MuLaN: Multilingual Label propagatioN for Word Sense Disambiguation

Edoardo Barba, **Luigi Procopio**, **Niccolò Campolungo**, **Tommaso Pasini** and **Roberto Navigli**
Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
{barba,procopio,campolungo,pasini,navigli}@di.uniroma1.it

Abstract

The knowledge acquisition bottleneck strongly affects the creation of multilingual sense-annotated data, hence limiting the power of supervised systems when applied to multilingual Word Sense Disambiguation. In this paper, we propose a semi-supervised approach based upon a novel label propagation scheme, which, by jointly leveraging contextualized word embeddings and the multilingual information enclosed in a knowledge base, projects sense labels from a high-resource language, i.e., English, to lower-resourced ones. Backed by several experiments, we provide empirical evidence that our automatically created datasets are of a higher quality than those generated by other competitors and lead a supervised model to achieve state-of-the-art performances in all multilingual Word Sense Disambiguation tasks. We make our datasets available for research purposes at https://github.com/SapienzaNLP/mulan.

1 Introduction

Recent years have witnessed an increasing ubiquity of Deep Learning, with state-of-the-art performances being steadily pushed up across virtually every branch of Artificial Intelligence, and Natural Language Processing (NLP) has been no exception. The Deep Learning paradigm, however, presents a major limitation that often hinders its applicability: it requires daunting amounts of data. In NLP this constraint is particularly cumbersome, especially when taking into account multiple languages: indeed, each language has to be manually annotated independently.

This situation is, moreover, aggravated still further in tasks where the high level of annotation expertise required poses an additional burden. A case in point is Word Sense Disambiguation (WSD), i.e., the task of associating a word in a context with its sense chosen from a fixed inventory [Navigli, 2009]. The fact that each word has its own set of senses increases both the sparsity of the problem and, as a consequence, the amount of manually annotated data required to reach satisfying performances, making the overall annotation process extremely demanding. Therefore, it comes as no surprise that, even in high-resource languages (i.e., English), we are still far from having large corpora of manually labeled data available [Pasini, 2020].

In order to cope with this shortage, a number of mitigation strategies have been devised. Some of these leverage parallel corpora, either by projecting annotations from English to lower-resourced languages [Lefever *et al.*, 2011], or by using them to ease sentence disambiguation [Delli Bovi *et al.*, 2017; Camacho-Collados *et al.*, 2016]. Others drop the parallel corpora requirement by relying, either upon already annotated data and label propagation techniques [Yuan *et al.*, 2016], or solely upon knowledge bases [Pasini and Navigli, 2020]. On the other hand, carefully chosen heuristics have also been shown to be capable of yielding high-quality sense-annotated data in multiple languages [Scarlini *et al.*, 2019].

An interesting alternative to silver datasets such as these comes from the so-called *zero-shot* framework. Thanks to underlying cross-lingual representations, classifiers can be trained on a language we do have labeled data for, and then tested on another language for which annotations are scarce [Bevilacqua and Navigli, 2020]. This framework has been drawing a considerable amount of interest thanks to the latest unsupervised deep multilingual models, such as Multilingual BERT [Devlin *et al.*, 2019] and XLM [Conneau and Lample, 2019; Conneau *et al.*, 2019], which in several tasks have attained performances almost on a par with their classic fully-supervised counterparts.

Even though these models are being studied extensively by the research community, to the best of our knowledge, no attempt has been made to use them to transfer sense annotations across languages. This paper frames itself in this very landscape and introduces a Multilingual Label propagatioN technique (MuLaN) tailored to WSD and capable of automatically producing sense-tagged training datasets in multiple languages. Our contributions are therefore as follows:

1. We present an alignment scheme that links together words of different corpora if they are used with the same meaning, regardless of their respective languages;
2. Leveraging such alignments, we project sense annotations and produce sense-annotated datasets;
3. We show that MuLaN achieves state-of-the-art performances on multilingual WSD. We also provide insights, identifying the possible reasons behind its success.

2 Related Work

The systems that tackle Word Sense Disambiguation are usually divided into two broad categories: knowledge-based and supervised. On the one hand, knowledge-based systems [Moro *et al.*, 2014; Agirre *et al.*, 2014; Scarlini *et al.*, 2020] were presented as a possible solution towards the paucity of sense-tagged corpora. Indeed, they infer senses by leveraging lexical-semantic resources such as dictionaries and graph-based knowledge bases [Miller *et al.*, 1990; Navigli and Ponzetto, 2012], or proximity-based algorithms such as the K-Nearest Neighbors. On the other hand, supervised systems rely on semantic annotations, either to learn sense embeddings [Loureiro and Jorge, 2019; Scarlini *et al.*, 2020], or to train neural architectures [Raganato *et al.*, 2017; Vial *et al.*, 2019; Kumar *et al.*, 2019; Huang *et al.*, 2019; Bevilacqua and Navigli, 2020] to model the probability of a sense given a word in its context.

While knowledge-based systems gracefully scale to rare senses and low-resource languages, their supervised alternatives have been consistently outperforming them whenever enough training data is available [Pilehvar and Navigli, 2014]. However, producing such data is utterly demanding as each and every word has its own set of labels, i.e., senses, hence making it time-consuming to create large amounts of examples for so many different concepts. Indeed, SemCor [Miller *et al.*, 1993], i.e., the largest manually-annotated corpus available, does not cover even half of the WordNet senses [Miller *et al.*, 1990] and is limited to English only. Other languages find themselves in an even worse spot as no gold standard training datasets are available.

