
Tommaso Pasini♠, Alessandro Raganato♦♥, Roberto Navigli♣
♠Sapienza NLP Group, Computer Science Department, Sapienza University of Rome
♥Department of Digital Humanities, University of Helsinki, Finland
♣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/.

1 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 (Liu, Lu, and Neubig 2018; Pu et al. 2018; 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 origi-
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, andNavigli (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.

2 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 andSalakhutdinov 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 andPonzetto 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, andSchwab 2019; Huang et al. 2019; Bevilacqua andNavigli 2020;Scarlini, Pasini, andNavigli 2020;Blevins andZettlemoyer 2020; Conia andNavigli 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, andNavigli 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 producesense distributions (Pasini, Scozzafava, andScarlini 2020) or sense-annotated data in languages other than English (Scozzafava et al. 2015; Delli Bovi et al. 2017; Scarlini, Pasini, andNavigli 2019; Pasini andNavigli 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 andSemEval competitions have focused mainly on English: Senseval-2 (Edmonds and Cotton 2001), Senseval-3 (Snyder andPalmer 2004), SemEval-07 Task 17 (Pradhan et al. 2007), SemEval-07 Task 7 (Navigli, Litkowski, andHargraves 2007), SemEval-10 Task 17 (Agirre et al. 2010), SemEval-13 Task 12 (Navigli, Jurgens, andVannella 2013) andSemEval-15 Task 13 (Moró andNavigli 2015), with only a few of them providing data for other languages too, i.e., SemEval-10 Task 17, SemEval-13 Task 12 andSemEval-15 Task 13. While the WSD framework proposed by Raganato, Camacho-Collados, andNavigli (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 (Liang et al. 2020; Hu et al. 2020; Lewis et al. 2020;Ponti et al. 2020; Raganato et al. 2020; Martelli et al. 2021, XGLUE, 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 or XGLUE 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, andNavigli (2017), by featuring coarse-grained datasets, and a larger training set.

3 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.

3.1 Sense Inventory

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

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, 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 lexicalized 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 $\mathcal{L}$ 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.

3.2 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 (Benitez 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 slid”\footnote{The stars of the Bornholm fishing industry are trout, salmon and herring.}, 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\footnote{The stars of the Bornholm fishing industry are trout, salmon and herring.} (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. German, Italian, 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, andNavigli (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

\url{https://babelnet.org}
senses from version 2.1; SemEval-13 and SemEval-15, instead, 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-Fine) including Senseval-2, Senseval-3, SemEval-10, SemEval-13 and SemEval-15, and a coarse-grained one (English-Coarse), i.e., SemEval-07 Task 17. As for development, we follow prior work (Raganato, Delli Bovi, and Navigli 2017; Kumar et al. 2019; Blevins and Zettlemoyer 2020) and retain SemEval-07 as development set (English-Dev). As a result, each language has a test and a development set in the same 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 language. As regards the test and development sets, most languages contain more than 1,000 and 200 gold annotations, respectively, 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.

**3.3 Training Data**

We now describe the sense-annotated training corpora that we create automatically and include in XL-WSD.

**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. 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 by leveraging the machine translation models made available by Tiedemann and Thottingal (2020, Opus-MT). 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, 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 \( \sigma_{EN}^w = w_1, \ldots, w_n \), its translation \( \sigma_T^w = w_T^1, \ldots, w_T^n \) and the synset annotation \( s \) for the word \( w_i \) in \( \sigma_{EN}^w \), we propagate the annotation to the word \( w_T^i \) in \( \sigma_T^w \) that appears as a synonym in \( s \). In the case that multiple annotations are associated with the same word \( w_T^j \), 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.

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.

**3.4 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 and development sets, most languages contain more than 1,000 and 200 gold annotations, respectively, 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.

---

---

We train all neural models for XLMR-Base, XLMR-Large (Conneau et al. 2020), BERT-Large, M-BERT10 (Devlin et al. 2019) and the language-specific versions of BERT (L-BERT) for each language11 available through the Huggingface library (Wolf et al. 2020). We train all neural models for 50 epochs with Adam optimizer12 and use the set of weights with the lowest loss on the development set for testing.


data types. 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 MULAN13 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
divided by the total number of word types. The word-type polysemy shows the ambiguity level of the words in each dataset, which 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 and development 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 different languages, 3M silver annotations from the T-SC+WNG datasets and more than 100K annotations for English.

