XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation

Tommaso Pasini[1*], **Alessandro Raganato**[2*], **Roberto Navigli**[1]
[1]Sapienza NLP Group, Computer Science Department, Sapienza University of Rome
[2] Department of Digital Humanities, University of Helsinki, Finland
{pasini,navigli}@di.uniroma1.it
alessandro.raganato@helsinki.fi

Abstract

Transformer-based architectures brought a breeze of change to Word Sense Disambiguation (WSD), improving models' performances by a large margin. The fast development of new approaches has been further encouraged by a well-framed evaluation suite for English, which has allowed their performances to be kept track of and compared fairly. However, other languages have remained largely unexplored, as testing data are available for a few languages only and the evaluation setting is rather matted. In this paper, we untangle this situation by proposing XL-WSD, a cross-lingual evaluation benchmark for the WSD task featuring sense-annotated development and test sets in 18 languages from six different linguistic families, together with language-specific silver training data. We leverage XL-WSD datasets to conduct an extensive evaluation of neural and knowledge-based approaches, including the most recent multilingual language models. Results show that the zero-shot knowledge transfer across languages is a promising research direction within the WSD field, especially when considering low-resourced languages where large pre-trained multilingual models still perform poorly. We make the evaluation suite and the code for performing the experiments available at https://sapienzanlp.github.io/xl-wsd/.

Introduction

Word Sense Disambiguation (WSD) is the task of associating words in context with their possible meanings contained in a pre-defined sense inventory (Navigli 2009). This task is central to the understanding of natural language (Navigli 2018), and it has received considerable attention over recent years as it can be beneficial for a variety of downstream tasks and applications, such as machine translation (Raganato, Scherrer, and Tiedemann 2019), information extraction (Delli Bovi, Telesca, and Navigli 2015), and text categorization (Shimura, Li, and Fukumoto 2019). The WSD task has been tackled with different approaches, which can be broadly divided into two main categories: knowledge-based (Moro, Raganato, and Navigli 2014; Agirre, de Lacalle, and Soroa 2014; Chaplot and Salakhutdinov 2018), which leverage computational lexicons and their structure, and supervised (Bevilacqua and Navigli 2020; Blevins and Zettlemoyer 2020; Conia and Navigli 2021), which train machine learning algorithms on sense-annotated data. This latter kind of approach attains state-of-the-art results in English WSD, constantly outperforming their knowledge-based counterparts (Raganato, Camacho-Collados, and Navigli 2017).

The evaluation in this field is usually carried out with the framework proposed by Raganato, Camacho-Collados, and Navigli (2017), which has set a level playing field among English WSD approaches, and has facilitated the fast development of models for this task (Raganato, Delli Bovi, and Navigli 2017; Luo et al. 2018; Huang et al. 2019; Bevilacqua and Navigli 2020; Bevilacqua, Maru, and Navigli 2020). Unfortunately, the same attention has not been devoted to multilingual WSD, which, in the last few years, has revolved around 4 European languages only, i.e., French, German, Italian and Spanish. Even though the research community has created both automatically sense-annotated corpora for different languages (Pasini 2020) and language-specific WordNet-like resources (Bond and Paik 2012; Navigli and Ponzetto 2012), the lack of reliable benchmarks in different languages remains the main limitation hampering the advancement of research in this field. Indeed, currently available multilingual gold standards use diverse data formats and outdated, or even unavailable, inventories of senses, making it hard to perform a fair comparison among systems and to draw reliable conclusions.

In this paper, we overcome the above problems and release what is, to the best of our knowledge, the first large-scale multilingual evaluation framework for WSD with a unified multilingual sense inventory covering 18 languages: Basque, Bulgarian, Catalan, Chinese, Croatian, Danish, Dutch, English, Estonian, French, Galician, German, Hungarian, Italian, Japanese, Korean, Slovenian, and Spanish from six families. On the one hand, we provide more than 70K new gold annotations across 13 non-English languages by leveraging the multilingual versions of WordNet. On the other hand, we standardise and unify the datasets available in another 4 languages from the past multilingual SemEval competitions, as well as the inventory of senses to be used across languages. This allows large multilingual models to be investigated through the semantics' lens, hence providing a new way of studying pre-trained contextualised word embeddings.

As for the English language, XL-WSD includes the original framework of Raganato, Camacho-Collados, and Navigli

[*]Authors marked with an asterisk (*) contributed equally.

(2017), further extending it, however, with data from: i) the fine-grained English WSD SemEval 2010 Task 17 (Agirre et al. 2010) and ii) the coarse-grained English WSD SemEval 2007 task 7 (Navigli, Litkowski, and Hargraves 2007). Moreover, XL-WSD features training data for the majority of its languages: SemCor (Miller et al. 1993) and the Princeton WordNet Gloss corpus[1] for English, and their automatically translated and annotated versions that we created for most of XL-WSD's other languages.

In summary, this paper makes the following novel contributions:

1. A multilingual WSD test suite in 18 languages from six language families, namely, Indo-European, Sino-Tibetan, Uralic, Japonica and Koreanic plus an isolated language, i.e., Basque. Our benchmark comprises 99,450 gold annotations in total, new automatically-produced training data for non-English languages and a unified multilingual inventory of concepts.
2. An extension of the fine-grained English WSD framework of Raganato, Camacho-Collados, and Navigli (2017) by including new training, development and testing data as well as a coarse-grained evaluation dataset.
3. Strong baselines based on large pre-trained multilingual language models and the first large-scale comparison among contextualised word embedding models and knowledge-based approaches on a monolingual and zero-shot cross-lingual setting.

