Non-Parametric Few-Shot Learning for Word Sense Disambiguation

Howard Chen, Mengzhou Xia, Danqi Chen
Princeton University
{howardchen, mengzhou, danqic}@cs.princeton.edu

Abstract

Word sense disambiguation (WSD) is a long-standing problem in natural language processing. One significant challenge in supervised all-words WSD is to classify among senses for a majority of words that lie in the long-tail distribution. For instance, 84% of the annotated words have less than 10 examples in the SemCor training data. This issue is more pronounced as the imbalance occurs in both word and sense distributions. In this work, we propose MetricWSD, a non-parametric few-shot learning approach to mitigate this data imbalance issue. By learning to compute distances among the senses of a given word through episodic training, MetricWSD transfers knowledge (a learned metric space) from high-frequency words to infrequent ones. MetricWSD constructs the training episodes tailored to word frequencies and explicitly addresses the problem of the skewed distribution, as opposed to mixing all the words trained with parametric models in previous work. Without resorting to any lexical resources, MetricWSD obtains strong performance against parametric alternatives, achieving a 75.1 F1 score on the unified WSD evaluation benchmark (Raganato et al., 2017b). Our analysis further validates that infrequent words and senses enjoy significant improvement.[1]

1 Introduction

Word sense disambiguation (WSD) (Navigli, 2009) is a widely studied problem that aims to assign words in text to their correct senses. Despite advances over the years, a major challenge remains to be the naturally present data imbalance issue. Models suffer from extreme data imbalance, rendering learning the long-tail examples a major focus. In the English all-words WSD task (Raganato et al., 2017b), 84% of the annotated words[2] have less than 10 occurrences in the training data and the most frequent sense (MFS) accounts for a large portion of the examples, resulting in a 65.2 test F1 score by simply predicting MFS (Figure 1).

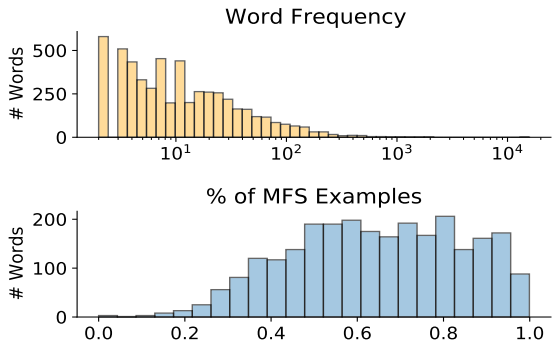

Figure 1: Top: the long-tail distribution of words in the training data (SemCor). Bottom: percentage of examples whose target senses are the most frequent sense for a given word (only words with ≥ 10 examples are considered). All single-sense words are excluded.

Recent approaches tackle this problem by resorting to extra sense information such as gloss (sense definition) and semantic relations to mitigate the issue of rare words and senses (Luo et al., 2018b,a; Kumar et al., 2019; Huang et al., 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020). However, most work sticks to the parametric models that share parameters between words and adopts standard supervised learning mixing all the words of different frequencies. We argue that this accustomed paradigm exposes a missing opportunity to explicitly address the data imbalance issue.

In this work, we propose MetricWSD, a simple non-parametric model coupled with episodic training to solve the long-tail problem, drawing inspiration from few-shot learning methods such as Prototypical Networks (Snell et al., 2017). Given a word, the model represents its senses by encoding a sampled subset (*support set*) of the training data and learns a distance metric between these sense repre-

[1] Our code is publicly available at: https://github.com/princeton-nlp/metric-wsd.

[2] Here we use "word" for simplicity. In WSD datasets, a word is a combination of its stem and part-of-speech tag.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1774–1781
June 6–11, 2021.

sentations and the representations from the remaining subset (*query set*). This lightens the load for a model by learning an effective metric space instead of learning a sense representation from scratch. By sharing only the parameters in the text encoder, the model will trickle the knowledge of the learned metric space down from high-frequency words to infrequent ones. We devise a sampling strategy that takes word and sense frequency into account and constructs support and query sets accordingly. In combination, this non-parametric approach naturally fits in the imbalanced few-shot problems, which is a more realistic setting when learning from a skewed data distribution as in WSD.

