Unifying Cross-Lingual Semantic Role Labeling with Heterogeneous Linguistic Resources

Simone Conia **Andrea Bacciu** **Roberto Navigli**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{first.lastname}@uniroma1.it

Abstract

While cross-lingual techniques are finding increasing success in a wide range of Natural Language Processing tasks, their application to Semantic Role Labeling (SRL) has been strongly limited by the fact that each language adopts its own linguistic formalism, from PropBank for English to AnCora for Spanish and PDT-Vallex for Czech, inter alia. In this work, we address this issue and present a unified model to perform cross-lingual SRL over heterogeneous linguistic resources. Our model implicitly learns a high-quality mapping for different formalisms across diverse languages without resorting to word alignment and/or translation techniques. We find that, not only is our cross-lingual system competitive with the current state of the art but that it is also robust to low-data scenarios. Most interestingly, our unified model is able to annotate a sentence in a single forward pass with all the inventories it was trained with, providing a tool for the analysis and comparison of linguistic theories across different languages. We release our code and model at https://github.com/SapienzaNLP/unify-srl.

1 Introduction

Semantic Role Labeling (SRL) – a long-standing open problem in Natural Language Processing (NLP) and a key building block of language understanding (Navigli, 2018) – is often defined as the task of automatically addressing the question "Who did what to whom, when, where, and how?" (Gildea and Jurafsky, 2000; Màrquez et al., 2008). While the need to manually engineer and fine-tune complex feature templates severely limited early work (Zhao et al., 2009), the great success of neural networks in NLP has resulted in impressive progress in SRL, thanks especially to the ability of recurrent networks to better capture relations over sequences (He et al., 2017; Marcheggiani et al., 2017). Owing to the recent wide availability of robust multilingual representations, such as multilingual word embeddings (Grave et al., 2018) and multilingual language models (Devlin et al., 2019; Conneau et al., 2020), researchers have been able to shift their focus to the development of models that work on multiple languages (Cai and Lapata, 2019b; He et al., 2019; Lyu et al., 2019).

A robust multilingual representation is nevertheless just one piece of the puzzle: a key challenge in multilingual SRL is that the task is tightly bound to linguistic formalisms (Màrquez et al., 2008) which may present significant structural differences from language to language (Hajic et al., 2009). In the recent literature, it is standard practice to sidestep this issue by training and evaluating a model on each language separately (Cai and Lapata, 2019b; Chen et al., 2019; Kasai et al., 2019; He et al., 2019; Lyu et al., 2019). Although this strategy allows a model to adapt itself to the characteristics of a given formalism, it is burdened by the non-negligible need for training and maintaining one model instance for each language, resulting in a set of monolingual systems.

Instead of dealing with heterogeneous linguistic theories, another line of research consists in actively studying the effect of using a single formalism across multiple languages through annotation projection or other transfer techniques (Akbik et al., 2015, 2016; Daza and Frank, 2019; Cai and Lapata, 2020; Daza and Frank, 2020). However, such approaches often rely on word aligners and/or automatic translation tools which may introduce a considerable amount of noise, especially in low-resource languages. More importantly, they rely on the strong assumption that the linguistic formalism of choice, which may have been developed with a specific language in mind, is also suitable for other languages.

In this work, we take the best of both worlds and propose a novel approach to cross-lingual SRL. Our contributions can be summarized as follows:

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 338–351
June 6–11, 2021.

- We introduce a unified model to perform cross-lingual SRL with heterogeneous linguistic resources;
- We find that our model is competitive against state-of-the-art systems on all the 6 languages of the CoNLL-2009 benchmark;
- We show that our model is robust to low-resource scenarios, thanks to its ability to generalize across languages;
- We probe our model and demonstrate that it implicitly learns to align heterogeneous linguistic resources;
- We automatically build and release a cross-lingual mapping that aligns linguistic formalisms from diverse languages.

We hope that our unified model will further advance cross-lingual SRL and represent a tool for the analysis and comparison of linguistic theories across multiple languages.

2 Related Work

End-to-end SRL. The SRL pipeline is usually divided into four steps: predicate identification, predicate sense disambiguation, argument identification, and argument classification. While early research focused its efforts on addressing each step individually (Xue and Palmer, 2004; Björkelund et al., 2009; Zhao et al., 2009), recent work has successfully demonstrated that tackling some of these subtasks jointly with multitask learning (Caruana, 1997) is beneficial. In particular, He et al. (2018) and, subsequently, Cai et al. (2018), Li et al. (2019) and Conia et al. (2020), indicate that predicate sense signals aid the identification of predicate-argument relations. Therefore, we follow this line and propose an end-to-end system for cross-lingual SRL.

Multilingual SRL. Current work in multilingual SRL revolves mainly around the development of novel neural architectures, which fall into two broad categories, syntax-aware and syntax-agnostic ones. On one hand, the quality and diversity of the information encoded by syntax is an enticing prospect that has resulted in a wide range of contributions: Marcheggiani and Titov (2017) made use of Graph Convolutional Networks (GCNs) to better capture relations between neighboring nodes in syntactic dependency trees; Strubell et al. (2018) demonstrated the effectiveness of linguistically-informed self-attention layers in SRL; Cai and Lapata (2019b) observed that syntactic dependencies often mirror semantic relations and proposed a model that jointly learns to perform syntactic dependency parsing and SRL; He et al. (2019) devised syntax-based pruning rules that work for multiple languages. On the other hand, the complexity of syntax and the noisy performance of automatic syntactic parsers have deterred other researchers who, instead, have found methods to improve SRL without syntax: Cai et al. (2018) took advantage of an attentive biaffine layer (Dozat and Manning, 2017) to better model predicate-argument relations; Chen et al. (2019) and Lyu et al. (2019) obtained remarkable results in multiple languages by capturing predicate-argument interactions via capsule networks and iteratively refining the sequence of output labels, respectively; Cai and Lapata (2019a) proposed a semi-supervised approach that scales across different languages.