Related Work

Word Sense Disambiguation has been tackled using various kinds of approach, from knowledge-based algorithms to fully supervised models. Knowledge-based methods (Chaplot and Salakhutdinov 2018; Maru et al. 2019) take advantage of the structural properties of a semantic network such as WordNet (Miller 1998), a manually-curated electronic dictionary for English, or BabelNet (Navigli and Ponzetto 2012), a large multilingual encyclopedic dictionary obtained by automatically merging various lexical resources (WordNet and Wikipedia, among others). While not relying on sense-annotated data, and hence being able to scale over different languages, knowledge-based approaches usually fall behind their supervised counterparts in terms of performance.

Supervised models (Vial, Lecouteux, and Schwab 2019; Huang et al. 2019; Bevilacqua and Navigli 2020; Scarlini, Pasini, and Navigli 2020; Blevins and Zettlemoyer 2020; Conia and Navigli 2021), by exploiting SemCor (Miller et al. 1993) – the largest manually-annotated corpus for English – have consistently attained state-of-the-art results on the English all-words WSD tasks (Raganato, Camacho-Collados, and Navigli 2017). However, their main drawback is that they have difficulty scaling over different languages, since no manually-curated training data is available to them. Automatic methods to produce sense distributions (Pasini, Scozzafava, and Scarlini 2020) or sense-annotated data in languages other than English (Delli Bovi et al. 2017; Scarlini, Pasini, and Navigli 2019; Pasini and Navigli 2020; Pasini 2020) have mitigated this limitation, thus allowing supervised approaches to be trained on different languages and to enter a field that was mainly dominated by knowledge-based methods.

Importantly, while multilingual word embeddings and, more recently, deep multilingual pre-trained neural language models have proven to perform zero-shot transfer from one language to another effectively, cross-lingual WSD research has been dramatically hampered by the lack of a clear and large-scale multilingual evaluation suite. Indeed, the evaluation benchmarks proposed over the years in the context of Senseval and SemEval competitions have focused mainly on English: Senseval-2 (Edmonds and Cotton 2001), Senseval-3 (Snyder and Palmer 2004), SemEval-07 Task 17 (Pradhan et al. 2007), SemEval-07 Task 7 (Navigli, Litkowski, and Hargraves 2007), SemEval-10 Task 17 (Agirre et al. 2010), SemEval-13 Task 12 (Navigli, Jurgens, and Vannella 2013) and SemEval-15 Task 13 (Moro and Navigli 2015), with only a few of them providing data for other languages too, i.e., SemEval-10 Task 17, SemEval-13 Task 12 and SemEval-15 Task 13. While the WSD framework proposed by Raganato, Camacho-Collados, and Navigli (2017) systematised and unified the datasets for the English fine-grained WSD task, it focused on English only and did not include any of the available multilingual datasets. As a result, WSD multilingual benchmarks today are still outdated, featuring old, language-specific or even unavailable sense inventories, which limits their use. This is in marked contrast to other NLP tasks where many efforts have been made to evaluate models across languages (Hu et al. 2020; Lewis et al. 2020; Ponti et al. 2020; Raganato et al. 2020; Martelli et al. 2021, XTREME, MLQA, XCOPA, XL-WiC, MCL-WiC, respectively),

To bridge this gap, we put forward a comprehensive multilingual WSD evaluation framework containing new gold development and test sets, as well as silver training data in 18 languages from 6 distinct language families, which ensures an easy and fair evaluation of WSD systems across languages. XL-WSD is similar to other multilingual evaluation benchmarks in terms of the number of instances, languages and linguistic families covered. Indeed, it is comparable to tasks like XTREME in terms of instances and covers more families than MLQA and more languages than XCOPA. Moreover, our framework also includes and enriches the original English test suite for WSD of Raganato, Camacho-Collados, and Navigli (2017), by featuring coarse-grained datasets, and a larger training set.

XL-WSD

In this Section, we detail the creation of XL-WSD. First, we define the unified multilingual sense inventory and introduce the new multilingual gold standards. Then, we present the new multilingual training data providing relevant statistics.

Sense Inventory

Sense inventories define the possible meanings for a word, and, while the Princeton WordNet (PWN) is the *de facto* standard sense inventory for English, there is no such convention in other languages.

[1] http://wordnetcode.princeton.edu/glosstag.shtml

Over the years, several efforts have been made to create WordNet-like resources in multiple languages and to link them to the PWN (Bond and Foster 2013). A superset of these lexical resources is BabelNet[2], a comprehensive multilingual encyclopedic dictionary that merges various resources (WordNet and Wikipedia, among others) into a unified multilingual repository. It provides a wide coverage of concepts across languages and several lexicalizations for each meaning, e.g., the machine meaning of the English word *computer* is lexicalised with *ordinateur* in French, *computadora* in Spanish, *calcolatore* in Italian, etc.

Therefore, we draw the sense inventory from BabelNet (version 4.0) and define the list of $117{,}659$ BabelNet synsets containing at least one sense from the Princeton WordNet (version 3.0) as our set of possible meanings $\mathcal{S}$. We constrain our synsets to contain at least one PWN sense, so as to allow training a model in English and testing in other languages and to ensure a wide coverage of meanings across many languages. Indeed, most non-English WordNets are created by, either translating PWN synsets into the target language (*extend mode*), or by linking newly created concepts to the PWN (*merge mode*) (Vossen 1998). Once the set $\mathcal{S}$ of synsets is defined, we extract the set of lemmas specific to a language L by collecting all lexicalisations of any synset in $\mathcal{S}$ in that language. We then associate each lemma and part-of-speech (POS) pair (l, p) with the set of its possible meanings $s \in \mathcal{S}$, i.e., all those synsets with POS tag p containing l among their lexicalisations.

