MultiMirror: Neural Cross-lingual Word Alignment for Multilingual Word Sense Disambiguation

Luigi Procopio , **Edoardo Barba** , **Federico Martelli** and **Roberto Navigli**
Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome
{luigi.procopio, edoardo.barba, federico.martelli, roberto.navigli}@uniroma1.it

Abstract

Word Sense Disambiguation (WSD), i.e., the task of assigning senses to words in context, has seen a surge of interest with the advent of neural models and a considerable increase in performance up to 80% F1 in English. However, when considering other languages, the availability of training data is limited, which hampers scaling WSD to many languages. To address this issue, we put forward MULTIMIRROR, a sense projection approach for multilingual WSD based on a novel neural discriminative model for word alignment: given as input a pair of parallel sentences, our model – trained with a low number of instances – is capable of jointly aligning, at the same time, all source and target tokens with each other, surpassing its competitors across several language combinations. We demonstrate that projecting senses from English by leveraging the alignments produced by our model leads a simple mBERT-powered classifier to achieve a new state of the art on established WSD datasets in French, German, Italian, Spanish and Japanese. We release our software and all our datasets at https://github.com/SapienzaNLP/multimirror.

1 Introduction

In the last few years, Word Sense Disambiguation (WSD), i.e., the task of determining the meaning of a word in a given context, has witnessed a remarkable improvement in performance due to increasingly refined neural approaches [Bevilacqua *et al.*, 2021]. However, the potential of these approaches is severely limited by the paucity of high-quality training corpora and by the demanding process, both in terms of expertise and amount of work, required to expand them. This problem, generally referred to as the knowledge acquisition bottleneck [Gale *et al.*, 1992], has long affected several word- and sentence-level semantic fields of NLP, and is particularly relevant to WSD [Pasini, 2020], owing to the fine granularity of senses, which are often difficult to distinguish from one another, and their Zipfian distribution in a corpus.

Although several attempts have been made to address the issue, this limitation is still far from being overcome, and, particularly in languages other than English, different lines of research have been explored that drop the requirement for vast amounts of labeled data. Among these, knowledge-based approaches cope with this limitation by utilizing the linguistic information contained in largely-interconnected knowledge bases such as WordNet [Miller, 1998] and BabelNet [Navigli and Ponzetto, 2012], typically by means of graph-based strategies such as the Personalized PageRank [Agirre *et al.*, 2014], or densest-subgraph algorithms [Moro *et al.*, 2014]. More recently, a valid alternative to these approaches has been identified in the so-called zero-shot paradigm. Thanks to the cross-lingual representational power of recent unsupervised embeddings such as mBERT [Devlin *et al.*, 2019], this paradigm proposes training supervised models on existing manually-annotated resources, which are typically in English, and applying them directly to other languages. Finally, a different research direction that has shown promising potential is annotation projection [Yarowsky *et al.*, 2001]. This technique consists in propagating manually-curated labels to unlabeled data and has not only been successfully used in WSD [Scarlini *et al.*, 2019; Barba *et al.*, 2020], but also in several other NLP tasks, such as Semantic Role Labeling [Aminian *et al.*, 2019] and Semantic Parsing [Blloshmi *et al.*, 2020].

However, despite significant steps forward, current approaches still suffer from two main shortcomings: i) they often depend heavily on the availability of high-quality and wide-coverage external resources in order to validate candidate projections, thus posing a serious limitation, particularly when dealing with low-resource languages; ii) they strive to transfer annotations onto untagged text, either from a dataset of a different nature (e.g., from news to Wikipedia), or by assigning sense tags automatically with clustering or induction techniques. In contrast, we transfer sense tags across parallel sentences, arguing that this is the most natural setting for label propagation, and an affordable one for several language pairs, thanks to the recent advances in multilingual generative approaches [Tang *et al.*, 2020].

To address the aforementioned issues, we introduce MULTIMIRROR, a novel approach for cross-lingual label projection based on an innovative neural architecture for word alignment. Our contribution is threefold:

1. We propose a cross-lingual sense projection approach based on word alignment, which enables a simple mBERT-powered classifier to achieve a new state of the art on established WSD datasets and proves to be effec-

tive also when projecting labels onto a distant language, i.e., Japanese.