We evaluate MetricWSD on the unified WSD evaluation benchmark (Raganato et al., 2017b), achieving a 75.1% test F1 and outperforming parametric baselines using only the annotated sense supervision. A further breakdown analysis shows that the non-parametric model outperforms the parametric counterparts in low-frequency words and senses, validating the effectiveness of our approach.

2 Related Work

Word sense disambiguation has been studied extensively as a core task in natural language processing. Early work computes relatedness through concept-gloss lexical overlap without supervision (Lesk, 1986; Banerjee and Pedersen, 2003). Later work designs features to build word-specific classifiers (*word expert*) (Zhong and Ng, 2010; Shen et al., 2013; Iacobacci et al., 2016). All-words WSD unifies the datasets and training corpora by collecting large scale annotations (Raganato et al., 2017b), which becomes the standard testbed for the WSD task. However, due to the naturally present long-tail annotation, word expert approaches fall short in utilizing information across different words.

Recent supervised neural approaches prevail word-independent classifiers by more effective sentence feature extraction and achieve higher performance (Kågebäck and Salomonsson, 2016; Raganato et al., 2017a). Approaches that use large pre-trained language models (Peters et al., 2018; Devlin et al., 2019) further boost the performance (Hadiwinoto et al., 2019). Recent work turns to incorporate gloss information (Luo et al., 2018b,a; Huang et al., 2019; Loureiro and Jorge, 2019; Blevins and Zettlemoyer, 2020). Other work explores more lexical resources such as knowledge graph structures (Kumar et al., 2019; Bevilacqua and Navigli, 2020; Scarlini et al., 2020b,a). All the above approaches mix words in the dataset and are trained under a standard supervised learning paradigm. Another close work to ours is Holla et al. (2020), which converts WSD into an N-way, K-shot few-shot learning problem and explores a range of meta-learning algorithms. This setup assumes disjoint sets of words between meta-training and meta-testing and deviates from the standard WSD setting.

3 Method

3.1 Task Definition

Given an input sentence $x = x_1, x_2, \ldots, x_n$, the goal of the *all-words* WSD task is to assign a sense y_i for every word x_i, where $y_i \in \mathcal{S}_{x_i} \subset \mathcal{S}$ for a given sense inventory such as the WordNet. In practice, not all the words in a sentence are annotated, and only a subset of positions are identified $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ to be disambiguated. The goal is to predict y_i for $i \in \mathcal{I}$.

We regard all the instances of a word $w \in \mathcal{W}$ as a classification task $\mathcal{T}_w$, since only the instances of word w share the output label set $\mathcal{S}_w$. We define input $\bar{x} = (x, t)$ where x is an input sentence, and $1 \leq t \leq n$ is the position of the target word and the output is y_t for x_t. A WSD system is a function f such that $y = f(\bar{x})$. Our method groups the training instances by word w: $\mathcal{A}(w) = \{(\bar{x}^{(i)}, y^{(i)}) : x^{(i)}_{t^{(i)}} = w\}_{i=1}^{\mathcal{N}(w)}$ where $\mathcal{N}(w)$ is the number of training instances for $\mathcal{T}_w$. It allows for word-based sampling as opposed to mixing all words in standard supervised training.

3.2 Episodic Sampling

We construct episodes by words with a tailored sampling strategy to account for the data imbalance issue. In each episode, all examples $\mathcal{A}(w)$ of a word w are split into a support set $\mathcal{S}(w)$ containing J distinct senses and a query set $\mathcal{Q}(w)$ by a predefined ratio r (splitting $r\%$ into the support set). When the support set is smaller than a predefined size K, we use the sets as they are. This split maintains the original sense distribution of the infrequent words as they will be used fully as support instances during inference. On the other hand, frequent words normally have abundant examples to form the support set. To mimic the few-shot behavior, we sample a *balanced* number of examples per sense in the support set for frequent words (referred to as the P_b strategy). We also compare to the strategy where the examples of all senses of