We are aware that limiting the conceptualisations of other languages to the English PWN may not define a faithful equivalent of a dictionary in other languages. However, doing so allows us to create a shared multilingual sense inventory, enabling a fair evaluation of models in the cross-lingual setting.

Gold Standards

WordNet datasets. The Princeton English WordNet organizes concepts in synsets, i.e., sets of synonyms, and provides, for each of them, one or more usage examples, i.e., sentences in which one of the synset's lexicalizations is used with that meaning. For example, the slope synset of *bank* contains the example "They pulled the canoe up on the *bank*", while the financial institution synset contains the example "He cashed a check at the *bank*". We leverage this structure that is common across WordNet-like resources and create new evaluation benchmarks from the following language-specific WordNets: Basque (Pociello et al. 2008), Bulgarian (Simov and Osenova 2010), Catalan (Benítez et al. 1998), Chinese (Huang et al. 2010), Croatian (Raffaelli et al. 2008), Danish (Pedersen et al. 2009), Dutch (Postma et al. 2016), Estonian (Vider and Orav 2002), Galician (Guinovart 2011), Hungarian (Miháltz et al. 2008), Japanese (Isahara et al. 2008), Korean (Yoon et al. 2009), and Slovenian (Fišer, Novak, and Erjavec 2012). The Galician, Catalan, and Basque WordNets are taken from the Multilingual Central Repository project (Gonzalez-Agirre, Laparra, and Rigau 2012), while the Bulgarian, Japanese, and Slovenian from the Open Multilingual WordNet project (Bond and Paik 2012).

In detail, given a synset s within a language-specific WordNet, and one of its usage examples $e = w_1, \ldots, w_n$, we select as target word the one having the same POS tag of s, and, as lemma, one of the lexicalisations of s. For example, given the nominal synset s for salmon, which contains the Danish word *laks* as one of its lexicalisations and the Danish example "Stjernerne i bornholmernes fiskerierhverv er ørred, laks og sild",[3] we mark *laks* as target word since it has been POS tagged with the same tag as s, and is a lexicalisation of s. If we find more than one word matching our criterion, we discard the sentence.

Finally, we leverage the available mapping from the language-specific WordNet to the English WordNet 3.0 and the mapping from WordNet 3.0 to BabelNet included in BabelNet itself, so as to tag each instance with the corresponding BabelNet synset within our sense inventory.

SemEval datasets. We consider all multilingual gold standards released in the past SemEval competitions, i.e., the Italian and Chinese datasets in SemEval-10 Task 17[4] (Agirre et al. 2010), French, German, Italian and Spanish datasets in SemEval-13 Task 12 (Navigli, Jurgens, and Vannella 2013), and Italian and Spanish datasets in SemEval-15 Task 13 (Moro and Navigli 2015).

The SemEval-10 dataset contains documents from the European Center for Nature Conservation and the Worldwide Wildlife Forum corpora. The SemEval-13 datasets contain 13 parallel documents from 2010, 2011 and 2012 editions of the Workshop on Statistical Machine Translation. French, German, and Spanish text data come directly from the Workshop, the Italian dataset, instead, was created by manually translating the English documents. As regards the datasets in SemEval-15, they were taken from the EMA (European Medicines Agency documents), KDEdoc (KDE manuals) and EUbookshop (documents from the EU bookshop) corpora. Originally, the SemEval-15 datasets were built for both all-words WSD and Entity Linking tasks. In this work, we use only the instances in the WSD split.

As for English, we consider all datasets in the Raganato, Camacho-Collados, and Navigli (2017) framework plus the English data from SemEval-10 Task 17 and the coarse-grained dataset from SemEval-07 Task 7 (Navigli, Litkowski, and Hargraves 2007, SemEval-07-Coarse). This latter contains documents extracted from the Wall Street Journal corpus, Wikipedia and the *Knights of the Art* book by Amy Steedman and is annotated with clusters of WordNet senses.

Data cleaning. Most datasets from the past SemEval competitions use different inventories. Specifically, Chinese and Italian datasets of SemEval-10 are tagged with WordNet 1.6; SemEval-07-Coarse is annotated with clusters of WordNet senses from version 2.1; SemEval-13 and SemEval-15, in-

[2]https://babelnet.org

[3]The stars of the Bornholm fishing industry are trout, salmon and herring.

[4]The SemEval-10 Dutch sense inventory is no longer available.

stead, use different versions of BabelNet as inventory, i.e., BabelNet 1.1.1 and BabelNet 2.5.1, respectively.

To standardise WordNet versions, we convert all the annotations from WordNet 1.6 and 2.1 to WordNet 3.0 utilising the automatically-generated mappings of Daude, Padro, and Rigau (2003), keeping the synsets with the highest confidence score only. As regards the instances tagged with BabelNet 1.1.1 and 2.5.1, we first map each annotation from its original BabelNet version to the latest available one (4.0), by using the corresponding BabelNet indices, and, then, retain only the instances tagged with a synset in our inventory. We finally remove all the instances that could not be mapped. All the other datasets, instead, have already been mapped to WordNet version 3.0, so we retrieve their corresponding BabelNet synset with the BabelNet API 4.0.1.

Evaluation split. We group all the datasets in the same language and randomly split their instances into two subsets, one for testing (80% of instances) and one for development (the remaining 20% of instances). As for English, instead, we provide 2 distinct test sets: a fine-grained one (English-F) including Senseval-2, Senseval-3, SemEval-10, SemEval-13 and SemEval-15, and a coarse-grained one (English-C), i.e., SemEval-07 Task 17. As for development, we follow prior work (Raganato, Delli Bovi, and Navigli 2017; Blevins and Zettlemoyer 2020) and use SemEval-07 (English-Dev). As a result, each language has a test and a development set in the same data format and tagged with the same unified inventory. Furthermore, we enrich the English benchmark of Raganato, Camacho-Collados, and Navigli (2017) with 3K more instances, covering different sense granularities.