To cope with this situation, several methods have been proposed over the years to automatically produce sense-annotated data in English, as well as in other languages [Pasini, 2020]. Taghipour and Ng [2015] and Yuan *et al.* [2016] focused on English only and, by exploiting parallel data and SemCor, respectively, produced new annotations for English sentences. However, while being useful for providing annotated examples for rare words and senses, these approaches did not address the paucity of data in other languages. To this end, parallel corpora have also been exploited to ease the disambiguation of words in parallel sentences [Delli Bovi *et al.*, 2017; Camacho-Collados *et al.*, 2016]. More recently, Pasini and Navigli [2017] and Scarlini *et al.* [2019] made progress towards relieving the burden of parallel corpora, producing large-scale sense-annotated data by leveraging knowledge bases and the structure of Wikipedia. Nevertheless, both these approaches focused on nominal concepts only, hindering as a consequence their applicability to all-words Word Sense Disambiguation tasks. None of the aforementioned approaches, however, leveraged the representational power of deep multilingual neural models, even though these latter were applied effectively to retrieve similar sentences across languages [Artetxe and Schwenk, 2019].

MuLaN stands out from current work as it is the first, to the best of our knowledge, to take advantage of contextualized word embeddings in order to transfer sense annotations across languages. Moreover, it does not require parallel corpora, nor does it have any restriction on either part-of-speech tags or the target corpus upon which the annotations are projected. Finally, its reliance on a knowledge base is limited solely to the multilingual lexicalizations contained therein.

3 MuLaN

In this Section we describe our proposed approach for automatically producing multilingual sense-annotated datasets along with the resources required.

Preliminaries. Our approach relies on a multilingual inventory $\mathbb{D}$ of synsets, i.e., sets of synonyms[1] in different languages. For example, $\mathbb{D}$ may contain the synset corresponding to the *fountain* meaning of *spring*, which has lexicalizations in different languages, including: $\textit{Quelle}_{DE}$, $\textit{spring}_{EN}$, $\textit{fountain}_{EN}$, $\textit{manantial}_{ES}$, $\textit{brollador}_{ES}$, $\textit{source}_{FR}$, $\textit{fonte}_{IT}$ and $\textit{sorgente}_{IT}$. We build this inventory by leveraging BabelNet[2] [Navigli and Ponzetto, 2012], a large multilingual semantic network whose nodes are concepts containing lexicalizations coming from various heterogeneous resources, including, inter alia, WordNet and Wikipedia. Thus, we define $\mathbb{D}$ as the set of synsets in BabelNet which contain at least one sense from WordNet and one or more senses in languages other than English. Then, we define $\mathbb{D}[s, l]$ as the set of senses in language l contained in the synset s, e.g., with s being the *fountain* meaning of *spring*, $\mathbb{D}[s, \text{EN}] = \{\textit{spring}_{\text{EN}}, \textit{fountain}_{\text{EN}}\}$, while $\mathbb{D}[s, \text{ES}] = \{\textit{manantial}_{\text{ES}}, \textit{brollador}_{\text{ES}}\}$.

For ease of reading, we also define a labeled corpus Γ as the set of instances $\gamma_i = (w, \sigma, s)$, i.e., a text span w in the context σ that has been manually-tagged with the synset s; and an unlabeled corpus Θ as the set of text spans $\theta_j = (w', \sigma')$, i.e., the context σ' containing the text span w' with this latter appearing as lexicalization of at least one of the meanings in $\mathbb{D}$.

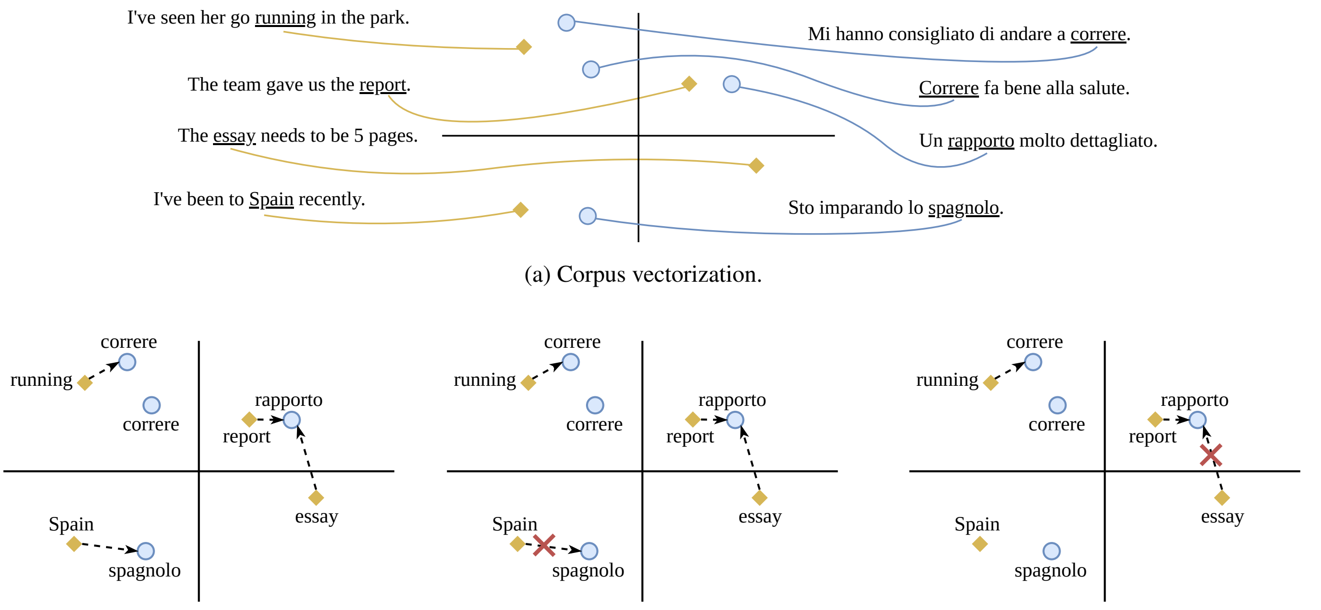
Propagating the Labels. We propose a Multilingual Label PropagatioN (MuLaN) approach which, by leveraging the semantic properties encoded in contextualized word embeddings [Reif *et al.*, 2019] and the unified inventory of concepts $\mathbb{D}$, aims at building a sense-annotated dataset in a given target language. To do this, MuLaN takes as input a labeled dataset Γ in the source language l_1 and an unlabeled corpus Θ[3] in the target language l_2 and applies the following steps:

