# Bridging the Gap in Multilingual Semantic Role Labeling: a Language-Agnostic Approach

**Simone Conia** and **Roberto Navigli**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{conia,navigli}@di.uniroma1.it

## Abstract

Recent research indicates that taking advantage of complex syntactic features leads to favorable results in Semantic Role Labeling. Nonetheless, an analysis of the latest state-of-the-art multilingual systems reveals the difficulty of bridging the wide gap in performance between high-resource (e.g., English) and low-resource (e.g., German) settings. To overcome this issue, we propose a fully language-agnostic model that does away with morphological and syntactic features to achieve robustness across languages. Our approach outperforms the state of the art in all the languages of the CoNLL-2009 benchmark dataset, especially whenever a scarce amount of training data is available. Our objective is not to reject approaches that rely on syntax, rather to set a strong and consistent language-independent baseline for future innovations in Semantic Role Labeling. We release our model code and checkpoints at https://github.com/SapienzaNLP/multi-srl.

## 1 Introduction

Semantic Role Labeling (SRL) – the task of automatically addressing "Who did What to Whom, How, When and Where?" (Gildea and Jurafsky, 2000; Màrquez et al., 2008) – is a long standing open problem in Natural Language Processing (NLP), and a central task required to complete the puzzle of Natural Lan-guage Understanding (Navigli, 2018). Its roots date back to several decades ago, to when Fillmore (1968) first theorized the existence of deep semantic relations between a predicate and other sentential constituents. Over the years, different linguistic formalisms and their corresponding predicate-argument structure inventories expanded Fillmore's seminal intuition (Dowty, 1991; Levin, 1993), yet the need to rely on manually designed complex feature templates severely limited early SRL models (Zhao et al., 2009). Fortunately, the recent great success of neural networks in NLP has drawn attention back to SRL and has led to considerable performance gains. This has been particularly the case for recurrent neural networks, thanks to their ability to better capture relations over sequences (Marcheggiani et al., 2017; He et al., 2017). The positive results obtained in SRL were rapidly extended to other fields where they proved to be beneficial to several downstream tasks, from Machine Translation (Marcheggiani et al., 2018) to Information Extraction (Christensen et al., 2011), Opinion Role Labeling (Zhang et al., 2019a), and Question Answering (He et al., 2015).

As researchers constantly explored new approaches to improve SRL, the exploitation of syntactic features soon emerged as a natural choice. Cai and Lapata (2019b) suggested that syntax ought to help semantic role labelers since i) a significant portion of the predicate-argument relations in a semantic dependency graph mirrors the edges that appear in a syntactic dependency graph, and ii) there is often a deterministic mapping from syntactic to semantic roles. Following this line of thought, numerous papers from major venues reported improvements in SRL by explicitly taking advantage of different properties in syntactic dependency trees to various extents (Wang et al., 2019; Zhang et al., 2019b; He et al., 2019).

However, if we step back to observe the larger picture, a significant gap among languages strikes the eye. Indeed, among the state-of-the-art multilingual SRL systems that reported their results on all the languages of the CoNLL-2009 benchmark dataset (Hajic et al., 2009), the currently best-performing system

*Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410
Barcelona, Spain (Online), December 8-13, 2020

(He et al., 2019) manifests a very significant discrepancy in performance between high- and low-resource languages (around 10% in $F_1$ score); the works of Chen et al. (2019) and Lyu et al. (2019), *inter alia*, also show the same behavior, with gaps in $F_1$ scores that fluctuate around 15% and 14%, respectively. These large differences from language to language suggest that recently proposed innovations do not seem to generalize consistently across different languages, especially when syntax plays a central role or the novelty is tested on out-of-domain data. Anyhow, and perhaps most importantly, these approaches do not address the ever-present performance gap between high-resource and low-resource languages, such as English and German, respectively, where the disparity in $F_1$ score still hovers around 10% in the in-domain evaluations of CoNLL-2009, and is even wider in the out-of-domain evaluations.

We argue that correctly harnessing contextual intra-sentence information can lead to large gains in performance on single languages and a more robust generalization capability when scaling across languages, especially when these belong to different linguistic families (e.g., English and Chinese). More precisely, the main contributions of this paper are as follows:

- We propose the first language-agnostic SRL model that achieves state-of-the-art performance across 6 languages[1] without the use of any morphological, part-of-speech or syntactic information.
- In the wake of the recent interest in cross-linguality, we report promising results in zero-shot cross-lingual SRL.
- We conduct an analysis to provide an empirical demonstration of the robustness of our approach in low-resource settings.
- We release our code and model checkpoints to allow easy reproduction of our experiments and facilitate the integration of future innovations on top of our model.

We stress that our objective is not to reject syntax in SRL. On the contrary, we strongly believe that clever integration of syntactic features is a promising avenue to advancing research. Here, however, our aim is to provide a strong language-agnostic baseline that is robust and consistent across languages, especially low-resource ones. We hope our effort can become a stepping stone for future developments of both syntax-agnostic and syntax-focused SRL.

## 2 Related Work

**Dependency or Span?** Currently, SRL is cast as either a span-based or a dependency-based labeling task. Given a predicate in a sentence, the main difference between the two settings is that, in the former, semantic role labels are assigned to the entire span of an argument, whereas, in the latter, semantic role labels are assigned only to the semantic head of the argument. Both span- and dependency-based SRL have continued to be developed and supported in parallel over the years with the organization of the CoNLL-2005 (Carreras and Màrquez, 2005) and CoNLL-2012 (Pradhan et al., 2012) tasks for span-based SRL, and the CoNLL-2008 (Surdeanu et al., 2008) and CoNLL-2009 tasks for dependency-based SRL. The debate over which representation is best is still open and subject to active investigation, with ongoing efforts aimed at merging the two into a unified formalism (Li et al., 2019). In this work, we focus mainly on dependency-based SRL as CoNLL-2009 includes the widest and most varied set of languages, but we also report results on the span-based CoNLL-2012 English benchmark.