Training Data

SemCor (SC). Introduced by Miller et al. (1993), this is the most used corpus for English Word Sense Disambiguation. It contains 37,176 sentences and 226,036 instances tagged with a sense in WordNet.

Princeton WordNet Gloss Corpus (WNG). A corpus created from the synset definitions and examples of WordNet.[5] Its annotations were carried out both manually and semi-automatically. By following Bevilacqua and Navigli (2020), given a gloss g for a sense s, we prepend to g the lemma of s and tag it with s so as to provide at least one annotated example for each concept. In total, it consists of 614,435 instances tagged with 117,653 different synsets.

Translated corpora (T-SC+WNG). We provide silver training data to train language-specific baselines for 15 non-English languages of our framework[6] by leveraging the machine translation models made available by Tiedemann and Thottingal (2020, Opus-MT).[7] The choice of these models is motivated, first, by the fact that both the English training corpora (SC and WNG) and the training data for the machine translation models (the OPUS parallel corpora collection (Tiedemann 2012)) are general-domain,[8] and, second, by considering that training several domain-specific models for each target language is resource expensive and beyond the scope of this work.

We create the language-specific training corpora by translating the English sentences of SC and WNG into the target languages, and, then, by transferring the sense annotations from the original English texts to their translations. In more detail, given an English sentence $\sigma^{EN} = w_1, \ldots, w_n$, its translation $\sigma^T = w_1^T, \ldots, w_m^T$ and the *synset* annotation s for the word w_i in σ^{EN}, we propagate the annotation to the word w_j^T in σ^T that appears as a synonym in s. In the case that multiple annotations are associated with the same word w_j^T, we discard all of them. To further refine the quality of the projections, we apply a part-of-speech tagger and a lemmatiser to both source and target languages, keeping only those senses in which both source and target words are tagged with the same part of speech.[9]

Our goal is not to create the best possible datasets, but rather to enable the training of monolingual baselines which can be used as a term of comparison for future work. Our approach, moreover, has the following advantages: i) it allows annotations to be automatically spread from one language to many others without human effort, ii) the sense distribution is potentially maintained across languages, iii) it produces annotations for virtually any word and language covered by BabelNet and for which a machine translation model exists.

Statistics

We report the general statistics for each dataset of XL-WSD in Table 1. The number of annotated instances in the training data varies across languages, ranging from more than 800K in English, to less than 25K in Japanese and cover from roughly 1,000 to 117K different synsets depending on the language. Even though the non-English training data, i.e., T-SC+WNG, are all created starting from the same source, i.e., SC+WNG, the number of transferred instances is affected by both translation quality and BabelNet's coverage of each specific language. As regards the test sets, most languages contain more than 1,000 gold annotations, with Bulgarian and Chinese containing even more test instances than English. Additionally, Table 1 shows the number of different word types for each language, the number of polysemous word types, i.e., words with more than one meaning, and the word-type polysemy measure, i.e., the total number of candidate synsets for each word type divided by the total number of word types. The word-type polysemy is similar across language-specific training sets as they all come from the translation of SC+WNG. On the other hand, the polysemy varies substantially across test sets, with Croatian having the least polysemous test set (1.24) and Spanish the most polysemous one (4.95).

In total, XL-WSD contains more than 99K semantically-tagged gold instances for testing and tuning across 18 differ-

[5] http://wordnetcode.princeton.edu/glosstag.shtml

[6] Chinese and Korean have no MT models at the time of writing.

[7] https://github.com/Helsinki-NLP/Opus-MT

[8] More details about the translation models and their translation quality are given at https://sapienzanlp.github.io/xl-wsd/.

[9] We use Stanza pre-trained neural models (Qi et al. 2020).

Language	Word Types		Polysemous Words		Word-Type Polysemy		Instances		Unique Synsets	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
English-F	106906	2882	24658	2199	1.458	3.689	840471	8062	117653	3469
English-C	-	980	-	750	-	4.255	-	1816	-	2190
Basque	12503	771	5294	525	2.331	3.224	197309	1580	16604	1423
Bulgarian	12413	2450	2412	1325	1.304	1.670	148479	9968	12600	2658
Catalan	18603	1276	8378	1107	2.291	3.940	331757	1947	25624	1767
Chinese	-	1786	-	1402	-	2.638	-	9568	-	2687
Croatian	6882	4389	1161	1652	1.268	1.244	94575	6333	6739	
Danish	15822	2623	3324	1318	1.338	1.722	234681	3502	16707	2693
Dutch	28351	2935	9121	2122	1.711	2.356	305692	4400	30490	2716
Estonian	10460	1615	1768	917	1.246	1.815	132240	1999	10462	1852
French	17850	549	5978	339	1.585	2.413	252756	1160	21510	584
Galician	8390	1244	3799	773	2.079	2.219	247379	2561	11821	1474
German	16213	421	2332	166	1.203	1.639	184952	862	16437	417
Hungarian	13234	3491	2908	1931	1.367	1.842	161119	4428	13297	4285
Italian	23773	985	9540	758	2.021	3.790	385248	2278	29869	1212
Japanese	1008	4338	581	2390	2.516	1.871	23217	7602	1141	5964
Korean	-	1886	-	920	-	1.373	-	3796	-	1452
Slovenian	7577	104	1296	93	1.245	3.519	128395	2032	7705	243
Spanish	22020	847	11784	696	2.811	4.955	393539	1851	32151	1103

Table 1: Statistics of XL-WSD training and test sets. "Train" column refers to SC+WNG for English and T-SC+WNG for the other languages.

ent languages, 3M silver annotations from the T-SC+WNG datasets and more than 100K annotations for English.

