On the Cross-lingual Transferability of Contextualized Sense Embeddings

Kiamehr Rezaee[1*], **Daniel Loureiro**[2*], **Jose Camacho-Collados**[3], **Mohammad Taher Pilehvar**[4]
[1] Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[2] LIAAD - INESC TEC, Department of Computer Science - FCUP, University of Porto, Portugal
[3] School of Computer Science and Informatics, Cardiff University, United Kingdom
[4] Tehran Institute for Advanced Studies, Khatam University, Tehran, Iran
k_rezaee@comp.iust.ac.ir, daniel.b.loureiro@inesctec.pt,
camachocolladosj@cardiff.ac.uk, mp792@cam.ac.uk

Abstract

In this paper we analyze the extent to which *contextualized sense embeddings*, i.e., sense embeddings that are computed based on contextualized word embeddings, are transferable across languages. To this end, we compiled a unified cross-lingual benchmark for Word Sense Disambiguation. We then propose two simple strategies to transfer sense-specific knowledge across languages and test them on the benchmark. Experimental results show that this contextualized knowledge can be effectively transferred to similar languages through pre-trained multilingual language models, to the extent that they can outperform monolingual representations learned from existing language-specific data.

1 Introduction

Word Sense Disambiguation (WSD) is an indispensable component of language understanding (Navigli, 2009); hence, it has been one of the most studied long-standing problems in lexical semantics. Currently, the dominant WSD paradigm is the supervised approach (Raganato et al., 2017), which highly relies on sense-annotated data. Similarly to many other supervised tasks, the amount of labeled (sense-annotated) data for WSD highly determines downstream performance. One of the factors that make WSD a challenging problem is that creating sense-annotated data is an expensive and arduous process, i.e., the so-called knowledge-acquisition bottleneck (Gale et al., 1992). Moreover, WSD research often focuses on the English language. While datasets for other languages exist (Petrolito and Bond, 2014a; Pasini and Camacho-Collados, 2020), these are generally automatically generated (Delli Bovi et al., 2017; Pasini et al., 2018; Scarlini et al., 2020a; Barba et al., 2020) or not large enough for training supervised WSD models (Navigli et al., 2013a; Moro and Navigli, 2015).[1]

However, recent contextualized embeddings have proven highly effective in English WSD (Peters et al., 2018; Loureiro and Jorge, 2019; Vial et al., 2019; Loureiro et al., 2021), as well as in capturing high-level linguistic knowledge that can be shared or transferred across different languages (Conneau et al., 2020; Cao et al., 2020). Therefore, cross-lingual transfer has opened new opportunities to circumvent the knowledge acquisition bottleneck for less-resourced languages. In this paper, we aim at investigating this opportunity. To this end, we build upon recent research on cross-lingual transfer to compute contextualized sense embeddings and verify if semantic distinctions in the English language are transferable to other languages.

The contributions are threefold: (1) We adapt existing datasets to build a unified benchmark for cross-lingual WSD based on WordNet; (2) we test the effectiveness of contextualized embeddings for cross-lingual transfer in the context of WSD; and (3) we establish relevant and simple baselines for future work in cross-lingual WSD.[2]

2 Related Work

This works lies at the intersection of two areas of NLP research: Word Sense Disambiguation and cross-lingual semantic representation. Hence, we cover the recent relevant work in the corresponding literature.

2.1 Word Sense Disambiguation

WSD techniques can be broadly put into two categories: knowledge-based and supervised. The main difference lies in that the latter makes use of sense-annotated data for its training phase, whereas the former exploits the encoded knowledge in sense in-

Authors marked with a star (*) contributed equally.

[1] An exception to this pattern is the recent XL-WSD benchmark (Pasini et al., 2021), contemporary to this paper.