**End-to-end approaches.** Regardless of the formalism of choice, SRL is traditionally divided into a set of simpler subtasks: predicate identification, predicate sense disambiguation, argument identification and argument classification. Early work used different sets of template features and statistical/neural models to tackle each predicate and argument subtask separately (Zhao et al., 2009), but the recent success of the multi-task learning paradigm (Caruana, 1997) prompted the development of end-to-end models that jointly address some of the subtasks (Cai et al., 2018; Li et al., 2019; He et al., 2019). Since the CoNLL-2009 shared task provides pre-identified predicates, systems – end-to-end approaches included – usually

[1] CoNLL-2009 originally included 7 languages (Catalan, Chinese, Czech, English, German, Japanese, Spanish), however we do not include Japanese in our studies as it is no longer available through LDC due to licensing problems.

process the same sentence $n_p$ times, where $n_p$ is the number of predicates in the sentence; at the cost of longer training times, this approach lets a system contextualize a sentence with respect to a specific pre-identified predicate. In our work we take another approach: our system contextualizes a sentence to all the predicates it contains in a single forward pass. This results in shorter inference time and a significant reduction in training time, because our model converges in less than 30 epochs compared to the 300 epochs required by the current state-of-the-art multilingual system of He et al. (2019).

**Syntax-agnostic SRL.** Marcheggiani et al. (2017) opened the door to the initial wave of syntax-agnostic models for SRL by efficiently employing a BiLSTM-based encoder to capture longer predicate-argument relations within an input sequence, thereby outperforming syntax-aware systems in the CoNLL-2009 English, Czech and Spanish evaluation sets for the first time. Cai et al. (2018) proposed the first full end-to-end syntax-agnostic SRL model to jointly learn to disambiguate predicate senses and recognize their corresponding semantic arguments, and further enhanced the model with an attentive biaffine scorer (Dozat and Manning, 2018) to better condition argument predictions on a given predicate in the input sentence. The combined contribution of these innovations realigned the performances of syntax-agnostic systems to the best syntax-aware systems. Most recently, Li et al. (2019) showed that the use of contextualized word embeddings such as ELMo (Peters et al., 2018) leads to further progress, thus lending support to the hunch that high-quality contextual information is key to enabling high-performing SRL systems. We stress that there is a subtle catch in the definition of syntax-agnostic: the foregoing approaches do not make use of sentence-level syntax, but do still consider lexical-level syntactic features, such as part-of-speech tags. As a result, their input is still language-dependent. In contrast, our approach does away with any lexical- or sentence-level syntactic feature and is therefore truly syntax-agnostic.

**Syntax-aware SRL.** Syntactic features have recently gained traction in SRL research, mostly due to the diversity of the available representations, the quality of the information they encode, and the wide range of techniques that can be used to take advantage of them. Notable work includes the use of graph convolutional networks to capture short-distance relations between neighbors in the syntactic dependency graph (Marcheggiani and Titov, 2017a), deriving argument pruning rules from syntactic dependency trees (He et al., 2018b), clustering dependency relations (Kasai et al., 2019), and syntax-based attention mechanisms (Strubell et al., 2018; Zhang et al., 2019b). However, most of the work that reported performance improvements in the CoNLL-2009 benchmark dataset only did so in a subset of its in- and out-of-domain evaluations of the 6 available languages, with the noteworthy exception of Lyu et al. (2019). This may fuel the suspicion that either there is a lack of real interest in syntax-aware multilingual SRL, or syntax-focused innovations do not scale immediately across languages. In any case, reporting fragmented results strongly limits full-fledged comparisons and, therefore, hinders progress in multilingual SRL.

## 3 Model Description

Building on top of recent successes in deep learning, our model learns to tackle predicate sense disambiguation and argument identification/classification jointly. In particular:

- our model features a word encoder which, in contrast to current work in SRL, exploits the internal states of a language model to compute the contextualized representation of a word (Section 3.1);
- while previous work uses a single sequence encoder to capture predicate-argument relations, we adopt a two-stage sequence encoding strategy:
    - a predicate-aware word encoder which re-contextualizes the word representations with respect to all the predicates in a sentence in a single forward pass (Section 3.2);
    - a predicate-argument encoder which specializes the representation of each argument to a single predicate, for each predicate in the input sentence (Section 3.3).

### 3.1 Contextualized Word Representation

The prior knowledge encoded in pretrained language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2018) and XLM-RoBERTa (Conneau et al., 2020) is currently showing significant benefits in an ever-increasing array of NLP tasks. Taking inspiration from recent work (Bevilacqua and Navigli, 2020), our model computes word-level representations by combining the different knowledge encoded at different hidden layers of a pretrained language. More formally, let $L : V^{|\mathbf{t}|} \rightarrow \mathbb{R}^{n \times h}$ be a language model with an input vocabulary $V$ that, given a sequence $\mathbf{t} = \langle t_1, t_2, \ldots, t_{|\mathbf{t}|} \rangle$ of tokens $t_i \in V$, returns a sequence $\mathbf{o} = \langle \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{|\mathbf{t}|} \rangle$ of dense vector representations $\mathbf{o}_i \in \mathbb{R}^h$. Also, let $L$ be structured in $K$ inner layers $l^1, l^2, \ldots, l^K$ such that $L(\mathbf{t}) = l^K(\cdots(l^1(\mathbf{t})))$, and that, for the $k$-th layer $l^k$, the corresponding output states $\mathbf{o}^k = l^k(\mathbf{o}^{k-1})$ are accessible. Then, given an input sentence $\mathbf{w} = \langle w_1, w_2, \ldots, w_n \rangle$ where each word $w_i$ can be tokenized into $m_i$ subwords belonging to the language model vocabulary, i.e., $w_i = \langle t_{i1}, t_{i2}, \ldots, t_{im_i} : t_{ij} \in V \rangle$, we define our input sequence $\mathbf{t}$ as:

$$\mathbf{t} = \big\langle \langle t_{\text{START}} \rangle,\ \langle t_{11}, t_{12}, \ldots, t_{1m_1} \rangle_1,\ \langle t_{21}, t_{22}, \ldots, t_{2m_2} \rangle_2, \ldots,\ \langle t_{n1}, t_{n2}, \ldots, t_{nm_n} \rangle_n, \langle t_{\text{END}} \rangle \big\rangle$$

where $t_{\text{START}}$ and $t_{\text{END}}$ are special tokens that indicate the beginning and the end of a sentence, respectively. We compute the contextualized word representation $\mathbf{e}_i$ of $w_i$ as the average of the Swish activation (Ramachandran et al., 2018) of the concatenated hidden states extracted from the upper $K'$ layers of $L$ for each subword $t_{ij} \in w_i$. More formally:

$$\begin{aligned}
\mathbf{c}_{ij} &= \mathbf{o}_{ij}^{K} \oplus \mathbf{o}_{ij}^{K-1} \oplus \cdots \oplus \mathbf{o}_{ij}^{K-(K'-1)} \\
\mathbf{c}'_{ij} &= \text{Swish}(\mathbf{W}^c \mathbf{c}_{ij} + \mathbf{b}^c) \\
\mathbf{e}_i &= \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbf{c}'_{ij}
\end{aligned}$$

where $\mathbf{o}_{ij}^k$ is the output state for the subword $t_{ij}$ from the $k$-th layer of $L$, $\mathbf{c}_{ij}$ is the concatenation of the output states for the subword $t_{ij}$ from the upper $K'$ layers of $L$, $\mathbf{W}^c \in \mathbb{R}^{d_w \times K' \cdot h}$ and $\mathbf{b}^c \in \mathbb{R}^{d_w}$.

While previous approaches often enriched their representations with lexical-level morphological or syntactic information, e.g., lemma and POS embeddings (Marcheggiani et al., 2017; Li et al., 2019; He et al., 2019), or with sentence-level syntactic information, e.g., syntactic dependency embeddings, GCNs over syntactic trees, and syntactic pruning rules (Marcheggiani and Titov, 2017b; He et al., 2019; Cai and Lapata, 2019b), we emphasize that our word encoder relies only on raw text as an input, and is therefore independent of any linguistic feature. However, while this word encoder provides a contextualized representation $\mathbf{e}_i$ of each word $w_i$ with respect to a sentence $\mathbf{w}$, the resulting representations are still unaware of the predicates appearing in $\mathbf{w}$.

### 3.2 Predicate-Aware Word Representation

As mentioned in Section 2, predicate-specific information is usually provided at the input level by means of a predicate indicator flag (Marcheggiani et al., 2017; Marcheggiani and Titov, 2017b; He et al., 2019), a predicate embedding (Cai et al., 2018), or by prepending or appending the predicate word to the input sentence (Shi and Lin, 2019). Instead, we adopt an opposite approach in which predicate-specific information is explicitly expressed only at the output level: for each sentence, the model is tasked, in a multi-objective fashion, with i) learning whether a word is a predicate, and ii) disambiguating the sense of each predicate. With this approach, predicate-specific information can back-propagate into a tailor-made sequence encoder to specialize the contextual word representations with respect to all the predicates in the input sentence simultaneously.

In particular, our model features a "fully-connected" stacked-BiLSTM sequence encoder where each BiLSTM layer $l^k$ is purposely tweaked to be directly connected not only to the layer $l^{k-1}$ immediately below $l^k$, but also to each underlying BiLSTM layer $l^{k-i}$ with $i \in \{1, 2, \ldots, k\}$. More formally, the contextualized word embeddings $\mathbf{E} = \langle \mathbf{e}_1, \ldots, \mathbf{e}_n \rangle$ are re-contextualized into predicate-aware word embeddings $\mathbf{H} = \langle \mathbf{h}_1, \ldots, \mathbf{h}_n \rangle$ as follows:

$$\begin{aligned}
\mathbf{H}^0 &= \mathbf{E} \\
\mathbf{H}^k &= \mathbf{H}^{k-1} \oplus \textsc{BiLSTM}(\textsc{LayerNorm}(\mathbf{H}^{k-1})) \\
\mathbf{H} &= \mathbf{H}^{K''} \\
\mathbf{s}_i^p &= \mathbf{W}^{s^p} \cdot \text{Swish}(\mathbf{W}^{h^p}\mathbf{h}_i + \mathbf{b}^{h^p}) + \mathbf{b}^{s^p} \\
\mathbf{s}_i^s &= \mathbf{W}^{s^s} \cdot \text{Swish}(\mathbf{W}^{h^s}\mathbf{h}_i + \mathbf{b}^{h^s}) + \mathbf{b}^{s^s}
\end{aligned}$$

where LAYERNORM stands for layer normalization (Ba et al., 2016) and $K''$ is the total number of layers in the sequence encoder, whereas $\mathbf{s}_i^p$ is the unnormalized log probability that word $w_i$ is a predicate and $\mathbf{s}_i^s$ is the unnormalized log probability over the predicate sense vocabulary. We highlight that, thanks to the direct input-output connections in our fully-connected BiLSTM encoder, predicate-specific information coming from $\mathbf{s}_i^p$ and $\mathbf{s}_i^s$ is immediately back-propagated into the $i$-th hidden state $\mathbf{h}_i^k$ of the $k$-th BiLSTM layer $l^k$.

