UNITED-SRL: A Unified Dataset for Span- and Dependency-Based Multilingual and Cross-Lingual Semantic Role Labeling

Rocco Tripodi[*]
University of Bologna
rocco.tripodi@unibo.it

Simone Conia
Sapienza University of Rome
conia@di.uniroma1.it

Roberto Navigli
Sapienza University of Rome
navigli@diag.uniroma1.it

Abstract

Multilingual and cross-lingual Semantic Role Labeling (SRL) have recently garnered increasing attention as multilingual text representation techniques have become more effective and widely available. While recent work has attained growing success, results on gold multilingual benchmarks are still not easily comparable across languages, making it difficult to grasp where we stand. For example, in CoNLL-2009, the standard benchmark for multilingual SRL, language-to-language comparisons are affected by the fact that each language has its own dataset which differs from the others in size, domains, sets of labels and annotation guidelines. In this paper, we address this issue and propose UNITED-SRL, a new benchmark for multilingual and cross-lingual, span- and dependency-based SRL. UNITED-SRL provides expert-curated parallel annotations using a common predicate-argument structure inventory, allowing direct comparisons across languages and encouraging studies on cross-lingual transfer in SRL. We release UNITED-SRL v1.0 at https://github.com/SapienzaNLP/united-srl.

1 Introduction

Semantic Role Labeling (SRL) – often considered to be a fundamental step towards Natural Language Understanding (Navigli, 2018) – consists in recovering the latent predicate-argument structure of a sentence by identifying the semantic relationship between a predicate and its arguments (Gildea and Jurafsky, 2002). SRL can be used to extract information from text and to provide a shallow semantic representation of sentences, finding applications in a wide range of Natural Language Processing (NLP) areas such as Machine Translation (Marcheggiani et al., 2018), Question Answering (Shen and Lapata, 2007; He et al., 2015), Visual Semantic Role Labeling (Silberer and Pinkal, 2018), Semantic Parsing (Banarescu et al., 2013) and Story Generation (Fan et al., 2018).

Given the popularity of this task, over the years several competitions have been organized within the Conference on Computational Language Learning (CoNLL) to evaluate SRL systems (Carreras and Màrquez, 2004, 2005; Surdeanu et al., 2008; Hajič et al., 2009; Pradhan et al., 2012). These shared tasks led to the release of several datasets that nowadays represent the *de facto* standard benchmarks for SRL, namely, CoNLL-2005, CoNLL-2008, CoNLL-2009 and CoNLL-2012.

However, despite their widespread use, the CoNLL datasets suffer from a considerable level of heterogeneity, which prevents systems from easily scaling across task formulations and languages: CoNLL-2005 (Carreras and Màrquez, 2005) is devised for span-based SRL where systems are required to identify and classify argument spans, whereas CoNLL-2009 (Hajič et al., 2009) is framed as a dependency-based task, where only the syntactic head of an argument has to be tagged. Moreover, when multiple languages are covered, for example in CoNLL-2009 and CoNLL-2012 (Pradhan et al., 2012), different inventories are used, such as English Propositional Bank (Palmer et al., 2005, PropBank), Chinese PropBank (Xue and Palmer, 2003) and AnCora (Taulé et al., 2008) for Spanish and Catalan, making it difficult to evaluate whether a system is able to generalize across languages and, if so, to what extent. In fact, these multilingual datasets are not aligned, they are considerably different in size and show significant dissimilarities in their domain distribution, strongly limiting language-to-language comparisons.

Some studies (Akbik et al., 2015; Daza and Frank, 2020) worked around this issue by electing the English PropBank as a universal semantic inventory and employing cross-lingual annotation projection techniques to produce annotations for other languages starting from English annotated

[*] Work partially carried out while at the Sapienza University of Rome.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2293–2305
November 7–11, 2021. ©2021 Association for Computational Linguistics

data. These approaches, however, tend to ignore the fact that ProbBank was conceived expressly for English predicates.

Forcing the semantics of different languages to adapt to English, without considering possible translation divergences in parallel sentences (Dorr, 1994), can lead to two distinct problems: i) incorrect projections, if divergent sentences are retained in the dataset; ii) elimination from the dataset of all the sources of linguistic divergences, if divergent sentence pairs are discarded. Another issue of SRL annotation projection techniques regards the use of PropBank as verbal resource. This, in fact, limits the informativeness of the annotations, since it does not mark semantic roles with semantically-consistent labels, leading to ambiguous or unclear annotations. For example, in *John is sleeping* and *John loves Mary*, *John* would be tagged ARG0 in both cases, but we argue that AGENT and EXPERIENCER are clearer and more fitting semantic roles to tag *John* with.

The consequence of all these limitations is that, in order to engage in this task, there is a plethora of features and settings to choose before selecting the proper dataset. To address the above-mentioned issues and encourage future work on cross-lingual approaches for SRL, we propose UNITED-SRL, a unified SRL dataset with the following features:

- The first manually-created parallel corpus with semantic role annotations in four different languages: Chinese, English, French and Spanish;
- The first manually-created dataset with gold parallel span- and dependency-based SRL annotations;
- We provide annotations with VerbAtlas (Di Fabio et al., 2019), a semantic resource explicitly designed to overcome the heterogeneous landscape of different predicate senses and semantic roles;
- Multi-domain training, development and test sets from 10 semantic domains derived from the taxonomy of the UN corpus;

We expect that the release of a parallel multilingual dataset will provide a fair evaluation for multilingual and cross-lingual SRL systems, making the results directly comparable from language to language. We release UNITED-SRL v1.0 at https://github.com/SapienzaNLP/united-srl.