While we follow the latter trend and develop a syntax-agnostic model, we underline that both the aforementioned syntax-aware and syntax-agnostic approaches suffer from a significant drawback: they require training one model instance for each language of interest. Their two main limitations are, therefore, that i) the number of trainable parameters increases linearly with the number of languages, and ii) the information available in one language cannot be exploited to make SRL more robust in other languages. In contrast, one of the main objectives of our work is to develop a unified cross-lingual model which can mitigate the paucity of training data in some languages by exploiting the information available in other, resource-richer languages.

Cross-lingual SRL. A key challenge in performing cross-lingual SRL with a single unified model is the dissimilarity of predicate sense and semantic role inventories between languages. For example, the multilingual dataset distributed as part of the CoNLL-2009 shared task (Hajic et al., 2009) adopts the English Proposition Bank (Palmer et al., 2005) and NomBank (Meyers et al., 2004) to annotate English sentences, the Chinese Proposition Bank (Xue and Palmer, 2009) for Chinese, the AnCora (Taulé et al., 2008) predicate-argument structure inventory for Catalan and Spanish, the German Proposition Bank which, differently from the other PropBanks, is derived from FrameNet (Hajic et al., 2009), and

PDT-Vallex (Hajic et al., 2003) for Czech. Many of these inventories are not aligned with each other as they follow and implement different linguistic theories which, in turn, may pose different challenges.

Padó and Lapata (2009), and Akbik et al. (2015, 2016) worked around these issues by making the English PropBank act as a universal predicate sense and semantic role inventory and projecting PropBank-style annotations from English onto non-English sentences by means of word alignment techniques applied to parallel corpora such as Europarl (Koehn, 2005). These efforts resulted in the creation of the Universal PropBank, a multilingual collection of semi-automatically annotated corpora for SRL, which is actively in use today to train and evaluate novel cross-lingual methods such as word alignment techniques (Aminian et al., 2019). In the absence of parallel corpora, annotation projection techniques can still be applied by automatically translating an annotated corpus and then projecting the original labels onto the newly created silver corpus (Daza and Frank, 2020; Fei et al., 2020), whereas Daza and Frank (2019) have recently found success in training an encoder-decoder architecture to jointly tackle SRL and translation.

While the foregoing studies have greatly advanced the state of cross-lingual SRL, they suffer from an intrinsic downside: using translation and word alignment techniques may result in a considerable amount of noise, which automatically puts an upper bound to the quality of the projected labels. Moreover, they are based on the strong assumption that the English PropBank provides a suitable formalism for non-English languages, and this may not always be the case. Among the numerous studies that adopt the English PropBank as a universal predicate-argument structure inventory for cross-lingual SRL, the work of Mulcaire et al. (2018) stands out for proposing a bilingual model that is able to perform SRL according to two different inventories at the same time, although with significantly lower results compared to the state of the art at the time. With our work, we go beyond current approaches to cross-lingual SRL and embrace the diversity of the various representations made available in different languages. In particular, our model has three key advantages: i) it does not rely on word alignment or machine translation tools; ii) it learns to perform SRL with multiple linguistic inventories; iii) it learns to link resources that would otherwise be disconnected from each other.

3 Model Description

In the wake of recent work in SRL, our model falls into the broad category of end-to-end systems as it learns to jointly tackle predicate identification, predicate sense disambiguation, argument identification and argument classification. The model architecture can be roughly divided into the following components:

- A universal sentence encoder whose parameters are shared across languages and which produces word encodings that capture predicate-related information (Section 3.2);
- A universal predicate-argument encoder whose parameters are also shared across languages and which models predicate-argument relations (Section 3.3);
- A set of language-specific decoders which indicate whether words are predicates, select the most appropriate sense for each predicate, and assign a semantic role to every predicate-argument couple, according to several different SRL inventories (Section 3.4).

Unlike previous work, our model does not require any preexisting cross-resource mappings, word alignment techniques, translation tools, other annotation transfer techniques, or parallel data, to perform high-quality cross-lingual SRL, as it relies solely on implicit cross-lingual knowledge transfer.

3.1 Input representation

Pretrained language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), *inter alia*, are becoming the *de facto* input representation method, thanks to their ability to encode vast amounts of knowledge. Following recent studies (Hewitt and Manning, 2019; Kuznetsov and Gurevych, 2020; Conia and Navigli, 2020), which show that different layers of a language model capture different syntactic and semantic characteristics, our model builds a contextual representation for an input word by concatenating the corresponding hidden states of the four top-most inner layers of a language model. More formally, given a word w_i in a sentence $\mathbf{w} = \langle w_0, w_1, \ldots, w_i, \ldots, w_{n-1} \rangle$ of n words and its hidden state $\mathbf{h}_i^k = l^k(w_i|\mathbf{w})$ from the k-th inner layer l^k of a language model with K layers, the model computes the word encoding $\mathbf{e}_i$

as follows:

$$\mathbf{h}_i = \mathbf{h}_i^K \oplus \mathbf{h}_i^{K-1} \oplus \mathbf{h}_i^{K-2} \oplus \mathbf{h}_i^{K-3}$$
$$\mathbf{e}_i = \text{Swish}(\mathbf{W}^w \mathbf{h}_i + \mathbf{b}^w)$$

where $\mathbf{x} \oplus \mathbf{y}$ is the concatenation of the two vectors $\mathbf{x}$ and $\mathbf{y}$, and $\text{Swish}(x) = x \cdot \text{sigmoid}(x)$ is a non-linear activation which was found to produce smoother gradient landscapes than the more traditional ReLU (Ramachandran et al., 2018).