### 3.3 Predicate-Argument Representation

Even if the above predicate-aware word representations encode valuable information about the position and the sense/meaning of the predicates in the input sentence, the same word may play a different semantic role for each predicate. While previous systems for dependency-based SRL relied on a single sequence encoder to capture predicate-argument relations, our model instead features a second sequence encoder dedicated to capturing predicate-argument relations. To obtain a predicate-argument representation, i.e., a specialization of a word representation with respect to a single predicate, our predicate-argument encoder first projects each predicate-aware word representation $\mathbf{h}_i$ obtained in the previous step for a word $w_i$ to two distinct vector spaces, a predicate-specific representation $\mathbf{p}_i$ and an argument-specific representation $\mathbf{a}_i$:

$$\begin{aligned}
\mathbf{p}_i &= \text{Swish}(\mathbf{W}^p\mathbf{h}_i + \mathbf{b}^p) \\
\mathbf{a}_i &= \text{Swish}(\mathbf{W}^a\mathbf{h}_i + \mathbf{b}^a)
\end{aligned}$$

Then, for each predicate $w_p$ with $1 \leq p \leq n$, we obtain a $w_p$-specific representation of the sentence by concatenating the predicate-specific representation $\mathbf{p}_p$ to each argument-specific representation $\mathbf{a}_i$ of every word $w_i$ in the input sentence:

$$(\mathbf{P} \oplus \mathbf{A})_p = \langle \mathbf{p}_p \oplus \mathbf{a}_1,\ \mathbf{p}_p \oplus \mathbf{a}_2,\ \ldots,\ \mathbf{p}_p \oplus \mathbf{a}_n \rangle$$

Finally, these representations are further refined by an argument-specific fully-connected BiLSTM encoder in order to capture the role each word $w_i$ plays with respect to the predicate $w_p$:

$$\begin{aligned}
\mathbf{Z}_p^0 &= \textsc{LayerNorm}((\mathbf{P} \oplus \mathbf{A})_p) \\
\mathbf{Z}_p^k &= \mathbf{Z}^{k-1} \oplus \textsc{LayerNorm}(\textsc{BiLSTM}(\mathbf{Z}^{k-1})) \\
\mathbf{Z}_p &= \mathbf{Z}^{K'''} \\
\mathbf{s}_{i|p}^r &= \mathbf{W}^r \cdot \text{Swish}(\mathbf{W}^z\mathbf{z}_{i|p} + \mathbf{b}^z) + \mathbf{b}^r
\end{aligned}$$

where $K'''$ is the number of layers in the argument-specific fully-connected BiLSTM encoder and $\mathbf{s}_{i|p}^r$ is the score distribution over the role vocabulary for a word $w_i$ with respect to a predicate $w_p$.

### 3.4 Multitask Training Objective

The model is trained to jointly minimize the sum of the categorical cross-entropy losses on predicate identification $\text{L}(\mathbf{s}^p)$ (see Section 3.2), predicate sense disambiguation $\text{L}(\mathbf{s}^s)$ (see Section 3.2), and argument identification/classification $\text{L}(\mathbf{s}^r)$ (see Section 3.3). The cumulative loss is the average of the individual subtask losses weighted on the number of training instances for each subtask:

$$\text{L} = \frac{N^p\text{L}(\mathbf{s}^p) + N^p\text{L}(\mathbf{s}^s) + N^r\text{L}(\mathbf{s}^r)}{N^p + N^p + N^r}$$

where $N^p$ and $N^r$ are the number of instances of predicates and arguments/roles in the training set, respectively.

## 4 Experiments

We evaluate our model in multilingual dependency-based SRL (CoNLL-2009) and English span-based SRL (CoNLL-2012). Hereafter, we provide a brief overview of the CoNLL-2009 and CoNLL-2012 benchmark datasets (Section 4.1), we describe our experimental setup (Section 4.2), and report results for each dataset and language (Section 4.3).

### 4.1 The CoNLL-2009 and CoNLL-2012 Benchmark Datasets

The CoNLL-2009 dataset is the most comprehensive multilingual SRL benchmark to date, featuring 6 languages from dissimilar linguistic families with 6 in-domain and 3 out-of-domain test sets. Predicates are already identified and, therefore, systems are evaluated on predicate sense disambiguation, argument identification and argument classification.

Performing multilingual SRL in CoNLL-2009 is not trivial since the annotation process and methodology vary considerably from language to language: German only contemplates verbal predicates, English also considers nominal predicates, Czech includes adjectival and adverbial predicates as well. Moreover, English uses separate predicate-argument structure inventories for verbal and nominal predicates, i.e., PropBank and NomBank, while Chinese uses a unified inventory. Finally, training sets vary dramatically in size, from more than 400,000 predicate instances in Czech to less than 20,000 in German. CoNLL-2009 is therefore the ideal benchmark for the ability of SRL models to scale multilingually.

While our focus is on multilingual SRL, we also evaluate our approach on CoNLL-2012, which is, to the best of our knowledge, the largest dataset with gold annotations for span-based SRL, and show that our model can also achieve state-of-the-art results on span-based English SRL.[2]

### 4.2 Experimental Setup

We implemented the model in PyTorch[3] and PyTorch Lightning[4], using the Transformers library[5] (Wolf et al., 2019) for language modeling. We selected the hyperparameter values according to the best $F_1$ score on the English development split; for any other language, the model was trained with the same hyperparameter configuration used for the English language. We trained each model configuration for at most 30 epochs using Adam (Kingma and Ba, 2015) with an initial linear learning rate warmup followed by a linear learning rate cooldown. Table 1b reports the hyperparameters for the best configuration. All the results reported in the following sections refer to the output of the official scoring scripts of CoNLL-2009[6] and CoNLL-2005[7], for dependency- and span-based SRL, respectively.

