Don't Neglect the Obvious:
On the Role of Unambiguous Words in Word Sense Disambiguation

Daniel Loureiro
LIAAD - INESC TEC
University of Porto, Portugal
dloureiro@fc.up.pt

Jose Camacho-Collados
School of Computer Science and Informatics
Cardiff University, United Kingdom
camachocolladosj@cardiff.ac.uk

Abstract

State-of-the-art methods for Word Sense Disambiguation (WSD) combine two different features: the power of pre-trained language models and a propagation method to extend the coverage of such models. This propagation is needed as current sense-annotated corpora lack coverage of many instances in the underlying sense inventory (usually WordNet). At the same time, unambiguous words make for a large portion of all words in WordNet, while being poorly covered in existing sense-annotated corpora. In this paper, we propose a simple method to provide annotations for most unambiguous words in a large corpus. We introduce the UWA (Unambiguous Word Annotations) dataset and show how a state-of-the-art propagation-based model can use it to extend the coverage and quality of its word sense embeddings by a significant margin, improving on its original results on WSD.

1 Introduction

There has been a lot of progress in word sense disambiguation (WSD) recently. This progress has been driven by two factors: (1) the introduction of large pre-trained Transformer-based language models and (2) propagation algorithms that extends the coverage of existing training sets. The gains due to pre-trained Neural Language Models (NLMs) such as BERT (Devlin et al., 2019) have been outstanding, helping reach levels close to human performance when training data is available. These models are generally based on a nearest neighbours strategy, where each sense is represented by a vector, exploiting the contextualized embeddings of these NLMs (Melamud et al., 2016; Peters et al., 2018; Loureiro and Jorge, 2019). However, training data for WSD is hard to obtain, and the most widely used training set nowadays, based on WordNet, dates back from the 90s (Miller et al., 1993,

SemCor). This lack of curated data produces the so-called knowledge-acquisition bottleneck (Gale et al., 1992; Navigli, 2009).

However, there is a key source of information that has been neglected so far in existing sense-annotated corpora and propagation methods, which is the presence of unambiguous words from the underlying knowledge resource. Strikingly, WordNet, which is known to be a comprehensive resource, is mostly composed of unambiguous entries (30k lemmas are ambiguous, compared to 116k unambiguous). While the lack of unambiguous annotations does not have a direct effect in WSD, the fact that these unambiguous words are part of the same semantic network means they can have an effect on ambiguous words via standard propagation algorithms. These propagation algorithms start from a seed of senses occurring in the training data (and therefore their embeddings can be directly computed) and then propagate to the whole sense inventory via the semantic network (Vial et al., 2018; Loureiro and Jorge, 2019). Consequently, computing sense embeddings for unambiguous words can increase the number of seeds and improve the whole process. Covering these unambiguous words, however, is not an arduous task, as unlabelled corpora may suffice. We explore this hypothesis by labeling a large amount of unambiguous words in corpora extracted from the web, using WordNet as our reference sense inventory. While we can certainly find usages of a word not covered by WordNet, we found that our approach can obtain accurate occurrences with simple heuristics.

The contribution of this paper is twofold. First, we devise a simple methodology to construct UWA (Unambiguous Word Annotations), a large and, most importantly, diverse sense-annotated corpus that focuses on WordNet unambiguous words. Second, we show that by leveraging UWA, we can significantly improve a state-of-the-art WSD model.

3514

2 Related Work

The knowledge-acquisition bottleneck has been frequently addressed by automatically constructing sense-annotated corpora. Recent works propose methods that exploit knowledge from Wikipedia, such as NASARI vectors (Camacho-Collados et al., 2016), for providing sense annotations for concepts and entities (Scarlini et al., 2019; Pasini and Navigli, 2019). In the case of Scarlini et al. (2019), and similarly to Raganato et al. (2016), their method requires hyperlinks and category information from Wikipedia, hence not extensible to other kinds of corpora.[1] Previous approaches relied on parallel corpora for two or more languages. The OMSTI corpus (Taghipour and Ng, 2015) was constructed by exploiting the alignments of an English-Chinese corpus. Similarly, Delli Bovi et al. (2017) presented EuroSense, a multilingual sense-annotated corpus using the Europarl parallel corpus for 21 languages as reference. In contrast to these approaches, we focus on unambiguous senses and, therefore, are not constrained to only nouns, knowledge from Wikipedia, or a specific type of corpus.