Algorithm 1 Episodic Sampling

K: maximum sample number for support set
r: support to query splitting ratio
P: sampling strategy $\in \{P_b, P_u\}$
Initialize empty dataset $\mathcal{D} = \emptyset$
for all $w \in \mathcal{W}$ **do**
  Retrieve $\mathcal{A}(w)$ and randomly split $\mathcal{A}(w)$ into $\tilde{\mathcal{S}}(w)$ and $\tilde{\mathcal{Q}}(w)$ with a ratio r.
  if $|\tilde{\mathcal{S}}(w)| \leq K$ **then**
    $\mathcal{S}(w) \leftarrow \tilde{\mathcal{S}}(w)$; $\mathcal{Q}(w) \leftarrow \tilde{\mathcal{Q}}(w)$
  else
    $J \leftarrow$ # of senses in $\tilde{\mathcal{S}}(w)$
    $\tilde{\mathcal{S}}_j(w) \leftarrow$ examples of sense j in $\tilde{\mathcal{S}}(w)$
    for $k = 1 \ldots |\tilde{\mathcal{S}}(w)|$ **do**
      $j \leftarrow$ the sense of k-th example
      $\alpha_k \leftarrow \begin{cases} \frac{1}{|\tilde{\mathcal{S}}_j(w)| \times J}, & \text{if } P = P_b \text{ (balanced)} \\ \frac{1}{|\tilde{\mathcal{S}}(w)|}, & \text{if } P = P_u \text{ (uniform)} \end{cases}$
    $\mathcal{S}(w) \leftarrow \textsc{RandChoice}(\tilde{\mathcal{S}}(w), K, \alpha)$
    $\mathcal{Q}(w) \leftarrow \tilde{\mathcal{Q}}(w) \cup (\tilde{\mathcal{S}}(w) \setminus \mathcal{S}(w))$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{\mathcal{S}(w), \mathcal{Q}(w)\}$
return $\mathcal{D}$

the word are *uniformly* sampled (referred to as the P_u strategy). We present the complete sampling strategy in Algorithm 1.

3.3 Learning Distance Metric

We use BERT-base (uncased) (Devlin et al., 2019) as the context encoder. We follow Blevins and Zettlemoyer (2020) closely and denote context encoding as $f_\theta(\bar{x}) = \texttt{BERT}(x)[t]$ where the context encoder is parameterized by θ. If a word x_t is split into multiple word pieces, we take the average of their hidden representations. In each episode, the model encodes the contexts in the support set $\mathcal{S}(w)$ and the query set $\mathcal{Q}(w)$, where the encoded support examples will be taken average and treated as the sense representations (*prototypes*). For word w, the prototype for sense j among the sampled J senses is computed from the support examples:

$$\mathbf{c}_j = \frac{1}{|\mathcal{S}_j(w)|} \sum_{(\bar{x},y) \in \mathcal{S}_j(w)} f_\theta(\bar{x}), \qquad (1)$$

where $\mathcal{S}_j(w) = \{(\bar{x}^{(i)}, y^{(i)}) : y^{(i)} = j\}_{i=1}^{|\mathcal{S}_j|} \subset \mathcal{S}(w)$. We compute dot product[3] as the scoring function $s(\cdot, \cdot)$ between the prototypes and the query representations to obtain the probability of predicting sense j given an example $(\bar{x}', y')$:

$$p(y = j \mid \bar{x}') = \frac{\exp(s(\mathbf{c}_j, f_\theta(\bar{x}')))}{\sum_k \exp(s(\mathbf{c}_k, f_\theta(\bar{x}')))}. \qquad (2)$$

[3]We experiment with negative squared l_2 distance as suggested in Snell et al. (2017) as the scoring function and find no improvement.

The loss is computed using negative log-likelihood and is minimized through gradient descent. During inference, we randomly sample $\min(I_S, |\mathcal{A}_j(w)|)$ examples in the training set for sense j as the support set, where I_S is a hyperparameter. We also experimented with a cross-attention model which learns a scoring function for every pair of instances, similar to the BERT-pair model in Gao et al. (2019); however, we didn't find it to perform better than the dual-encoder model.

3.4 Relation to Prototypical Networks

Our non-parametric approach is inspired and closely related to Prototypical Networks (Snell et al., 2017) with several key differences. First, instead of using disjoint tasks (i.e., words in our case) for training and testing, MetricWSD leverages the training data to construct the support set during inference. Second, we control how to sample the support set using a tailored sampling strategy (either balanced or uniform sense distribution). This encourages learning an effective metric space from frequent examples to lower-frequency ones, which is different from adapting between disjoint tasks as in the typical meta-learning setup.