- **Vectorization**, which projects each instance $\gamma_i \in \Gamma$ and each text span $\theta_j \in \Theta$ into a shared latent space (Section 3.1);
- **Candidate Production**, which associates each instance $\gamma_i \in \Gamma$ with the closest text spans $\theta_j \in \Theta$ according to the cosine similarity of their projected vectors (Section 3.2);
- **Dataset Generation**, which finally, given a synset s, collects all the sentences in Θ containing a text span that

[1] In what follows, we use *lexicalization* and *sense* interchangeably when referring to a word with a specific meaning.

[2] https://babelnet.org

[3] Both corpora are preprocessed by applying a sentence splitter, a POS tagger, a lemmatizer and an additional module that takes care of removing duplicate sentences.


I've seen her go running in the park.
The team gave us the report.
The essay needs to be 5 pages.
I've been to Spain recently.
Mi hanno consigliato di andare a correre.
Correre fa bene alla salute.
Un rapporto molto dettagliato.
Sto imparando lo spagnolo.
correre
running
correre
rapporto
report
essay
Spain
spagnolo

(a) Corpus vectorization.

(b) K-NN retrieval.

(c) Knowledge-base filtering.

(d) Backward test.

Figure 1: The MuLaN process: (a) gold annotations γ (◆) are projected in the same space as raw candidates θ (○); (b) k_1 nearest θs of each γ are retrieved (in the example, $k_1 = 1$); (c) invalid instances are discarded with support coming from the knowledge base (e.g., $(\text{Spain}, \text{spagnolo})$); (d) (γ, θ) pairs failing the backward test are removed (e.g., $(\text{essay}, \text{rapporto})$).

was previously associated with an instance tagged with s, and retains only the top-ranked ones according to the marginalized cosine similarity measure (Section 3.3).

3.1 Vectorization

This first step aims at creating comparable representations for the instances $\gamma_i \in \Gamma$ and the text spans $\theta_j \in \Theta$. To this end, we used a multilingual contextualized word embedding model, i.e., Multilingual BERT [Devlin *et al.*, 2019] (henceforth m-BERT), which has been pretrained on the concatenation of Wikipedia in 104 languages[4]. Formally, we compute the representation v_w^σ of a word w in the context σ as follows: $v_w^\sigma = \text{m-BERT}(\sigma, w)$ where w uniquely identifies the corresponding word in σ and $\text{m-BERT}(\sigma, w)$ returns the hidden vector of the last layer of m-BERT corresponding to the word w. In the case that w is split into multiple sub-words, we take their average and, if w is a multi-word, we first average the sub-words of each word within the compound and then take their average as representation for w.

As illustrated in Figure 1a, given the instances in Γ and the text spans in Θ, we project them into the same shared space by leveraging the contexts they appear in.

3.2 Candidate Production

The second step aims at associating each $\gamma_i \in \Gamma$ with a list of text spans $\theta_j \in \Theta$ that might express the same meaning. To this end, we leverage cosine similarity between a given γ_i and any $\theta_j \in \Theta$ as an indicator of whether or not θ_j is used with the same synset of γ_i. More formally, given an instance $\gamma_i \in \Gamma$, we compute its k_1[5] nearest neighbors $\mathcal{N}_{k_1}(\gamma_i) = \theta_1^{\gamma_i}, \ldots, \theta_{k_1}^{\gamma_i}$ among all the $\theta_j \in \Theta$ according to the cosine similarity[6]. As shown in Figure 1b, we select the text span *correre* from the sentence "Mi hanno consigliato di andare a correre."[7] as the closest candidate for the instance *running* from the sentence "I've seen her go running in the park.".

However, since no constraint is imposed on the neighbors of an instance, we might end up having a non-negligible amount of noise. For example, consider the instance $\gamma = (\textit{Spain}, \text{"I've been to Spain recently."}, \{\text{Spain, Kingdom of Spain, Espana}\})$ in Figure 1b: its closest text span in Θ is (*spagnolo*, "Sto imparando lo spagnolo."[8]), however, *spagnolo* and *Spain* refer to two different meanings, i.e., the Spanish language and the country, respectively. We cope with this issue by considering as valid text span candidates for a given instance $\gamma_i = (w, \sigma, s)$ only those text spans $\theta_j = (w', \sigma')$ such that $w' \in \mathbb{D}[s, l_2]$, i.e., w' is a possible lexicalization in language l_2 for synset s. Hence, in our example, we can discard *spagnolo* from the list of candidates for γ. At the end of this step, each instance $\gamma_i = (w, \sigma, s) \in \Gamma$ is associated with a filtered list $\mathcal{N}^{kb}(\gamma_i) = \{\theta_1^{\gamma_i}, \ldots, \theta_m^{\gamma_i}\}$ with $m \leq k_1$ of candidates containing text spans that are likely to express the

[4] While other multilingual models could have been used, our aim is to show that multilingual contextualized representations can be utilized to transfer semantic labels from one language to another rather than to provide an extensive evaluation of pretrained models for contextualized embeddings.