2. We put forward a novel neural architecture for word alignment which is capable of aligning all source and target tokens at the same time with just a few hundred training sentences, achieving a new state of the art and considerably reducing the processing speed.
3. Additionally, as a by-product of our experiments, we also release four new manually-annotated datasets for word alignment in the following language combinations: English-French, English-German, English-Italian, and English-Spanish.

2 Related Work

Several efforts have been made to cope with the scarcity of sense-tagged data via annotation projection. The earliest works focused on harvesting training instances by exploiting WordNet's lexical relations and particularly monosemous relatives of polysemous words [Leacock *et al.*, 1998; Agirre and Martinez, 2004]. Although these approaches proved their ability to produce new training instances, they failed however in scaling across languages and domains.

Subsequent research addressed these drawbacks by leveraging parallel corpora [Delli Bovi *et al.*, 2017], or multilingual knowledge bases such as BabelNet. For instance, Pasini and Navigli [2017] proposed a language-independent approach, which, given as input a multilingual knowledge base and an unlabeled corpus, is capable of generating sense-annotated data in a fully-automatic fashion. Following this line of research, Scarlini *et al.* [2019] automatically assigned senses to words in Wikipedia sentences building upon the *One Sense per Wikipedia Category* assumption, that is, all occurrences of a given word in Wikipedia pages associated with the same Wikipedia category share the same meaning. Recently, Barba *et al.* [2020] proposed MuLaN, a label projection approach using unsupervised multilingual embeddings to automatically generate sense-annotated data. Taking as input a labeled dataset and an unlabeled corpus, MuLaN projects all labeled and unlabeled instances into a shared vector space and propagates labels by adopting the k-nearest neighbors algorithm with an additional candidate validation step.

In other areas of NLP, cross-lingual label propagation has often been implemented via word alignment, i.e., the task of identifying translation correspondences between words in parallel sentences. For instance, Xi and Hwa [2005] automatically generated data for Chinese POS Tagging, Padó and Lapata [2009] exploited word alignment to project semantic roles and, more recently, Stengel-Eskin *et al.* [2019] proposed a neural discriminative architecture for word alignment and applied it to label propagation for Chinese NER. However, despite its success in a wide range of NLP tasks, to the best of our knowledge, no attempt has been made previously to leverage neural word alignment to project word senses across languages.

3 MultiMirror

In this Section, we describe MULTIMIRROR, our approach for cross-lingual sense projection. We first explain the novel word alignment model we put forward and compare it to the currently best-performing system (Section 3.1). Then, we detail how, using this model, we automatically generate sense-tagged corpora (Section 3.2).

3.1 Cross-lingual Word Alignment Model

We now introduce our discriminative word alignment model, whose structure is reported in Figure 1. It takes as input two parallel sentences $U = u_1, \dots, u_i, \dots, u_l$ and $V = v_1, \dots, v_j, \dots, v_k$, where u_i is the i-th token of U and v_j is the j-th token of V. In order to obtain continuous representations of each token, we employ a pretrained contextualized embedding, namely multilingual BERT [Devlin *et al.*, 2019, mBERT]. Following the reference paper, we concatenate the two sentences, separating them with a special token, i.e. [SEP], and surround the whole sequence with two additional tokens, namely [CLS] and [SEP]. In order to match mBERT input format, we further split the input tokens into subwords.

Once fed to mBERT, we take the output of its final layer and use it to compute a representation for each token by averaging the vectors associated with the subwords that token was split into. Thanks to this procedure, subwords can attend to each other and generate a representation that is not only contextualized on the sentence in which they appear, but also on its translation. Finally, we leverage an additional 6-layer Transformer Encoder, fully resembling the architecture of mBERT,[1] so as to enable a token-level contextualization; formally, it takes as input the current representations of each token and outputs a sequence of $l + k$ vectors $token\_level\_out = h_{u_1}, \dots, h_{u_l}, h_{v_1}, \dots, h_{v_k}$ of dimension 768 each.[2]

As we now have a fully contextualized representation for each token, we classify each possible alignment separately. To this end, we first compute the tensor $H \in \mathbb{R}^{l \times k \times 768}$, where H_{ij} is the vector resulting from the element-wise product of h_{u_i} and h_{v_j}. Then, the word alignment matrix $A \in \mathbb{R}^{l \times k}$, with A_{ij} containing the probability of aligning u_i to v_j, is obtained as follows:

$$
\begin{aligned}
A &= Sigmoid(R_2 W_3) \\
R_2 &= Relu(R_1 W_2) \\
R_1 &= Relu(H W_1)
\end{aligned}
$$

where W_1 and W_2 are both matrices $\in \mathbb{R}^{768 \times 768}$ and $W_3 \in \mathbb{R}^{768 \times 1}$.