3.2 Universal sentence encoder

Expanding on the seminal intuition of Fillmore (1968), who suggests the existence of deep semantic relations between a predicate and other sentential constituents, we argue that such semantic relations may be preserved across languages. With this reasoning in mind, we devise a universal sentence encoder whose parameters are shared across languages. Intuitively, the aim of our universal sentence encoder is to capture sentence-level information that is not formalism-specific and spans across languages, such as information about predicate positions and predicate senses. In our case, we implement this universal sentence encoder as a stack of BiLSTM layers (Hochreiter and Schmidhuber, 1997), similarly to Marcheggiani et al. (2017), Cai et al. (2018) and He et al. (2019), with the difference that we concatenate the output of each layer to its input in order to mitigate the problem of vanishing gradients. More formally, given a sequence of word encodings $\mathbf{e} = \langle \mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_{n-1} \rangle$, the model computes a sequence of timestep encodings $\mathbf{t}$ as follows:

$$\mathbf{t}_i^j = \begin{cases} \mathbf{e}_i & \text{if} \quad j = 0 \\ \mathbf{t}_i^{j-1} \oplus \text{BiLSTM}_i^j(\mathbf{t}^{j-1}) & \text{otherwise} \end{cases}$$
$$\mathbf{t} = \langle \mathbf{t}_0^{K'}, \mathbf{t}_1^{K'}, \ldots, \mathbf{t}_{n-1}^{K'} \rangle$$

where $\text{BiLSTM}_i^j(\cdot)$ is the i-th timestep of the j-th BiLSTM layer and K' is the total number of layers in the stack. Starting from each timestep encoding $\mathbf{t}_i$, the model produces a predicate representation $\mathbf{p}_i$, which captures whether the corresponding word w_i is a predicate, and a sense representation $\mathbf{s}_i$ which encodes information about the sense of a predicate at position i:

$$\mathbf{p}_i = \text{Swish}(\mathbf{W}^p \mathbf{t}_i + \mathbf{b}^p)$$
$$\mathbf{s}_i = \text{Swish}(\mathbf{W}^s \mathbf{t}_i + \mathbf{b}^s)$$

We stress that the vector representations obtained for each timestep, each predicate and each sense lie in three spaces that are shared across the languages and formalisms used to perform SRL.

3.3 Universal predicate-argument encoder

In the same vein, and for the same reasoning that motivated the design of the above universal sentence encoder, our model includes a universal predicate-argument encoder whose parameters are also shared across languages. The objective of this second encoder is to capture the relations between each predicate-argument couple that appears in a sentence, independently of the input language. Similarly to the universal sentence encoder, we implement this universal predicate-argument encoder as a stack of BiLSTM layers. More formally, let w_p be a predicate in the input sentence $\mathbf{w} = \langle w_0, w_1, \ldots, w_p, \ldots, w_{n-1} \rangle$, then the model computes a sequence of predicate-specific argument encodings $\mathbf{a}$ as follows:

$$\mathbf{a}_i^j = \begin{cases} \mathbf{t}_p \oplus \mathbf{t}_i & \text{if} \quad j = 0 \\ \mathbf{a}_i^{j-1} \oplus \text{BiLSTM}_i^j(\mathbf{a}^{j-1}) & \text{otherwise} \end{cases}$$
$$\mathbf{a} = \langle \mathbf{a}_0^{K''}, \mathbf{a}_1^{K''}, \ldots, \mathbf{a}_{n-1}^{K''} \rangle$$

where $\mathbf{t}_i$ is the i-th timestep encoding from the universal sentence encoder and K'' is the total number of layers in the stack. Starting from each predicate-specific argument encoding $\mathbf{a}_i$, the model produces a semantic role representation $\mathbf{r}_i$ for word w_i:

$$\mathbf{r}_i = \text{Swish}(\mathbf{W}^r \mathbf{a}_i + \mathbf{b}^r)$$

Similarly to the predicate and sense representations $\mathbf{p}$ and $\mathbf{s}$, since the predicate-argument encoder is one and the same for all languages, the semantic role representation $\mathbf{r}$ obtained must draw upon cross-lingual information in order to abstract from language-specific peculiarities.

3.4 Language-specific decoders

The aforementioned predicate encodings $\mathbf{p}$, sense encodings $\mathbf{s}$ and semantic role encodings $\mathbf{r}$ are shared across languages, forcing the model to learn from semantics rather than from surface-level features such as word order, part-of-speech tags and syntactic rules, all of which may differ from language to language. Ultimately, however, we want our model to provide semantic role annotations according to an existing predicate-argument structure inventory, e.g., PropBank, AnCora, or PDT-Vallex. Our model, therefore, includes a set of linear decoders that indicate whether a word w_i is

a predicate, what the most appropriate sense for a predicate w_p is, and what the semantic role of a word w_r with respect to a specific predicate w_p is, for each language l:

$$\sigma^p(w_i|l) = \mathbf{W}^{p|l}\mathbf{p}_i + \mathbf{b}^{p|l}$$
$$\sigma^s(w_p|l) = \mathbf{W}^{s|l}\mathbf{s}_i + \mathbf{b}^{s|l}$$
$$\sigma^r(w_r|w_p, l) = \mathbf{W}^{r|l}\mathbf{r}_i + \mathbf{b}^{r|l}$$

Although we could have opted for more complex decoding strategies, in our case linear decoders have two advantages: 1) they keep the language-specific part of the model as simple as possible, pushing the model into learning from its universal encoders; 2) they can be seen as linear probes, providing an insight into the quality of the cross-lingual knowledge that the model can capture.

3.5 Training objective

The model is trained to jointly minimize the sum of the categorical cross-entropy losses on predicate identification, predicate sense disambiguation and argument identification/classification over all the languages in a multitask learning fashion. More formally, given a language l and the corresponding predicate identification loss $\mathcal{L}^{p|l}$, predicate sense disambiguation loss $\mathcal{L}^{s|l}$ and argument identification/classification loss $\mathcal{L}^{r|l}$, the cumulative loss $\mathcal{L}$ is:

$$\mathcal{L} = \sum_{l \in L} \left(\mathcal{L}^{p|l} + \mathcal{L}^{s|l} + \mathcal{L}^{r|l}\right)$$

where L is the set of languages – and the corresponding formalisms – in the training set.

4 Experiments

We evaluate our model in dependency-based multilingual SRL. The remainder of this Section describes the experimental setup (Section 4.1), provides a brief overview of the multilingual dataset we use for training, validation and testing (Section 4.2), and shows the results obtained on each language (Section 4.3).

