Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration

Simone Conia **Roberto Navigli**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{conia,navigli}@di.uniroma1.it

Abstract

Recent studies treat Word Sense Disambiguation (WSD) as a single-label classification problem in which one is asked to choose only the best-fitting sense for a target word, given its context. However, gold data labelled by expert annotators suggest that maximizing the probability of a single sense may not be the most suitable training objective for WSD, especially if the sense inventory of choice is fine-grained. In this paper, we approach WSD as a multi-label classification problem in which multiple senses can be assigned to each target word. Not only does our simple method bear a closer resemblance to how human annotators disambiguate text, but it can also be extended seamlessly to exploit structured knowledge from semantic networks to achieve state-of-the-art results in English all-words WSD.

1 Introduction

Word Sense Disambiguation (WSD) is traditionally framed as the task of associating a word in context with its correct meaning from a finite set of possible choices (Navigli, 2009). Following this definition, recently proposed neural models were trained to maximize the probability of the most appropriate meaning while minimizing the probability of the other possible choices (Huang et al., 2019; Vial et al., 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020). Although this training objective proved to be extremely effective and even led to Bevilacqua and Navigli (2020) reaching the estimated upper bound of inter-annotator agreement for WSD performance on the unified evaluation framework of Raganato et al. (2017b), adhering to it underplays a fundamental aspect of how human annotators disambiguate text. Indeed, past studies have observed that it is not uncommon for a word to have multiple appropriate meanings in a given context, meanings that can be used interchangeably under some circumstances because their boundaries are not clear cut (Tuggy, 1993; Kilgarriff, 1997; Hanks, 2000; Erk and McCarthy, 2009). This is especially evident if the underlying sense inventory is fine-grained, as the complexity, and therefore performance, of WSD is tightly coupled to sense granularity (Lacerra et al., 2020). The difficulty an annotator faces in choosing the most appropriate meaning from a fine-grained sense inventory becomes clear from an analysis of gold standard datasets: a non-negligible 5% of the target words are annotated with two or more sense labels in several gold standard datasets, including Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Snyder and Palmer, 2004), SemEval-2007 (Pradhan et al., 2007), SemEval-2013 (Navigli et al., 2013), and SemEval-2015 (Moro and Navigli, 2015). Therefore, we follow Erk and McCarthy (2009), Jurgens (2012), and Erk et al. (2013), and argue that forcing a system to treat WSD as a single-label classification problem and learn that only one sense is correct for a word in a given context does not reflect how human beings disambiguate text.

In contrast to recent work, we approach WSD as a soft multi-label classification problem in which multiple senses can be assigned to each target word. We show that not only does this simple method bring significant improvements at low or no additional cost in terms of training and inference times and number of trainable parameters, but it can also be seamlessly extended to integrate senses from relational knowledge in structured form, e.g., similarity, hypernymy and hyponymy relations from semantic networks such as WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2012). While structured knowledge has been naturally utilized by graph-based algorithms for WSD (Agirre and Soroa, 2009; Moro et al., 2014; Scozzafava et al., 2020), the incorporation of such information into

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3269–3275
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

neural approaches has recently been garnering significant attention. However, currently available models can only take advantage of this knowledge with purposely-built layers (Bevilacqua and Navigli, 2020) that require additional complexities and/or trainable parameters. To the best of our knowledge, the work presented in this paper is the first to integrate structured knowledge into a neural architecture at negligible cost in terms of training time and number of parameters, while at the same time attaining state-of-the-art results in English all-words WSD.

2 Method

Single-label vs multi-label. WSD is the task of selecting the best-fitting sense s among the possible senses S_w of a target word w in a given context $c = \langle w_1, w_2, \ldots, w_n \rangle$, where S_w is a subset of a predefined sense inventory S. Abstracting away from the intricacies of any particular supervised model for WSD, the output of a WSD system provides a probability y_i for each sense $s_i \in S_w$. Recently proposed machine learning models – Kumar et al., 2019; Barba et al., 2020; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020, *inter alia* – are trained to maximize the probability of the single most appropriate sense $\hat{s}$ by minimizing the cross-entropy loss $\mathcal{L}_{\text{CE}}$:

$$\mathcal{L}_{\text{CE}}(w, \hat{s}) = -\log(y_{\hat{s}}) \tag{1}$$

We observe that this loss function is only suitable for single-label classification problems. In the case of WSD, this is equivalent to assuming that there is just a single appropriate sense $\hat{s} \in S_w$ for the target word w in the given context c, that is, $\hat{s}$ is clearly dissimilar from any other sense in S_w. Indeed, minimizing the cross-entropy loss in order to maximize the probability of two or more senses generates conflicting training signals; at the same time, choosing to ignore one of the correct senses results in a loss of valuable information.