Experimental Setup

Architecture details. We follow Bevilacqua and Navigli (2020) and employ a Transformer-based text encoder (Vaswani et al. 2017) followed by a 2-layer feedforward network with *swish* activation function and batch-normalization. We stack on top of it an unbiased softmax linear layer for classification. We represent each sub-token by summing the outputs of the last four layers of the text encoder and each word by averaging its sub-token representations. Finally, we apply a linear transformation and feed the resulting vectors to a linear layer for classification. As text encoders, we use XLMR-B, XLMR-L (Conneau et al. 2020), BERT-L, BERT-M[10] (Devlin et al. 2019) and the language-specific versions of BERT (LS-BERT) for each language[11] available through the Huggingface library (Wolf et al. 2020). We train all neural models for 50 epochs with Adam optimizer[12] and use the set of weights with the lowest loss on the development set for testing.

Evaluation measure. As standard in the literature, we adopt the F1 score, i.e., the harmonic mean between Precision and Recall. We note that Precision, Recall and F1 score are the same, as our models always provide an answer.

Data. As for training, we use SemCor and WordNet Gloss (SC+WNG) for English and their translations, i.e., T-SC+WNG, for the monolingual experiments in other languages. As a term of comparison, we also report the results attained by training our baseline models on MULAN[13] datasets (Barba et al. 2020). Differently from our approach, MULAN leverages the multilingual contextualised word representations of BERT to pair manually-tagged examples in English with their most similar sentences in a corpus of raw texts, e.g., Wikipedia, and then transfer the sense annotations.

Model	English-F	English-C	ALL
			SemCor
LMMS ◇	-	-	75.40
GlossBERT ◇	-	-	77.00
BERT_{GLU}	-	-	74.10
BEM◇	-	-	**79.00**
XLMR-B	71.29	86.01	73.21
BERT-M	69.19	84.80	61.54
XLMR-L	72.46	86.12	74.24
BERT-L	**72.66**	**86.51**	74.33
			SC+WNG
EWISER	-	-	**80.10**
XLMR-B	74.50	91.02	76.24
BERT-M	72.40	89.70	74.10
XLMR-L	76.28	91.30	78.11
BERT-L	**76.77**	**91.57**	78.36

Table 2: Comparison on the English datasets. ○ indicates that the model is an ensemble, while ◇ indicates that the model leverages raw sense definitions.

[10]The base multilingual-cased version of BERT.

[11]More details at https://sapienzanlp.github.io/xl-wsd/.

[12]Gradient clipping = 1.0; learning rate = $2 \cdot 10^{-5}$; patient = 3.

[13]https://github.com/SapienzaNLP/mulan

	∅-Shot (SC+WNG)			Language-Specific (MULAN)		Language-Specific (T-SC+WNG)		Knowledge-Based		
Dataset	XLMR-L	XLMR-B	BERT-M	XLMR-L	LS-BERT	XLMR-L	LS-BERT	SyntagRank	Babelfy	MCS
English-F	76.28	74.50	72.40	-	-	76.28	**76.77**	69.96	64.09	63.37
English-C	91.30	91.02	89.70	-	-	91.30	**91.57**	83.78	82.54	80.23
Basque	**47.15**	43.80	42.41	-	-	41.96	43.04	42.91	36.65	32.72
Bulgarian	**72.00**	71.59	68.78	-	-	58.18	57.85	61.10	60.39	58.16
Catalan	**49.97**	47.77	47.35	-	-	36.00	36.98	43.98	36.65	27.17
Chinese	**51.62**	49.77	48.99	-	-	-	-	41.23	34.94	29.62
Croatian	**72.29**	72.13	70.65	-	-	63.15	62.89	68.35	63.75	62.88
Danish	**80.61**	79.18	76.04	-	-	78.67	76.41	72.93	71.33	64.33
Dutch	**59.20**	58.77	56.64	-	-	57.27	56.64	56.00	44.27	44.61
Estonian	**66.13**	64.82	64.33	-	-	50.78	51.23	56.31	49.62	46.87
French	**83.88**	82.33	81.64	81.98	80.78	71.38	71.12	69.57	67.41	59.31
Galician	66.28	64.79	**68.07**	-	-	56.18	56.95	67.56	64.17	60.85
German	83.18	82.13	80.63	**83.29**	82.13	73.78	73.78	75.99	77.84	75.99
Hungarian	67.64	**68.38**	65.24	-	-	52.60	52.17	57.98	51.99	47.29
Italian	77.66	76.73	76.16	74.10	73.88	**77.70**	75.68	69.57	64.22	52.77
Japanese	**61.87**	61.46	60.34	-	-	50.55	50.16	57.46	51.91	48.71
Korean	**64.20**	63.65	63.37	-	-	-	-	50.29	51.95	52.48
Slovenian	**68.36**	66.34	62.16	-	-	51.13	49.66	52.25	35.38	36.71
Spanish	75.85	76.55	74.66	73.47	74.77	**77.26**	74.88	68.58	64.07	55.65
AVG	**65.66**	64.82	62.84	-	-	-	-	57.68	52.85	49.31

Table 3: F1 scores of supervised and knowledge-based approaches as well as language-specific BERT models (LS-BERT) and the Most Common Sense (MCS) baseline on the test splits. As for the ∅-Shot columns, models are trained and tuned in English only and tested in all the other languages. As for the Language-Specific columns, models are trained, tuned and tested on either MULAN or T-SC+WNG language-specific datasets. The AVG row shows the micro F1 across all languages but English.