[5] We set $k_1 = 1000$.

[6] We used FAISS [Johnson *et al.*, 2019] in order to cope with the large number of comparisons to perform.

[7] They encouraged me to go running.

[8] I am learning Spanish.

same meaning s of γ_i.

3.3 Dataset Generation

The third step aims at producing the final sense-annotated dataset featuring examples in the target language l_2. To this end, we further refine the list of candidates $\mathcal{N}^{kb}(\gamma_i) = \{\theta_1^{\gamma_i}, \dots, \theta_m^{\gamma_i}\}$ by applying a backward compatibility test. That is, we retain a candidate $\theta_j^{\gamma_i}$ only if $\gamma_i \in^9 \mathcal{N}_{k_1}(\theta_j^{\gamma_i})$, i.e., the instance γ_i is among the k_1 nearest neighbors of $\theta_j^{\gamma_i}$. We therefore define the new list of candidates $\mathcal{N}^{bw}(\gamma_i)$ as the list of neighbors $\theta_j^{\gamma_i}$ of γ_i that have passed the backward compatibility test. Intuitively, consider the example in Figure 1d. As can be seen, while *essay* from the sentence "The essay needs to be 5 pages." has *rapporto*[10] from the sentence "Un rapporto molto dettagliato."[11] as closest neighbor, *rapporto* is closer to *report* than to *essay*. Thus, since in the example $k_1 = 1$, we remove *rapporto* from the candidate list of *essay*.

Once we have further refined the candidate lists, we proceed to create the final dataset in language l_2. For each synset s that appears in Γ, we create $\mathcal{C}(s) = \{\theta_1^{\gamma_1}, \dots, \theta_{|\mathcal{C}(s)|}^{\gamma_r}\}$, i.e., the union of the text-span candidates $\theta_j^{\gamma_i}$ for each instance $\gamma_i = (w, \sigma, s) \in \Gamma$ that is tagged with s. In order to select the most suitable sentences for the synset s from $\mathcal{C}(s)$, we score each element $\theta_j^{\gamma_i}$ by means of the marginalized cosine similarity (*mcos*) [Artetxe and Schwenk, 2019] calculated according to its corresponding instance γ_i. Formally, given γ_i and $\theta_j^{\gamma_i}$, we compute *mcos* as follows:

$$mcos(\gamma_i, \theta_j^{\gamma_i}) = \frac{cos(\gamma_i, \theta_j^{\gamma_i})}{\mathcal{M}(\gamma_i) + \mathcal{M}(\theta_j^{\gamma_i})}$$

$$\mathcal{M}(x) = \sum_{z \in \mathcal{N}_{k_2}(x)} \frac{cos(x, z)}{2k_2}$$

where $\mathcal{N}_{k_2}(x)$ is the list of the k_2[12] closest candidates for x without applying any filtering. The intuition behind this measure revolves around *contextualizing* the cosine similarity by taking into account how close the nearest neighbors of both arguments are.

Finally, once the elements in $\mathcal{C}(s)$ have been ranked, we compute t_s, i.e., the number of occurrences of s in Γ, select the top t_s text spans $\theta_j^{\gamma_i} = (w', \sigma')$ in $\mathcal{C}(s)$, collect their corresponding sentences $\sigma'_1, \dots, \sigma'_{t_s}$ and tag the words $w'_1, \dots, w'_{t_s}$ therein with s. Should a candidate be associated with two instances γ_i and γ_j, we retain the association with the instance that maximizes the marginalized cosine similarity score. As the outcome of this last step, we have a new dataset with examples in language l_2, each containing at least one word tagged with a synset from the inventory $\mathbb{D}$.

4 Experimental Setup

We assess the quality of our corpora in the Word Sense Disambiguation task by comparing the performance of a transformer-based architecture [Vaswani *et al.*, 2017] trained on MuLaN with several competitors.

4.1 Reference WSD Model

In order to carry out the evaluation, we employ a transformer-based classifier as our WSD reference model. Specifically, we use m-BERT to encode the input word pieces into latent vectors[13] which, in their turn, are fed into a fully-connected layer with a softmax activation function. When a text span is split into multiple word pieces, we follow Devlin *et al.* [2019] and use the first word-piece hidden vector as the representation for the whole text span. During training, rather than fine-tuning all the model parameters, we keep the BERT weights fixed and let the gradient flow through the last layer only.

The model is trained for 50 epochs, with early stopping technique set with a patience parameter of 3; we used the Adam optimizer with learning rate fixed at $2 \cdot 10^{-5}$ and a cross-entropy loss criterion. As validation set, due to the lack of any publicly available sets in the languages being considered, we reserved a small random percentage from the training set for this purpose only. We note that we also applied this same strategy when training the classifier on the data produced by our competitors.

4.2 Test Bed

We report the performances on multilingual datasets for all-words WSD which were made available in the context of the past SemEval competitions, namely, SemEval-13 [Navigli *et al.*, 2013], containing nominal instances in French, German, Italian and Spanish, and SemEval-15 [Moro and Navigli, 2015], comprising Italian and Spanish datasets. We perform the evaluation using the same SemEval-13 and SemEval-15 versions used by our competitors in their respective original papers; we will refer to these versions as SemEval-13* and SemEval-15*. Furthermore, we also use the revised version of the evaluation datasets[14] (WordNet split), which is updated to be consistent with the 4.0.1 release of BabelNet. As a result, we can test on a larger number of instances than was previously possible.