2 Related Work

Multilingual SRL Datasets. Due to the complexity of the task, SRL annotations are expensive to produce as they require expert annotators who are comfortable with the linguistic theories behind the predicate-argument inventory of choice. This makes the creation of multilingual SRL datasets even more difficult. Perhaps the largest effort in this direction was made on the occasion of the CoNLL-2009 shared task (Hajič et al., 2009). The CoNLL-2009 multilingual dataset for dependency-based SRL originally featured 7 languages: English, Chinese, Czech, German, Catalan, Spanish and Japanese.[1] However, each of these datasets was annotated separately, starting from different corpora and using different predicate-argument structure inventories, e.g., the English PropBank (Palmer et al., 2005) for English, PDT-Vallex (Hajic et al., 2003) for Czech, AnCora (Taulé et al., 2008) for Spanish and Catalan. Universal Propositional Bank[2] (Akbik et al., 2015; Akbik and Li, 2016) adds SRL annotations on top of the Universal Dependency corpus (de Marneffe et al., 2014). The limitations of this dataset are that, even if it covers 8 languages, the sentences in it are not parallel and were annotated automatically.

We argue that this heterogeneity inhibits, or at least slows down, further advances in multilingual and cross-lingual SRL.

Cross-lingual SRL Datasets. To the best of our knowledge, only silver datasets exist for cross-lingual SRL. These datasets are based mainly on annotation projection, an approach that, starting from gold annotated data in a language, allows annotations to be transferred to parallel sentences in other languages. Many works that use this approach for cross-lingual SRL have been presented over the years (Padó and Lapata, 2009; van der Plas et al., 2011; Aminian et al., 2019) both for FrameNet (Baker, 2014) and PropBank (Palmer et al., 2005).

Annotation projection is based on the Direct Semantic Transfer (van der Plas et al., 2011, DST) assumption, which states that, given two sentences S and T, predicate-argument relations $R(x_S, y_S)$ can be transferred to $R(x_T, y_T)$ only if there exists a word alignment between x_S and x_T and between

[1] Japanese is no longer available due to copyright issues.

[2] https://github.com/System-T/UniversalPropositions

	English				Spanish				French				Chinese			
	Sents	Preds	Args	P/S	Sents	Preds	Args	P/S	Sents	Preds	Args	P/S	Sents	Preds	Args	P/S
Train	$5,503$	$12,549$	$23,468$	2.28	464	$1,074$	$2,098$	2.31	464	$1,121$	$2,236$	2.41	$3,645$	$11,822$	$20,358$	3.24
Dev	$1,027$	$2,554$	$4,776$	2.48	$1,027$	$2,495$	$4,736$	2.43	$1,026$	$2,555$	$4,991$	2.49	957	$3,495$	$6,012$	3.65
Test	$1,027$	$2,609$	$4,916$	2.54	$1,027$	$2,531$	$4,802$	2.46	$1,027$	$2,561$	$5,008$	2.49	952	$3,419$	$5,992$	3.59

Table 1: Overall statistics of the UNITED-SRL dataset. Number of sentences (Sents), annotated predicates (Preds) and arguments (Args) and average number of predicates per sentence (P/S) in each split and for each language.

y_S and y_T, where x and y are predicates and arguments, respectively. Another constraint of this model is that x_T has to be a verb (verbal predicate). Even if, thanks to the progress made in machine translation, multilingual sentence embedding and multilingual language modeling the task of aligning spans of text in different languages has become quite effective (Dou and Neubig, 2021; Lacerra et al., 2021; Procopio et al., 2021), annotation projection techniques, in the specific case of SRL, retain parallel annotations only when they satisfy specific constraints (e.g., same predicate-argument structure). These limitations can hinder the evaluation of cross-lingual transfer learning techniques on this task, since the benchmarks created contain only examples for which it is already known, by means of the DST assumption behind annotation projection techniques, that the same features are present in the source and the target languages, thereby omitting cases of translation divergences altogether (Dorr, 1994; Blloshmi et al., 2020) and evaluating only on a subset of cases for which the transfer from one sentence to another is direct.

On this line of research lies X-SRL, a recently introduced dataset proposed by Daza and Frank (2020). X-SRL is a multilingual parallel SRL corpus that is based on the English part of CoNLL-2009 (Hajič et al., 2009) for in-domain dependency-based SRL. The gold annotations of CoNLL-2009 in English have been translated using high-quality machine translation services into three target languages, namely, French, German and Spanish. Once a machine-translated parallel corpus has been created, mBERT (Devlin et al., 2019) is used to produce vector representations of text and to compute the similarity between source and target tokens. These embeddings are then used to align tokens and to transfer annotations across different languages. The annotation of the training and development sets of this dataset is automatic while the annotation of the test set is semi-automatic. In particular, annotators were asked to validate translations and to mark in the target sentence tokens that can express the same meaning of predicates and arguments of the English gold annotations.