As for development and testing, we use the language-specific data that we previously introduced. We also consider the ALL dataset in the Raganato, Camacho-Collados, and Navigli (2017) framework, which comprises Senseval-2, Senseval-3, SemEval-07, SemEval-13 and SemEval-15 to compare our baselines against the state of the art.

Results

English Benchmark

As a preliminary experiment, in Table 2 we compare our baselines with the most recent WSD models in the literature on the English datasets, to give an idea about how our models compare against the state of the art.

As one can see, our baselines perform in the same ballpark as most of the other approaches. When using SemCor only for training, BEM is the best system across the board, however, it requires the finetuning of two distinct BERT-base models and leverages raw WordNet glosses. When using the SC+WNG dataset, instead, both BERT-L and XLMR-L perform less than 2 F1 points lower than EWISER, which, however, leverages additional information from sense embeddings and the topology of a knowledge graph.

Therefore, since our multilingual baselines attain results that are comparable with the current best performing models for WSD, we employ them to carry out the evaluation.

Multilingual Evaluation

Table 3 shows the performance on the proposed multilingual benchmark, reporting the results attained by our reference models trained and tuned, i) on English data only, i.e., SC+WNG, ii) on the automatically-translated language-specific training data, i.e., T-SC+WNG, and iii) on MULAN. Additionally, we consider two knowledge-based approaches: Babelfy (Moro, Raganato, and Navigli 2014), which is based on a densest sub-graph algorithm, and SyntagRank (Scozzafava et al. 2020), which relies on the Personalized PageRank algorithm and leverages the collocational relations in SyntagNet (Maru et al. 2019). We also show the results of the Most Common Sense (MCS) baseline, which tags each word with its most common sense according to BabelNet.

Zero-shot setting. As one can see from Table 3, XLMR-L achieves the best results across the board, with a big gap with respect to knowledge-based systems. Interestingly enough, supervised models trained on English data only (zero-shot columns) almost always outperform their language-specific counterparts, i.e., either multilingual models trained on language-specific training sets (Language-Specific / XLMR-L columns) or language-specific models trained on language-specific data (LS-BERT columns). We note the same behaviour for French, German, Italian and Spanish, where MULAN training data are also available.

These results are in line with the most recent findings, i.e., that large multilingual language models play a key role in

Dataset	ALL	N	V	A	R
English-F	76.28	77.92	65.74	81.47	86.71
English-C	91.30	92.72	88.64	89.55	91.75
Basque	47.15	47.15	-	-	-
Bulgarian	72.00	70.69	86.04	74.07	-
Catalan	49.97	49.28	54.84	52.89	-
Chinese	51.62	57.92	45.47	47.01	84.48
Croatian	72.29	71.85	70.37	85.03	-
Danish	80.61	80.32	79.66	83.63	-
Dutch	59.20	56.08	63.56	-	-
Estonian	66.13	68.81	49.66	74.63	68.14
French	83.88	83.88	-	-	-
Galician	66.28	71.43	-	65.97	-
German	83.18	83.18	-	-	-
Hungarian	67.64	70.41	50.41	-	-
Italian	77.66	77.91	71.89	81.58	77.27
Japanese	61.87	67.87	52.72	56.39	71.29
Korean	64.20	64.47	46.43	-	-
Slovenian	68.36	68.34	-	-	-
Spanish	75.85	76.72	66.83	77.88	85.00

Table 4: XLMR-L F1 on the zero-shot setting by POS tags, i.e., nouns (N), verbs (V), adjectives (A) and adverbs (R).

making up for the paucity of annotated data in non-English languages (Conneau et al. 2020), and therefore represent a promising approach towards mitigating the knowledge-acquisition bottleneck problem in WSD. Furthermore, while the multilingual WSD task has so far usually been addressed with knowledge-based approaches, it is now clear that thanks to large multilingual pre-trained language models, neural networks can compete in this task too. Indeed, despite the fact that SyntagRank manages to outperform several language-specific models trained on T-SC+WNG, it performs 5 and 7 points lower than BERT-M and XLMR-B, respectively, and falls behind XLMR-L trained on English by 8 points on average. These results corroborate previous English-focused artificially large-scale findings on the robustness of supervised WSD approaches (Pilehvar and Navigli 2014).

Language-specific setting. Overall, pre-trained language-specific BERT models perform equal or lower than their multilingual counterparts. This is mainly due to the difference in the model size, indeed, XLMR-L has roughly 200M more parameters than most of the language-specific models, which are based on BERT-B. Interestingly, our newly introduced training data, i.e., T-SC+WNG, despite being a baseline, proves to lead the neural models to attain higher performance than when trained on MULAN, i.e., high-quality silver data, in Italian and Spanish. This is explained by the fact that Italian and Spanish datasets contain the highest number of transferred labels, as shown in Table 1.

Discussion. Overall, XLMR-L is the best performing model scoring 65.66 on average on all non-English languages. Its extensive pre-training and the number of parameters play a crucial role in achieving such high scores. Nevertheless, we note that it still performs poorly in some languages, i.e., Basque, Catalan, and Chinese. This is due to the fact that a large portion of the test instances are annotated with synsets occurring only a few times in the training data, thus making these datasets particularly challenging. This also highlights that representing the least frequent meanings remains an open issue even for large pre-trained language models.