Following the literature, we show Precision, Recall and F1 scores, i.e., the harmonic mean of Precision and Recall. Furthermore, we report the statistical significance as computed by the McNemar's test [McNemar, 1947] over the Precision measure with $\alpha = 0.01$ between the results attained by the reference WSD model trained on each of our datasets and the best possible competitor in the same setting.

4.3 Comparison Systems

We compare MuLaN with the following alternative corpora (marked with †) and models (marked with $^\diamond$):

- **Most Common Sense (MCS)**$^\diamond$: a baseline in Word Sense Disambiguation where each pair (lemma, part of speech) is tagged with its most common sense (note that, compared to the traditional Most Frequent Sense baseline, we do not have sense frequencies for non-English

[9] We extend the $\in$ operator to work with lists as well.

[10] Report, technical report.

[11] A very detailed report.

[12] By following the original paper, we use $k_2 = 4$.

[13] We use the concatenation of the last 4 layers' outputs to represent each word piece.

[14] https://github.com/SapienzaNLP/mwsd-datasets.

	SemCor+WNG	MuLaN				
	EN	IT	ES	FR	DE	4L
# instances	723k	415k	452k	310k	245k	1424k
# senses	91k	44k	57k	29k	22k	141k
# synsets	70k	33k	43k	25k	19k	50k

Table 1: Number of annotated instances, unique senses and unique synsets contained in the concatenation of SemCor and WNG and in the MuLaN datasets.

languages, therefore we follow the BabelNet ranking, which is based on the reliability and frequency within the underlying resources);

- **OneSeC**† [Scarlini *et al.*, 2019]: a fully automatic knowledge-based method for creating sense-annotated corpora for nominal instances only (OneSeC). We also consider for comparison OneSeC$_{4L}$, i.e., the concatenation of its datasets created for Italian, Spanish, French and German. In what follows, we report the results attained by the reference WSD model when trained on each of the aforementioned datasets;
- **∅-Shot**$^\diamond$: we test our reference WSD model in the ∅-Shot setting, i.e., when trained on English data and tested on other languages. We compare against the WSD model trained on the following two English datasets: i) SemCor [Miller *et al.*, 1993], i.e., the *de facto* standard training set for WSD, which features roughly $40K$ sentences and more than $200K$ annotations, and ii) SemCor+WNG, i.e., the concatenation of SemCor and the Princeton WordNet Gloss Corpus [Langone *et al.*, 2004], which contains glosses and examples for WordNet synsets that were disambiguated both manually and automatically;
- **SensEmBERT**$^\diamond$ [Scarlini *et al.*, 2020]: a knowledge-based approach for producing BERT-based embeddings of senses by exploiting the lexical-semantic information in BabelNet and Wikipedia. It currently attains state-of-the-art results in multilingual WSD for nouns;
- **UKB+SyntagNet**$^\diamond$ [Maru *et al.*, 2019]: a pre-BERT knowledge-based approach which applies Personalized PageRank to the WordNet graph enriched with manually-annotated free association and collocation edges between WordNet senses (SyntagNet[15]).

4.4 Source and Target Corpora

As labeled corpus Γ, we use the concatenation of SemCor and WNG since it is the largest available corpus annotated with senses. As unlabeled corpus Θ, on the other hand, we use Wikipedia, since it covers most domains of human knowledge and is available in several languages, with the additional benefit that it maintains the same writing style across languages, i.e., a descriptive one. Thus, we use MuLaN to generate a sense-annotated dataset (see Section 3) for each language used in the test sets under consideration, i.e., Italian, French, Spanish and German, and we show their statistics in Table 1.

4.5 Test Configurations

Together with the results attained when training the reference model on MuLaN monolingual datasets, we also report its performance when trained on: i) our datasets restricted to the nominal instances only, i.e., MuLaN^N, ii) MuLaN datasets in all the four languages together, i.e., MuLaN$_{4L}$, and iii) when considering all the languages and focusing on the nominal instances only MuLaN$_{4L}^N$. The variants where we restrict to nouns only are needed in order to compare fairly with OneSeC and SensEmBERT. Indeed, we recall from Section 4.3 that OneSeC cannot provide annotations to anything but nouns, hence reducing the overall task complexity. On the other hand, as shown in Table 1, each monolingual corpus covers, on average, less than half of the synsets originally available in SemCor+WNG. This phenomenon is largely caused by, either language-specific shortcomings in BabelNet and m-BERT, or the simple fact that some senses are harder to project towards some languages rather than towards others. The $4L$ variants cope with this issue by concatenating all monolingual datasets together, providing a wider coverage of synsets, as shown in Table 1 (last column).

5 Results

5.1 Multilingual Word Sense Disambiguation

In this Section we compare the results attained by the WSD reference model (see Section 3) when trained on MuLaN datasets with those achieved by our competitors.

As a first result, we show the performance of the WSD reference model in the zero-shot setting. As one can see in Table 2, both ∅-shot-SemCor and ∅-shot-SemCor+WNG achieve competitive performance in comparison to the other supervised and knowledge-based approaches, i.e., UKB+SyntagNet, OneSeC and SensEmBERT. Specifically, ∅-shot-SemCor+WNG outperforms all our competitors in most datasets, with the additional advantage of not having to rely on language-specific sense-annotated data or on large multilingual knowledge bases. This result highlights, for the first time in the context of Word Sense Disambiguation, the multilingual capabilities of m-BERT and its ability to encode the semantics of words regardless of their language.