<table>
<thead>
<tr>
<th>Language</th>
<th>Word Types</th>
<th>Polysynous Words</th>
<th>Word-Type Polysemy</th>
<th>Instances</th>
<th>Unique Synsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Fine</td>
<td>106906</td>
<td>8822</td>
<td>328</td>
<td>1458</td>
<td>3689</td>
</tr>
<tr>
<td>English-Coarse</td>
<td>-</td>
<td>-</td>
<td>750</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the training, test and development sets comprised in XL-WSD. The Train column refers to SC+WNG for English and to T-SC+WNG for all the other languages.

**4 Experimental Setup**

In this Section, we describe the baseline models, the evaluation measures and the data used in our experiments.

Architecture details. We follow Bevilacqua andNavigli (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-Base, XLMR-Large (Conneau et al. 2020), BERT-Large, M-BERT10 (Devlin et al. 2019) and the language-specific versions of BERT (L-BERT) for each language11 available through the Huggingface library (Wolf et al. 2020). We train all neural models for 50 epochs with Adam optimizer12 and use the set of weights with the lowest loss on the

---

10 The base multilingual-cased version of BERT.
11 A detailed list of the language-specific models is available at https://sapienzanlp.github.io/xl-wsd/.
12 Gradient clipping = 1.0; learning rate = 2 · 10−5; patient = 3.
Table 3: F1 scores of supervised and knowledge-based approaches as well as language-specific BERT models (L-BERT) and the Most Common Sense (MCS) baseline on the test splits. As for the Ø-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+WNGLanguage-specific datasets. The Micro AVG row shows the micro F1 across all languages but English.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ø-Shot (SC+WNGL)</th>
<th>Language-Specific (MuLAN)</th>
<th>Language-Specific (T-SC+WNGL)</th>
<th>Knowledge-Based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>XLMR-Large</td>
<td>XLMR-Base</td>
<td>XLMR-Large</td>
<td>XLMR-Large</td>
</tr>
<tr>
<td>English-Fine</td>
<td>76.28</td>
<td>74.50</td>
<td>72.40</td>
<td>76.28</td>
</tr>
<tr>
<td>English-Coarse</td>
<td>91.30</td>
<td>91.02</td>
<td>89.70</td>
<td>91.30</td>
</tr>
</tbody>
</table>

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 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 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 52.55 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

Micro AVG    **65.66** 64.82 62.84  -   -   -   57.68 52.85 49.31

Table 5 Results

5.1 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+WNGLanguage-specific training data, instead, both BERT-Large and XLMR-Large 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.

5.2 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+WNGLanguage-specific training data, ii) on the automatically-translated language-specific training data, i.e., T-SC+WNGLanguage-specific data, 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 (Scozafava et al. 2020), which relies on the Personalized PageRank algorithm and leverages the collocational relations in SyntagNet (Muru 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-Large achieves the best results across the board, showing a big gap when comparing its performance with 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-Large columns) or language-specific models trained on language-specific data (L-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 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
we analyse the frequency of the test set instances' labels which are based on BERT-Base. Interestingly, our newly XLMR-Large trained on English by a margin of 8 points on average on all non-English languages. Indeed, despite the fact that SyntagRank manages to outperform several language-specific models trained on TSC+WNG, it performs 5 and 7 points lower, on average, than M-BERT and XLMR-Base, respectively, and falls behind XLMR-Large trained on English by a margin of 8 points on average. These results corroborate previous English-focused artificially large-scale findings on the robustness of supervised WSD approaches (Pilehvar andNavigli 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-Large has roughly 200M more parameters than most of the language-specific models, which are based on BERT-Base. Interestingly, our newly introduced training data, i.e., TSC+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-Large 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. To investigate these results, we analyse the frequency of the test set instances’ labels that also appear within the training sets. A large portion of the test instances are annotated with synsets occurring only a few times in the training data, hence 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, and this requires more investigation.

Table 4: F1 scores breakdown of XLMR-Large model on the zero-shot setting by POS tags.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ALL</th>
<th>NOUN</th>
<th>VERBS</th>
<th>ADJ</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Fine</td>
<td>76.28</td>
<td>77.92</td>
<td>65.74</td>
<td>81.47</td>
<td>86.71</td>
</tr>
<tr>
<td>English-Coarse</td>
<td>91.30</td>
<td>92.72</td>
<td>88.64</td>
<td>89.55</td>
<td>91.75</td>
</tr>
</tbody>
</table>

| 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 |

In Table 4 we provide further insights by showing the results breakdown on each POS tag of the best performing model, i.e., XLMR-Large. 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-Large 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.

6 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, andNavigli 2014; Vannella et al. 2014), 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, and FoTran No. 771113 and the ELEXIS project No. 731015 under the European Union’s Horizon 2020 research and innovation programme. Authors also thank the CSC - IT Center for Science (Finland) for the computational resources.

References


Bivelacqua, M.; andNavigli, 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.


