relation types. We also define $\overline{\Delta\rho}$ as the arithmetic average over them while taking favorable positive/negative directions into account.

$$\begin{aligned}
\Delta\rho_{\mathcal{S}^P} &= \rho_{\mathcal{S}^P} - \rho_{\mathcal{W}^{\mathrm{gt}}}\\
\Delta\rho_{\mathcal{S}^U} &= \rho_{\mathcal{S}^U} - \rho_{\mathcal{W}^{\mathrm{gt}}}\\
\Delta\rho_{\mathcal{S}^N} &= \rho_{\mathcal{S}^N} - \rho_{\mathcal{W}^{\mathrm{gt}}}\\
\overline{\Delta\rho} &= \frac{1}{3}(\Delta\rho_{\mathcal{S}^P} - \Delta\rho_{\mathcal{S}^U} - \Delta\rho_{\mathcal{S}^N})
\end{aligned} \tag{17}$$

Element	Example
Sense (sense key)	computer%1:06:00::
Lemma	*computer*
Synset	computer.n.01
Definition sentence	*a machine for performing calculations automatically*
Example	Not Available
Synonym lemmas	*computer, computing device, data processor, ...*
Related senses	computing_device%1:06:00:: (synonym), analog_computer%1:06:00:: (hyponym), compute%2:31:00:: (derivative), ...
Different senses	computer%1:18:00::
unrelated senses (randomly chosen)	goldfish%1:05:00::, chef%1:18:01::, ...

Table 6: Example of WordNet lexical resources used in the proposed method.

Category	Relation names
Sense key	pertainyms, antonyms
Synset	hyponyms, hypernyms, part_holonyms, part_meronyms, member_holonyms, member_meronyms, entailments, attributes, similar_tos, causes, substance_holonyms, substance_meronyms, usage_domains, also_sees

Table 7: WordNet semantic relation names used for collecting related senses.

E Implementation Details

We implemented the transformation functions using PyTorch library[16]. We trained them using single NVIDIA 2080Ti GPU. It took approximately two hours for a single run. We precomputed BERT embeddings for training and evaluation dataset and saved them to temporary files for computation efficiency.

[16] https://pytorch.org/