[2] Data and code are available at https://github.com/danlou/Zero-MWSD

	Train - SemCor		Test - SemEval-15			Test - SemEval-13					Test - FN
	EN	IT	EN	IT	ES	EN	FR	DE	ES	IT	FA
Nouns	87,002	43,058	512	515	512	1,637	1,438	958	1,176	1,448	3,063
Verbs	88,334	25,164	252	233	260	–	–	–	–	–	29
Adj.	31,753	16,029	136	159	119	–	–	–	–	–	366
Adv.	18,947	7,951	82	25	53	–	–	–	–	–	40
ALL	226,036	92,202	982	932	944	1,637	1,438	958	1,176	1,448	3,498
RAW	226,036	92,202	1,175	1,151	1,155	1,931	1,656	1,467	1,481	1,706	4,272

Table 1: Number of sense-annotated instances in the benchmark datasets after cleaning and unification. RAW counts correspond to number of instances in the original datasets before cleaning and unification.

ventories such as WordNet (e.g. semantic relations, sense glosses, distributions, etc.) for inference.

For the last decade, the supervised approach has been the dominant paradigm for WSD (Raganato et al., 2017), either the conventional feature-based systems (Zhong and Ng, 2010; Iacobacci et al., 2016), LSTM-driven techniques (Melamud et al., 2016; Yuan et al., 2016), or the more recent trend empowered by pre-trained language models (Loureiro and Jorge, 2019; Scarlini et al., 2020b). In the latter approaches, feature extraction strategies where sense embeddings are determined by averaging a word's contextualised representations have proven surprisingly effective (Loureiro et al., 2021), even in multilingual settings (Bevilacqua and Navigli, 2020; Raganato et al., 2020). We extend this simple idea to the cross-lingual setting, showing that the vanilla contextualized sense embeddings achieving outstanding results in the monolingual setting can also be effective for transferring knowledge across languages.

2.2 Cross-lingual representation

WSD performance is largely dependent on the availability of large amounts of manually-curated sense annotations. However, as usual in NLP, most of the sense-annotated corpora are dedicated to the English language only. Nonetheless, recent work on cross-lingual word embeddings has shown that it is possible to reliably align monolingual semantic spaces with minimal or no supervision (Artetxe et al., 2017; Zhang et al., 2017; Conneau et al., 2018). Moreover, pre-trained language models, like BERT (Devlin et al., 2019), have been shown to be effective in transferring knowledge across languages (Lample and Conneau, 2019; Pires et al., 2019; Artetxe et al., 2020). In this paper we build on these ideas to take the best of both worlds. Instead of transferring static word embeddings, the main idea is to learn sense embeddings (learned using multilingual language models) that can be shared across languages.

3 Cross-lingual WSD Benchmark

In order to develop a unified benchmark for cross-lingual Word Sense Disambiguation, we opted for Princeton **WordNet** (Fellbaum, 1998, PWN) as our reference sense inventory. Thanks to its completeness (covering different parts of speech) and open nature, PWN is regarded as the de facto sense inventory for WSD in English. Moreover, the multilingual efforts from Open Multilingual WordNet (Bond and Foster, 2013), linked to the English PWN, make this resource prompt to extensions for sense-annotated corpora in other languages. We use PWN v3.0 for all our experiments, including the unified cross-lingual benchmark, converting all datasets to both the same XML schema of Raganato et al. (2017) and a practical *json* format. In the following we describe the datasets used to build the cross-lingual WSD benchmark, with Table 1 summarizing their main statistics.

3.1 SemCor training sets

As training corpora we used SemCor (Miller et al., 1993), which consists of a collection of English documents annotated with PWN senses. Despite its age, SemCor remains the standard training corpus for WSD due to its large number of manual sense annotations. There have been several efforts towards providing sense annotations for translated versions of SemCor (Petrolito and Bond, 2014a). Consequently, we also considered the Italian version of SemCor included in MultiSemCor (Bentivogli and Pianta, 2005), which is the language with most PWN annotations available. MultiSemCor includes sense annotations from PWN v1.6; hence, we used the mappings from Daudé et al. (2000) to convert these into PWN v3.0 annotations.