3 The UNITED-SRL Dataset

UNITED-SRL consists of two parallel training sets in Chinese and English with $5,503$ and $3,645$ sentences, respectively. It also includes $2,000$ parallel sentences for Chinese, English, French and Spanish ($1,000$ for the development and $1,000$ for the test set of each language). The overall statistics of the dataset including the number of sentences, the number of predicates, the number of roles and the average number of annotated predicates per sentence are presented in Table 1.

The sentences of our dataset have been selected from the UN Parallel Corpus[3] (Ziemski et al., 2016), a multilingual collection of official records and parliamentary reports of the United Nations. This corpus contains over 11 million sentences per language across $86,000$ documents, organized in 18 semantic domains. We selected this corpus because it consists of multilingual human-translated documents, it is available for free and it contains a large number of documents from different semantic domains. This choice allowed us to create a multi-domain dataset that can be used to test the generalization capabilities of SRL systems, avoiding their specialization to a specific domain (like the financial domain of CoNLL-2009 in-domain English dataset). To ensure the heterogeneity of the textual material in UNITED-SRL, we sampled documents belonging to the 10 most frequent domains of the UN corpus, following the domain distribution in the corpus.

One of the main novelties of our dataset is the use of a new verbal resource: VerbAtlas[4] (Di Fabio et al., 2019). This resource contains $13,767$ synsets from BabelNet[5] (Navigli and Ponzetto, 2012; Navigli et al., 2021) manually clustered in around 400

[3] https://conferences.unite.un.org/uncorpus

[4] http://verbatlas.org

[5] https://babelnet.org

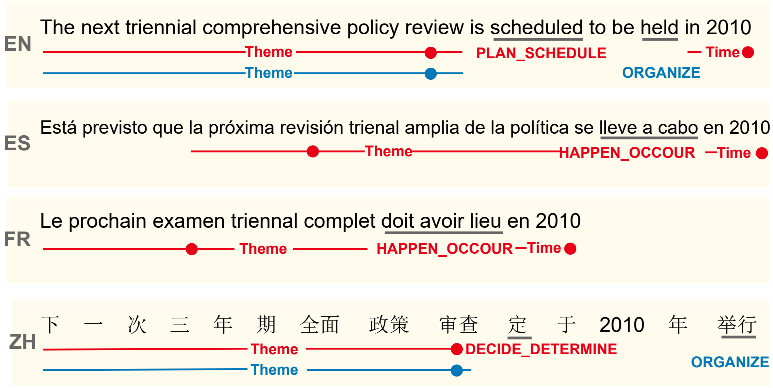

EN
The next triennial comprehensive policy review is scheduled to be held in 2010
Theme
PLAN_SCHEDULE
Time
Theme
ORGANIZE
ES
Está previsto que la próxima revisión trienal amplia de la política se lleve a cabo en 2010
Theme
HAPPEN_OCCOUR
Time
FR
Le prochain examen triennal complet doit avoir lieu en 2010
Theme
HAPPEN_OCCOUR
Time
ZH
下 一 次 三 年 期 全面 政策 审查 定 于 2010 年 举行
Theme
DECIDE_DETERMINE
Theme
ORGANIZE

Figure 1: A cross-lingual annotation example, for a sentence in English (EN), Spanish (ES), French (FR) and Chinese (ZH).

semantically-coherent frames. Each frame is provided with an argument structure composed of explicit semantic roles, such as AGENT, THEME or BENEFICIARY, and with a list of possible lexicalizations in different languages, each of which is connected to a particular synset, making the resource inherently multilingual.

An example of a multilingual annotation from UNITED-SRL is shown in Figure 1, where predicate-argument structures are indicated with different colors (red and blue), argument spans are indicated with straight lines and dependencies are marked with a circle on the corresponding span line. The example reported in Figure 1 is paradigmatic and illustrates the nature of our dataset very clearly. It shows how a source sentence in English can be translated (by experts) in a way in which the same information is conveyed with different predicate-argument structures. In particular, we want to show that the frames of the English sentence are not present in the French and the Spanish ones, which in their turn have the same predicate-argument structures, and that in Chinese the first predicate needs a different frame, DECIDE_DETERMINE instead of PLAN_SCHEDULE. These divergences would have caused annotation projection techniques to discard or wrongly annotate the aforementioned sentences, whereas in our dataset they are maintained and can be easily compared and evaluated.

Another advantage of our annotation lies in the employment of VerbAtlas as verbal resource. In fact, a VerbAtlas frame includes all the synsets with a meaning related to a particular concept, for example the ORGANIZE frame is connected to lexicalizations of words such as *organize*, *prepare*, *arrange*, *plan* and *coordinate*, in different languages. All these lexicalizations feature the same argument structure and are particularly suited for the annotations of parallel linguistic units in different languages.