4 Experiments

We evaluate our approach with the WSD framework proposed by Raganato et al. (2017b). We train our model on SemCor 3.0 and use SemEval-2007 (SE07) for development and the rest: Senseval-2 (SE02), Senseval-3 (SE03), SemEval-2013 (SE13), and SemEval-2015 (SE15) for testing. Following standard practice, we report performance on the separate test sets, the concatenation of all test sets, and the breakdown by part-of-speech tags. For all the experiments, we use the BERT-base (uncased) model as the text encoder.

Baselines We first compare to two simple baselines: WordNet S1 always predicts the first sense and MFS always predicts the most frequent sense in the training data. We compare our approach to BERT-classifier: a linear classifier built on top of BERT (all the weights are learned together). As opposed to our non-parametric approach, the BERT-classifier has to learn the output weights from scratch. We compare to another supervised baseline using contextualized word representations that extends the input context text with its surrounding sentences in the SemCor dataset (Hadiwinoto et al., 2019). We also compare to a non-parametric nearest neighbor baseline BERT-kNN, which obtains

	Gloss?	Dev	Test Datasets				Concatenation of Test Datasets				
		SE07	SE02	SE03	SE13	SE15	Nouns	Verbs	Adj.	Adv.	ALL
WordNet S1	✗	55.2	66.8	66.2	63.0	67.8	67.6	50.3	74.3	80.9	65.2
Most frequent sense (MFS)	✗	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
Bi-LSTM (Raganato et al., 2017a)	✗	64.8	72.0	69.1	66.9	71.5	71.5	57.5	75.0	83.8	69.9
BERT-kNN	✗	64.6	74.7	73.5	70.3	73.9	74.7	61.6	77.7	85.3	72.6
BERT-classifier	✗	68.6	75.2	74.7	70.6	75.2	74.6	63.6	78.6	**87.0**	73.5
1sent (Hadiwinoto et al., 2019)	✗	67.0	75.0	71.6	69.7	74.4	-	-	-	-	72.7
1sent+1sur† (Hadiwinoto et al., 2019)	✗	69.3	75.9	73.4	70.4	75.1	-	-	-	-	73.7
MetricWSD (ours)	✗	**71.4**	**77.3**	**75.6**	**71.9**	**76.6**	**77.1**	**64.9**	**79.9**	85.3	**75.1**
EWISE (Kumar et al., 2019)	✓	67.3	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1	71.8
GlossBERT (Huang et al., 2019)	✓	72.5	77.7	75.2	76.1	80.4	79.8	67.1	79.6	87.4	77.0
EWISER (Bevilacqua and Navigli, 2020)	✓	71.0	78.9	**78.4**	78.9	79.3	**81.7**	66.3	81.2	85.8	78.3
BEM (Blevins and Zettlemoyer, 2020)	✓	**74.5**	**79.4**	77.4	**79.7**	**81.7**	81.4	**68.5**	**83.0**	**87.9**	**79.0**

Table 1: F1-scores for fine-grained all-words WSD task. We compare our non-parametric model against models without access to gloss information. We take results from EWISER where only supervised data and gloss are used but not WordNet examples. †: surrounding sentences are used as extra context. Our system uses the sampling strategy in Algorithm 1 with $K = 40, r = 0.4, I_S = 30, P = P_b$.

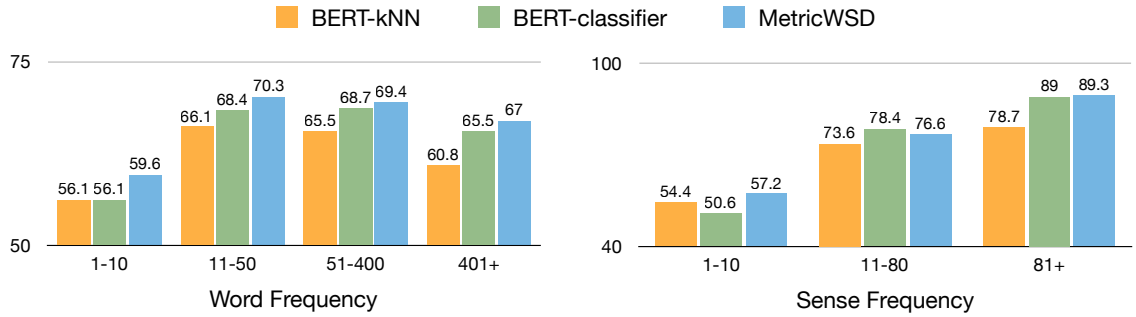

Figure 2: Accuracy breakdown by word frequency and sense frequency on the test set.