**Language Models.** We compare our model to current state-of-the-art multilingual SRL systems which, however, use contextualized word representations from different pretrained language models. For a fair comparison, we report the performance of our model when the underlying input representation is ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), and multilingual BERT (m-BERT hereafter). We also compare the SRL performance when the weights of the underlying language model are kept frozen or fine-tuned during training. Lastly, we also include the results of XLM-RoBERTa (Conneau et al., 2020), a recent language model trained with an explicit cross-lingual objective.

### 4.3 Results

**English Results.** The results on the CoNLL-2009 English in-domain test sets are outlined in Table 1a, where, for completeness, we report the scores achieved by our model both when the underlying

[2] For span-based SRL, we use the BIO format to convert span-level tags to token-level tags.

[3] `https://pytorch.org`

[4] `https://pytorchlightning.ai`

[5] `https://huggingface.co/transformers`

[6] `https://ufal.mff.cuni.cz/conll2009-st/scorer.html`

[7] `http://www.lsi.upc.es/~srlconll/srl-eval.pl`

| CoNLL-2009 – English | | *P* | *R* | *F*$_1$ |
|---|---|---|---|---|
| *ELMo models* | | | | |
| Cai and Lapata (2019b) | ⊙ | 90.9 | 89.1 | 90.0 |
| Lyu et al. (2019) | ⊙ | – | – | 90.1 |
| Kasai et al. (2019) | ⊙ | 90.3 | 90.0 | 90.2 |
| Li et al. (2019) | ⊘ | 89.6 | 91.2 | 90.4 |
| Chen et al. (2019) | ⊘ | 90.7 | 91.4 | 91.1 |
| Cai and Lapata (2019a) | ⊘ | 91.7 | 90.8 | 91.2 |
| This work | ⊘ | 91.3 | 91.7 | **91.4** |
| *BERT models (frozen)* | | | | |
| He et al. (2019) | ⊙ | 90.4 | 91.3 | 90.9 |
| This work | ⊘ | 91.2 | 91.8 | **91.5** |
| *BERT models (fine-tuned)* | | | | |
| Shi and Lin (2019) | ⊘ | 92.4 | 92.3 | 92.4 |
| This work | ⊘ | 92.5 | 92.7 | **92.6** |
| *XLM-RoBERTa models* | | | | |
| This work$_{\text{frozen}}$ | ⊘ | 91.5 | 92.1 | 91.8 |
| This work$_{\text{fine-tuned}}$ | ⊘ | 92.2 | 92.6 | **92.4** |

(a) Results on the English in-domain test set of the CoNLL-2009 shared task. Best results in F$_1$ score in **bold**. ⊙: syntax-aware system. ⊘: syntax-agnostic system.

| Hyperparameter | | Value |
|---|---|---|
| $K'$ | Upper lang. model layers | 4 |
| $d_w$ | Word emb. size | 512 |
| $K''$ | Pred.-aware encoder layers | 3 |
| $d_h$ | Pred.-aware state size | 512 |
| $d_p$ | Pred.-specific emb. size | 512 |
| $d_a$ | Arg.-specific emb. size | 512 |
| $K'''$ | Pred.-arg. encoder layers | 2 |
| $d_z$ | Pred.-arg. emb. size | 256 |
| $d_{s_p}$ | Pred. id. emb. size | 32 |
| $d_{s_s}$ | Pred. dis. emb. size | 256 |
| $d_{s_a}$ | Arg. class. emb. size | 512 |
| | Batch size | 32 |
| | Batch size when fine-tuning | 64 |
| | Max learning rate | $10^{-3}$ |
| | Min learning rate | $10^{-5}$ |
| | Max lr for LM fine-tuning | $10^{-5}$ |
| | Min lr for LM fine-tuning | $10^{-8}$ |
| | Warmup epochs | 1 |
| | Cooldown epochs | 15 |
| | Training epochs | 30 |

(b) Best hyperparameter values for the English CoNLL-2009 dev split. We use the same values for all the other languages.

Table 1: Results on the English in-domain test of the CoNLL-2009 benchmark (a) and corresponding model configuration (b).

| CoNLL-2012 - English | | P | R | F$_1$ |
|---|---|---|---|---|
| He et al. (2018a) | ⊘ | — | — | 85.5 |
| Ouchi et al. (2018) | ⊘ | **87.1** | 85.3 | 86.2 |
| Li et al. (2019) | ⊙ | 85.7 | 86.3 | 86.0 |
| Shi and Lin (2019) $_{\text{BERT-base}}$ | ⊘ | 85.7 | 86.7 | 86.2 |
| Shi and Lin (2019) $_{\text{BERT-large}}$ | ⊘ | 85.9 | 87.0 | 86.5 |
| This work $_{\text{BERT-base}}$ | ⊘ | 86.4 | 86.9 | 86.7 |
| This work $_{\text{BERT-large}}$ | ⊘ | 86.9 | **87.7** | **87.3** |

Table 2: Precision (P), Recall (R), and F$_1$ scores of recent non-ensemble systems on the CoNLL-2012 English benchmark. ⊙: syntax-aware system. ⊘: syntax-agnostic system.