Since there is a not insignificant number of instances where multiple similar senses of the target word w fit the given context c (see Section 1), we frame WSD as a multi-label classification problem in which a machine learning model is trained to predict whether a sense $s \in S_w$ is appropriate for a word w in a given context c, independently of the other senses in S_w. This is simply equivalent to minimizing the binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ on the probabilities of the candidate senses S_w:

$$\mathcal{L}_{\text{BCE}}(w, \hat{S}_w) = -\sum_{\hat{s} \in \hat{S}_w} \log(y_{\hat{s}}) - \sum_{s \in S_w \setminus \hat{S}_w} \log(1 - y_s) \tag{2}$$

where $\hat{S}_w \subseteq S_w$ is the set of appropriate senses for the target word w in the given context c. We note that this simple yet fundamental change in paradigm does not come with an increased computational complexity as $|S_w|$ is usually small. Moreover, it is independent of the underlying model used to calculate the output probabilities and, therefore, it does not increase the number of trainable parameters.

Knowledge integration. If our model benefits from learning to assign multiple *similar* senses to a target word in a given context, then it makes sense that the very same model may also benefit from learning what *related* senses can be assigned to that word. For example, in the sentence *"the quick brown fox jumps over the lazy dog"*, our model may formulate a better representation of *fox* if it is also trained to learn that any fox is a canine (hypernymy relation) or that the fox species includes arctic foxes, red foxes, and kit foxes (hyponymy relations). In this way, not only would the model learn that canines, foxes and arctic foxes are closely related, but it would also learn that canines and arctic foxes may have the ability to jump, and this could act as a data augmentation strategy especially for those senses that do not appear in the training set.

There is a growing interest in injecting relational information from knowledge bases into neural networks but, so far, recent attempts have required purposely-designed strategies or layers. Among others, Kumar et al. (2019) aid their model with a gloss encoder that uses the WordNet graph structure; Vial et al. (2019) adopt a preprocessing strategy aimed at clustering related senses to decrease the number of output classes; Bevilacqua and Navigli (2020) introduce a logit aggregation layer that takes into account the neighboring meanings in the WordNet graph.

In contrast, our multi-labeling approach to WSD can be seamlessly extended to integrate relational knowledge from semantic networks such as WordNet without any increase in architectural complexity, training time, and number of trainable param-

eters. We simply relax the definition of the set of possible senses S_w for a word w to include all the senses related to a sense in S_w. More formally, let $G = (S, R)$ be a semantic network where S is a sense inventory and R is the set of semantic connections between any two senses. Then we define S_w^+ to also include every sense s_j that is connected to any sense $s_i \in S_w$ by an edge $(s_i, s_j) \in R$, that is, $S_w^+ = S_w \cup \{s_j : (s_i, s_j) \in R, s_i \in S_w\}$. The loss function is updated accordingly to maximize not only the probability of the correct senses, but also the probability of their related senses:

$$\mathcal{L}_{\text{BCE}}(w, \hat{S}_w^+) = -\sum_{\hat{s} \in \hat{S}_w^+} \log(y_{\hat{s}}) - \sum_{s \in S_w^+ \setminus \hat{S}_w^+} \log(1 - y_s) \quad (3)$$

where $\hat{S}_w^+ = \hat{S}_w \cup \{s_j : (\hat{s}_i, s_j) \in R, \hat{s}_i \in \hat{S}_w\}$. We note that the increase of the number of possible choices ($|S_w^+| \geq |S_w|$) and correct meanings ($|\hat{S}_w^+| \geq |\hat{S}_w|$) does not hinder the learning process since each probability is computed independently of the others. Finally, we stress that our approach to structured knowledge integration is completely model-agnostic, as it is independent of the architecture of the underlying supervised model.