sense representations by averaging BERT encoded representations from training examples of the same sense. It predicts the nearest neighbor of the input among the sense representations. The BERT weights are frozen which, different from our approach, does not learn the metric space. Models using only supervised WSD data fall back to predicting the most frequent sense (MFS) when encountering unseen words. For reference, we also list the results of recent state-of-the-art methods that incorporate gloss information including EWISE (Kumar et al., 2019), EWISER (Bevilacqua and Navigli, 2020), GlossBERT (Huang et al., 2019), and BEM (Blevins and Zettlemoyer, 2020). More implementation details are given in Appendix A.

Overall results Table 1 presents the overall results on the WSD datasets. Comparing against systems without using gloss information, MetricWSD achieves strong performance against all baselines. In particular, MetricWSD outperforms BERT-classifier by 1.4 points and BERT-kNN by 2.5 points respectively in F1 score on the test set. Using gloss information boosts the performance by a large margin especially for unseen words, where systems without access to gloss can only default to the first sense. We believe adding gloss has the potential to enhance the performance for our non-parametric approach and we leave it to future work.

Performance on infrequent words and senses The performance breakdown for words and senses of different frequency groups is given in Figure 2. The non-parametric methods (both MetricWSD and BERT-kNN) are better at handling infrequent words and senses. In particular, our approach outperforms BERT-classifier 3.5% for the words with ≤ 10 occurrences and 6.6% for the senses with ≤ 10 occurrences. It demonstrates the effectiveness of MetricWSD to handle scarce examples.

Ablation on sampling strategies We provide an ablation study for the sampling strategy on the development set. The system using the balanced strategy (P_b) achieves a 71.4 F1 on the development set and drops to 69.2 F1 when the uniform strategy (P_u) is used. Balancing the sampled

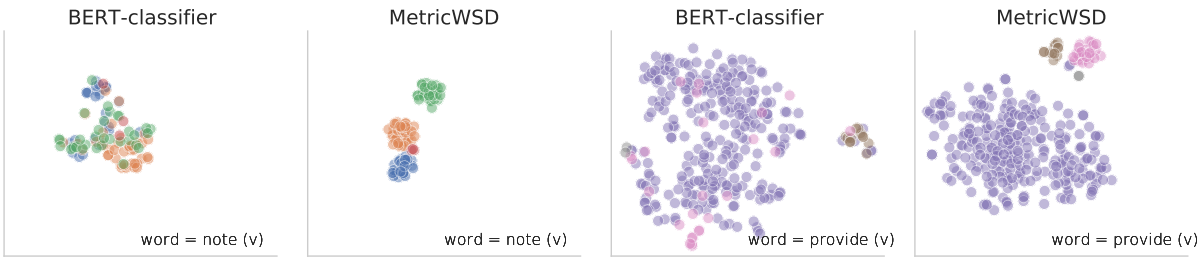

Figure 3: t-SNE visualization of the learned representations $f_\theta(\bar{x})$ for the examples of *note* (v) and *provide* (v) in the SemCor dataset. It shows that MetricWSD is better than BERT-classifier in grouping different senses.

Context	BERT-classifier prediction	MetricWSD prediction
The ***art*** of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world.	art%1:06:00:: (freq = 48) the products of human creativity; works of art collectively	art%1:09:00:: (freq = 6) a superior skill that you can learn by study and practice and observation
Eyes that were ***clear***, but also bring with a strange intensity, a sort of cold fire burning behind him.	clear%3:00:00:: (freq = 45) readily apparent to the mind	clear%3:00:02:: (freq = 4) allowing light to pass through
And, according to ***reports*** from US broadcaster CNBC, Citigroup is also planning to replay the state support.	report%1:10:03:: (freq = 50) a written document describing the findings of some individual or group	report%1:10:00:: (freq = 9) a short account of the news

Table 2: Examples of contexts and model predictions. The bold italic words in the contexts are disambiguated by BERT-classifier and MetricWSD. We present the predicted sense key, the corresponding sense definition, and the sense frequency in the training set.

senses achieves significantly higher performance than sampling with the uniform distribution and this observation is consistent across different hyperparameter settings.