We train our model by minimizing the Binary Cross Entropy loss between A and $\hat{A}$, i.e., the alignment reference matrix containing the gold standard annotations, and such that $\hat{A}_{ij} = 1$ if the tokens u_i and v_j are aligned and 0 otherwise. The final loss for each sentence is thus computed as:

[1] With the exception of the embedding layer, which we omit as we already have continuous representations for each token.

[2] The hidden size of the final layer of multilingual BERT.

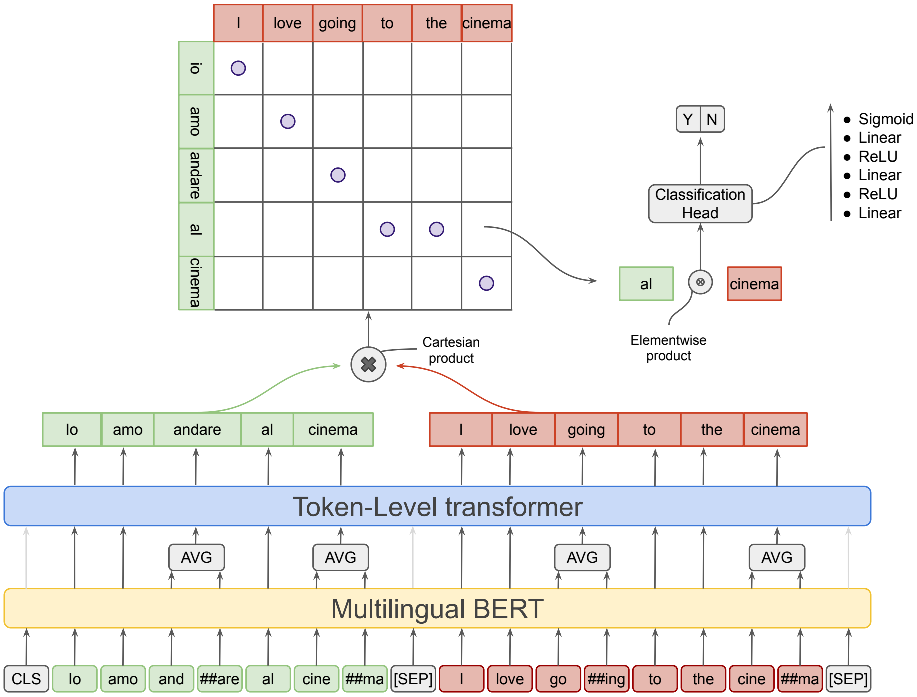

Figure 1: Depiction of our proposed word alignment model. The model takes as input the concatenation of two sentences and computes the alignment matrix. Best seen in color.

$$\mathcal{L}(A,\hat{A}) = -\frac{1}{l}\frac{1}{k}\sum_{i=1}^{l}\sum_{j=1}^{k}\Big(\hat{A}_{ij}\cdot\log(A_{ij}) \\ + (1-\hat{A}_{ij})\cdot(1-\log(A_{ij}))\Big)$$

Finally, as the datasets we employ were annotated by different annotators and, more in general, the consistency of the annotations is hampered by human factors, we apply a special masking over the loss function. Indeed, if two tokens u_i and v_j in two parallel sentences are not aligned to any other token, i.e., $\nexists\ m \in \{1,\ldots,l\} \mid \hat{A}_{mj} = 1$ and $\nexists\ r \in \{1,\ldots,k\} \mid \hat{A}_{ir} = 1$, the prediction of the model is not considered when computing the loss function; thus, the model is only updated when the predictions consider at least one token that has been manually aligned between the two parallel sentences. We train our model with a learning rate of 2×10^{-5}, Adam [Kingma and Ba, 2015] as the optimizer and a token batch size of 512.