Earlier works exploiting unambiguous words (Leacock et al., 1998; Mihalcea, 2002; Agirre and Martinez, 2004) and especially the subsequent extension by Martinez et al. (2008) are the most directly related to our paper. Martinez et al. (2008) retrieved example sentences with monosemous nouns from web search snippets and used them towards improved performance on WSD by leveraging WordNet relations. However, the WSD methods analyzed were sensitive to frequency bias, leading their collection effort to collect a large number of examples for fewer senses (and only nouns). In contrast, our solution is designed for all monosemous words, retrieving examples from web texts instead of snippets, attaining performance gains with even a single example per word.

3 Methodology

In this section we first explain our method to construct a corpus with unambiguous word annotations (Section 3.1). Then, we explain current models based on language models for WSD (Section 3.2) and describe a propagation method to infer additional OOV sense representations (Section 3.3).

3.1 Unambiguous Word Annotations (UWA)

In order to properly test our hypothesis, we first require a sizable compilation of unambiguous words in context, particularly words that correspond to lemmas covered by WordNet. The extensiveness of WordNet means that most of its lemmas occur very rarely, and thus require processing large volumes of texts to achieve a high coverage. As such, in this work we develop the Unambiguous Word Annotations (UWA) corpus based on Open-WebText (Gokaslan and Cohen, 2019) and English Wikipedia (November 2019), processing over 53GB of texts from the web.

Each text is annotated for lemmas and part-of-speech using the Stanford CoreNLP toolkit (Manning et al., 2014). The annotations are filtered so that we only consider lemma/part-of-speech pairs that are present in WordNet, and correspond to a single sense (hence unambiguous), e.g., 'keypad/noun'. Naturally, some lemma/part-of-speech pairs may have additional meanings not covered in WordNet. For example, in "*Inception* was a box-office hit.", *Inception* makes reference to a movie and not to the unambiguous word *inception* from WordNet. To mitigate this issue, we applied Named Entity Recognition (NER) tagging, using spaCy (Honnibal and Montani, 2017), to discard lemmas that are recognized as entities but do not correspond to an entity in their WordNet sense. To this end, we leverage the entity annotations of WordNet synsets available in BabelNet (Navigli and Ponzetto, 2012). To keep the corpus at a reasonable size, we cut-off the maximum number of associated sentences (examples henceforth) per sense at 100.

Statistics. UWA covers a total of 98,494 senses, where 56.7% have 100 examples, and 81.2% have at least 10 examples. In Table 1 we show that UWA covers most senses for unambiguous words and, combined with SemCor, includes most senses in WordNet. This contrasts with other automatically-constructed datasets such as OMSTI (Taghipour and Ng, 2015) or T-o-M (Pasini and Navigli, 2019). These sense-annotated corpora, not aimed specifically at unambiguous words, have limited coverage in this respect, as they are mainly composed of annotations for senses already available in SemCor.

3.2 Neural Language Models for WSD

Recent NLMs, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), have been used with a high degree of success on WSD. They have

---

[1] Pasini and Camacho-Collados (2020) provide a more detailed overview of existing sense-annotated corpora.

Corpus	# Instances		Avg	Coverage (w/ SC)		
	Amb	Unamb	# Exs	Amb	Unamb	Total
SemCor	198,153	27,883	6.8	26.2	7.4	16.1
OMSTI	909,830	1,304	244.7	26.8	7.4	16.4
T-o-M	719,888	114,580	152.4	28.5	7.5	17.2
UWA(1)	0	98,494	1.0	26.2	82.9	56.7
UWA(10)	0	804,861	8.8	26.2	82.9	56.7
UWA(all)	0	6,111,453	54.1	26.2	82.9	56.7

Table 1: Number of instances, average number of examples per word sense, and coverage percentage (including SemCor) of various sense-annotated corpora.

been used differently depending on the nature of the disambiguation task: as feature providers for other neural architectures (Vial et al., 2019), simple classifiers after fine-tuning (Wang et al., 2019), or as generators of contextual embeddings to be matched through nearest neighbours (Melamud et al., 2016; Peters et al., 2018; Loureiro and Jorge, 2019; Reif et al., 2019, 1NN). Our experiments in this paper will focus on improving the latter type of approach. In particular, we will investigate the state-of-the-art LMMS model (Loureiro and Jorge, 2019). This model learns sense embeddings based on BERT states. These embeddings are then propagated through WordNet's ontology to infer additional senses, effectively providing a full coverage. While Loureiro and Jorge (2019) proposed variants of LMMS that combined propagation with gloss embeddings, or static embeddings, this paper is only concerned with the propagation method.