4.1 Experimental Setup

We implemented the model in PyTorch[1] and PyTorch Lightning[2], and used the pretrained language models for multilingual BERT (m-BERT) and XLM-RoBERTa (XLM-R) made available by the Transformers library (Wolf et al., 2020). We trained each model configuration for 30 epochs using Adam (Kingma and Ba, 2015) with a "slanted triangle" learning rate scheduling strategy which linearly increases the learning rate for 1 epoch and then linearly decreases the value for 15 epochs. We did not perform hyperparameter tuning and opted instead for standard values used in the literature; we provide more details about our model configuration and its hyperparameter values in Appendix A. In the remainder of this Section, we report the F_1 scores of the best models selected according to the highest F_1 score obtained on the validation set at the end of a training epoch.[3]

4.2 Dataset

To the best of our knowledge, the dataset provided as part of the CoNLL-2009 shared task (Hajic et al., 2009) is the largest and most diverse collection of human-annotated sentences for multilingual SRL. It comprises 6 languages[4], namely, Catalan, Chinese, Czech, English, German and Spanish, which belong to different linguistic families and feature significantly varying amounts of training samples, from 400K predicate instances in Czech to only 17K in German; we provide an overview of the statistics of each language in Appendix B. CoNLL-2009 is the ideal testbed for evaluating the ability of our unified model to generalize across heterogeneous resources since each language adopts its own linguistic formalism, from English PropBank to PDT-Vallex, from Chinese PropBank to AnCora. We also include VerbAtlas (Di Fabio et al., 2019), a recently released resource for SRL[5], with the aim of understanding whether our model can learn to align inventories that are based on "distant" linguistic theories; indeed, VerbAtlas is based on clustering WordNet synsets into frames that share similar semantic behavior, whereas PropBank-based resources enumerate and define the possible senses of a lexeme.

As a final note, we did not evaluate our model on Universal PropBank[6] since i) it was semi-automatically generated through annotation pro-

[1] https://pytorch.org
[2] https://www.pytorchlightning.ai

[3] Hereafter, all the results of our experiments are computed by the official scorer of the CoNLL-2009 shared task, available at https://ufal.mff.cuni.cz/conll2009-st/scorer.html.
[4] The CoNLL-2009 shared task originally included a seventh language, Japanese, which is not available anymore on LDC due to licensing issues.
[5] We build a training set for VerbAtlas using the mapping from PropBank available at http://verbatlas.org.
[6] https://github.com/System-T/UniversalPropositions

CoNLL-2009 - Multilingual - In Domain		CA	CZ	DE	EN	ES	ZH
CoNLL-2009 ST best	⊙	80.3	85.4	79.7	85.6	80.5	78.6
Marcheggiani et al. (2017)	⊘	—	86.0	—	87.7	80.3	81.2
Chen et al. (2019)	⊘	81.7	88.1	76.4	91.1	81.3	81.7
Cai and Lapata (2019b)	⊙	—	—	82.7	90.0	81.8	83.6
Cai and Lapata (2019a)	⊙	—	—	83.8	91.2	82.9	85.0
Lyu et al. (2019)	⊙	80.9	87.5	75.8	90.1	80.5	83.3
He et al. (2019)	⊙	86.0	89.7	81.1	90.9	85.2	86.9
This work $_{\text{m-BERT frozen / monolingual}}$	⊘	86.2	90.0	85.2	90.5	85.0	86.4
This work $_{\text{m-BERT / monolingual}}$	⊘	86.8	90.3	85.8	90.7	85.3	86.9
This work $_{\text{m-BERT / cross-lingual}}$	⊘	87.1	90.8	86.5	91.0	85.6	87.3
This work $_{\text{XLM-R frozen / monolingual}}$	⊘	86.8	90.4	86.5	90.8	85.2	86.9
This work $_{\text{XLM-R / monolingual}}$	⊘	87.8	**91.6**	87.6	91.6	86.0	87.5
This work $_{\text{XLM-R / cross-lingual}}$	⊘	**88.0**	91.5	**88.0**	**91.8**	**86.3**	**87.7**

Table 1: F_1 scores on the in-domain evaluation CoNLL-2009 with gold pre-identified predicates. "CoNLL-2009 ST best" refers to the best results obtained (by different systems) during the Shared Task. We include all the systems that reported results in at least 4 languages. ⊙: syntax-aware system. ⊘: syntax-agnostic system.

CoNLL-2009 - OOD	CZ	DE	EN
CoNLL-2009 ST best	85.4	65.9	73.3
Zhao et al. (2009)	82.7	67.8	74.6
Marcheggiani et al. (2017)	87.2	—	77.7
Li et al. (2019)	—	—	81.5
Chen et al. (2019)	—	—	82.7
Lyu et al. (2019)	86.0	65.7	82.2
This work $_{\text{m-BERT / mono}}$	90.4	72.6	84.6
This work $_{\text{m-BERT / cross}}$	91.0	73.0	**85.0**
This work $_{\text{XLM-R / mono}}$	90.8	73.9	83.7
This work $_{\text{XML-R / cross}}$	**91.1**	**74.2**	84.3

Table 2: F_1 scores on the out-of-domain evaluation of CoNLL-2009 with gold pre-identified predicates.

jection techniques, and ii) it uses the English PropBank for all languages, which goes against our interest in capturing cross-lingual knowledge over heterogeneous inventories.

4.3 Results

Cross-lingual SRL. Table 1 compares the results obtained by our unified cross-lingual model against the state of the art in multilingual SRL, including both syntax-agnostic and syntax-aware architectures, on the in-domain test sets of CoNLL-2009 when using gold pre-identified predicates, rather than the predicates identified by the model itself, as standard in the CoNLL-2009 shared task. While proposing a state-of-the-art architecture is not the focus of this work, we believed it was important to build our cross-lingual approach starting from a strong and consistent baseline. For this reason, Table 1 includes the results obtained when training a separate instance of our model for each language, using the same strategy adopted by current multilingual systems (Cai and Lapata, 2019a; He et al., 2019; Lyu et al., 2019) and showing results that are competitive with He et al. (2019), *inter alia*. Remarkably, thanks to its universal encoders shared across languages and formalisms, our unified cross-lingual model outperforms our state-of-the-art baseline in all the 6 languages at a fraction of the cost in terms of number of trainable parameters (a single cross-lingual model against six monolingual models, each trained on a different language). Similar results can be seen in Table 2 where our cross-lingual approach improves over the state of the art on the out-of-domain evaluation of CoNLL-2009, especially in the German and English test sets which were purposely built to include predicates that do not appear in the training set. These results confirm empirically our initial hunch that semantic role labeling relations are deeply rooted beyond languages, independently of their surface realization and their predicate-argument structure inventories.