Datasets The datasets we select for our experiments are those described by Nagata *et al.* [2020]. We use 4 manually-annotated datasets for word alignment in the following language combinations: English-French (En-Fr), German-English (De-En), Japanese-English (Ja-En) and Romanian-English (Ro-En). The En-Fr dataset and the Ro-En dataset were proposed during the the HLT-NAACL-2003 workshop on Building and Using Parallel Texts [Mihalcea and Pedersen, 2003]. The De-En dataset is derived from [Vilar *et al.*, 2006] and, finally, the Ja-En dataset was provided by Neubig [2011] and obtained by manually translating Wikipedia pages regarding Kyoto from Japanese into English. We split the sentences into training and test data as detailed in Nagata *et al.* [2020].

Comparison System We compare against the system with the best reported performances on these datasets, that is, the architecture proposed by Nagata *et al.* [2020]. The authors present a novel SQuAD-style approach in which every alignment is formulated as a query with respect to a context: the query represents a word in a sentence, the context is the translation of that sentence and the answer is the query word's corresponding translation. As in the authors' experimental setup, we consider here their *bidi-average* symmetrization strategy.

Results We show in Table 1 how our model fares when compared to our competitor in terms of F1 score over the alignments. Following previous works, we do not report the AER measure since it has been debated to be skewed towards precision [Fraser and Marcu, 2007].

As a first result, we note how our model performs considerably better in all settings and, interestingly, that the difference is most marked on a distant language, namely Japanese, where MULTIMIRROR surpasses Nagata *et al.* [2020] by almost 2 F1 points. Second, with the sole exception of En-Fr, our approach features a recall that is always consistently higher, with the closest gap on De-En amounting to 1.9 points. This finding is particularly relevant to our setting as MULTIMIRROR features filtering heuristics that validate the proposed alignments, thus discarding any spurious ones that might occur.

However, these results do not fully convey the benefits of our formulation. Indeed, thanks to it: i) we do not require

Languages	Method	P	R	F1
Ja-En	Nagata *et al.* [2020]	77.3	78.0	77.6
	MultiMirror Word Aligner	78.3	80.5	**79.4**
De-En	Nagata *et al.* [2020]	89.9	81.7	85.6
	MultiMirror Word Aligner	90.1	83.6	**86.7**
En-Fr	Nagata *et al.* [2020]	79.6	93.9	86.2
	MultiMirror Word Aligner	81.5	92.7	**86.8**
Ro-En	Nagata *et al.* [2020]	90.4	85.3	87.8
	MultiMirror Word Aligner	90.6	88.5	**89.1**

Table 1: Comparison between our architecture for word alignment and the SQuAD-style formulation of Nagata *et al.* [2020].

any symmetrization strategy as our predictions are symmetrical by design; ii) we enable the alignment of non-contiguous spans, which is not straightforward in SQuAD-style formulations; iii) our architecture is significantly faster since tokens are aligned jointly and in a single pass, in contrast to Nagata *et al.* [2020] where, instead, a different query for each source token has to be performed.

3.2 Sense Projection

Based on the output of our word alignment model, we now show how MULTIMIRROR projects senses across languages. Taking as input a source corpus, i.e., a list of sense-tagged sentences $s^1, \dots, s^n$ in language L_1 and their respective translations $t^1, \dots, t^n$ in language L_2, MULTIMIRROR automatically projects sense annotations from each s^i to the corresponding t^i, thus producing an automatically-tagged dataset as output.

Let $s^i = s^i_1 \dots s^i_{|s^i|}$ be the i-th source sentence and t^i the corresponding target sentence defined analogously. Given a sense tag $l_{a,b}$ for a contiguous span $s^i_a \dots s^i_b$ in s^i,[3] our aim is to identify the target span $t^i_{a'} \dots t^i_{b'}$ onto which sense tag $l_{a,b}$ should be projected. To this end, we utilize the alignment matrix A obtained from our cross-lingual word alignment model (Section 3.1) for the pair (s^i, t^i). However, the alignments in A are token-level and potentially non-contiguous. We therefore bring together such alignments to produce a list of non-overlapping continuous target spans:

$$T(a,b) = \{t^i_{a'} \dots t^i_{b'} | \forall j \in [a', b'] \exists k \in [a,b] \; s.t. \; A_{k,j} > 0.5\}$$

whose elements are sorted by their start index a'. For each label $l_{a,b}$ in the set of sense annotations for s^i, we select the first target span $\tau \in T(a,b)$ which does not overlap with any of the previously-chosen target spans for any of the preceding labels and such that the part of speech of any of the tokens in τ matches with the one of the sense $l_{a,b}$ we are projecting.