In our case, we essentially follow LMMS's layer pooling method to generate contextual embeddings for each sense occurrence in context (from a training set), and derive sense embeddings from the average of all corresponding contextual embeddings.

3.3 Network Propagation for Full-Coverage

The propagation method used in LMMS exploits the WordNet ontology to obtain a full coverage of sense embeddings from an initial set of embeddings based on a manually sense-annotated corpus like SemCor. This method explores different abstraction levels represented in WordNet: sets of synonyms (synsets), Is-A relations (hypernyms) and categorical groupings (lexnames[2]).

Initial sense embeddings are first used to compute synset embeddings as the average of all corresponding senses (analogously to how sense embeddings are computed from contextual embeddings).

---

[2]Lexnames are also known as supersenses in the literature (Flekova and Gurevych, 2016; Pilehvar et al., 2017).

From that point, missing senses are represented by their corresponding synset embeddings. The remaining unrepresented senses are inferred from their hypernym and lexname embeddings, computed by averaging their neighbour synset embeddings. Note that this propagation process does not follow transitive relations in WordNet, i.e., a single synset's hypernym is considered, while the subsequent hypernyms along the root paths are ignored.

Since lexname embeddings can always be computed, this process can reach a full-coverage of WordNet starting with just the initial set of embeddings produced using SemCor. However, the set of SemCor embeddings only covers 16.1% of WordNet, so many of the inferred representations are redundant and therefore not entirely meaningful.

4 Evaluation

For our experiments we are interested in verifying the impact of using UWA to improve WSD performance. In particular, we test the unambiguous annotations of UWA as a complement of existing sense-annotated training data. To this end, as explained in Section 3, we make use of the state-of-the-art WSD model LMMS (Loureiro and Jorge, 2019). In addition to the original version using BERT, we also provide results with RoBERTa (Liu et al., 2019) for completeness. We use the 24-layer models for both BERT and RoBERTa.[3]

4.1 Word Sense Disambiguation (WSD)

Table 2 shows the WSD results on the standard evaluation framework of Raganato et al. (2017) for LMMS trained on the concatenation of SemCor and automatically-constructed corpora. In the table we include UWA with two different maximum number of examples per unambiguous word, i.e., 1 and 10. For comparison, we also include the results of EWISE (Kumar et al., 2019) and GlossBERT (Huang et al., 2019), which attempt to overcome the limited coverage of SemCor by exploiting textual definitions. As can be observed, the concatenation of our UWA corpus and SemCor provides the best overall results, regardless of the number of examples cut-off. Perhaps surprisingly, our corpus is the only one that provides improvements over the baseline (SemCor-only). These improvements are statistically significant on the full test set (i.e. ALL) for both BERT and RoBERTa with p < 0.0005, based on a t-test with respect to the

---

[3]Commonly referred to as *large* models.

	Corpus	SE-2	SE-3	SE07	SE13	SE15	ALL
LMMS-BERT	SC-noProp.	70.2	71.1	64.7	65.5	70.2	69.0
	SC-only	75.5	74.2	**66.8**	72.9	75.3	74.0
	OMSTI	73.7	68.8	63.5	**73.2**	74.8	71.9
	T-o-M	69.9	66.1	62.4	64.8	74.2	67.9
	UWA (1)	77.0	**74.2**	66.2	73.1	75.4	**74.5**
	UWA (10)	**77.3**	74.1	66.2	72.7	**75.7**	**74.5**
LMMS-RoBERTa	SC-noProp.	70.7	70.6	66.7	65.1	70.5	69.2
	SC-only	76.0	73.6	**69.2**	72.3	**75.9**	74.1
	OMSTI	73.4	70.1	66.6	71.5	74.6	71.9
	T-o-M	70.3	65.9	64.8	65.8	74.0	68.4
	UWA (1)	**77.8**	73.6	68.8	72.0	75.3	74.5
	UWA (10)	77.6	**73.7**	68.8	**72.7**	75.3	**74.6**
SOTA	SC‡ LMMS+	76.3	75.6	68.1	75.1	77.0	75.4
	SC† Vial et al.	76.6	76.9	69.0	73.8	75.4	75.4
	SC‡† EWISE	73.8	71.1	67.3*	69.4	74.5	71.8
	SC‡† GlossBERT	77.7	75.2	72.5*	76.1	80.4	77.0

Table 2: F1 performance on the unified WSD evaluation framework. All corpora marked are concatenated with SemCor (SC). SOTAs reported for reference but not directly comparable due to use of definitions (‡) or not using a 1NN approach (†). All reported SOTAs are based on BERT trained on SC. Results in datasets that were used as development are marked with *.