language model weights are kept frozen and when they are fine-tuned during training. Independently of the pretrained language model used to represent the input, our syntax-agnostic model outperforms the best scores reported by the most recent state-of-the-art systems, both syntax-agnostic and syntax-aware. In particular, among ELMo-based models, our approach is slightly more effective than the work of Cai and Lapata (2019a) (+0.2% in F$_1$), despite this latter approach being self-supervised on additional training data. Our approach achieves a new state of the art, not only when using frozen BERT weights where the previous best-performing approach was the syntax-aware system of He et al. (2019) (+0.6%) – which takes advantage of purposely-built pruning rules based on syntactic dependency trees – but also when fine-tuning BERT, where it surpasses the syntax-agnostic system of Shi and Lin (2019) (+0.2%). Our model is also able to achieve a new state of the art in span-based English SRL, as shown in Table 2.

| CoNLL-2009 - Multilingual - In Domain | | CA | CZ | DE | ES | ZH |
|---|---|---|---|---|---|---|
| CoNLL-2009 ST best | ⊙ | 80.3 | 85.4 | 79.7 | 80.5 | 78.6 |
| Marcheggiani et al. (2017) | ⊘ | — | 86.0 | — | 80.3 | 81.2 |
| Chen et al. (2019) | ⊘ | 81.7 | 88.1 | 76.4 | 81.3 | 81.7 |
| Cai and Lapata (2019a) | ⊙ | — | — | 83.8 | 82.9 | 85.0 |
| Cai and Lapata (2019b) | ⊙ | — | — | 82.7 | 81.8 | 83.6 |
| Lyu et al. (2019) | ⊙ | 80.9 | 87.5 | 75.8 | 80.5 | 83.3 |
| He et al. (2019) | ⊙ | 86.0 | 89.7 | 81.1 | 85.2 | 86.9 |
| This work $_{\text{m-BERT frozen}}$ | ⊘ | 86.2 | 90.1 | 86.5 | 85.3 | 87.3 |
| This work $_{\text{m-BERT + fine-tuning}}$ | ⊘ | 87.4 | 91.1 | 87.1 | 85.9 | 88.0 |
| This work $_{\text{XLM-R frozen}}$ | ⊘ | 87.4 | 91.4 | 88.7 | 86.0 | 88.4 |
| This work $_{\text{XLM-R + fine-tuning}}$ | ⊘ | 88.3 | 92.1 | 89.1 | 86.9 | 89.1 |

Table 3: $F_1$ scores on the in-domain evaluation CoNLL-2009. "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 3 languages. ⊙: syntax-aware system. ⊘: syntax-agnostic system.

**Multilingual Results.** While English is certainly the most studied language in SRL, the core of our contribution lies in bridging the gap in performance between high-resource languages and low-resource languages. Indeed, Table 3 shows that high-resource languages, e.g., Chinese, witnessed comparatively larger improvements over the years, whereas lower-resource languages lagged behind. However, as Table 3 also shows, we find that our language-agnostic approach is able to bring larger improvements on low-resource languages such as German, with a 5.4% absolute improvement in $F_1$ score over the state-of-the-art multilingual system of He et al. (2019) when both systems use m-BERT to build a contextualized representation of the input sentences. We also find that our model benefits greatly from using XLM-RoBERTa, a language model pretrained with an explicit cross-lingual objective: our best result on German is remarkably close to the performance obtained on high-resource languages (89.1% vs 92.4% in $F_1$ score on German and English, respectively). We argue that improving the results on languages where a large margin for progress still exists should not be dismissed as an easier task, and that closing the gap between high-resource and low-resource settings should not be regarded as a foregone conclusion: as a matter of fact, among the systems that report results on all 6 languages in Table 3, the narrowest performance gap between English and German hovers around 10% in $F_1$. Our model significantly closes this gap to just 3% in $F_1$, while advancing the state of the art across the board.

The robustness of our model is particularly evident when considering the results on the CoNLL-2009 out-of-domain test sets (Table 4a), where, once again, it outperforms the state of the art in all 3 languages. Up until now, supervised techniques have not been able to surpass the performance of the best system based on manual feature engineering in the challenging out-of-domain German test set, since the sentences included in this split were specifically designed to contain a large number of infrequent predicates (Hajic et al., 2009). To the best of our knowledge, our approach results in the first supervised model capable of trumping manual feature engineering in such a setting.

**Zero-shot Cross-Lingual Results.** Having parity in the quality and quantity of data across languages is often an unrealistic expectation, especially whenever the task requires expert annotators (Pasini, 2020): this is the reason why, over the last few years, cross-lingual transfer learning techniques and benchmarks have garnered attention in NLP (Barba et al., 2020; Blloshmi et al., 2020; Conia and Navigli, 2020; Hu et al., 2020). While transfer learning techniques are becoming increasingly popular, their application to SRL is not straightforward. Indeed, as mentioned in Section 4.1, in CoNLL-2009 different languages use different, non-overlapping inventories for predicate senses and their argument structures. The only exceptions in CoNLL-2009 are the Catalan and Spanish languages which both use the AnCora inventory (Taulé et al., 2008). We take advantage of these two languages to evaluate how our system performs in

| CoNLL-2009 - OOD | | CZ | DE | EN |
|---|---|---|---|---|
| CoNLL-2009 ST best | $\odot$ | 85.4 | 65.9 | 73.3 |
| Zhao et al. (2009) | $\odot$ | 82.7 | 67.8 | 74.6 |
| Marcheggiani et al. (2017) | $\oslash$ | 87.2 | — | 77.7 |
| Li et al. (2019) | $\oslash$ | — | — | 81.5 |
| Chen et al. (2019) | $\oslash$ | — | — | 82.7 |
| Lyu et al. (2019) | $\odot$ | 86.0 | 65.7 | 82.2 |
| This work $_{\text{BERT frozen}}$ | $\oslash$ | 90.6 | 73.1 | 84.6 |
| This work $_{\text{BERT + ft}}$ | $\oslash$ | 91.3 | 72.5 | 85.9 |
| This work $_{\text{XLM-R frozen}}$ | $\oslash$ | 91.4 | 74.6 | 83.7 |
| This work $_{\text{XLM-R + ft}}$ | $\oslash$ | 92.1 | 74.6 | 85.2 |