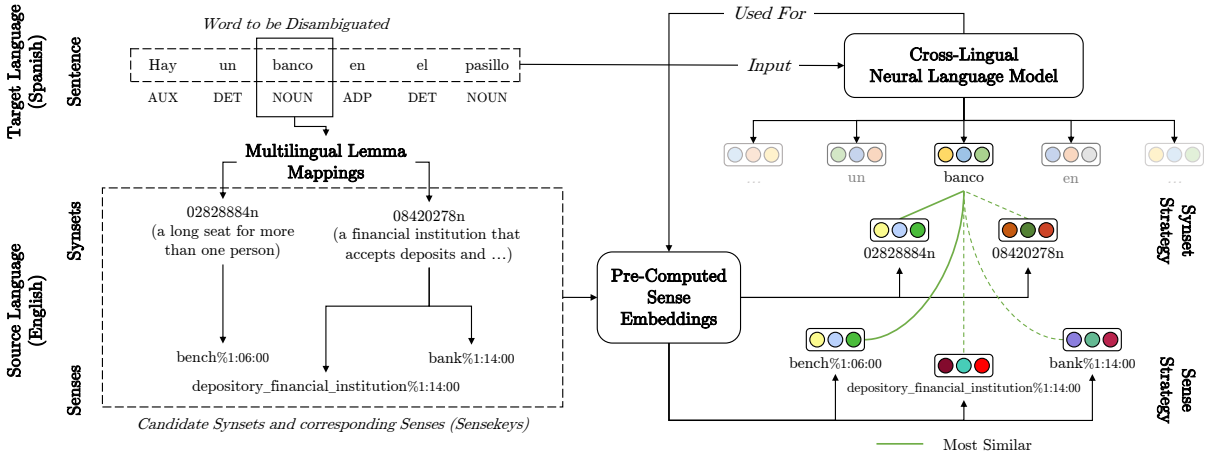

Figure 1: Overview of the proposed method for multilingual zero-shot word sense disambiguation. The example sentence presented ('There is a bench in the hall', in English) is using a different language (target) than our sense inventory (source), but using multilingual language models and lemma mappings, we demonstrate how it's still possible to perform disambiguation, using either variations of our method (sense and synset strategies).

3.2 Multilingual SemEval test sets

We considered two multilingual datasets: SemEval 2013 (Navigli et al., 2013a) available for English, French and Italian, and SemEval 2015 (Moro and Navigli, 2015), available for English, French, Italian, Spanish and German. These datasets were annotated with BabelNet (Navigli and Ponzetto, 2012), a resource that contains WordNet, among other linked sense inventories. Therefore, from each dataset we simply considered those disambiguated instances that could be mapped to PWN 3.0, while the rest of instances were removed. We also rely on BabelNet to gather a representative set of candidate senses for any given target word.

3.3 FarsNet

To extend the evaluation set beyond European languages, we first performed an exhaustive search for available WSD datasets that could be integrated into our benchmark, unsuccessfully.[3] To fill this gap, we constructed an evaluation set for a distant low-resource language: Farsi. This dataset was constructed based on example sentences provided with the FarsNet project (Shamsfard et al., 2010). As the largest Farsi WordNet available, FarsNet is constructed in a semi-automatic manner. In its latest version (v3.0), the lexical resource covers more than 100K lexical entries in around 40K synsets. FarsNet synsets are aligned with those in PWN 3.0, whenever a link could be established. This allows utilizing the resource in cross-lingual applications.

Many of the synsets in FarsNet are provided with a usage example sentence for one of the terms in the corresponding Farsi synset. We take this as the basis for the construction of the Farsi dataset. Specifically, there are more than 30K Farsi usage examples that are linked to the corresponding synsets in PWN. From these, we extract a set of 4,272 sentences for 3,498 unique target words, after discarding monosemous words and filtering. The distribution of instances over the four parts of speech can be found in Table 1.