Corpus	BERT			RoBERTa		
	F1	P@5	MRR	F1	P@5	MRR
OMSTI	50.2	66.0	57.5	44.1	59.9	51.7
T-o-M	45.8	62.1	53.3	42.1	60.7	50.2
UWA (10)	**54.9**	**74.1**	**63.5**	62.1	80.2	70.1

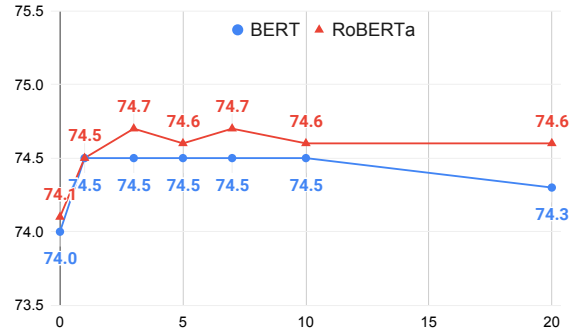Table 3: Performance comparison in the uninformed setting. Each corpus is concatenated with SemCor.

Figure 1: WSD performance (F1 on the ALL test set) with different numbers of UWA examples.

accuracy scores (equal to F1 in this setting). This can be explained by the fact that our corpus is the only one that significantly extends the coverage of SemCor, as explained in Section 3.1.

4.2 Uninformed Sense Matching (USM)

In standard WSD benchmarks, models are given the advantage of knowing the pre-defined set of possible senses before-hand. This is because gold PoS tags and lemmas are provided in these datasets. However, to better understand how robust a 1NN WSD model is, we can test it in an uninformed setting, i.e., where PoS tags and lemmas are not given and the model does not have access to the list of candidate senses. Instead, the model has to match senses from the whole sense inventory, unconstrained. Therefore, in this Uninformed Sense Matching (USM) setting we can use information retrieval ranking metrics with the model predictions (i.e. MRR or P@K) in addition to the standard F1. In line with the WSD results, Table 3 shows that UWA also substantially improves performance in the USM setting when comparing against currently available alternatives.

5 Analysis

In this section, we provide an analysis based on the number of examples (Section 5.1) and a visualization of the embedding space (Section 5.2).

5.1 Number of Examples

When compiling examples for learning sense representations, a natural question that arises is: how many examples are required to learn effective representations? The answer to this question can not only guide collection efforts, but also help clarify the requirements for learning effective representations in the simplest setting. To that end, we analyse the impact of using different number of examples from UWA on LMMS's WSD and USM performance. In Figure 1, we show the WSD performance trend using different number of examples per sense. As can be seen, performance improves substantially with only one example, and then stops improving after just two examples.

Similarly to our findings for WSD, Table 4 shows that a low number of examples, such as 2, already achieves the best overall results in the USM setting for BERT. Likewise, RoBERTa does not benefit from more than 5 examples. More generally, in USM the differences with respect to SemCor are more marked in comparison to the regular WSD setting. This is expected as the propagation algorithm has a stronger effect in this setting where all sense embeddings are considered.

5.2 Visualization of the Embedding Space

The propagation method used in LMMS is designed to backoff to increasingly abstract repre-

Corpus	BERT			RoBERTa		
	F1	P@5	MRR	F1	P@5	MRR
SemCor	52.5	67.1	59.2	58.0	72.8	64.7
UWA (1)	55.1	74.1	63.5	61.3	79.8	69.5
UWA (2)	**55.5**	**74.6**	**64.0**	61.8	**80.3**	70.0
UWA (3)	55.4	74.5	63.9	61.9	**80.3**	70.0
UWA (5)	55.4	74.4	63.8	**62.1**	**80.3**	**70.1**
UWA (7)	55.2	74.1	63.7	61.9	**80.3**	70.0
UWA (10)	54.9	74.1	63.5	**62.1**	80.2	**70.1**
UWA (20)	54.9	73.7	63.3	**62.1**	79.9	70.0

Table 4: USM performance of the LMMS model using SemCor and UWA with different example thresholds. Models tested on the concatenation of all WSD datasets of Raganato et al. (2017). As before, UWA is concatenated with SC in this experiment.

sentation levels, from synsets, to hypernyms, to supersenses (see Section 3.3 of the main paper). This naturally leads to a clustering effect, where many senses are represented with very similar, or equal, embeddings. In fact, we find that only 22% of sense embeddings learned from SemCor, and propagated following LMMS, are actually unique (remaining are shared by two or more senses). The addition of UWA increases this percentage to 68%.