(a) $F_1$ scores on the out-of-domain evaluation of CoNLL-2009.

| 0-Shot Cross-Lingual | SRC-TGT | $F_1$ |
|---|---|---|
| This work $_{\text{BERT frozen}}$ | CA-ES | 77.7 |
| This work $_{\text{BERT + ft}}$ | CA-ES | 79.1 |
| This work $_{\text{XLM-R frozen}}$ | CA-ES | 80.3 |
| This work $_{\text{XLM-R + ft}}$ | CA-ES | 81.7 |
| This work $_{\text{BERT frozen}}$ | ES-CA | 79.2 |
| This work $_{\text{BERT + ft}}$ | ES-CA | 79.8 |
| This work $_{\text{XLM-R frozen}}$ | ES-CA | 81.2 |
| This work $_{\text{XLM-R + ft}}$ | ES-CA | 81.8 |

(b) $F_1$ scores in the zero-shot cross-lingual setting between Catalan and Spanish, and vice versa.

Table 4: $F_1$ scores on the out-of-domain evaluation of the CoNLL-2009 Shared Task, and in the zero-shot cross-lingual setting. $\odot$: syntax-aware system. $\oslash$: syntax-agnostic system.

zero-shot cross-lingual SRL. Table 4b reports the results obtained when our model is trained on Catalan and evaluated on Spanish, and vice versa. While our model shows a significant drop in performance when compared to the original setting (trained and evaluated on the same language), we note that both training sets are relatively small (less than 15,000 sentences) and that the results are still promising (above 80% in both CA-ES and ES-CA), providing an estimate of the performance our model would obtain when trained on a high-resource language and evaluated on a low-resource language. To the best of our knowledge, we are the first to report the results in zero-shot cross-lingual SRL: we hope that our effort can be a basis for further development in this direction.

## 5 Analysis and Discussion

**Ablation Study.** We conducted an ablation study of our model architecture to better understand the individual contribution of its main components. As shown in Table 5, a baseline model, where the predicate-aware sequence encoder (see Section 3.2) is completely removed, shows a large drop in performance on the English development split of CoNLL-2009 (86.3% against 90.4% in $F_1$ score). Indeed, this encoder is central to our architecture and represents around 70% of the total parameter count in the model (pretrained language model excluded). Rather than removing the entire predicate-aware sequence encoder, we also evaluated the difference in performance when using simple stacked-BiLSTM layers instead of the fully-connected stacked-BiLSTM layers explained in Section 3.2 (+2.0 and +3.2% in $F_1$ score over the baseline, respectively). Finally, we found it beneficial to rely on features and mechanisms that do not exploit language-specific peculiarities, demonstrating that it is possible, not only to achieve and surpass the state of the art without syntax, but also, and most importantly, to bridge the gap between languages: for example, we perform layer normalization on the output of each BiLSTM layer (+0.3%), adopt Swish in place of the more traditional ReLU (+0.1%), and add predicate identification as a secondary task (+0.5%).

**Do We Need More Training Data?** Our model significantly narrows the gap between high- and low-resource languages, nevertheless one may wonder whether the amount of annotated data is still too limited for some of the languages in CoNLL-2009. In particular, German should be the most affected by this bottleneck, as it comes with the smallest number of annotated predicates (see Appendix A). However, surprisingly perhaps, Figure 1 shows that greatly reducing the already limited German training set does not degrade performance as much as one would expect: our model is still able to achieve state-of-the-art results in German on the in-domain evaluation with just 50% of the training data, and on the out-of-domain evaluation with just 25% of the training data.

| Ablation Study | $F_1$ |
|---|---|
| Baseline | 86.3 |
| + predicate-aware encoder | 88.3 |
| + fully-connected BiLSTM layers | 89.5 |
| + layer normalization | 89.8 |
| + Swish activation | 89.9 |
| + predicate identification | 90.4 |

Table 5: Results of the ablation study on the English development split.

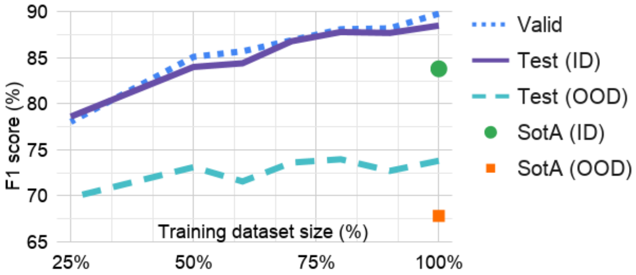

Figure 1: Our model outperforms the state of the art in German even when trained on a much smaller dataset.

## 6 Conclusion and Future Work

Recently, research in Semantic Role Labeling has revolved predominantly around syntax, with several studies showing the benefits of integrating syntactic features into existing models. Syntax-based innovations, however, have struggled to transfer improvements from high- to low-resource languages, as they appear to require a substantial amount of high-quality annotations.