When considering MuLaN, instead, we note that it is the only approach surpassing ∅-shot-SemCor+WNG on most datasets, while beating the other alternatives in all but two benchmarks, i.e., the French and German test sets of SemEval-13*, where it is surpassed by SensEmBERT and OneSeC, respectively. However, both these latter approaches are inherently restricted to performing WSD on nominal instances only.

When moving our focus to the best setting of our approach, we note that MuLaN$_{4L}$ leads the WSD reference model to surpass the state of the art in all tasks but German, where it lags behind OneSeC by 3.4 F1 points. However, as we will show in more detail in Section 5.2, this difference is in the main due to a bias towards the most common sense that both OneSeC and the German SemEval-13* test set have, and not to a better quality of the training data. Furthermore, while MuLaN provides tagged instances for all the open-class part-of-speech tags, OneSeC covers only nouns. In this regard, we

[15] http://syntagnet.org

	SemEval-13*												SemEval-15*					
	IT			ES			FR			DE			IT			ES		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MCS	69.25	66.92	68.07	70.26	67.97	69.09	72.39	66.84	69.50	75.70	74.60	75.14	60.77	60.05	60.41	56.25	55.81	56.03
∅-shot-SemCor	74.03	72.64	73.33	75.77	75.17	75.47	70.70	70.35	70.52	74.05	73.55	73.80	69.13	68.72	68.93	64.82	64.17	64.49
∅-shot-SemCor+WNG	77.40	75.94	76.66	75.07	74.70	74.88	74.97	74.60	74.78	74.47	74.11	74.29	70.69	70.27	70.48	67.64	66.96	67.29
UKB+SyntagNet	74.20	74.20	74.20	73.40	73.40	73.40	72.70	72.70	72.70	66.90	66.90	66.90	65.00	65.00	65.00	61.20	61.20	61.20
SensEmBERT	69.60	69.60	69.60	74.60	74.60	74.60	78.00	78.00	78.00	78.00	78.00	78.00	-	-	-	-	-	-
OneSeC	73.26	71.88	72.57	74.67	74.08	74.37	74.47	74.40	74.59	81.02	80.47	80.75	-	-	-	-	-	-
OneSeC$_{4L}$	76.92	75.47	76.19	77.56	76.95	77.25	77.35	76.97	77.16	81.82	81.27	81.54	-	-	-	-	-	-
MuLaN	$\underline{79.13}$	77.64	78.38	77.36	76.75	77.06	78.15	77.76	77.95	79.31	78.77	79.04	71.41	70.98	71.19	66.85	66.18	66.51
MuLaN^N	$\underline{79.32}$	77.83	78.56	$\underline{77.67}$	77.15	77,40	78.36	77.81	78.08	79.54	79.00	79.26	-	-	-	-	-	-
MuLaN$_{4L}$	$\underline{79.71}$	78.20	78.95	$\underline{79.76}$	79.12	79.44	79.84	79.44	79.64	78.40	77.86	78.13	$\underline{72.12}$	71.70	**71.91**	68.88	68.19	**68.53**
MuLaN$_{4L}^N$	$\underline{82.40}$	80.84	**81.61**	$\underline{82.05}$	81.40	**81.72**	$\underline{81.13}$	80.73	**80.93**	82.97	82.40	**82.68**	-	-	-	-	-	-

Table 2: Comparison of MuLaN against its competitors on SemEval-13* and SemEval-15* multilingual WSD datasets. Underlined results are statistically significant with respect to their best performing competitor according to McNemar's test, $\alpha = 0.01$.

	SemEval-13				SemEval-15	
Model	IT	ES	FR	DE	IT	ES
MCS	44.20	37.10	53.20	70.20	44.60	39.60
∅-shot-SemCor	74.93	76.70	79.54	81.13	69.41	65.68
∅-shot-SemCor+WNG	76.70	77.10	79.64	81.64	70.54	68.67
UKB+SyntagNet	72.14	74.12	70.32	76.39	68.95	63.37
SensEmBERT	69.80	73.40	77.80	79.20	-	-
OneSeC	63.45	61.59	65.10	75.84	-	-
OneSeC$_{4L}$	60.26	64.13	64.87	76.30	-	-
MuLaN	77.45	77.70	80.12	82.09	70.31	68.73
MuLaN^N	77.15	77.45	78.00	80.10	-	-
MuLaN$_{4L}$	**77.85**	**81.11**	**81.64**	**82.34**	**71.80**	**69.42**
MuLaN$_{4L}^N$	77.65	80.95	80.95	82.09	-	-

Table 3: Comparison of MuLaN against its competitors on the new versions of SemEval-13 and SemEval-15 multilingual datasets.

recall from Section 4.5 that comparing MuLaN with OneSeC is unfair with regard to our approach, as providing annotations solely for nouns inevitably leads to a lower level of confusion in the training data and, consequently, in the supervised model as well. Therefore, in order to set a level playing field between the two systems, we computed the nouns-only version of MuLaN, i.e., MuLaN$_{4L}^N$, and tested it along with the other datasets. As shown in Table 2, MuLaN$_{4L}^N$ attains the best results across the board, surpassing its closest competitor, i.e., OneSeC$_{4L}$, by an overall significant margin.