To better understand this clustering effect, we used T-SNE (Maaten and Hinton, 2008) to visualize the WordNet synset embedding space. In Figure 2 we show synset embeddings learned from the SemCor+UWA(10) dataset, and learned from SemCor alone, both based on RoBERTa. While the same number of synset embeddings are learned in both cases, SemCor+UWA embeddings are better distributed across the vector space. This, in turn, causes a substantial reduction of high-density clusters, which stand in opposition to a rich distributional representation of senses.[4]

6 Conclusion

Unambiguous words are a surprisingly large portion of existing knowledge resources like Word-Net. At the same time, their coverage in existing sense-annotated corpora is very limited. In this paper, we proposed a simple method which exploits sense annotations of unambiguous words from unlabeled corpora, thereby effectively extending existing sense-annotated corpora with low-effort. By leveraging a state-of-the-art BERT-based WSD sys-

---

[4]We share interactive visualizations focusing on each of the 45 supersense groups (e.g. noun.communication) from WordNet at our UWA release website.

Figure 2: T-SNE comparison of synset embeddings for whole WordNet learned from SC+UWA10 (top), or just SC (bottom). Colors represent source of annotations for embeddings (● SC ● UWA ● Propagation).

tem that propagates sense embeddings across Word-Net, we have shown that these unambiguous words provide an excellent bridge to reach a wider range of OOV senses. This translates, in turn, into improving results for WSD. For future work it would be interesting to test these sense embeddings in a wider range of applications outside WSD. Since the embedding space is clearly more diversified, as shown in Figure 2, this may lead to improvements in other downstream tasks.

Moreover, one of the most surprising findings from this paper is that a single occurrence of OOV unambiguous words is enough to improve the performance of WSD models. This is relevant because (1) it is not always easy to retrieve a large number of examples for unambiguous words, and (2) it facilitates a cheaper manual verification, if required.

Finally, we openly release UWA, a large corpus annotated with unambiguous words, together improved BERT and RoBERTa-based sense embeddings, model predictions and visualizations at http://danlou.github.io/uwa.

References

Eneko Agirre and David Martinez. 2004. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Barcelona, Spain. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. 2017. EuroSense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600, Vancouver, Canada. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucie Flekova and Iryna Gurevych. 2016. Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2029–2041, Berlin, Germany. Association for Computational Linguistics.

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

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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)*, pages 3507–3512, Hong Kong, China. Association for Computational Linguistics.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy. Association for Computational Linguistics.

Claudia Leacock, Martin Chodorow, and George A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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 Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

D. Martinez, O. Lopez de Lacalle, and E. Agirre. 2008. On the Use of Automatically Acquired Examples for All-Nouns Word Sense Disambiguation. *Journal of Artificial Intelligence Research*, 33:79–107.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Rada F. Mihalcea. 2002. Bootstrapping large sense tagged corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308, Plainsboro, N.J.

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

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 and Jose Camacho-Collados. 2020. A short survey on sense-annotated corpora. In *Proceedings of the International Conference on Language Resources and Evaluation*, Marseille, France.

Tommaso Pasini and Roberto Navigli. 2019. Train-o-matic: Supervised word sense disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103215.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. 2017. Towards a seamless integration of word senses into downstream NLP applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1857–1869, Vancouver, Canada. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2016. Automatic Construction and Evaluation of a Large Semantically Enriched Wikipedia. In *Proceedings of IJCAI*, pages 2894–2900, New York City, NY, USA.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, pages 8592–8600.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "OneSeC" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709, Florence, Italy. Association for Computational Linguistics.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Beijing, China. Association for Computational Linguistics.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. Improving the coverage and the generalization ability of neural word sense disambiguation through hypernymy and hyponymy relationships. *arXiv preprint arXiv:1811.00960*.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. In *Proceedings of the 10th Global WordNet Conference*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.