5 Analysis

Qualitative analysis Table 2 shows the examples which are correctly predicted by our method but incorrectly predicted by BERT-classifier. We see that MetricWSD is able to correctly predict the sense *art%1:09:00::* (a superior skill that you can learn by study and practice and observation), which has only 6 training examples. The BERT-classifier model incorrectly predicts the sense *art%1:06:00::* (the products of human creativity; works of art collectively) that has many more training examples.

Visualization of learned representations We conduct a qualitative inspection of the learned representations for the BERT-classifier model and MetricWSD. Figure 3 shows the encoded representations of all 105 examples in the SemCor dataset of the word *note* (with part-of-speech tag *v*). We see that the BERT-classifier model fails to learn distinct grouping of the senses while MetricWSD forms clear clusters. Note that even for the sense (red) with only few examples, our method is able to learn representations that are meaningfully grouped. Similarly, MetricWSD separates senses more clearly than BERT-classifier for the word *provide* (with part-of-speech tag *v*, especially on the rare sense (pink).

6 Conclusion

In this work, we introduce MetricWSD, a few-shot non-parametric approach for solving the data imbalance issue in word sense disambiguation. Through learning the metric space and episodic training, the model learns to transfer knowledge from frequent words to infrequent ones. MetricWSD outperforms previous methods only using the standard annotated sense supervision and shows significant improvements on low-frequency words and senses. In the future, we plan to incorporate lexical information to further close the performance gap.

Acknowledgement

We thank the members of the Princeton NLP group and the anonymous reviewers for their valuable comments and feedback. We also thank Terra Blevins at University of Washington for providing code and checkpoints for the baselines. Both HC and MX are supported by a Graduate Fellowship at Princeton University.

Ethical Considerations

We identify areas where the WSD applications and our proposed approach will impact or benefit users. WSD systems are often used as an assistive submodule for other downstream tasks, rendering the risk of misuse less pronounced. However, it might still exhibit risk when biased data incurs erroneous disambiguation. For example, the word "shoot" might have a higher chance to be interpreted as a harmful action among other possible meanings when the context contains certain racial or ethnic groups that are biasedly presented in training data. Our proposed method does not directly address this issue. Nonetheless, we identify the opportunity for our approach to alleviate the risk by providing an easier way to inspect and remove biased prototypes instead of making prediction using learned output weights that are hard to attribute system's biased behavior. We hope future work extends the approach and tackles the above problem more explicitly.

References

Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

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 *Association for Computational Linguistics (ACL)*, pages 2854–2864.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*, pages 4171–4186.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu1, Peng Li, Maosong Sun, and Jie Zhou. 2019. Fewrel 2.0: Towards more challenging few-shot relation classification. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 5297–5306.

Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation. In *Findings of the Empirical Methods in Natural Language Processing (EMNLP Findings)*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 3500–3505.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Association for Computational Linguistics (ACL)*.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Association for Computational Linguistics (ACL)*, pages 5670–5681.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional lstm. In *International Conference on Computational Linguistics (COLING)*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Conference on Systems Documentation (SIGDOC)*.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In *Association for Computational Linguistics (ACL)*.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1402–1411.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018b. Incorporating glosses into neural word sense disambiguation. In *Association for Computational Linguistics (ACL)*, pages 2473–2482.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Association for Computational Linguistics (NAACL)*, volume 1, pages 2227–2237.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural sequence learning models for word sense disambiguation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1156–1167.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Conference on Artificial Intelligence (AAAI)*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Hui Shen, Razvan Bunescu, and Rada Mihalcea. 2013. Coarse to fine grained sense disambiguation in wikipedia. In *Conference on Lexical and Computational Semantics (SEM)*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4077–4087.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Association for Computational Linguistics (ACL)*.

A Appendix

Implementation Details

When constructing support and query sets, we skip words that contain only one sense or words that have only one example per sense since they cannot be split into meaningful support and query sets (e.g., only two senses and one example each for a word). During inference, all the aforementioned examples are used as supports.

For BERT-classifier, all parameters are fine-tuned during supervised training. We use batch size of 4 sentences, and runs for 20 epochs. Our non-parametric model constructs an episode for each word type, accumulate gradients for 5 episodes before performing gradient update, runs for 100 epochs. All models use the AdamW optimizer with learning rate 1e-5.