When testing the systems on the latest available versions of SemEval-13 and SemEval-15 (Table 3), MuLaN remains the best performing model across the board. OneSeC has a large drop in performance, which is mainly due to limitations of the resource it draws on, i.e., NASARI. Indeed, OneSeC cannot provide annotated examples for several synsets that appear as gold answers within the datasets because they are not associated with any NASARI vector. For future comparisons, we strongly encourage the community to consider the results reported in Table 3 as they are computed on the latest and updated versions of the datasets.

5.2 Most Common Sense Analysis

In this Section we investigate the ability of MuLaN to also produce examples for Least Common Senses, i.e., for all but the most common sense for each word, and compare its performance with OneSeC. Indeed, senses are known to follow a Zipfian distribution [McCarthy *et al.*, 2007], hence making it easy to achieve competitive results in Word Sense Disambiguation by simply tagging each word with its Most Common Sense (MCS). Therefore, in order to analyze the extent to which MuLaN is biased towards the Most Common Sense, we create the datasets IT$_{LFS}$, ES$_{LFS}$, FR$_{LFS}$ and DE$_{LFS}$ for each dataset in SemEval-13* and SemEval-13* by retaining only those instances of the original test sets that are tagged with one of their Least Common Senses, and then use these datasets to carry out the evaluation. As shown in Table 4, MuLaN proves to be considerably less biased towards the MCS than OneSeC. Most importantly, Table 4 provides interesting insights on the German dataset of SemEval-13*, that is, the only setting where OneSeC outperforms MuLaN (Table 2). Indeed, as one can see, when focusing on Least Common Senses, OneSeC's performance drops to 14.4 points, i.e., the lowest score across the board. Therefore, by considering the performance reported in Table 2, we note that OneSeC is very effective in providing sense annotations for the most common senses in German, whereas it is considerably less accurate in modeling other senses. Thus, when also taking into account the unusually high results of the MCS baseline in German, we argue that the lower performances of MuLaN compared to OneSeC are most likely due to the strong bias towards the Most Common Sense of both OneSeC and the test set, rather than to the lower quality of our dataset.

	SemEval-13*				SemEval-15*	
Model	IT$_{LFS}$	ES$_{LFS}$	FR$_{LFS}$	DE$_{LFS}$	IT$_{LFS}$	ES$_{LFS}$
OneSeC	31.12	27.30	27.75	14.40	-	-
OneSeC$_{4L}$	38.20	33.84	37.14	16.00	-	-
MuLaN	57.14	51.51	45.74	27.34	**56.69**	48.94
MuLaN$_{4L}$	59.09	55.30	53.44	29.23	55.76	**51.58**
MuLaN$_{4L}^N$	**63.00**	**59.69**	**54.87**	**40.62**	-	-

Table 4: Results of the WSD reference model trained on both OneSeC and MuLaN and tested on the instances of each dataset that are tagged with one of their Least Common Senses.

This difference is even more marked when comparing OneSeC$_{4L}$ and MuLaN in a fair setting, i.e., when we restrict

	SemEval-13*				SemEval-15*	
Model	IT$_{US}$	ES$_{US}$	FR$_{US}$	DE$_{US}$	IT$_{US}$	ES$_{US}$
MuLaN	72.07	71.96	76.66	67.61	66.61	63.49
MuLaN$_{4L}$	74.64	74.49	78.47	68.81	68.12	65.80

Table 5: Precision on synsets in the test sets with a lexicalization that is not present in the training set.

our dataset to its nominal instances only (MuLaN$^{N}_{4L}$). In this scenario, while consistently achieving results that are on average 20 points higher than those of OneSeC$_{4L}$, MuLaN$^{N}_{4L}$ also shows an unmatched ability to provide examples for infrequent senses across different languages.

5.3 Unseen Senses

Finally, we move our focus to analyzing how different languages can help our WSD reference model to generalize over unseen senses. To this end, we create six new datasets, namely, IT$_{US}$, ES$_{US}$, FR$_{US}$ and DE$_{US}$ for SemEval-13*, and IT$_{US}$ and ES$_{US}$ for SemEval-15*. Each of these contains all the tagged instances (l, s) from the original datasets in SemEval-13* and SemEval-15*, where l is a lemma and s is a synset, such that l never appears tagged with s in the MuLaN training set of the corresponding language and s appears as the label for at least another lemma l' therein. We then leverage these datasets to compare the performance of our WSD reference model when provided with either monolingual or multilingual training data. As shown in Table 5, providing data in multiple languages always proves to be beneficial. Indeed, MuLaN$_{4L}$ leads the WSD reference model to attain on average 2 F1 points higher on all test sets than its counterpart trained on monolingual corpora, i.e., MuLaN. This is mainly due to the multilingual nature of the sense inventory of our reference model. Indeed, this model exploits synsets lexicalized in different languages in the output vocabulary, which enables it to effectively leverage the annotations available for a given meaning regardless of their language.

6 Conclusion

In this work, we presented MuLaN, a novel Multilingual Label propagatioN technique for creating sense-annotated datasets in multiple languages. Our approach enables the annotation effort to be focused on English-only, while at the same time automatically projecting the manually-produced sense labels to corpora in other languages. Our experiments show that MuLaN outperforms all its competitors by several points on the multilingual WSD tasks when jointly leveraging its automatically produced data in all languages. Furthermore, when considering instances tagged with their Least Common Senses only, MuLaN also shows an unmatched ability to provide high-quality examples for rare synsets – one that is out of reach for its alternatives.

At https://github.com/SapienzaNLP/mulan we release about $800K$ sentences with more than $1.4M$ sense-tagged instances in Italian, Spanish, French and German. As future work, we aim at extending the sense coverage of our approach to those meanings not included in the source corpus, and at generalizing MuLaN so as to enable it to transfer labels for other tasks as well.