4 Experimental Setup

In this Section, we present the experimental setting we adopt to evaluate MULTIMIRROR in the Word Sense Disambiguation task. First, we detail the corpora we use in our sense projection process to automatically create training data (Section 4.1). Second, we illustrate the test bed (Section 4.2) and the WSD reference model employed (Section 4.3). Finally, we describe the comparison systems in Section 4.4.

4.1 Datasets

Alignment Datasets In order to perform sense projection, we need data on which to train our alignment model. To this end, we use the same reference alignment data described in Section 3.1. As those datasets cover only alignments between English and three other languages, i.e., French, German and Japanese, we produce datasets of comparable size for Italian and Spanish by manually aligning for each of them 300 sentence pairs (with around 4,000 aligned tokens) from WikiMatrix [Schwenk *et al.*, 2021], of which 50 are reserved for development.

However, our alignment data is based on Wikipedia text, while the reference data from Section 3.1 comes from different genres, including constitution laws, news text, etc. To determine whether different genres and domains might impact on the performance of our word alignment model (and, consequently, on WSD), we also create analogous alignment datasets in two of the languages for which reference data was already available (cf. Section 3.1), i.e., French and German.

Overall, we report slightly inferior word alignment performances when training on our newly annotated datasets and testing on the reference test sets, respectively 84.6 for French (-2.2) and 83.2 for German (-3.5). This result is partially motivated by the different source distribution of the two datasets and the smaller number of training instances caused by the fact that the sentences in WikiMatrix have a shorter length on average. In Section 5, we further investigate this difference and evaluate how both our training data and the reference training data fare when used for sense projection.

Source and Target Corpora As source corpus, following Barba *et al.* [2020], we use the concatenation of the two largest English corpora tagged with WordNet senses, i.e., SemCor [Miller *et al.*, 1993] and the Princeton WordNet Gloss Corpus[4] (WNG). This concatenation amounts to roughly $720k$ sense annotations. As reference translations of this data are not available, we generate the target corpus by automatically translating the source corpus. To this end, we employ the multilingual translation model released by Tang *et al.* [2020] and use it to perform the translation towards the languages included in our evaluation test bed, namely: French, German, Italian, Japanese, and Spanish.

Sense Inventory Following Barba *et al.* [2020], we use BabelNet [Navigli and Ponzetto, 2012] as our sense inventory, whose multilingual synsets include WordNet ones. This choice enables a seamless projection of senses from WordNet-annotated corpora, like ours, to any other language, since the projected synset remains the same. We note that, in stark contrast to other dataset creation approaches, including MuLaN, we propagate sense annotations based solely on the resulting span alignment and independently of whether BabelNet contains the target span as a lexicalization in the synset at hand.

[3] $s^i_a \dots s^i_b$ corresponds to a word if $a = b$ and a multiword expression otherwise.

[4] https://wordnetcode.princeton.edu/glosstag.shtml

Produced Datasets We report coverage statistics in Table 2 for the datasets generated by both MULTIMIRROR and our strongest competitor, i.e., MuLaN [Barba *et al.*, 2020], over each of the 5 target languages we consider. As Japanese was not among the languages originally included in Barba *et al.* [2020], we execute their projection pipeline ourselves[5], transferring annotations from the concatenations of SemCor and WNG towards Japanese Wikipedia[6]. First of all, we note how MULTIMIRROR steadily transfers a significantly larger number of instances, the only exception being Japanese, highlighting the better coverage our system attains. However, arguably more interesting is the difference in the number of senses and synsets transferred; indeed, not only does our approach increase the number of projected synsets by as few as $2k$, but also pushes up the amount of projected senses by an astounding margin, almost quadrupling them in Japanese. This seems to suggest that, for a given synset, most transfers MuLaN performs are skewed towards a specific sense.