4 Methodology

We describe two WSD strategies based on contextualized embeddings (§4.1) and propose adaptations to the cross-lingual setting (§4.2). Figure 1 provides an overview of the overarching methodology.

4.1 Contextualized Embeddings for WSD

One of the most effective, yet simple, solutions for WSD is matching contextualized embeddings against precomputed sense embeddings learned using the same Neural Language Model (NLM). This approach has been used by earlier works (Melamud et al., 2016; Peters et al., 2018; Loureiro and Jorge, 2019), achieving state-of-the-art results. In these works, sense embeddings (or what we refer to as *contextualized sense embeddings*) are computed from the average of all corresponding contextual embeddings in sense-annotated corpora. One limi-

[3] Our active search efforts are described in the appendix. In general, existing WSD datasets were either under a restrictive license, not available anymore, or not linkable to English PWN.

tation of contextualized sense embeddings is that they only cover those senses that are present in the underlying training corpus. This issue can be alleviated by exploiting the structure of the semantic network. For this, we leverage the simple graph-based propagation method described in Loureiro and Jorge (2019), which allows for a coverage of the entire sense inventory of PWN.

With this strategy we can easily test whether contextualized sense embeddings can be transferred across languages with a solution purely based on matching nearest neighbors only, without any other artifacts. Below we describe the sense-based strategy generally used in the literature for monolingual WSD (sense strategy), and propose a new strategy that can be directly used in a cross-lingual setting (synset strategy).

Sense strategy. This strategy is the standard approach used in most WSD methods. After computing our sense embeddings from sense-annotated corpora, we disambiguate target words during testing based on a nearest neighbour strategy using their contextualized embeddings.

Synset strategy. In this alternative strategy, we learn synset embeddings by converting the annotated sense labels to synset labels, thus learning representations from multiple word senses that refer to the same concept, and becoming less reliant on lexical features.

4.2 Cross-lingual Adaptations

In cross-lingual experiments we are given sense-annotated corpora in a source language but not in the target language. Therefore, a multilingual NLM such as mBERT (Devlin et al., 2019) or XLM-R (Lample and Conneau, 2019) is required.

We adapt the sense strategy to the cross-lingual setting as follows. Given the lemma and part-of-speech of a word in the target language, we first gather all the candidate synsets in the source language from Babelnet. Then each candidate synset is associated with one or more senses. For example, we can find two candidate PWN senses for the word *presente* (*present* in Spanish). The first sense corresponds to the PWN synset "intermediate between past and future" and the second to "being or existing in a specified place". Finally, we compute the cosine distance from the contextualized embeddings of target word (*presente*) to all the candidate contextualized sense embeddings from the source language (*present%3:00:01::* and *present%3:00:02::*, respectively), and select the closest candidate sense. Note that the synset strategy does not require adaptation because synsets are language-independent.

5 Evaluation

We evaluate the proposed strategies in two different settings, using WSD as our test bed: monolingual (Section 5.1) and cross-lingual (Section 5.2).

Experimental setting. We use **RoBERTa** (Liu et al., 2019) for monolingual experiments, and **XLM-R** (Conneau et al., 2020) for cross-lingual experiments.[4] These two models are state-of-the-art for English and multilingual tasks while sharing a very similar architecture. The most important difference between these models is that while RoBERTa is pre-trained only on English texts, XLM-R is pre-trained on 100 languages (unevenly distributed). We use the large variants of these models (355M parameters for each).[5] All results are measured according to the F-measure.

5.1 Experiment 1: Monolingual WSD

Before delving into the cross-lingual experiments, we present monolingual results in English and Italian (languages with training data) in Table 2. The aim of this experiment is twofold: (1) compare the effectiveness of the disambiguation strategies, and (2) compare the performance of monolingual (RoBERTa) and multilingual models (XLM-R).