Finally, for completeness, Appendix E includes the results of our system on the individual subtasks, namely, predicate identification and predicate sense

CoNLL-2009 - In Domain	CA	CZ	DE	EN	ES	ZH
This work XLM-R / monolingual / 10% training	52.7	79.9	60.2	81.7	49.2	72.9
This work XLM-R / cross-lingual / 10% training	**78.2**	**84.0**	**69.9**	**84.3**	**76.1**	**78.6**
This work XLM-R / monolingual / 1-shot learning	44.5	21.8	40.9	67.4	46.5	72.1
This work XLM-R / cross-lingual / 1-shot learning	63.2	28.9	50.1	70.2	62.6	73.6
This work XLM-R / cross-lingual / 1-shot learning / 100% EN	**66.4**	**29.6**	**55.5**	91.6*	**64.3**	**76.7**

Table 3: F_1 scores on the in-domain evaluation CoNLL-2009 with gold pre-identified predicates for low-resource (top) and one-shot learning (bottom) scenarios. *: the result in EN on the last line is not directly comparable with those above as we use the full English training set.

disambiguation.

Low-resource cross-lingual SRL. We evaluate the robustness of our model in low-resource cross-lingual SRL by artificially reducing the training set of each language to 10% of its original size. Table 3 (top) reports the results obtained by our model when trained separately on the reduced training set of each language (monolingual), and the results obtained by the same model when trained on the union of the reduced training sets (cross-lingual). The improvements of our cross-lingual approach compared to the more traditional monolingual baseline are evident, especially in lower-resource scenarios, with absolute improvements in F_1 score of 25.5%, 9.7% and 26.9% on the Catalan, German and Spanish test sets, respectively. This is thanks to the ability of the model to use the knowledge from a language to improve its performance on other languages.

One-shot cross-lingual SRL. An interesting open question in SRL is whether a system can learn to model the semantic relations between a predicate sense s and its arguments, given a limited number of training samples in which s appears. In particular in our case, we are interested in understanding how the model fares in a synthetic scenario where each sense appears at most once in the training set, that is, we evaluate our model in a one-shot learning setting. As we can see from Table 3 (bottom), our cross-lingual approach outperforms its monolingual counterpart trained on each synthetic dataset separately by a wide margin, once again providing strong absolute improvements – 18.7% in Catalan, 9.2% in German and 16.1% in Spanish in terms of F_1 score – for languages where the number of training instances is smaller.

It is not uncommon for supervised cross-lingual tasks to feature different amounts of data for each language, depending on how difficult it is to get manual annotations for each language of interest. We simulate this setting in SRL by training our model on 100% of the training data available for the English language, while keeping the one-shot learning setting for all the other languages. As Table 3 (bottom) shows, non-English languages exhibit further improvements as the number of English training samples increases, lending further credibility to the idea that SRL can be learnt across languages even when using heterogeneous resources. Not only do these results suggest that a cross-lingual/cross-resource approach might mitigate the need for a large training set in each language, but also that reasonable cross-lingual results may be obtained by maintaining a single large dataset for a high-resource language, together with several small datasets for low-resource languages.

5 Analysis and Discussion

Cross-formalism SRL. In contrast to existing multilingual systems, a key benefit of our unified cross-lingual model is its ability to provide annotations for predicate senses and semantic roles in any linguistic formalism. As we can see from Figure 1 (left), given the English sentence "the cat threw its ball out of the window", our language-specific decoders produce predicate sense and semantic role labels not only according to the English PropBank inventory, but also for all the other resources, as it correctly identifies the agentive and patientive constituents independently of the formalism of interest. And this is not all, our model may potentially work on any of the 100 languages supported by the underlying language model (m-BERT or XLM-RoBERTa), e.g., in Italian, as shown in Figure 1 (right). This is vital for those languages for which a predicate-argument structure inventory has not yet been developed – an endeavor that may take

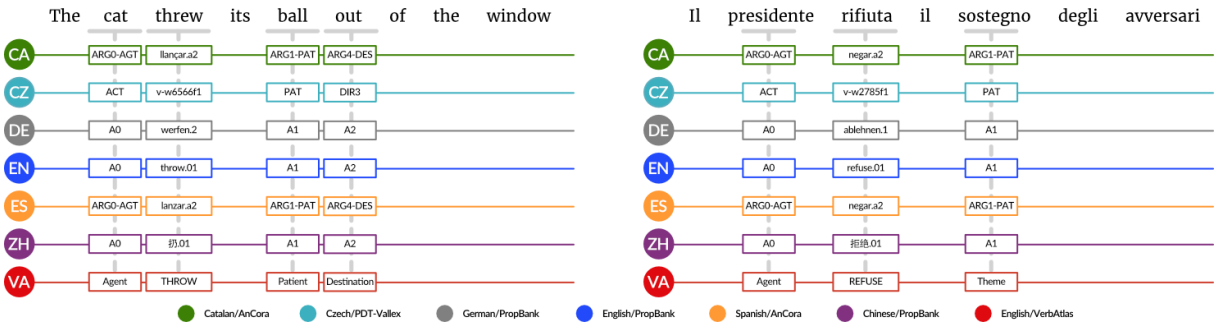

Figure 1: Thanks to its universal encoders, our unified cross-lingual model is able to provide predicate sense and semantic role labels according to several linguistic formalisms. Left: SRL labels for an English input sentence. Right: SRL labels for an Italian input sentence, which can be translated into English as "The president refuses the help of the opponents". Notice that Italian is not among the languages in the training set.