3.1 Data Annotation and Quality

UNITED-SRL was annotated using a dedicated web interface by six annotators, four for English, two for Chinese and one for French and Spanish. The annotators for each language were selected from native speakers and expert translators with experience in linguistic annotation tasks. They were instructed with annotation guidelines (see Appendix B) and weekly meetings in which all the annotators discussed common problems and proposed solutions for them. The average annotation time was around 10 sentences per hour for the SRL layers (both span- and dependency-based) and for the sense annotation layer.

An additional annotator for English and Chinese was employed to check the quality of the annotations. They were asked to annotate a random sample of 100 sentences at two different times: after the first $1,000$ annotated sentences and at the end of the annotation. To compute the inter-annotator agreement between two annotations A_1 and A_2 we considered different layers of annotation for each sentence, i.e., predicate identification A_i^{preds}, predicate sense disambiguation A_i^{dis}, argument structure identification A_i^{args}, and, for each argument, span selection A_i^{span}, and dependency identification A_i^{dep}. We compared two sets of annotations, A_1 and A_2, and computed the agreement among them as the number of identical annotations divided by the total number of annotations in all layers. More formally, we computed the annotation overlap as:

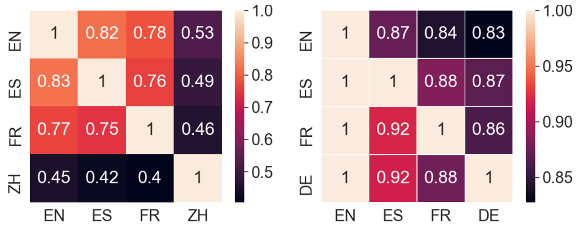
$$O = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}.$$

With this measure we were able to easily identify and interpret disagreements.

The first round of evaluation had an annotation overlap of 0.83 for English and 0.72 for Chinese, which correspond to 0.67 Cohen's κ (Cohen, 1960) for English and 0.60 for Chinese on predicate identification and disambiguation. This evaluation served to identify and correct some idiosyncrasies in the dataset, e.g., in some cases due to part-of-speech tagging errors some adjectives were tagged as verbs (past participle) and the annotators in some cases annotated them and in some cases did not. With the analysis of the disagreements, we were able to refine the annotation guidelines and to correct annotations that were wrong due to part-of-speech tagging.

Carefully checking the annotations, we also noticed that in most cases the disagreement on predicate sense disambiguation was on short sentences, where the lack of context made the disambiguation difficult and in some cases arbitrary. For example, in the sentence *The Committee notes the importance of the role of traditional education, particularly in remote island communities*, it is not clear which frame to use for the verb *note*. It can be SPEAK (make mention of), PERCEIVE (notice or perceive) or SEE (observe with care or pay close attention to), *inter alia*. For this reason, we decided to discard these sentences from the dataset. Thanks to these actions the annotation overlap scores for the second round of annotations rose to 0.86 and 0.76 for English and Chinese, respectively, which correspond to 0.80 Cohen's κ for English and 0.69 for Chinese on predicate identification and disambiguation.

In addition to the agreement computation, in order to ensure the quality of the annotated data, different automatic checks were used to verify not only that the annotations were well-formed but also that they respected the VerbAtlas structure, i.e., to ensure that predicates were annotated with coherent frames and that roles were selected only among those admitted by the VerbAtlas predicate-argument structures.



(a) UNITED-SRL (b) X-SRL

Figure 2: Heatmaps reporting the fraction of frames in the test sets of UNITED-SRL and X-SRL that are shared among languages. Each cell indicates the fraction of frames in a source language (vertical axis) that are also present in a target language (horizontal axes) on a sentence-by-sentence basis.

3.2 Predicate-argument Structures across Languages

As already mentioned, the frame and the role inventory of UNITED-SRL are independent of the language: as a result, our semantic annotation will enable the study of predicate-argument semantics across languages and the investigation of how the information of these structures can be transferred from one language to another (see §4). To this end, we analyzed the predicates annotated in our test set and compared them with those in X-SRL (Daza and Frank, 2020), a dataset that uses annotation projection to transfer predicate-argument structures from English to other languages.

This analysis is presented in Figure 2, where we indicate the fraction of predicates of a source language (reported on the vertical axis of the heatmaps) that has been annotated with the same frames in a target language (reported on the horizontal axis of the heatmaps). As we can see from Figure 2a, the directions EN→ES and EN→FR share 82% and 78% of the frames in UNITED-SRL, while the fraction for Chinese is much lower (53%). This large discrepancy is justified by the fact that Chinese and English are two genetically distant languages; indeed, we also observed in our dataset that in many cases nominal, adjectival and prepositional expressions in other languages are featured in Chinese using verbal phrases (e.g., the Chinese parallel sentence of *The treaty on the prohibition of nuclear weapons* can be translated literally as *The treaty that forbids nuclear weapons*). Different annotations in our parallel sentences are not due to inconsistencies in the annotations. Our annotation guidelines (see Appendix B) require that the annotators in languages other than English always have to check the English parallel annotations and, if it is possible, they have to maintain the same frame annotations in other languages.

Figure 2b shows that the EN→FR and EN→ES fraction in X-SRL are slightly higher than those reported in our dataset, by 5% and 6%, respectively. We can also notice from this figure that all the languages covered by X-SRL share a large number of annotations.