Baselines. As additional baselines, we add the results of the original BERT-based **LMMS** model (Loureiro and Jorge, 2019) and Context2Vec (**C2V**) (Melamud et al., 2016), which is also based on a simple nearest neighbors strategy, in this case with an LSTM instead of a transformer model.

Results. Table 2 shows the results of this English monolingual experiment.[6] We report results in all datasets from the unified WSD evaluation framework of Raganato et al. (2017).[7] As expected, the

[4] Following Loureiro and Jorge (2019), we consider token-level embeddings as the average of sub-token embeddings, which is computed as the sum of embeddings from the last 4 layers of the corresponding NLM.

[5] Our code is based on the Fairseq toolkit (Ott et al., 2019). We run our experiments on a single RTX 2070, with a runtime under 2 hours for generating all embeddings used in this work.

[6] We experimented with NLMs trained exclusively on Italian (i.e. UmBERTo-CC and dbmdz-IT-XXL), but found that the senses learned using those models do not consistently outperform the MFS baseline on both test sets.

[7] Datasets of the unified WSD framework: Senseval-2 (Edmonds and Cotton, 2001), Senseval-3 (Mihalcea et al., 2004),

Model	Strategy	SensEval-2	SensEval-3	SemEval-2007	SemEval-2013	SemEval-2015	ALL
RoBERTa	Sense	**75.5**	73.5	**69.2**	72.2	75.9	**73.9**
	Synset	74.4	**74.1**	68.4	72.1	**76.0**	73.6
XLM-R	Sense	71.0	67.8	61.5	70.1	72.1	69.5
	Synset	70.0	67.2	60.9	69.7	72.3	69.0
LMMS	Sense	*75.4*	*74.0*	*66.4*	***72.7***	*75.3*	*73.8*
C2V	Sense	*71.8*	*69.1*	*61.3*	*65.6*	*71.9*	*69.0*
MFS	–	*65.6*	*66.0*	*54.5*	*63.8*	*67.1*	*64.8*

Table 2: English monolingual F1 results on the evaluation framework of Raganato et al. (2017) for the two strategies: Sense (Se) and Synset (Sy).

purely monolingual model performs better than the multilingual one. As for the strategies, the usual sense strategy shows better performance. Nonetheless, the language-independent synset strategy attains competitive results, clearly outperforming a strong baseline such as Context2Vec, for example.

5.2 Experiment 2: Cross-lingual Transfer

We experimented with a pure zero-shot cross-lingual setting where senses are learned in one language (in our case English or Italian) and directly evaluated on another language. We employ the two strategies explained in Section 4.

Baselines. As baselines we include a random baseline (i.e., randomly picking a sense/synset from the target language's inventory), and a system that relies on the **static** word embeddings from XLM-R (i.e., input layer embeddings of XLM-R without making use of the context), instead of the **contextualized** sense embeddings obtained as described in Section 4.1.[8] The reason behind the possibility of using static word embeddings in this WSD setting lies in the fact that different senses of a word may be translated into different words in another language. This baseline makes use of the same sense and synset strategies.

Results. Table 3 shows the results for the zero-shot cross-lingual WSD experiment.[9] XLM-R outperforms the baselines by a large margin and proves to be robust in the cross-lingual setting (with performance in the same ballpark as in the monolingual setting). In particular, the fact that XLM-R outperforms the static embedding baseline by a large margin reinforces the idea that contextualized sense embedding are indeed transferable across languages (at least similar ones) to a large extent, which in turn opens up interesting avenues for future work on cross-lingual WSD. Nonetheless, as with English WSD, there are still many open questions as to what extent the fine granularity of PWN can be captured by automatic models.

Finally, as expected, learning representations using the larger English SemCor provides consistently better results than the smaller Italian counterpart, except for the distant language Farsi. More interestingly, XLM-R senses learned from English data can outperform the senses from the same model learned from language-specific data in the Italian test sets. Nonetheless, the simple synset strategy on Italian clearly surpasses the static baselines as well.