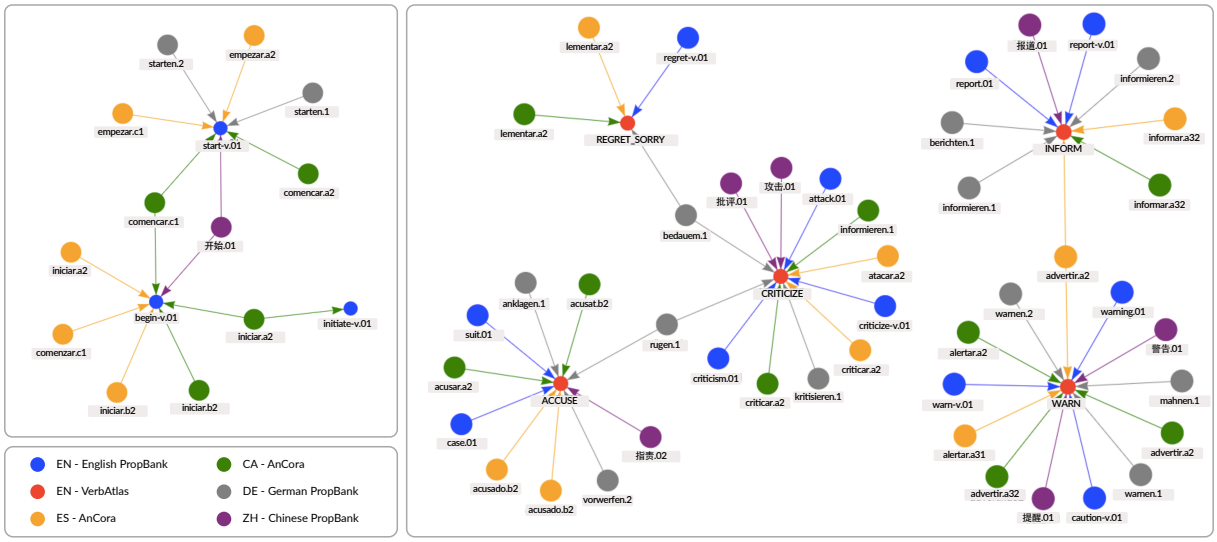

Figure 2: Visualization of the cross-resource mapping learnt by our model. Left: Mapping from Chinese PropBank, German PropBank and AnCora (both Catalan and Spanish) to English PropBank. Right: Mapping from English PropBank, German PropBank, Chinese PropBank and AnCora (both Spanish and Catalan) to VerbAtlas.

years to come to fruition – and, therefore, manually annotated data are unavailable. Thus, as long as a large amount of pretraining data is openly accessible, our system provides a robust cross-lingual tool to compare and analyze different linguistic theories and formalisms across a wide range of languages, on the one hand, and to overcome the issue of performing SRL on languages where no inventory is available, on the other.

Aligning heterogeneous resources. As briefly mentioned previously, the universal encoders in the model architecture force our system to learn cross-lingual features that are important across different formalisms. A crucial consequence of this approach is that the model learns to implicitly align the resources it is trained on, without the aid of word aligners and translation tools, even when these resources may be designed around specific languages and, therefore, present significant differences. In order to bring to light what our model implicitly learns to align in its shared cross-lingual space (see Sections 3.2 and 3.3), we exploit its language-specific decoders to build a mapping from any source inventory, e.g., AnCora, to a target inventory, e.g., the English PropBank. In particular, we use our cross-lingual model to label a training set originally tagged with a source inventory to produce silver annotations according to a target inventory, similarly to what is shown in Figure 1. While producing the silver annotations, we keep track of the number of times each predicate sense in the source inventory is associated by the model with a predicate sense of the target inventory. As a result, we produce a weighted directed graph in which the nodes are predicate senses and an edge

(a, b) with weight w indicates that our model maps the source predicate sense a to the target predicate sense b at least w times. A portion of this graph is displayed in Figure 2 where, for visualization purposes, we show the most frequent alignments for each language, i.e., the top-3 edges with largest weight from the nodes of each inventory to the nodes of the English PropBank (Figure 2, left) and to the nodes of VerbAtlas (Figure 2, right).[7]

For example, Figure 2 (left) shows that our model learns to map the Spanish AnCora sense *empezar.c1* and the German PropBank sense *starten.2* to the English PropBank sense *start.01*, but also that, depending on the context, the Chinese PropBank sense 开始*.01* can correspond to both *start.01* and *begin.01*. Figure 2 (right) also shows that our model learns to map senses from different languages and formalisms to the coarse-grained senses of VerbAtlas, even though the latter formalism is quite distant from the others as its frames are based on clustering WordNet synsets – sets of synonymous words – that share similar semantic behavior, rather than enumerating and defining all the possible senses of a lexeme as in the English and Chinese PropBanks. To the best of our knowledge, our unified model is the first transfer-based tool to automatically align diverse linguistic resources across languages without relying on human supervision.

6 Conclusion and Future Work

On one hand, recent research in multilingual SRL has focused mainly on proposing novel model architectures that achieve state-of-the-art results, but require a model instance to be trained on and for each language of interest. On the other hand, the latest developments in cross-lingual SRL have revolved around using the English PropBank inventory as a universal resource for other languages through annotation transfer techniques. Following our hunch that semantic relations may be deeply rooted beyond the surface realizations that distinguish one language from another, we propose a new approach to cross-lingual SRL and present a model which learns from heterogeneous linguistic resources in order to obtain a deeper understanding of sentence-level semantics. To achieve this objective, we equip our model architecture with "universal" encoders which share their weights across languages and are, therefore, forced to learn knowledge that spans across varying formalisms.

Our unified cross-lingual model, evaluated on the gold multilingual benchmark of CoNLL-2009, outperforms previous state-of-the-art multilingual systems over 6 diverse languages, ranging from Catalan to Czech, from German to Chinese, and, at the same time, also considerably reduces the amount of trainable parameters required to support different linguistic formalisms. And this is not all. We find that our approach is robust to low-resource scenarios where the model is able to exploit the complementary knowledge contained in the training set of different languages.

Most importantly, our model is able to provide predicate sense and semantic role labels according to 7 predicate-argument structure inventories in a single forward pass, facilitating comparisons between different linguistic formalisms and investigations about interlingual phenomena. Our analysis shows that, thanks to the prior knowledge encoded in recent pretrained language models and our focus on learning from cross-lingual features, our model can be used on languages that were never seen at training time, opening the door to alignment-free cross-lingual SRL on languages where a predicate-argument structure inventory is not yet available. Finally, we show that our model implicitly learns to align heterogeneous resources, providing useful insights into inter-resource relations. We leave an in-depth qualitative and quantitative analysis of the learnt inter-resource mappings for future work.