Acknowledgments

This work has been supported by the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

This work has also been supported by the PerLIR project (Personal Linguistic resources in Information Retrieval) funded by the MIUR Progetti di ricerca di Rilevante Interesse Nazionale programme (PRIN 2017).

The authors wish to thank Babelscape (http://babelscape.com) for providing the computing facilities that made it possible for this work to be carried out.

References

[Agirre *et al.*, 2014] Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84, 2014.

[Artetxe and Schwenk, 2019] Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proc. of ACL*, pages 3197–3203, 2019.

[Bevilacqua and Navigli, 2020] Michele Bevilacqua and Roberto Navigli. 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*, 2020.

[Camacho-Collados *et al.*, 2016] José Camacho-Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. A Large-Scale Multilingual Disambiguation of Glosses. In *Proc. of LREC*, pages 1701–1708, 2016.

[Conneau and Lample, 2019] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances in NeurIPS*, pages 7059–7069. 2019.

[Conneau *et al.*, 2019] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *arXiv preprint arXiv:1911.02116*, 2019.

[Delli Bovi *et al.*, 2017] Claudio Delli Bovi, Jose Camacho-Collados, Alessandro Raganato, and Roberto Navigli. Eurosense: Automatic harvesting of multilingual sense annotations from parallel text. In *Proc. of ACL*, volume 2, pages 594–600, 2017.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186, June 2019.

[Huang *et al.*, 2019] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *Proc. of EMNLP-IJCNLP*, pages 3500–3505, 2019.

[Johnson *et al.*, 2019] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019.

[Kumar *et al.*, 2019] Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. Zero-shot Word Sense Disambiguation using sense definition embeddings. In *Proc. of ACL*, pages 5670–5681, July 2019.

[Langone *et al.*, 2004] Helen Langone, Benjamin R Haskell, and George A Miller. Annotating WordNet. In *Proc. of the Workshop Frontiers in Corpus Annotation*, 2004.

[Lefever *et al.*, 2011] Els Lefever, Véronique Hoste, and Martine De Cock. ParaSense or how to use parallel corpora for Word Sense Disambiguation. In *Proc. of ACL, 2011*, pages 317–322, 2011.

[Loureiro and Jorge, 2019] Daniel Loureiro and Alípio Jorge. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In *Proc. of ACL*, pages 5682–5691, 2019.

[Maru *et al.*, 2019] Marco Maru, Federico Scozzafava, Federico Martelli, and Roberto Navigli. SyntagNet: Challenging Supervised Word Sense Disambiguation with Lexical-Semantic Combinations. In *Proc. of EMNLP-IJCNLP*, pages 3525–3531, 2019.

[McCarthy *et al.*, 2007] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Unsupervised acquisition of predominant word senses. *Comput. Linguist.*, 33(4):553–590, 2007.

[McNemar, 1947] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[Miller *et al.*, 1990] George A. Miller, R.T. Beckwith, Christiane D. Fellbaum, D. Gross, and K. Miller. WordNet: an online lexical database. *Int. J. Lexicogr*, 3(4):235–244, 1990.

[Miller *et al.*, 1993] George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. A semantic concordance. In *Proc. of the Workshop on Human Language Technology*, pages 303–308, Plainsboro, N.J., 1993.

[Moro and Navigli, 2015] Andrea Moro and Roberto Navigli. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proc. of SemEval-2015*, pages 288–297, 2015.

[Moro *et al.*, 2014] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity Linking meets Word Sense Disambiguation: a unified approach. *TACL*, 2:231–244, 2014.

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

[Navigli *et al.*, 2013] Roberto Navigli, David Jurgens, and Daniele Vannella. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Proc. of SemEval-2013*, volume 2, pages 222–231, 2013.

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

[Pasini and Navigli, 2017] Tommaso Pasini and Roberto Navigli. Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data. In *Proc. of EMNLP*, pages 78–88, 2017.

[Pasini and Navigli, 2020] Tommaso Pasini and Roberto Navigli. Train-O-Matic: Supervised Word Sense Disambiguation with no (manual) effort. *Artificial Intelligence*, 279:103–215, 2020.

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

[Pilehvar and Navigli, 2014] Mohammad Taher Pilehvar and Roberto Navigli. A large-scale pseudoword-based evaluation framework for state-of-the-art Word Sense Disambiguation. *Comput. Linguist.*, 40(4):837–881, 2014.

[Raganato *et al.*, 2017] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proc. of EMNLP*, pages 1156–1167, 2017.

[Reif *et al.*, 2019] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and Measuring the Geometry of BERT. In *Advances in NeurIPS*, pages 8592–8600, 2019.

[Scarlini *et al.*, 2019] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. Just "OneSeC" for Producing Multilingual Sense-Annotated Data. In *Proc. of ACL*, pages 699–709, 2019.

[Scarlini *et al.*, 2020] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In *Proc. of AAAI*, 2020.

[Taghipour and Ng, 2015] Kaveh Taghipour and Hwee Tou Ng. One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. In *Proc. of CoNLL*, pages 338–344, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in NeurIPS*, pages 6000–6010, 2017.

[Vial *et al.*, 2019] Loïc Vial, Benjamin Lecouteux, and Didier Schwab. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Proc. of GWC*, 2019.

[Yuan *et al.*, 2016] Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. Semi-supervised Word Sense Disambiguation with neural models. In *Proc. of COLING*, pages 1374–1385, 2016.