		IT	ES	FR	DE	JA
MULTIMIRROR	# instances	519k	552k	387k	318k	301k
	# senses	77k	92k	62k	68k	98k
	# synsets	37k	50k	29k	22k	25k
	# multiwords	28k	38k	19k	19k	40k
MuLaN	# instances	415k	452k	310k	245k	310k
	# senses	44k	57k	29k	22k	27k
	# synsets	33k	43k	25k	19k	21k
	# multiwords	18k	22k	20k	6k	552

Table 2: Statistics of training sets from MULTIMIRROR (top) and MuLaN (bottom) in terms of number of sense-tagged instances, unique senses, unique synsets and number of multiwords.

4.2 Test Bed

We evaluate our automatically-generated sense-tagged corpora against the well-established SemEval-13 [Navigli *et al.*, 2013] and SemEval-15 [Moro and Navigli, 2015] datasets[7], the former containing nominal instances in German, French, Italian and Spanish, and the latter covering all open-class parts of speech for Italian and Spanish.

Furthermore, so as to assess the extent to which our approach and its alternatives scale to distant languages, we include Japanese in our experiments and evaluate MULTIMIRROR on the Japanese section of XL-WSD, a cross-lingual Word Sense Disambiguation framework recently proposed by Pasini *et al.* [2021];[8] the resources the authors released for this language include both a validation set and a test set, featuring 1,901 and 7,602 instances, respectively, and covering all parts of speech.

As in previous works, we report the F1 score between reference and predicted labels. We further show, in each setting, whether the results we attain are statistically significant with respect to the strongest competitor by performing McNemar's test [Dietterich, 1998].

4.3 Reference WSD Model

To carry out the evaluation, we employ a simple linear classifier on top of *multilingual BERT* (mBERT) [Devlin *et al.*, 2019] as our reference model. Specifically, given as input a list of words, which we tokenize into word pieces as required by mBERT, we first encode them, taking the concatenation of the last 4 layers as the representation of each word piece, and then feed the resulting vectors into a fully-connected layer with a softmax activation function; should a text span be split into multiple word pieces, we use the vector associated with the first word piece as the representation of the whole span. We explore the usage of 2 different strategies at training time: i) *fine-tuned (FT)*: where we fine-tune the whole model, with mBERT being updated along with the linear classifier, and ii) *feature-based (FB)*: where, instead, we freeze weights and update only the linear classifier.

In all our experiments, we train our models on a single NVIDIA RTX 2080 Ti for 50 epochs, monitoring validation accuracy for early stopping (patience $p = 3$) and using the Adam optimizer with learning rate 2×10^{-5}. As no official validation sets exist for the 4 SemEval-13 and 2 SemEval-15 tasks, we reserve 1000 sentences from our training set and use them for development only; conversely, we use the official validation set when dealing with Japanese.

4.4 Comparison Systems

We consider as our baselines the Most Common Sense, computed as described in Pasini *et al.* [2021] for Japanese and Barba *et al.* [2020] for the other languages, and two neural settings in which our model is trained on English datasets and tasked to *zero-shot* over the test languages: i) $\emptyset$-shot-SemCor, where the learning procedure is performed over SemCor, and ii) $\emptyset$-shot-SemCor+WNG, where, instead, the concatenation of SemCor and WNG is used. In both cases, we use the English SemEval-2007 dataset for validation.

As competitors, we consider the following systems: UKB+SyntagNet [Maru *et al.*, 2019], a knowledge-based approach which applies Personalized Page Rank over the WordNet graph further enriched with collocational edges, ARES [Scarlini *et al.*, 2020], a semi-supervised approach for producing sense embeddings that lie in a space comparable to that of multilingual contextualized embeddings, and MuLaN, an annotation projection technique that currently holds the state of the art in most tasks of our experimental setting.

5 Results

We now investigate how MULTIMIRROR fares when compared with current state-of-the-art alternatives in multilingual WSD. To this end, we train our reference model in all 5 languages, in both *FB* and *FT* configurations, which we denote as MULTIMIRROR^{FB} and MULTIMIRROR^{FT}, respectively. Furthermore, we also report the results on the *4L* setting, the

[5] We utilized the original code that the authors released at https://github.com/SapienzaNLP/mulan.