6 Conclusions

In this paper, we analyzed to what extent contextualized embeddings can be transferred across languages, using WSD as our test bed. To this end, we developed a unified framework that can be used for evaluating cross-lingual models. The first results are encouraging, as they show that multilingual language models can learn contextualized sense embeddings that can be effectively transferred from one language to another, attaining competitive results in WSD with no access to annotated data in the target language or external resources. One limitation of this work is in the nature of languages evaluated, which are all Indo-European for which test data was available. As future work it will be interesting to extend this benchmark to languages

SemEval-2007 (Agirre et al., 2007), SemEval-2013 (Navigli et al., 2013b), and SemEval-2015 (Moro and Navigli, 2015).

[8] We also tried to include Babelfy (Moro et al., 2014) as a multilingual knowledge-based baseline, without success. The latest public API of Babelfy misses over 50% of the instances in our benchmark and therefore the recall was suboptimal.

[9] English monolingual results slightly differ from those in Table 2, as in this case we focused on the BabelNet portion of the SemEval datasets (as explained in Section 3.2). For Italian we only report the synset strategy, as we could not have access to all the senses in MultiWordNet (Pianta et al., 2002).

Language	Strategy	Type	SemEval-13					SemEval-15			FN
			EN	FR	DE	ES	IT	EN	IT	ES	FA
English	**Se**	Static	*38.1*	28.0	50.1	38.7	41.0	*41.5*	45.2	40.6	48.7
		Contextualized	***67.4***	**60.7**	57.5	**69.7**	**66.1**	***71.7***	69.1	**68.0**	55.4
	Sy	Static	*41.0*	30.6	48.2	39.4	43.2	*39.8*	46.2	43.6	48.2
		Contextualized	*66.3*	59.0	**58.3**	66.8	64.9	*71.4*	**69.5**	67.2	**56.4**
Italian	**Sy**	Static	51.3	41.0	55.1	54.4	48.3	54.7	54.6	56.8	46.7
		Contextualized	63.2	55.9	55.4	65.9	*62.9*	67.1	*67.2*	65.4	**56.4**
Random baseline			37.8	25.8	47.9	38.7	38.6	41.7	44.5	38.0	49.5

Table 3: F1 results using zero-shot cross-lingual transfer (XLM-R) with English or Italian annotations. Two different types of sense embedding: static (S) and contextualized (C). Monolingual setting (i.e., learn and test in the same language) is also included for completeness.

from different families, for which cross-lingual embedding transfer has been shown to be more challenging (Glavaš et al., 2019; Doval et al., 2020).

References

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of SemEval-2007*.

Ali Alkhatlan, Jugal Kalita, and Ahmed Alhaddad. 2018. Word sense disambiguation for arabic exploiting arabic wordnet and word embedding. *Procedia Computer Science*, 142:50 – 60. Arabic Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. Mulan: Multilingual label propagation for word sense disambiguation. In *Proc. of IJCAI*, pages 3837–3844.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR*.

J. Daudé, L. Padró, and G. Rigau. 2000. Mapping WordNets using structural information. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 504–511, Hong Kong. Association for Computational Linguistics.

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.

Mona Diab, Musa Alkhalifa, Sabry ElKateb, Christiane Fellbaum, Aous Mansouri, and Martha Palmer. 2007. SemEval-2007 task 18: Arabic semantic labeling. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 93–98, Prague, Czech Republic. Association for Computational Linguistics.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2020. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4013–4023, Marseille, France. European Language Resources Association.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

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

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. SemEval-2007 task 05: Multilingual Chinese-English lexical sample. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 19–23, Prague, Czech Republic. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

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. *CoRR*, abs/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.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, pages 1–55.

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 Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28.

Rada Mihalcea and Phil Edmonds. 2004. Senseval-3: Third international workshop on the evaluation of systems for the semantic analysis of text. In *Association for Computational Linguistics*, volume 4.