If we invert the direction of the comparison, i.e., ES→EN or FR→EN, we can see that in UNITED-SRL the fractions are consistent with their inverted direction counterparts, while in X-SRL the annotations in other languages share all the annotations with the English annotation. This is due to the fact that with annotation projection techniques the annotations in other languages can only cover a subset of the English ones and suggests that approximately 15% of predicate annotations in the X-SRL dataset (in languages other than English) are missing. This reinforces the suspicion that the evaluation of the language transfer capabilities of a model on annotation projection datasets may be overestimated.

4 Experiments

One of the main objectives of UNITED-SRL is to allow past and future systems and their results to be directly comparable across diverse languages without having to deal with or take into account heterogeneous linguistic resources, different domains and varying dataset sizes. To this end, we use UNITED-SRL to train and evaluate a recently proposed state-of-the-art SRL system, showing how our dataset provides interesting insights into the cross-lingual transferability of predicate senses and their argument structures.

4.1 Experimental Setup

For our experiments we use the state-of-the-art SRL system proposed by Conia and Navigli (2020), CN-20 hereafter, which performs on par with recently introduced models for end-to-end SRL (Blloshmi et al., 2021; Conia et al., 2021a,b). CN-20 represents the input sentence using a pretrained language model and then feeds these representations into a stack of BiLSTM layers to disambiguate predicate senses and label their arguments. The advantage of using CN-20 is that i) it is syntax-agnostic, i.e., it does not require any syntactic feature at the input level, ii) it can easily be used on top of different language models, and iii) it has been shown to attain state-of-the-art results in both dependency- and span-based SRL. In the following, we evaluate the performance of this system with two different pretrained language models, multilingual-BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), distinguishing between the results when their weights are left frozen or fine-tuned together with the rest of the system during training.

We train each model configuration for 20 epochs using Adam (Kingma and Ba, 2015) with a learning rate that is initially warmed up to 10^{-5} for 1 epoch and then cooled down to 10^{-6} in 15 epochs. We leave the rest of the hyperparameter values as in the original paper of Conia and Navigli (2020). All the experimental details are reported in Appendix C.

4.2 Results on Sense Disambiguation

In the following, we describe and discuss the results of CN-20 on predicate sense disambiguation, that is, the task of assigning the most appropriate sense to a predicate in context. We first focus on zero-shot cross-lingual predicate sense disambiguation and then show how even a small language-specific sample leads to significant improvements.

0-shot Cross-lingual Sense Disambiguation. Table 2 shows the results obtained by CN-20 on predicate sense disambiguation in different training settings. We observe that, even though the training splits of UNITED-SRL may be considered relatively small in comparison to other currently available datasets such as CoNLL-2009 and CoNLL-2012, CN-20 is still able to attain remarkable results on this subtask in both English and Chinese when trained on their respective training sets, achieving 88.4% and 78.0% in accuracy. Unsurprisingly, the results on predicate sense disambiguation show a significant drop when CN-20 is trained on a language, e.g., English, and evaluated on another language, e.g., Chinese. We stress that, since the development and test sets are parallel, the results are directly comparable across any two languages, meaning that the drop in performance is primarily caused by the linguistic differences between the two languages considered (see §3.2). In general, CN-20 seems to perform similarly with multilingual-BERT and XLM-RoBERTa when evaluated on the languages it was trained on, e.g., training and evaluating on the English dataset. However, XLM-RoBERTa shows stronger knowl-

		100%		500 sentences				Sense Accuracy (%)			
	FT?	EN	ZH	EN	ZH	ES	FR	EN	ZH	ES	FR
multilingual-BERT	–	✔	–	–	–	–	–	84.9	50.6	50.3	48.6
	✔	✔	–	–	–	–	–	87.4	54.6	52.5	50.7
	–	–	✔	–	–	–	–	45.3	76.9	30.4	31.0
	✔	–	✔	–	–	–	–	48.9	77.7	37.0	31.5
	–	✔	–	–	✔	✔	✔	85.1	73.4	61.2	57.8
	✔	✔	–	–	✔	✔	✔	**87.9**	75.9	**65.0**	**60.2**
	–	–	✔	✔	–	✔	✔	70.1	77.2	57.5	53.3
	✔	–	✔	✔	–	✔	✔	76.9	77.9	61.2	56.1
	–	✔	✔	–	–	–	–	84.6	77.8	52.8	51.8
	✔	✔	✔	–	–	–	–	86.8	**78.2**	53.9	51.7
	–	✔	✔	–	–	✔	✔	84.4	78.1	62.4	59.4
	✔	✔	✔	–	–	✔	✔	86.9	77.8	64.0	59.7
XLM-RoBERTa	–	✔	–	–	–	–	–	87.0	54.7	59.1	55.3
	✔	✔	–	–	–	–	–	**88.4**	60.5	61.2	55.7
	–	–	✔	–	–	–	–	63.6	78.0	53.3	44.8
	✔	–	✔	–	–	–	–	69.6	76.7	54.0	48.5
	–	✔	–	–	✔	✔	✔	87.0	76.0	67.1	62.4
	✔	✔	–	–	✔	✔	✔	87.9	75.8	67.1	62.8
	–	–	✔	✔	–	✔	✔	77.7	77.7	63.6	60.3
	✔	–	✔	✔	–	✔	✔	80.5	77.6	66.1	61.3
	–	✔	✔	–	–	–	–	87.4	77.9	62.1	58.2
	✔	✔	✔	–	–	–	–	**88.4**	78.0	63.0	59.5
	–	✔	✔	–	–	✔	✔	86.8	78.1	66.3	62.9
	✔	✔	✔	–	–	✔	✔	88.1	**78.5**	**69.0**	**63.3**