[6] The authors used Wikipedia as the target corpus in all their experiments.

[7] We use the latest version of these datasets available at https://github.com/SapienzaNLP/mwsd-datasets.

[8] The choice of Japanese is primarily motivated by the fact that it was the only language not already included in the standard evaluation suite for which we already had at disposal its alignment data.

			SemEval-13				SemEval-15		XL-WSD
	Model	Alignment Data	IT	ES	FR	DE	IT	ES	JA
Baselines	MCS	-	44.20	37.10	53.20	70.20	44.60	39.60	48.71
	∅-shot-SemCor	-	74.63	78.25	80.06	79.09	70.21	65.77	55.20
	∅-shot-SemCor+WNG	-	77.30	79.30	81.40	80.00	72.00	68.20	58.10
1L	UKB+SyntagNet	-	72.14	74.12	70.32	76.39	68.95	63.37	-
	ARES	-	77.00	75.30	81.20	79.60	71.40	70.10	-
	MuLaN	-	77.45	77.70	80.12	82.09	70.31	68.73	57.59
	MultiMirror^{FB}	Nagata *et al.* [2020]	-	-	81.09	82.16	-	-	58.34
	MultiMirror^{FT}	Nagata *et al.* [2020]	-	-	81.78	**83.18**	-	-	**62.60**
	MultiMirror^{FB}	Ours	78.59	79.68	80.81	81.13	**73.49**	69.03	-
	MultiMirror^{FT}	Ours	**79.53**	**81.83**	**83.44**	82.81	72.89	69.42	-
4L	MuLaN$_{4L}$	-	77.85	81.11	81.64	82.34	71.80	69.42	-
	MultiMirror$^{FB}_{4L}$	Best	78.59	81.67	81.64	82.43	73.39	69.42	-
	MultiMirror$^{FT}_{4L}$	Best	**79.60**	**82.17**	**83.64**	**83.71**	**73.69**	**70.42**	-

Table 3: Comparison of MultiMirror against its competitors on SemEval-13, SemEval-15, and the Japanese test set released by Pasini *et al.* [2021]. Underlined: the first statistically significant results against their best performing competitor according to McNemar's test, $p < 0.01$. Bold: best system in its category.

new configuration introduced by Barba *et al.* [2020] where the concatenation of the datasets generated for the 4 European languages is used as the training set. We show the F1 scores that MultiMirror attains in Table 3.

First of all, our results suggest that the word alignment data we manually annotated is indeed compliant with their reference counterpart (Section 3.1), with performances almost in the same ballpark on German and significantly better on French, most likely thanks to the increased number of *correct alignments*[9]; indeed, as pointed out by Nagata *et al.* [2020], the reference French dataset contains a considerable number of noisy *possible alignments*.

More interestingly, we point out that MultiMirror achieves state-of-the-art results, with MultiMirror^{FB} outperforming its main competitor, namely MuLaN, in all settings. In particular, since MultiMirror^{FB} features the very same WSD model MuLaN employs, this result clearly highlights the quality of our automatically-generated datasets. Once we switch to MultiMirror^{FT}, this trend becomes even more marked:[10] MultiMirror^{FT} outperforms all previous *monolingual*[11] systems in virtually all settings, with the only exception of the Spanish SemEval-15 dataset.

However, we believe that the most interesting result concerns Japanese, i.e., our distant language. This setting shows the largest gap between our approach and MuLaN, which actually performs worse than direct zero-shot itself, i.e., ∅-shot-SemCor+WNG. This finding suggests that there might be possible limitations in applying MuLaN on distant languages. Conversely, MultiMirror^{FT} scales remarkably well and outperforms ∅-shot-SemCor+WNG by more than 4 F1 points.

Finally, we consider the *4L* setting. Similarly to Barba *et al.* [2020], we report significant improvements over the other systems across the board, including outperforming ARES on the Spanish SemEval-15 dataset.