Model description. In order to assess the benefits of our multi-labeling approach and avoid improvements that may not be related to the overall objective of this paper, we conduct our experiments with a simple WSD model. Similarly to Bevilacqua and Navigli (2020), this model is simply composed of BERT (large-cased, frozen), a non-linear layer, and a linear classifier. Thus, given a word w in context we build a contextualized representation $\mathbf{e}_w \in \mathbb{R}^{d_{\text{BERT}}}$ of the word w as the average of the corresponding hidden states of the last four layers of BERT, apply a non-linear transformation to obtain $\mathbf{h}_w \in \mathbb{R}^{d_{\text{h}}}$ with $d_{\text{h}} = 512$, and finally a linear projection to $\mathbf{o}_w \in \mathbb{R}^{|S|}$ to compute the sense scores. More formally:

$$\mathbf{e}_w = \text{BatchNorm}\left(\frac{1}{4}\sum_{i=1}^{4} \mathbf{b}_w^{-i}\right)$$
$$\mathbf{h}_w = \text{Swish}(\mathbf{W}_h \mathbf{e}_w + \mathbf{b}_h)$$
$$\mathbf{o}_w = \mathbf{W}_o \mathbf{h}_w + \mathbf{b}_o$$

where $\mathbf{b}_w^{-i}$ is the hidden state of the i-th layer of BERT from the topmost one, $\text{BatchNorm}(\cdot)$ is the batch normalization operation, and $\text{Swish}(x) = x \cdot \text{sigmoid}(x)$ is the Swish activation function (Ramachandran et al., 2017).

3 Experiments and Results

Experimental setup. We train our models in different configurations to assess the individual contribution of several factors. First, we compare our baseline model trained with a single-label objective (Equation 1) to the same model trained with a multi-label objective (Equation 2). Then, we gradually include structured knowledge in the form of WordNet relations using Equation 3, starting from similarity relations (similar-to, also-see, verb-group, and derivationally-related-form), and incrementally including generalization and specification relations (hypernymy, hyponymy, instance-hypernymy, instance-hyponymy). In order to keep a level playing field with single-label systems, we choose only the meaning with highest probability for our multi-label models.

Datasets. We evaluate the models on the Unified Evaluation Framework for English all-words WSD proposed by Raganato et al. (2017b). This evaluation includes five gold standard datasets, namely, Senseval-2, Senseval-3, SemEval-2007, SemEval-2013, and SemEval-2015. Following standard practice we use the smallest gold standard as our development set, SemEval-2007, and the remaining ones as test sets. We distinguish between two settings: closed and open. In the former setting, we include systems that only use SemCor (Miller et al., 1994) as the training corpus, while in the latter we also include those systems that use WordNet glosses and examples and/or Wikipedia.

Hyperparameters. We use the pretrained version of BERT-large-cased (Devlin et al., 2019) available on HuggingFace's Transformers library (Wolf et al., 2020) to build our contextualized embeddings (Section 2). BERT is left frozen, that is, its parameters are not updated during training. Each model is trained for 25 epochs using Adam (Kingma and Ba, 2015) with a learning rate of 10^{-4}. We avoid hyperparameter tuning and opt for values that are close to the ones reported in the literature so as to have a fairer comparison.

Comparison systems. In order to have a comprehensive comparison with the current state of the art in WSD, we include the work of:

							Concatenation of ALL datasets				
		SE2	**SE3**	**SE07**	**SE13**	**SE15**	**Nouns**	**Verbs**	**Adj**	**Adv**	**ALL**
SemCor only	Raganato et al. (2017a)	72.0	69.1	64.8	66.9	71.5	71.5	57.5	75.0	83.8	69.9
	$\text{BERT}_{\text{Large}}$	76.3	73.2	66.2	71.7	74.1	–	–	–	–	73.5
	Hadiwinoto et al. (2019)	75.5	73.6	68.1	71.1	76.2	–	–	–	–	73.7
	Peters et al. (2019)	–	–	–	–	–	–	–	–	–	75.1
	Vial et al. (2019)	–	–	–	–	–	–	–	–	–	75.6
	Vial et al. (2019) - *Ensemble*	77.5	77.4	69.5	76.0	**78.3**	79.6	65.9	79.5	85.5	76.7
	This work	**78.4**	**77.8**	**72.2**	**76.7**	78.2	**80.1**	**67.0**	**80.5**	**86.2**	**77.6**
SemCor + definitions / examples	Loureiro and Jorge (2019)	76.3	75.6	68.1	75.1	77.0	78.0	64.0	80.7	84.5	75.4
	Scarlini et al. (2020a)	–	–	–	78.7	–	80.4	–	–	–	–
	Conia and Navigli (2020)	77.1	76.4	70.3	76.2	77.2	78.7	65.6	81.1	84.7	76.4
	Bevilacqua et al. (2020)	78.0	75.4	71.9	77.0	77.6	79.9	64.8	79.2	86.4	76.7
	Huang et al. (2019)	77.7	75.2	72.5	76.1	80.4	–	–	–	–	77.0
	Scarlini et al. (2020b)	78.0	77.1	71.0	77.3	83.2	80.6	68.3	80.5	83.5	77.9
	Blevins and Zettlemoyer (2020)	79.4	77.4	74.5	79.7	81.7	81.4	68.5	83.0	**87.9**	79.0
	Bevilacqua and Navigli (2020)	**80.8**	**79.0**	75.2	80.7	81.8	**82.9**	69.4	82.9	87.6	80.1
	This work	80.4	77.8	**76.2**	**81.8**	**83.3**	**82.9**	**70.3**	**83.4**	85.5	**80.2**

Table 1: WSD results in F_1 scores on Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), SemEval-2015 (SE15), and the concatenation of all the datasets (ALL). Top: closed setting (only SemCor allowed as the training corpus without definitions and/or examples). Bottom: open setting (WordNet glosses and examples are also used for training).

WSD	Sim	See	Rel	Vrb	Hpe	Hpo	Hpe_I	Hpo_I	**SE07**	**ALL**
SL	–	–	–	–	–	–	–	–	69.0	74.7
ML	–	–	–	–	–	–	–	–	69.2	75.7
ML	✔	✔	✔	✔	–	–	–	–	70.6	76.6
ML	✔	✔	✔	✔	✔	–	–	–	71.0	77.0
ML	✔	✔	✔	✔	–	✔	–	–	**72.5**	77.4
ML	✔	✔	✔	✔	✔	✔	–	–	72.2	**77.6**
ML	✔	✔	✔	✔	✔	✔	✔	✔	72.2	**77.6**

Table 2: WSD results in F_1 scores on SemEval-2007 (SE07) and the concatenation of all the datasets (ALL). SL/ML: single-label/multi-label. Sim: similar-to. See: also-see. Rel: derivationally-related-forms. Vrb: verb-groups. Hpe: hypernymy. Hpo: hyponymy. Hpe_I: instance-hypernyms. Hpo_I: instance-hyponyms.

- Raganato et al. (2017a) which was one of the first to propose a neural sequence model for WSD based on a stack of BiLSTM layers;
- $\text{BERT}_{\text{large}}$, a simple 1-neareast-neighbor approach based on the last hidden state of the BERT-large-cased model (Loureiro and Jorge, 2019);
- Hadiwinoto et al. (2019) which was among the first to exploit pretrained contextualized models for WSD;
- Peters et al. (2019) which incorporated WSD knowledge directly into the training process of BERT;
- Huang et al. (2019) which tasked the model to learn which gloss is the most appropriate for a word in context;
- Bevilacqua et al. (2020) which tackled WSD as a gloss generation problem;
- Loureiro and Jorge (2019) and Conia and Navigli (2020) which created and enhanced sense embeddings with relational knowledge from WordNet and BabelNet;
- Scarlini et al. (2020a) which proposed nominal sense embeddings built by exploiting BabelNet to automatically retrieve sense-specific context;
- Scarlini et al. (2020b) which extended the above approach to non-nominal senses and multiple languages;
- alongside the aforementioned work of Vial et al. (2019), Blevins and Zettlemoyer (2020), and Bevilacqua and Navigli (2020).

The systems are divided into two groups in Table 1: in the upper part we compare our approach against those systems that do not take advantage of information coming from WordNet glosses and/or examples, while in the lower part we also include those systems that make use of such knowledge.