We hope that our work can set a stepping stone for future developments towards the unification of heterogeneous SRL. We release the code to reproduce our experiments and the checkpoints of our best models at https://github.com/SapienzaNLP/unify-srl.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University.

[7] We release the full alignment and the corresponding graph at https://github.com/SapienzaNLP/unify-srl.

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual Semantic Role Labeling. In *Proceedings of ACL*.

Alan Akbik, Vishwajeet Kumar, and Yunyao Li. 2016. Towards semi-automatic generation of proposition banks for low-resource languages. In *Proceedings of EMNLP*.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2019. Cross-lingual transfer of semantic roles: From raw text to semantic roles. In *Proceedings of IWCS*.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual Semantic Role Labeling. In *Proceedings of CoNLL*.

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings COLING*.

Rui Cai and Mirella Lapata. 2019a. Semi-supervised Semantic Role Labeling with cross-view training. In *Proceedings of EMNLP*.

Rui Cai and Mirella Lapata. 2019b. Syntax-aware Semantic Role Labeling without parsing. *Transactions of ACL (TACL)*, 7:343–356.

Rui Cai and Mirella Lapata. 2020. Alignment-free cross-lingual Semantic Role Labeling. In *Proceedings of EMNLP*.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Xinchi Chen, Chunchuan Lyu, and Ivan Titov. 2019. Capturing argument interaction in Semantic Role Labeling with capsule networks. In *Proceedings of EMNLP*.

Simone Conia, Fabrizio Brignone, Davide Zanfardino, and Roberto Navigli. 2020. InVeRo: Making Semantic Role Labeling accessible with intelligible verbs and roles. In *Proceedings of EMNLP*.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of COLING*.

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

Angel Daza and Anette Frank. 2019. Translate and label! an encoder-decoder approach for cross-lingual semantic role labeling. In *Proceedings of EMNLP*, pages 603–615.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual Semantic Role Labeling dataset. In *Proceedings of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of NAACL*.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to Semantic Role Labeling. In *Proceedings of EMNLP*.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of ICLR*.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual Semantic Role Labeling with high-quality translated training corpus. In *Proceedings of ACL*.

Charles J. Fillmore. 1968. The case for case. *Universals in Linguistic Theory*.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of ACL*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*.

Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Stepánek, Pavel Stranák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*.

Jan Hajic, J. Panevová, Zdenka Uresová, Alevtina Bémová, V. Kolárová, and P. Pajas. 2003. PDT-Vallex: Creating a large-coverage valency lexicon for treebank annotation.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural Semantic Role Labeling. In *Proceedings of ACL*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep Semantic Role Labeling: What works and what's next. In *Proceedings of ACL*.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual Semantic Role Labeling. In *Proceedings of EMNLP*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8).

Jungo Kasai, Dan Friedman, Robert Frank, Dragomir R. Radev, and Owen Rambow. 2019. Syntax-aware neural Semantic Role Labeling with supertags. In *Proceedings of NAACL*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedigns of ICLR*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of EMNLP*.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform semantic role labeling. In *Proceedings of AAAI*.

Chunchuan Lyu, Shay B. Cohen, and Ivan Titov. 2019. Semantic Role Labeling with iterative structure refinement. In *Proceedings of EMNLP*.

Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. A simple and accurate syntax-agnostic neural model for dependency-based Semantic Role Labeling. In *Proceedings of CoNLL*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for Semantic Role Labeling. In *Proceedings of EMNLP*.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic Role Labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *Proceedings of the Workshop Frontiers in Corpus Annotation*.

Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith. 2018. Polyglot Semantic Role Labeling. In *Proceedings of ACL*.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (present and future) use. In *Proceedings of IJCAI*.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *J. Artif. Intell. Res.*, 36:307–340.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2018. Searching for activation functions. In *Proceedings of ICLR*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for Semantic Role Labeling. In *Proceedings of EMNLP*.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of LREC*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of EMNLP*.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the chinese treebank. *Nat. Lang. Eng.*, 15(1):143–172.

Hai Zhao, Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Multilingual dependency learning: Exploiting rich features for tagging syntactic and semantic dependencies. In *Proceedings of CoNLL*.

A Model Hyperparameters

Table 4 reports the hyperparameter values we choose for our model configuration and experiments.

Hyperparameter		Value
d_w	Size of $\mathbf{e}_i$	512
K'	Universal sentence encoder layers	3
d_t	Size of $\mathbf{t}_i$	512
K''	Universal pred.-arg. encoder layers	1
d_z	Size of $\mathbf{a_i}$	256
d_{s_p}	Size of $\mathbf{p}_i$	32
d_{s_s}	Size of $\mathbf{s}_i$	512
d_{s_a}	Size of $\mathbf{r}_i$	512
	Batch size	32
	Batch size when fine-tuning	128
	Max learning rate	10^{-3}
	Min learning rate	10^{-5}
	Max *lr* for LM fine-tuning	10^{-5}
	Min *lr* for LM fine-tuning	10^{-6}
	Warmup epochs	1
	Cooldown epochs	15
	Training epochs	30

Table 4: Hyperparameter values for our model architecture. We use the same hyperparameter values for our monolingual and cross-lingual experiments.

B Data Statistics

Tables 5, 6 and 7 provide an overview of the training sets provided as part of the CoNLL-2009 shared task, with statistics about sentences, predicates and arguments.

C Hardware Infrastructure

All the experiments were performed on a x86-64 architecture with 64GB of RAM, an 8-core CPU running at 3.60GHz, and a single Nvidia RTX 2080Ti with 11GB of VRAM.

D Training Details

Training was performed using half-precision via Apex.[8] Training times varied considerably depending on the experiment setting: the shorter experiment lasted 26 minutes (training m-BERT on 10% of the Catalan training set), whereas the longest experiment lasted for 46 hours (training XLM-RoBERTa on the union of all the datasets of all the languages).

E Other Results

Predicate identification. In Table 8 we report the results of our model on predicate identification.

Predicate sense disambiguation. In Table 9 we report the results of our model on predicate sense disambiguation.

F Alignment Examples

Figure 3 provides two more examples, one in French (left), the other in Catalan (right). We remark that the training set of CoNLL-2009 does not include sentences in French, however, our cross-lingual model correctly outputs SRL tags according to the other seven language-specific decoders.