Table 2: Accuracy on predicate sense disambiguation on the test sets in English (EN), Chinese (ZH), Spanish (ES) and French (FR). We report the results obtained when using multilingual-BERT and XLM-RoBERTa, finding a consistent behavior between the two language models. **FT?:** is the language model fine-tuned for the task? **100%:** trained on 100% of the data available for that language (5, 500 sentences in English, 3, 500 sentences in Chinese). 500 **sentences:** trained on 500 sentences for that language. Best results are in **bold**.

edge transfer capabilities, providing double-digit improvements on zero-shot cross-lingual predicate sense disambiguation in Spanish and French.

Cross-lingual Sense Disambiguation. Table 2 also includes the results of CN-20 when trained on more than one language at the same time. In particular, we carry out experiments with several combinations of languages in order to assess the capability of a state-of-the-art system to model different language-specific linguistic properties. Contrary to our expectations, our results show that training CN-20 jointly on English and Chinese, two very distant languages linguistically, does not hamper the results on predicate sense disambiguation; in fact, adding the Chinese training set to the English one actually leads to an improvement on Spanish and French. Moreover, including less than 500 annotated sentences in Spanish and French brings a further significant improvement. We highlight that these additional sentences in Spanish and French do not provide additional coverage as they can be found translated in the English and the Chinese datasets, suggesting that CN-20 takes advantage of such sentences for language adaptation (Ruder et al., 2019).

4.3 Results on Argument Labeling

In what follows, instead, we report and analyze the results of CN-20 on argument labeling, that is, the task of identifying the arguments of a pred-

		100%		500 sentences				Dependency F1				Span F1			
	FT?	EN	ZH	EN	ZH	ES	FR	EN	ZH	ES	FR	EN	ZH	ES	FR
multilingual-BERT	–	✔	–	–	–	–	–	83.7	33.6	56.8	58.7	75.8	37.4	54.2	54.5
	✔	✔	–	–	–	–	–	83.8	37.0	60.6	62.6	77.2	39.5	55.2	55.8
	–	–	✔	–	–	–	–	47.1	74.7	36.9	39.0	45.0	68.4	33.8	35.7
	✔	–	✔	–	–	–	–	49.3	75.3	42.6	40.4	47.4	70.0	34.9	37.5
	–	✔	–	–	✔	✔	✔	83.6	68.1	70.7	69.7	75.0	60.5	62.7	61.3
	✔	✔	–	–	✔	✔	✔	**84.5**	70.9	72.3	70.3	77.5	64.5	64.9	63.0
	–	–	✔	✔	–	✔	✔	75.4	74.3	68.6	67.7	65.6	67.8	58.3	57.3
	✔	–	✔	✔	–	✔	✔	77.9	**75.8**	70.7	68.3	68.6	69.1	61.8	59.0
	–	✔	✔	–	–	–	–	83.8	75.0	59.3	60.9	75.7	69.2	54.1	54.8
	✔	✔	✔	–	–	–	–	84.3	**75.8**	63.9	65.0	77.8	70.5	54.8	56.5
	–	✔	✔	–	–	✔	✔	83.4	74.3	71.2	69.8	75.4	68.5	63.4	60.9
	✔	✔	✔	–	–	✔	✔	**84.5**	75.1	**73.0**	**70.4**	**78.3**	**70.3**	**65.7**	**63.3**
XLM-RoBERTa	–	✔	–	–	–	–	–	83.8	22.8	62.9	63.4	77.3	40.5	60.3	58.3
	✔	✔	–	–	–	–	–	84.7	31.9	67.2	66.6	78.5	42.9	62.6	58.8
	–	–	✔	–	–	–	–	55.7	75.4	44.3	45.7	51.6	69.3	39.3	38.0
	✔	–	✔	–	–	–	–	57.7	76.9	51.4	49.4	55.7	71.9	48.6	45.9
	–	✔	–	–	✔	✔	✔	84.4	69.8	73.2	71.4	76.8	62.9	65.5	63.3
	✔	✔	–	–	✔	✔	✔	85.3	72.2	**74.8**	73.0	**79.3**	67.7	**67.5**	65.3
	–	–	✔	✔	–	✔	✔	77.9	75.6	70.6	68.3	68.8	68.7	63.0	60.3
	✔	–	✔	✔	–	✔	✔	80.8	76.7	73.1	71.2	71.4	70.4	64.9	63.2
	–	✔	✔	–	–	–	–	84.5	75.8	64.0	64.0	77.6	70.6	60.4	58.3
	✔	✔	✔	–	–	–	–	85.3	**77.2**	69.6	69.8	78.8	71.9	63.7	60.6
	–	✔	✔	–	–	✔	✔	84.3	75.7	72.9	71.4	77.0	70.8	66.1	63.8
	✔	✔	✔	–	–	✔	✔	**85.5**	**77.2**	74.6	**73.1**	78.8	**72.0**	**67.5**	**65.4**