Results. The first two rows of Table 2 show the results of switching from a single-label to a multi-label approach for WSD: this single change already brings a significant improvement in performance (+1.0% in F_1 score, significant with $p < 0.1$, χ^2 test). Not only that, increasing the number and variety of WordNet relations further increases the performance of the model, with hyponyms being particularly beneficial (+0.8% in F_1 score). Unfortunately, including instance hypernyms and instance hyponyms does not bring further improvements; this may be due to the relatively low number of instances that can take advantage of such relations in SemCor.

Nonetheless, the results obtained set a new state of the art among single and ensemble systems trained only on SemCor without the use of additional training data or resources external to WordNet such as Wikipedia, surpassing the previous state-of-the-art non-ensemble system of Vial et al. (2019) by 2.0% in F_1 score (significant with $p < 0.05$, χ^2 test), as shown in Table 1. When further trained on the WordNet glosses and examples, our model attains state-of-the-art results (+1.2% and +0.1% in F_1 score compared to the systems of Blevins and Zettlemoyer (2020) and Bevilacqua and Navigli (2020), respectively) despite being simpler than most of the techniques it is compared against.

4 Conclusion

WSD is a key task in Natural Language Understanding with several open challenges and with the granularity of sense inventories being undoubtedly the most pressing issue (Navigli, 2018). We departed from recent work on WSD and investigated the effect of tackling the task as a multi-label classification problem. Not only is our approach simple and model-agnostic, but it can also be seamlessly extended to integrate relational knowledge in structured form from semantic networks such as WordNet, and at no extra cost in terms of architectural complexity, training times, and number of parameters.

Our experiments show that our method, thanks to its more comprehensive notion of loss over equally valid and structurally-related senses, achieves state-of-the-art results in English all-words WSD, especially when there is a lower amount of annotated text available. These results open the path to further research in this direction, from exploring more complex models and richer knowledge bases to exploiting multiple labels in innovative disambiguation settings which can overcome the fine granularity of sense inventories. Not only that, our knowledge integration approach could potentially be applied to address the knowledge acquisition bottleneck in multilingual WSD (Pasini, 2020; Pasini et al., 2021). Finally, with the rise of ever more complex general and specialized pretrained models, we believe that our simple model-agnostic approach can be another step towards knowledge-based (self-)supervision.

We release our software and model checkpoints at `https://github.com/SapienzaNLP/multilabel-wsd`.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*.

Edoardo Barba, Luigi Procopio, Niccolò Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLaN: Multilingual label propagation for Word Sense Disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or: "How we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

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, ACL 2020, Online, July 5-10, 2020.

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, ACL 2020, Online, July 5-10, 2020*.

Simone Conia and Roberto Navigli. 2020. Conception: Multilingually-enhanced, human-readable concept vector representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online).

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)*, Minneapolis, Minnesota.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL@ACL 2001, Toulouse, France, July 5-6, 2001*.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved Word Sense Disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*.

Patrick Hanks. 2000. Do word meanings exist? *Comput. Humanit.*, 34(1-2):205–215.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for Word Sense Disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*.

David Jurgens. 2012. An evaluation of graded sense disambiguation using word sense induction. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM 2012, June 7-8, 2012, Montréal, Canada*.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha P. Talukdar. 2019. Zero-shot Word Sense Disambiguation using sense definition embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*.

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, New York, NY, USA, February 7-12, 2020*.

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, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*.

George A. Miller. 1995. WordNet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words Sense Disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2.

Roberto Navigli. 2009. Word Sense Disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.

Roberto Navigli. 2018. Natural language understanding: Instructions for (present and future) use. In *IJCAI*, pages 5697–5702.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual Word Sense Disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*.

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

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, IJCAI-20*, 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 *Proc. of AAAI*.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007*.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural sequence learning models for Word Sense Disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*.

Alessandro Raganato, José Camacho-Collados, and Roberto Navigli. 2017b. 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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2017. Searching for activation functions. *arXiv*, abs/1710.05941.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-enhanced sense embeddings for multilingual 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, New York, NY, USA, February 7-12, 2020*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual Word Sense Disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), demos*, Online.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, SENSEVAL@ACL 2004, Barcelona, Spain, July 25-26, 2004*.

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4:273–290.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of WordNet for neural Word Sense Disambiguation.

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. Huggingface's transformers: State-of-the-art natural language processing.