[8] https://github.com/NVIDIA/apex

	Sentences			Predicates		Arguments	
	Total$_s$	Annotated	Avg. Len.	Total$_p$	Senses	Total$_a$	Roles
CoNLL-2009							
CA	13,200	12,873	30.2	37,431	3,554	84,367	38
CZ	38,727	38,578	16.9	414,237	9,135	365,255	60
DE	36,020	14,282	22.2	17,400	1,271	34,276	10
EN	39,279	37,847	25.0	179,014	8,237	393,699	52
ES	14,329	13,835	30.7	43,824	4,534	99,054	43
ZH	22,277	21,071	28.5	102,813	12,587	231,869	36

Table 5: Overview of the CoNLL-2009 training sets. For each dataset we report the number of sentences (*Total$_s$*), the number of sentences with at least an annotated predicate (*Annotated*), the average number of tokens per sentence (*Avg. Len.*), the number of predicates (*Total$_p$*) and predicate senses (*Senses*), and also the number of arguments (*Total$_a$*) and argument roles (*Roles*).

	Sentences			Predicates		Arguments	
	Total$_s$	Annotated	Avg. Len.	Total$_p$	Senses	Total$_a$	Roles
CoNLL-2009							
CA	1,724	1,675	31.5	5,105	1,436	11,529	34
CZ	5,228	5,210	16.9	55,517	3,467	49,071	54
DE	2,000	532	19.7	588	255	1,169	9
EN	1,334	1,283	25.7	6,390	1,990	13,865	32
ES	1,655	1,588	31.4	5,076	1,565	11,600	36
ZH	1,762	1,663	29.5	8,103	2,535	18,554	24

Table 6: Overview of the CoNLL-2009 development datasets. For each dataset we report the number of sentences (*Total$_s$*), the number of sentences with at least an annotated predicate (*Annotated*), the average number of tokens per sentence (*Avg. Len.*), the number of predicates (*Total$_p$*) and predicate senses (*Senses*), and also the number of arguments (*Total$_a$*) and argument roles (*Roles*).

	Sentences			Predicates		Arguments	
	Total$_s$	Annotated	Avg. Len.	Total$_p$	Senses	Total$_a$	Roles
CoNLL-2009							
CA	1,862	1,802	29.4	5,001	1,425	11,275	32
CZ	4,213	4,196	16.8	44,585	3,018	39,223	55
DE	2,000	506	20.1	550	238	1,073	8
EN	2,000	1,913	25.0	8,987	2,254	19,949	35
ES	1,725	1,663	30.2	5,175	1,623	11,824	33
ZH	2,556	2,400	30.1	12,282	3,458	27,712	26

Table 7: Overview of the CoNLL-2009 testing datasets. For each dataset we report the number of sentences (*Total$_s$*), the number of sentences with at least an annotated predicate (*Annotated*), the average number of tokens per sentence (*Avg. Len.*), the number of predicates (*Total$_p$*) and predicate senses (*Senses*), and also the number of arguments (*Total$_a$*) and argument roles (*Roles*).

CoNLL-2009 - Predicate identification	CA	CZ	DE	EN	ES	ZH
This work m-BERT frozen / monolingual	97.9	98.6	90.5	93.8	97.8	94.3
This work m-BERT / monolingual	98.3	98.9	91.4	94.3	98.4	95.0
This work m-BERT / cross-lingual	98.3	99.0	91.6	94.4	98.4	95.1
This work XLM-R frozen / monolingual	97.9	98.9	90.5	93.9	98.0	94.7
This work XLM-R / monolingual	98.3	99.2	91.5	94.3	98.4	95.2
This work XLM-R / cross-lingual	**98.5**	**99.3**	**91.9**	**94.6**	**98.6**	**95.4**

Table 8: F_1 scores on the predicate identification subtask which is not part of the CoNLL-2009 shared task setting.

CoNLL-2009 - Predicate disambiguation	CA	CZ	DE	EN	ES	ZH
This work m-BERT frozen / monolingual	90.0	93.2	86.9	96.8	87.3	94.9
This work m-BERT / monolingual	90.3	93.5	87.3	**97.2**	87.5	95.0
This work m-BERT / cross-lingual	90.3	93.5	87.3	**97.2**	87.6	95.3
This work XLM-R frozen / monolingual	90.1	93.6	86.8	96.8	87.4	95.2
This work XLM-R / monolingual	90.4	93.7	87.3	97.1	87.6	95.6
This work XLM-R / cross-lingual	**90.5**	**93.9**	**87.5**	**97.2**	**87.8**	**95.8**

Table 9: Accuracy on the predicate sense disambiguation subtask computed by the official CoNLL-2009 scorer which, by default, takes into account only the sense numbers, e.g., *01* of *eat.01*.

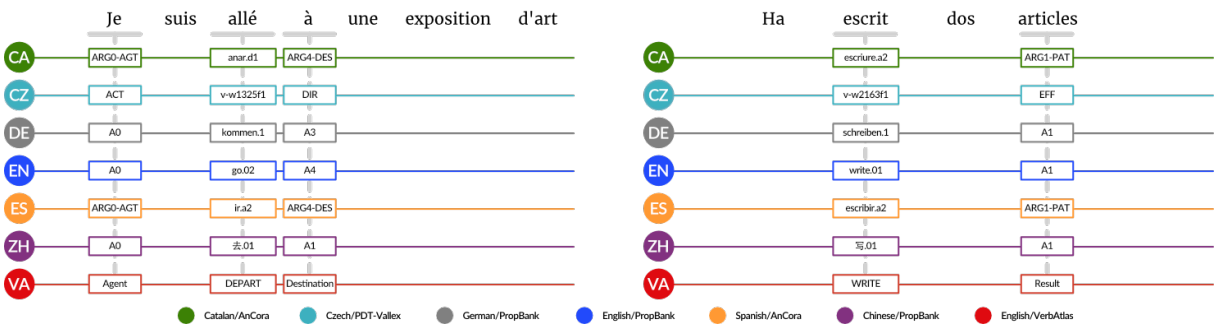

Figure 3: Output of our cross-lingual system for a French (left) and a Catalan (right) sentence.