Word Sense Disambiguation: A Survey on Recent Methods Since 2020

Li-Kuang Chen
National Tsing Hua University
Hsinchu, Taiwan
lkchen@nlplab.cc

1 Approaches

All the works mentioned are evaluated on Raganato et al. (2017), a unified evaluation benchmark composed of several WSD test datasets.

1.1 As Classification

GlossBERT (Huang et al. 2019) treats WSD as a sentence pair classification problem, where a pair consists of (1) the *context*: the sentence containing the target word, and (2) the *gloss*: one of the definitions of the target word.

At training, each (context, gloss) pair is labelled with binary labels (*yes*, *no*), representing whether the gloss is the correct sense definition for the target word. They append a feed-forward classification layer after BERT's final layer, and fine-tune BERT with the added layer using the sentence pairs.

At inference, the classifier predicts the label (*yes*, *no*) for each (context, gloss) pairs for each gloss of the target word given a context sentence. The gloss with the highest probability for the label *yes* is chosen as the predicted sense.

Their also show that adding "weak supervision", i.e. symbols and tokens marking the target word, helps boost the classifier's performance.

Song et al. (2021) uses a similar architecture to GlossBERT, but they concatenate additional information to the gloss during training, including WordNet examples, synonyms, and hypernym glosses of each gloss synset. They also replace BERT with RoBERTa, reporting that the latter gives much better performance. These modifications allows their method to considerably surpass GlossBERT and is competitive to concurrent SoTAs on not only popular benchmarks, but also unseen words and less frequent senses. (TODO: verify)

Unlike GlossBERT, which encodes the context and the gloss as one single embedding, **BEM** (a bi-encoder model, Blevins and Zettlemoyer (2020)) separately encodes the gloss and the context with two encoders. The dot product of the gloss and the context embeddings are taken as the similarity score, and the gloss with the highest score is chosen as the predicted sense.

In this approach, the two encoders are jointly optimized to embed each input token near the representation of its correct word sense.

RWTE (Zhang et al. 2023) refines the embeddings of target word glosses and the context such that the attention focuses on features in the target word embedding that are relevant to the correct gloss embedding. This involves calculating the attention score of the gloss embeddings in terms of the context embedding, and vice versa in the next round, alternating for multiple rounds.

After the "refinement" process, a logit representing the glosses is produced, and the gloss with the highest probability after taking the softmax is taken as the predicted sense.

Scarlini et al. (2020) proposes to improve sense embeddings by training the embeddings on a set of contet extracted from existing knowledge bases such as WordNet. Senses are predicted by taking the sense whose embedding is closest to that of the target word sense.

Conia and Navigli (2021) find that treating WSD as a multi-label classification task helps improve performance on multiple benchmarks over those that train on a single-label classification objective.

SACE (Wang and Wang 2021), like ConSec (Barba et al. (2021b), to be introduced in the next subsection), includes senses of nearby words of the target to inform the disambiguation process, and iteratively . In addition, they exploit relevant sentences in the document (in their case, the SemCor dataset) as the context, and uses a try-again mechanism (Wang and Wang 2020) at evaluation.

1.2 As Sequence Tagging

ESCHER (Barba et al. 2021a) takes as input a sentence concatenated with all the target word's definitions and outputs two indices indicating the start and end token in the input text as the target word definition. During training, the target word is marked with special tags in the context sentence, and the context sentence is separated from the glosses with another set of special tags. BART-large is finetuned for this task, though the authors also show that models whose number of parameters is in the same order of magnitude (including RoBERTa-large and XLNET-large) work equally well.

1.3 Generative Approaches

Wahle et al. (2021) proposes Languages Model Gloss Classification (LMGC), which determines the most likely gloss-context pair at once from the aggregated representation of each gloss-context pair of the target word. They also show that training a masked language model with LMGC as an additional objective helps reach competitive WSD performance while maintaining performance on other tasks such as classification.

However, the fine-tuned model provided by the authors is based on a pre-trained transformer (a sequence-to-sequence T5) not mentioned in the paper (which are all run in classification mode). Furthermore, the input format described in the paper and in the source code also differ from that in the demonstration link. Thorough evaluation on the provided model is advised before using it as-is.

ConSec (Barba et al. 2021b) explicitly incorporates the definitions of words surrounding the target word with the context during training, building upon relative positional embedding (Huang et al. 2018), (Yang et al. 2019) to insert the definitions into the context. DeBERTa-large is fine-tuned to extract the correct sense given the context sentence and the candidate senses.

Although both Escher and ConSec formulate WSD as a text extraction task, ConSec is provided with the additional information of the surrounding words, and generates definition predictions directly instead of predicting the start-end indices.

1.4 Zero-Shot / Few-Shot WSD with LLMs

Despite the popularity of large language models (LLMs) in recent years, with which tasks are mainly performed few-shot or zero-shot, little research about their capabilities on WSD is published.

Kang et al. (2023) first prompts the LM to translate the target word in context into a target language, L_t. In the word sense inventory of L_t, the candidate senses of the target word are taken and ranked with another prompt. The intersection of the top-scored senses in L_t and the candidate senses in the original language is taken to be the prediction in the first step. The authors use an ensemble of multiple target languages to obtain better performance.