6 Conclusions

In this work, we presented MultiMirror, a novel approach for cross-lingual sense projection based on word alignment. Leveraging an innovative neural aligner that requires just a few hundred sentences for its training, we tackle the creation of sense-tagged datasets in multiple languages by projecting sense annotations across the produced alignments. We find that MultiMirror achieves high sense and synset coverage in many languages and that its automatically-generated datasets lead a simple transformer-based classifier to reach state-of-the-art results in established WSD benchmarks against both strong alternatives for silver data creation and state-of-the-art multilingual systems. Most interestingly, we report significantly better results when projecting onto a distant language, namely Japanese.

As a by-product of our experiments, we release four novel datasets for word alignment, each containing 300 sentences, in English and one of the following languages: French, German, Italian and Spanish. As future work, we plan to assess the scalability of this framework when considering comparable rather than parallel sentences.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the ELEXIS project No. 731015 under the European Union's Horizon 2020 research and innovation programme. This work was partially supported by the MIUR under the grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of the Sapienza University of Rome.

[9] Here, we use *correct alignment*, as opposed to *possible alignment*.

[10] We also explored training MuLaN data on this architecture but did not observe any significant gain.

[11] Here we use *monolingual* as opposed to *4L*.

References

[Agirre and Martinez, 2004] Eneko Agirre and David Martinez. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proc. of EMNLP*, pages 25–32, 2004.

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

[Aminian *et al.*, 2019] Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. Cross-lingual transfer of semantic roles: From raw text to semantic roles. In *Proc. of COLING*, 2019.

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

[Bevilacqua *et al.*, 2021] Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. Recent trends in Word Sense Disambiguation: A survey. In *Proc. of IJCAI*, 2021.

[Blloshmi *et al.*, 2020] Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. Enabling cross-lingual amr parsing with transfer learning techniques. In *Proc. of EMNLP*, 2020.

[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*, 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, 2019.

[Dietterich, 1998] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[Fraser and Marcu, 2007] Alexander Fraser and Daniel Marcu. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303, 2007.

[Gale *et al.*, 1992] William A Gale, Kenneth W Church, and David Yarowsky. Using bilingual materials to develop word sense disambiguation methods. In *Proc. of Conference on Theoretical and Methodological Issues in MT*, 1992.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[Leacock *et al.*, 1998] Claudia Leacock, Martin Chodorow, and George A Miller. Using corpus statistics and wordnet relations for sense identification. *Comput. Ling.*, 24(1):147–165, 1998.

[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*, 2019.

[Mihalcea and Pedersen, 2003] Rada Mihalcea and Ted Pedersen. An evaluation exercise for word alignment. In *Proc. of HLT-NAACL*, pages 1–10, 2003.

[Miller *et al.*, 1993] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In *Proc. of HLT*, pages 303–308, 1993.

[Miller, 1998] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[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.

[Moro *et al.*, 2014] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the ACL*, 2, 2014.

[Nagata *et al.*, 2020] Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proc. of EMNLP*, pages 555–565, November 2020.

[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*, pages 222–231, 2013.

[Neubig, 2011] Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

[Padó and Lapata, 2009] Sebastian Padó and Mirella Lapata. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340, 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 *et al.*, 2021] Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proc. of AAAI*, 2021.

[Pasini, 2020] Tommaso Pasini. The knowledge acquisition bottleneck problem in multilingual word sense disambiguation. In *Proceedings of IJCAI*, 2020.

[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, July 2019.

[Scarlini *et al.*, 2020] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proc. of EMNLP*, pages 3528–3539, 2020.

[Schwenk *et al.*, 2021] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from wikipedia. In *Proc. of EACL*, 2021.

[Stengel-Eskin *et al.*, 2019] Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. A discriminative neural model for cross-lingual word alignment. In *Proc. of EMNLP-IJCNLP*, pages 909–919, 2019.

[Tang *et al.*, 2020] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.

[Vilar *et al.*, 2006] David Vilar, Maja Popović, and Hermann Ney. Aer: Do we need to “improve” our alignments? In *Proc. of Workshop on Spoken Language Translation*, 2006.

[Xi and Hwa, 2005] Chenhai Xi and Rebecca Hwa. A backoff model for bootstrapping resources for non-english languages. In *Proc. of HLT-EMNLP*, pages 851–858, 2005.

[Yarowsky *et al.*, 2001] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. Technical report, Johns Hopkins Univ., 2001.