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.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. *Proceedings of SemEval-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)*, 2:231–244.

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

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013a. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013b. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proceedings of SemEval 2013*, pages 222–231.

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.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. SemEval-2010 task: Japanese WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 69–74, Uppsala, Sweden. Association for Computational Linguistics.

Zeynep Orhan, Emine Çelik, and Demirgüç Neslihan. 2007. SemEval-2007 task 12: Turkish lexical sample task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 59–63, Prague, Czech Republic. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Tommaso Pasini and Jose Camacho-Collados. 2020. A short survey on sense-annotated corpora. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5759–5765, Marseille, France. European Language Resources Association.

Tommaso Pasini, Francesco Elia, and Roberto Navigli. 2018. Huge automatically extracted training-sets for multilingual word SenseDisambiguation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

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.

Tommaso Petrolito and Francis Bond. 2014a. A survey of WordNet annotated corpora. In *Proceedings of the Seventh Global Wordnet Conference*, pages 236–245, Tartu, Estonia. University of Tartu Press.

Tommaso Petrolito and Francis Bond. 2014b. A survey of wordnet annotated corpora. In *Proceedings Global WordNet Conference, GWC-2014*, pages 236–245.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. 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 EACL*, pages 99–110, Valencia, Spain.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5905–5911.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proceedings of the Thirty-Fourth Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

Mehrnoush Shamsfard, Akbar Hesabi, Hakimeh Fadaei, Niloofar Mansoory, Payam Noor, Ali Reza Gholi Famian, Somayeh Bagherbeigi, Elham Fekri, and Maliheh Monshizadeh. 2010. Semi automatic development of FarsNet; the Persian WordNet. In *Proceedings of 5th global WordNet conference*.

Kiyoaki Shirai. 2002. Construction of a word sense tagged corpus for SENSEVAL-2 Japanese dictionary task. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

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*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of COLING*, pages 1374–1385.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.

A Appendix: Compilation of WordNet-based WSD Datasets

In addition to the datasets included in our benchmark, mostly composed of European languages, we made an effort to retrieve and compile datasets for other languages. In Table 4 we provide details about our unsuccessful attempts and issues to integrate existing WSD datasets in the literature, mainly taken from DKPro[10] and the survey of Petrolito and Bond (2014b). Not only are most of these datasets unavailable, we also learn from their respective publications that the sense inventories used often aren't based on WordNet, and thus would require manual remapping of annotations for integration into our benchmark.

[10] https://dkpro.github.io/dkpro-wsd/corpora/

Resource	Language	# Instances	Inventory	Availability	License
Senseval-2 (Shirai, 2002)	Japanese	10,000	IKJ	N/A	N/A
Senseval-2 (Edmonds and Cotton, 2001)	Korean	N/A	N/A	N/A	N/A
Senseval-3 (Mihalcea and Edmonds, 2004)	Chinese	1,204	HowNet	Publicly Avail.	Public Domain
SemEval-2007 Task 5 (Jin et al., 2007)	Chinese	3,621	CSD	N/A	N/A
SemEval-2007 Task 11 (Orhan et al., 2007)	Turkish	5,385	TKD	N/A	N/A
SemEval-2007 Task 18 (Diab et al., 2007)	Arabic	888	AWN	N/A	N/A
SemEval-2010 Task 16 (Okumura et al., 2010)	Japanese	2,500	IKJ	On Request	Restrictive
OntoNotes (Weischedel et al., 2011)	Arabic	200K	Coarse WN	For Members	Restrictive
OntoNotes (Weischedel et al., 2011)	Chinese	800K	Coarse WN	For Members	Restrictive
Alkhatlan et al. (2018)	Arabic	240	AWN	N/A	N/A

Table 4: Details about the various WSD datasets covering non-European languages surveyed in our work (N/A: Not Available; WN: WordNet; AWN: Arabic WordNet; we refer to respective papers for remaining acronyms).