The method yields promising result on a multi-lingual WSD test set. The authors also enphasize the method's ability to out-perform specialized architectures fine-tuned on WSD datasets under a zero-shot setting. It should be noted, however, that multiple calls to the (L)LM are required to disambiguate one word, which should be taken into consideration if interence costs and latency is a concern.

1.5 The Current State-of-the-Art

In terms of performance on public benchmarks, Barba et al. (2021b), Zhang et al. (2023) and Wang and Wang (2021) all achieve similarly competitive scores as of Februrary 2024.

2 The Long Tail of Infrequent Words

Blevins and Zettlemoyer (2020) have pointed out that "WSD systems show a strong bias towards predicting the most frequent sense (MFS) of a word regardless of the surrounding context". It is important that WSD systems learn to correctly predict both common and uncommon senses.

Su et al. (2022) proposes to scale the cross-entropy loss of the target words based on the number of senses it has. This is shown to close the gap between the performance of the more-common senses and the less-common senses, though later models such as Escher and ConSec show that they are able to achieve stronger results on less-common and unseen senses.

3 Taking Efficiency Into Consideration

It is hard to directly estimate which model requires more compute, both at training and at inference other than empirically. Escher (Barba et al. 2021a) and ConSec (Barba et al. 2021b) report their training hardware and hours. Sace (Wang and Wang 2021) directly compare their method to BEM (Blevins and Zettlemoyer 2020), where the former takes 1/140 the GPU hours to train than the latter. RWTE (Zhang et al. 2023) report that their architecture takes less time to train compared to ConSec and Sace.

4 Related Keywords and Applications

- MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation (Tanner and Hoffman 2023)
- Metaphorical Polysemy Detection: Conventional Metaphor Meets Word Sense Disambiguation (Maudslay and Teufel 2022)
- Exemplification modelling (making example sentences from word senses) (Harvill et al. 2023)
- Existing APIs (Orlando et al. 2022)

References

Barba E, Pasini T, Navigli R (2021a) ESC: Redesigning WSD with Extractive Sense Comprehension. In: Toutanova K, Rumshisky A, Zettlemoyer L, et al. (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Online, pp 4661–4672

Barba E, Procopio L, Navigli R (2021b) ConSeC: Word Sense Disambiguation as Continuous Sense Comprehension. In: Moens M-F, Huang

X, Specia L, Yih S W-t (eds) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 1492–1503

Blevins T, Zettlemoyer L (2020) Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In: Jurafsky D, Chai J, Schluter N, Tetreault J (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 1006–1017

Conia S, Navigli R (2021) Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration. In: Merlo P, Tiedemann J, Tsarfaty R (eds) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, pp 3269–3275

Harvill J, Hasegawa-Johnson M, Yoon H S, et al (2023) One-Shot Exemplification Modeling via Latent Sense Representations. In: Can B, Mozes M, Cahyawijaya S, et al. (eds) Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023). Association for Computational Linguistics, Toronto, Canada, pp 303–314

Huang C-Z A, Vaswani A, Uszkoreit J, et al (2018) Music Transformer

Huang L, Sun C, Qiu X, Huang X (2019) GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 3509–3514

Kang H, Blevins T, Zettlemoyer L (2023) Translate to Disambiguate: Zero-shot Multilingual Word Sense Disambiguation with Pretrained Language Models. arXiv preprint arXiv:230413803

Maudslay R H, Teufel S (2022) Metaphorical Polysemy Detection: Conventional Metaphor Meets Word Sense Disambiguation. In: Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp 65–77

Orlando R, Conia S, Faralli S, Navigli R (2022) Universal Semantic Annotator: the First Unified API for WSD, SRL and Semantic Parsing. In: Calzolari N, Béchet F, Blache P, et al. (eds) Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 2634–2641

Raganato A, Camacho-Collados J, Navigli R (2017) Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In: Lapata M, Blunsom P, Koller A (eds) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics, Valencia, Spain, pp 99–110

Scarlini B, Pasini T, Navigli R (2020) With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 3528–3539

Song Y, Ong X C, Ng H T, Lin Q (2021) Improved Word Sense Disambiguation with Enhanced Sense Representations. In: Moens M-F, Huang X, Specia L, Yih S W-t (eds) Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 4311–4320

Su Y, Zhang H, Song Y, Zhang T (2022) Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 4713–4723

Tanner J, Hoffman J (2023) MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation. In: Bouamor H, Pino J, Bali K (eds) Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp 181–193

Wahle J P, Ruas T, Meuschke N, Gipp B (2021) Incorporating Word Sense Disambiguation in Neural Language Models. CoRR

Wang M, Wang Y (2021) Word Sense Disambiguation: Towards Interactive Context Exploitation from Both Word and Sense Perspectives. In: Zong C, Xia F, Li W, Navigli R (eds) 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). Association for Computational Linguistics, Online, pp 5218–5229

Wang M, Wang Y (2020) A Synset Relation-enhanced Framework with a Try-again Mechanism for Word Sense Disambiguation. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 6229–6240

Yang Z, Dai Z, Yang Y, et al (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach H, Larochelle H, Beygelzimer A, et al. (eds) Advances in Neural Information Processing Systems. Curran Associates, Inc., p

Zhang X, Zhang R, Li X, et al (2023) Word Sense Disambiguation by Refining Target Word Embedding. Association for Computing Machinery, Austin, TX, USA, p 1405