Table 3: F1 scores on dependency- and span-based argument labeling on the test sets in English (EN), Chinese (ZH), Spanish (ES) and French (FR). We report the results obtained when using multilingual-BERT and XLM-RoBERTa, finding a consistent behavior between the two language models. **FT?:** is the language model fine-tuned for the task? 100%**:** trained on 100% of the data available for that language ($5,500$ sentences in English, $3,500$ sentences in Chinese). 500 **sentences:** trained on 500 sentences for that language. Best results are in **bold**.

icate and assigning the most appropriate role to each predicate-argument relation. Similarly to the previous Section, we will first provide an overview of our results on zero-shot cross-lingual argument labeling and then focus on the improvements that language-specific data can bring.

0-shot Cross-lingual SRL. Table 3 shows the results obtained by CN-20 on dependency- and span-based argument labeling in the same training settings as those devised for our experiments on predicate sense disambiguation. Similarly to what we found in our predicate sense disambiguation experiments, CN-20 is able to perform dependency- and span-based SRL with remarkable results when trained and evaluated on the same language (84.7% and 78.5% in F1 score on dependency- and span-based English argument labeling, respectively), especially considering the complexity of the task and the relatively small size of the training sets. As expected, the drop in performance in zero-shot cross-lingual argument labeling is more pronounced than that which we saw for the predicate sense disambiguation subtask. Indeed, the position of a semantic head and the start/end of a span are more affected by the linguistic differences between English, Chinese, Spanish and French. Interestingly, the results on zero-shot cross-lingual argument labeling are very similar between Spanish and French, both in dependency- and span-based SRL, probably owing to the fact that they are both neo-Latin languages.

Cross-lingual SRL. Table 3 also includes the results of CN-20 when it is trained jointly on multiple languages. Similarly to what is shown in Table 2 for the subtask of predicate sense disambiguation, the reported results provide an empirical demonstration that an automatic system can indeed benefit from learning over multiple languages at the same time. In particular, adding the Chinese training set to the English training set brings an improvement, albeit small, to the results on both English and Chinese dependency- and span-based argument labeling; including a further 500 sentences in French and Spanish to the training set makes CN-20 remarkably strong in those languages as well. These results provide additional empirical confirmation that it is not necessary to annotate large amounts of text in each language of interest, but that convincing performance can be achieved by annotating a large dataset for just a single language (e.g. English) supported by several small datasets that allow a system to learn the peculiarities of each language.

We highlight that training CN-20 jointly on multiple languages is not only beneficial from a performance point of view, but also relieves researchers and downstream users from having to train and maintain multiple instances of the same model for each and every language, i.e., train and use one system for English inputs, another system for Chinese inputs, and so on. In general, our results lend credibility to the idea that cross-lingual data annotated with predicate sense and semantic role labels from a single inventory shared across languages could open the door to the development of more robust cross-lingual SRL systems.

5 Conclusion

In this paper, we presented the first version of UniteD-SRL, a unified dataset for span- and dependency-based multilingual and cross-lingual SRL. Our dataset can be used as test bed to investigate different aspects of multilingual and cross-lingual Semantic Role Labeling. The same models can be evaluated for both span- and dependency-based SRL on different languages and on the same verb inventory. These features allow us to have a realistic view of the transfer learning and language adaptation capabilities of past, current and future SRL systems, thereby enabling studies on cross-lingual transfer also in this task.

We conducted an extensive evaluation of a state-of-the-art SRL model on UniteD-SRL, through which we were able to validate different hypotheses on individual languages and across multiple languages. Thanks to the shared verbal inventory employed by UniteD-SRL we were able to train with examples in different languages and to test the effect that this has on the performances of the model. The results obtained with our evaluation reinforce the idea that cross-lingual annotated data with predicate sense and semantic role labels from a single inventory shared across languages could open the door to the development of more robust cross-lingual SRL systems.

One of our most important findings was that with just 500 training examples in a language the performances of a model evaluated in a 0-shot setting was raised by more than 10 points in accuracy, encouraging the study of language adaptation techniques and the development of other small parallel datasets not only for other languages but also for other tasks. Indeed, as future work, we plan to extend the number of languages covered by UniteD-SRL, starting with Arabic and Russian, which are already part of the UN Corpus, and then moving on to integrate low-resourced languages from other parallel corpora.

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.

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 the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Alan Akbik and Yunyao Li. 2016. POLYGLOT: Multilingual semantic role labeling with unified labels. In *Proceedings of ACL-2016 System Demonstrations*, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2019. Cross-lingual transfer of semantic roles: From raw text to semantic roles. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 200–210, Gothenburg, Sweden. Association for Computational Linguistics.

Collin F. Baker. 2014. FrameNet: A knowledge base for natural language processing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 1–5, Baltimore, MD, USA. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Rexhina Blloshmi, Simone Conia, Rocco Tripodi, and Roberto Navigli. 2021. Generating Senses and RoLes: An end-to-end model for dependency- and span-based Semantic Role Labeling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3786–3793. Main Track.