In Table 4 we provide further insights by showing the results breakdown on each POS tag of the best performing model, i.e., XLMR-L. As one can see, verbs represent the most challenging instances in most languages with an average F1 10 points lower than on nouns. Bulgarian, Catalan and Dutch are the only languages where the model performs better on verbs than on nouns. This is because verb instances in Bulgarian and Dutch are in general less polysemous than nouns. As for Catalan, instead, while verbs are more polysemous than nouns, the test set contains only 31 verbal instances, hence making the test on verbs not significant.

Overall, there is still large room for improvement in multilingual and zero-shot Word Sense Disambiguation. Specifically, our benchmarks show that the gap between English and other languages is in general wide, with XLMR-L performing, on average, 10 points lower than on English. This highlights the fact that word meanings are still not well captured by state-of-the-art language models, which struggle both on low-resourced languages, such as Catalan or Basque, as well as on resource-rich languages, e.g., Chinese.

Conclusion

In this paper, we presented XL-WSD, a large-scale evaluation benchmark for Word Sense Disambiguation in 18 different languages. On the one hand, XL-WSD features 34 new gold datasets for testing and tuning in 17 non-English languages and 15 silver datasets for training, which we built automatically by translating manually-annotated data into the target languages. On the other hand, it includes and enriches the previously available standard evaluation framework for English (Raganato, Camacho-Collados, and Navigli 2017) with the addition of two test sets, i.e., SemEval-10 and SemEval-07-Coarse. All datasets share a common format and, more importantly, a unified multilingual sense inventory, thus allowing a fair and easy comparison among systems that was previously out of reach. Furthermore, we provided strong baselines for the multilingual WSD task and, for the first time, a large-scale evaluation of different contextualised word embedding models on a task with explicit semantics, comparing their results across languages against those achieved by knowledge-based models. XL-WSD stands, therefore, as a key semantic benchmark not only in terms of size (i.e. the number of test instances), but also in terms of coverage (i.e. the number of covered languages), thereby fostering research in multilingual WSD and cross-lingual transfer.

As future work, we plan to further extend our framework by validating the data, manually annotating datasets in new languages and providing standard splits for testing a model on instances tagged with senses having different frequencies.

XL-WSD code and data are freely available for research purposes at https://sapienzanlp.github.io/xl-wsd/.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grants MOUSSE No. 726487, FoTran No. 771113, and the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme. Authors also thank the IT Center for Science (Finland).

References

Agirre, E.; de Lacalle, O. L.; and Soroa, A. 2014. Random Walks for Knowledge-Based Word Sense Disambiguation. *CL* 40(1).

Agirre, E.; Lopez de Lacalle, O.; Fellbaum, C.; Hsieh, S.-K.; Tesconi, M.; Monachini, M.; Vossen, P.; and Segers, R. 2010. SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain. In *Proc. of SemEval*.

Barba, E.; Procopio, L.; Campolungo, N.; Pasini, T.; and Navigli, R. 2020. MuLaN: Multilingual Label propagatioN for Word Sense Disambiguation. In *Proc. of IJCAI*.

Benítez, L.; Cervell, S.; Escudero, G.; López, M.; Rigau, G.; and Taulé, M. 1998. Methods and tools for building the Catalan WordNet. *Proc. of ELRA* .

Bevilacqua, M.; Maru, M.; and Navigli, R. 2020. Generationary or:"How We Went beyond Word Sense Inventories and Learned to Gloss". In *Proc. of EMNLP*, 7207–7221.

Bevilacqua, M.; and Navigli, R. 2020. Breaking through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information. In *Proc. of ACL*.

Blevins, T.; and Zettlemoyer, L. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders. In *Proc. of ACL*.

Bond, F.; and Foster, R. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proc. of ACL*.

Bond, F.; and Paik, K. 2012. A Survey of WordNets and their Licenses. In *Proc. of GWC 2012*, volume 8.

Chaplot, D. S.; and Salakhutdinov, R. 2018. Knowledge-based word sense disambiguation using topic models. In *Proc. of AAAI*.

Conia, S.; and Navigli, R. 2021. Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration. In *Proc. of EACL*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-Lingual Representation Learning at Scale. In *Proc. of ACL*.

Daude, J.; Padro, L.; and Rigau, G. 2003. Validation and tuning of wordnet mapping techniques. In *Proc. of RANLP*.

Delli Bovi, C.; Camacho-Collados, J.; Raganato, A.; and Navigli, R. 2017. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *ACL*.

Delli Bovi, C.; Telesca, L.; and Navigli, R. 2015. Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis. *TACL* 3.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*.

Edmonds, P.; and Cotton, S. 2001. SENSEVAL-2: Overview. In *Proc. of SENSEVAL-2*.

Fišer, D.; Novak, J.; and Erjavec, T. 2012. SloWNet 3.0: development, extension and cleaning. In *Proc. of GWC*.

Gonzalez-Agirre, A.; Laparra, E.; and Rigau, G. 2012. Multilingual Central Repository version 3.0. In *Proc. of LREC*.

Guinovart, X. G. 2011. Galnet: WordNet 3.0 do galego. *Linguamática* 3(1).

Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; and Johnson, M. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proc. of ICML*.

Huang, C.-R.; Hsieh, S.-K.; Hong, J.-F.; Chen, Y.-Z.; Su, I.-L.; Chen, Y.-X.; and Huang, S.-W. 2010. Chinese Wordnet: Design, Implementation and Application of an Infrastructure for Cross-Lingual Knowledge Processing. *Journal of Chinese Information Processing* 24(2): 14.

Huang, L.; Sun, C.; Qiu, X.; and Huang, X. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proc. of EMNLP*.

Isahara, H.; Bond, F.; Uchimoto, K.; Utiyama, M.; and Kanzaki, K. 2008. Development of the Japanese WordNet. In *Proc. of LREC*.