In this paper, we have gone against the flow and proposed a model that puts both sentence- and lexical-level syntax aside, in order to avoid relying on noisy language-specific features. Our truly syntax-agnostic approach surpasses the previous state of the art among all syntax-agnostic and syntax-aware systems in the in- and out-of-domain evaluations of all 6 languages in the CoNLL-2009 benchmark dataset. Most crucially, our model seamlessly scales across languages, finally bridging the long-standing gap between high- and low-resource settings. Our analysis delineates the strengths of our approach and highlights its exceptional robustness: our model only needs 50% and 25%, respectively, of the training data to surpass the previous best-performing system in the in-domain and out-of-domain German test sets. To the best of our knowledge, we are the first to evaluate a model on zero-shot cross-lingual SRL, where we obtain results that are promising with a view to further developments in transfer learning techniques. Finally, we demonstrate that, not only is our model able to achieve state-of-the-art results in dependency-based SRL, but it also surpasses current syntax-agnostic and syntax-aware techniques in span-based SRL on CoNLL-2012.

As previously stated, our objective has not been to dismiss the undeniable importance of syntax-based innovation in SRL, but rather to establish a launch pad from which future syntactic developments can take off. In order to encourage future work on joint syntactic and semantic dependency parsing (Cai and Lapata, 2019b), the use of more powerful or cleverly trained language models (Lewis et al., 2020), the integration of SRL into other cross-lingual semantics-first tasks such as Semantic Parsing (Blloshmi et al., 2020) and Word Sense Disambiguation (Scarlini et al., 2020), and the exploitation and integration of newly available knowledge from recently released resources, such as VerbAtlas (Di Fabio et al., 2019) and Conception (Conia et al., 2020), we make available not only the code for our SRL model and experiments, but also our model checkpoints and training/validation logs at `https://github.com/SapienzaNLP/multi-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.

## References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv*, abs/1607.06450.

Edoardo Barba, Luigi Procopio, Niccolo Campolungo, Tommaso Pasini, and Roberto Navigli. 2020. MuLaN: Multilingual label propagation for Word Sense Disambiguation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3837–3844.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling Cross-Lingual AMR parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020.*

Rui Cai and Mirella Lapata. 2019a. Semi-supervised Semantic Role Labeling with cross-view training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1018–1027.

Rui Cai and Mirella Lapata. 2019b. Syntax-aware Semantic Role Labeling without parsing. *Trans. Assoc. Comput. Linguistics*, 7:343–356.

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 of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2753–2765.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL 2005, Ann Arbor, Michigan, USA, June 29-30, 2005*, pages 152–164.

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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5414–5424.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the 6th International Conference on Knowledge Capture (K-CAP 2011), June 26-29, 2011, Banff, Alberta, Canada*, pages 113–120.

Simone Conia and Roberto Navigli. 2020. Conception: Multilingually-enhanced, human-readable concept vector representations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020.*

Simone Conia, Fabrizio Brignone, Davide Zanfardino, and Roberto Navigli. 2020. InVeRo: Making Semantic Role Labeling accessible with intelligible verbs and roles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020.*

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 the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 627–637.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 484–490.

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

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, China, October 1-8, 2000*, pages 512–520.

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 the Thirteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2009, Boulder, Colorado, USA, June 4, 2009*, pages 1–18.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven Semantic Role Labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 643–653.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep Semantic Role Labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 473–483.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018a. Jointly predicting predicates and arguments in neural Semantic Role Labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 364–369, Melbourne, Australia, July.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018b. Syntax for Semantic Role Labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2061–2071.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual Semantic Role Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5349–5358.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv*, abs/2003.11080.

Jungo Kasai, Dan Friedman, Robert Frank, Dragomir R. Radev, and Owen Rambow. 2019. Syntax-aware neural Semantic Role Labeling with supertags. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 701–709.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6730–6737.

Chunchuan Lyu, Shay B. Cohen, and Ivan Titov. 2019. Semantic Role Labeling with iterative structure refinement. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1071–1082.

Diego Marcheggiani and Ivan Titov. 2017a. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1506–1515.

Diego Marcheggiani and Ivan Titov. 2017b. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1506–1515.

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 the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 411–420.

Diego Marcheggiani, Jasmijn Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 486–492.

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.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (present and future) use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5697–5702, 7.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for Semantic Role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1630–1642, Brussels, Belgium, October-November.

Tommaso Pasini. 2020. The knowledge acquisition bottleneck problem in multilingual Word Sense Disambiguation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4936–4942.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012, July 13, 2012, Jeju Island, Korea*, pages 1–40.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2018. Searching for activation functions. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and Semantic Role Labeling. *arXiv preprint arXiv:1904.05255*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for Semantic Role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5027–5038.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL 2008, Manchester, UK, August 16-17, 2008*, pages 159–177.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

Yufei Wang, Mark Johnson, Stephen Wan, Yifang Sun, and Wei Wang. 2019. How to best use syntax in semantic role labelling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5338–5343.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Meishan Zhang, Peili Liang, and Guohong Fu. 2019a. Enhancing opinion role labeling with semantic-aware word representations from Semantic Role Labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 641–646.

Yue Zhang, Rui Wang, and Luo Si. 2019b. Syntax-enhanced self-attention-based Semantic Role Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 616–626.

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 the Thirteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL 2009, Boulder, Colorado, USA, June 4, 2009*, pages 61–66.

|  | 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 |
| *CoNLL-2012* |  |  |  |  |  |  |  |
| EN | 115,812 | 90,856 | 20.5 | 253,070 | 5,287 | 852,053 | 67 |

Table 6: Overview of the CoNLL-2009 and CoNLL-2012 training 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*).

.

## A Data Statistics

Table 6 shows the composition of the training sets of CoNLL-2009 and CoNLL-2012. While German is not usually considered a low-resource language, in the CoNLL-2009 dataset it is the language with the lowest amount of annotated predicate instances. As shown in the table, German has less than 50% of the predicate instances of Catalan and less than 10% of the predicate instances of English.

## B Computing Infrastructure

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