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)*, pages 2487–2500, Online. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 89–97, Boston, Massachusetts, USA. Association for Computational Linguistics.

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)*, pages 152–164, Ann Arbor, Michigan.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021a. Unifying cross-lingual Semantic Role Labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351, Online.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Simone Conia, Riccardo Orlando, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021b. InVeRo-XL: Making cross-lingual Semantic Role Labeling accessible with intelligible verbs and roles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2020. X-SRL: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings* *of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, 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 (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado.

Jan Hajic, Jarmila Panevová, Zdenka Urešová, Alevtina Bémová, Veronika Kolárová, and Petr Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of the second workshop on treebanks and linguistic theories*, volume 9, pages 57–68.

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*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

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*.

Caterina Lacerra, Tommaso Pasini, Rocco Tripodi, and Roberto Navigli. 2021. ALaSca: an Automated approach for Large-Scale Lexical Substitution. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3836–3842. Main Track.

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, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.

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

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. Survey Track.

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

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):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.

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 EMNLP and CoNLL - Shared Task*, CoNLL '12, pages 1–40, Stroudsburg, PA, USA.

Luigi Procopio, Edoardo Barba, Federico Martelli, and Roberto Navigli. 2021. MultiMirror: Neural cross-lingual word alignment for multilingual Word Sense Disambiguation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3915–3921. Main Track.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic. Association for Computational Linguistics.

Carina Silberer and Manfred Pinkal. 2018. Grounding semantic roles in images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2616–2626, Brussels, Belgium.

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 *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177, Manchester, England. Coling 2008 Organizing Committee.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 299–304, Portland, Oregon, USA. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese Treebank. In *Proceedings of the Second Workshop on Chinese Language Processing, SIGHAN 2003, Sapporo, Japan, July 11-12, 2003.*

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia.

A Data Format

UNITED-SRL's annotations will be provided using the CoNLL-style column organization to guarantee the compatibility with this standard. Table 4 shows an example of annotation for an English sentence. Each column represents a layer of annotation organized as follows: column A indicates the token ID. Columns B-F, namely, the inflected form, its lemma, the universal POS, the syntactic dependency relation it is involved in and the head of the relation, follow the same formalism adopted in the Stanza (Qi et al., 2020) NLP library (we used this tool to preprocess the documents). An underscore at the end of a token in column B indicates that the token is part of a multi-word. Column G indicates the VerbAtlas frame of the corresponding verb. Finally, there are as many columns as the number of predicates in the sentence, e.g., H in Table 4. Asterisks in column H indicate head words for dependency-based SRL.

B Annotation Guidelines

Annotators were provided with a dedicated interface for the annotation. We preprocessed the documents of the UN corpus (Ziemski et al., 2016) using the Stanza (Qi et al., 2020) NLP library and provided annotators with sentences annotated with morphological and syntactic information (columns A-F in Table 4).

The predicates to be annotated are already marked in the interface. The steps that the annotators have to follow are:

1. check if the part-of-speech annotation of the sentence is correct;
2. check if there are missing marked verbs;
3. check if there are tokens erroneously marked as verbs;
4. if the sentence is not in English, the annotator should look at the English parallel annotation and try to see if it is possible to annotate predicates in the current sentence with the same frames selected for the English one;
5. select the first verb in the sentence and collect the possible VerbAtlas frames for the lemma of the verb;
6. disambiguate the selected predicate using the collected list of possible frames;
7. collect the argument structure for the selected frame;
8. mark the span of text in the sentence that contains an argument from the selected predicate-argument structure;
9. mark the head of the span (syntactic dependency of the span);
10. repeat the previous 2 steps for each role in the sentence;
11. repeat the last 5 steps (from point 5 to point 10) for all the verbs in a sentence.

Additional guidelines regard the annotation of phrasal verbs that have to be connected with an underscore if they are adjacent or the specific token-ids of distant elements have to be inserted in a specific field of the interface. Named entities are also connected with an underscore. Auxiliary verbs are not annotated.

C Experimental Details

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.

A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
doc1sent1tok1	He	He	PRON	NSUBJ	3	_	AGENT*
doc1sent1tok2	had	have	AUX	AUX	3	_	_
doc1sent1tok3	established	establish	VERB	ROOT	0	ESTABLISH	_
doc1sent1tok4	the	the	DET	DET	5	_	THEME
doc1sent1tok5	post	post	NOUN	OBJ	3	-	THEME*
doc1sent1tok6	of	of	ADP	CASE	7	_	THEME
doc1sent1tok7	Secretary_	Secretary	PROPN	NMOD	5	_	THEME
doc1sent1tok8	of_	of	ADP	CASE	9	_	THEME
doc1sent1tok9	State	State	PROPN	NMOD	7	_	THEME
doc1sent1tok10	.	.	PUNCT	PUNCT	3	_	_

Table 4: An example of annotation for an English sentence. A: token ID. B: word form. C: lemmatized token. D-F: syntactic labels. G: VerbAtlas frames. H: semantic roles for the predicates.

The total number of configurations that we used and reported in the paper is 48. The maximum time for training the XLM-R fine-tuned model is 2 hours.