Lewis, P.; Oguz, B.; Rinott, R.; Riedel, S.; and Schwenk, H. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proc. of ACL*.

Luo, F.; Liu, T.; Xia, Q.; Chang, B.; and Sui, Z. 2018. Incorporating Glosses into Neural Word Sense Disambiguation. In *Proc. of ACL*.

Martelli, F.; Kalach, N.; Tola, G.; and Navigli, R. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proc. of SemEval*.

Maru, M.; Scozzafava, F.; Martelli, F.; and Navigli, R. 2019. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In *Proc. of EMNLP*.

Miháltz, M.; Hatvani, C.; Kuti, J.; Szarvas, G.; Csirik, J.; Prószéky, G.; and Váradi, T. 2008. Methods and Results of the Hungarian WordNet Project. In *Proc. of GWC*.

Miller, G. A. 1998. *WordNet: An electronic lexical database*. MIT press.

Miller, G. A.; Leacock, C.; Tengi, R.; and Bunker, R. T. 1993. A semantic concordance. In *Proc. of HLT*.

Moro, A.; and Navigli, R. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval*.

Moro, A.; Raganato, A.; and Navigli, R. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *TACL* .

Navigli, R. 2009. Word Sense Disambiguation: A Survey. *ACM computing surveys (CSUR)* 41(2).

Navigli, R. 2018. Natural Language Understanding: Instructions for (Present and Future) Use. In *Proc. of IJCAI*.

Navigli, R.; Jurgens, D.; and Vannella, D. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Proc. of SemEval*.

Navigli, R.; Litkowski, K. C.; and Hargraves, O. 2007. SemEval-2007 Task 07: Coarse-Grained English All-Words Task. In *Proc. of SemEval*.

Navigli, R.; and Ponzetto, S. P. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intel* .

Pasini, T. 2020. The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation. In *Proc. of IJCAI*.

Pasini, T.; and Navigli, R. 2020. Train-O-Matic: Supervised Word Sense Disambiguation with no (manual) effort. *Artif. Intel.* 279.

Pasini, T.; Scozzafava, F.; and Scarlini, B. 2020. CluBERT: A cluster-based approach for learning sense distributions in multiple languages. In *Proc. of ACL*, 4008–4018.

Pedersen, B. S.; Nimb, S.; Asmussen, J.; Sørensen, N. H.; Trap-Jensen, L.; and Lorentzen, H. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation* 43.

Pilehvar, M. T.; and Navigli, R. 2014. A Large-Scale Pseudoword-Based Evaluation Framework for State-of-the-Art Word Sense Disambiguation. *Comput. Linguistics* 40(4).

Pociello, E.; Gurrutxaga, A.; Agirre, E.; Aldezabal, I.; and Rigau, G. 2008. WNTERM: Enriching the MCR with a Terminological Dictionary. In *Proc. of LREC*.

Ponti, E. M.; Glavaš, G.; Majewska, O.; Liu, Q.; Vulić, I.; and Korhonen, A. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proc. of EMNLP*.

Postma, M.; van Miltenburg, E.; Segers, R.; Schoen, A.; and Vossen, P. 2016. Open Dutch WordNet. In *Proc. of GWC*.

Pradhan, S.; Loper, E.; Dligach, D.; and Palmer, M. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proc. of SemEval*.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proc. of ACL*.

Raffaelli, I.; Tadić, M.; Bekavac, B.; and Agić, Ž. 2008. Building croatian wordnet. In *Proc. of GWC*.

Raganato, A.; Camacho-Collados, J.; and Navigli, R. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proc. of EACL*.

Raganato, A.; Delli Bovi, C.; and Navigli, R. 2017. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proc. of EMNLP*.

Raganato, A.; Pasini, T.; Camacho-Collados, J.; and Pilehvar, M. T. 2020. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In *Proc. of EMNLP*.

Raganato, A.; Scherrer, Y.; and Tiedemann, J. 2019. The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proc. of WMT*, 470–480.

Scarlini, B.; Pasini, T.; and Navigli, R. 2019. Just "OneSeC" for Producing Multilingual Sense-Annotated Data. In *Proc. of ACL*.

Scarlini, B.; Pasini, T.; and Navigli, R. 2020. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In *Proc. of EMNLP*.

Scozzafava, F.; Maru, M.; Brignone, F.; Torrisi, G.; and Navigli, R. 2020. Personalized PageRank with Syntagmatic Information for Multilingual Word Sense Disambiguation. *Proc. of ACL Demos* .

Shimura, K.; Li, J.; and Fukumoto, F. 2019. Text Categorization by Learning Predominant Sense of Words as Auxiliary Task. In *Proc. of ACL*.

Simov, K.; and Osenova, P. 2010. Constructing of an Ontology-based Lexicon for Bulgarian. In *Proc. of LREC*.

Snyder, B.; and Palmer, M. 2004. The English all-words task. In *Proc. of Senseval*.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proc. of LREC*.

Tiedemann, J.; and Thottingal, S. 2020. OPUS-MT — Building open translation services for the World. In *Proc. of EAMT*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in NeurIPS*.

Vial, L.; Lecouteux, B.; and Schwab, D. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proc. of GWC*.

Vider, K.; and Orav, H. 2002. Estonian WordNet and Lexicography. In *Proc. of 11th Int. Symposium on Lexicography*.

Vossen, P. 1998. A multilingual database with lexical semantic networks. *Dordrecht,* .

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proc. of EMNLP*.

Yoon, A.-S.; Hwang, S.-H.; Lee, E.-R.; and Kwon, H.-C. 2009. Construction of Korean WordNet. *Journal of KIISE: Software and Applications* 36(1).