On the Benefit of Syntactic Supervision for Cross-lingual Transfer in Semantic Role Labeling

Zhisong Zhang, Emma Strubell, Eduard Hovy
Language Technologies Institute, Carnegie Mellon University
zhisongz@cs.cmu.edu, strubell@cmu.edu, hovy@cmu.edu

Abstract

Although recent developments in neural architectures and pre-trained representations have greatly increased state-of-the-art model performance on fully-supervised semantic role labeling (SRL), the task remains challenging for languages where supervised SRL training data are not abundant. Cross-lingual learning can improve performance in this setting by transferring knowledge from high-resource languages to low-resource ones. Moreover, we hypothesize that annotations of syntactic dependencies can be leveraged to further facilitate cross-lingual transfer. In this work, we perform an empirical exploration of the helpfulness of syntactic supervision for cross-lingual SRL within a simple multitask learning scheme. With comprehensive evaluations across ten languages (in addition to English) and three SRL benchmark datasets, including both dependency- and span-based SRL, we show the effectiveness of syntactic supervision in low-resource scenarios.

1 Introduction

The task of semantic role labeling (SRL) annotates predicate-argument structures in text and is thus a desirable output of natural language processing (NLP) pipelines designed to extract information from text (Gildea and Jurafsky, 2002; Palmer et al., 2010). Recent developments in neural architectures (Vaswani et al., 2017) and pre-trained contextualized representations (Devlin et al., 2019; Liu et al., 2019) have greatly improved the performance of SRL systems (Zhou and Xu, 2015; He et al., 2017; Tan et al., 2018; Shi and Lin, 2019). However, most previous work focuses on high-resource English SRL scenarios, and it remains a challenge to extend these approaches, which require plentiful supervised examples, to other languages where training resources may be limited.

A popular approach addressing this challenge is cross-lingual learning: leveraging the shared structures across human languages to transfer knowledge from high-resource languages to low-resource ones. Model transfer, where an SRL model is directly transferred across languages using shared representations (Kozhevnikov and Titov, 2013, 2014; Fei et al., 2020b), is a particularly promising approach thanks to recent developments in multilingual contextualized representations (Lample and Conneau, 2019; Conneau et al., 2020), which have proven effective for cross-lingual transfer (Wu and Dredze, 2019; Pires et al., 2019).

Another common strategy for improving SRL model performance in both high- and low-resource scenarios is incorporating syntactic information. Syntactic analysis was until recently considered a prerequisite for most SRL systems (Gildea and Palmer, 2002; Punyakanok et al., 2008) and has been shown to benefit recent neural models as well (Marcheggiani and Titov, 2017; He et al., 2018; Swayamdipta et al., 2018; Strubell et al., 2018). Despite much work exploring cross-lingual learning and incorporating syntactic information into SRL systems, most such previous work explores these two avenues separately, though there are numerous reasons that carefully incorporating syntax into a cross-lingual system for SRL could provide further benefits: First, whereas SRL annotations are limited to only about a dozen languages, much richer resources are available for syntax, thanks to the development of the Universal Dependencies (UD) framework and accompanying corpora (Nivre et al., 2016b, 2020), which defines syntactic annotations that are consistent across languages, with treebanks in over 100 languages to date. Second, UD treebanks in particular have the potential to increase beneficial sharing of information across languages by providing a unified syntactic structure to ground cross-lingual representations.

Most previous work utilizing syntax for cross-lingual SRL have incorporated syntactic information only as an input to the model, either as sparse

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6229–6246
November 7–11, 2021. ©2021 Association for Computational Linguistics

features (Kozhevnikov and Titov, 2013; Pražák and Konopík, 2017) or as structures for tree encoders (Fei et al., 2020b). These strategies require syntactic pre-processing by an additional model and can suffer from error propagation. In this work, we explore an alternative approach that has yet to be explored in the cross-lingual setting: adopting syntactic annotations as auxiliary supervision and performing multitask learning (Caruana, 1997) together with SRL (Swayamdipta et al., 2018; Strubell et al., 2018; Cai and Lapata, 2019).

To evaluate the extent to which syntactic supervision can help facilitate cross-lingual transfer in SRL, we perform a comprehensive empirical analysis on three SRL benchmark datasets, covering ten languages (in addition to English). We evaluate our models in both zero-shot and semi-supervised scenarios, and on both dependency- and span-based SRL. Highlights of our findings include:

- Training SRL models with syntactic supervision is consistently helpful in low-resource SRL scenarios. (§3.2, §3.3, §3.4, §3.5)
- When lacking direct syntactic annotations for the target language, available treebanks from related languages can be used instead to improve SRL performance (§3.4)
- For span-based SRL, a syntax-aware SRL decoder out-performs BIO-tagging when combined with syntactic training. (§3.5)

Our implementation is available at `https://github.com/zzsfornlp/zmsp/`.

2 Model

We adopt the typical encoder-decoder paradigm for multi-task learning to perform syntactic dependency parsing and SRL together in one model. A shared encoder gives the hidden representations for the input words and each task has its own decoder that takes those shared representations as inputs and predicts task-specific labels. We hypothesize that syntactic training can provide helpful signals for SRL through the shared encoder.

2.1 Encoder

We adopt multilingual pre-trained contextualized models as our encoder, following previous work reporting strong performance for SRL (Shi and Lin, 2019; He et al., 2019; Conia and Navigli, 2020) and cross-lingual learning (Wu and Dredze, 2019; Pires et al., 2019). For an input sequence of words $w_1, \ldots, w_n$, the encoder produces their contextualized representations $h_1, \ldots, h_n$. These pre-trained models take sub-word tokens as input, but our SRL and syntactic data have word-level annotations, so we take the first sub-token of a word as its representation. These representations are then provided to task-specific decoders.

2.2 Syntax Decoder

For the syntactic (dependency) parsing task, we ignore the single-head constraints in training and view it as a pairwise labeling task into the space of dependency labels $\mathcal{R}_d$:

$$p(r_d|w_H, w_M) = \frac{e^{\text{score}_{r_d}(h_H,h_M)}}{\sum_{r' \in \mathcal{R}_d \cup \{\epsilon\}} e^{\text{score}_{r'}(h_H,h_M)}}$$

where $p(r_d|w_H, w_M)$ denotes the probability that the head w_H has a dependency relation r_d to the modifier w_M (or ϵ, which means no syntactic relation). Following Dozat and Manning (2017), we use biaffine modules for the scoring (score_{r_d}), which take the encoder representations and produce relation scores. For training, we use cross-entropy as the objective. Notice that although this type of pairwise formulation is not widely used for syntactic dependencies, it has been shown effective for semantic dependency parsing (Dozat and Manning, 2018). Our main motivation[1] to utilize it here is to make the syntactic task more similar to SRL. In our syntactic parsing evaluation, we find that this method obtains similar results to the head-selection method.

2.3 SRL Decoder

We focus on the end-to-end SRL task, which extracts both the predicates and their arguments (i.e. we do not assume gold predicates unless otherwise noted). For argument extraction, we explore two categories of SRL formalism: dependency-based SRL, which only requires labeling the syntactic head word of an argument, and span-based SRL, which requires labeling full argument spans.

2.3.1 Predicate Identification

Predicate identification is cast as a binary classification task. We use a linear scorer over each word's

[1] Another potential benefit is that certain parameters of the output layers may be shareable between syntactic and SRL decoders. Though in preliminary experiments we did not find obvious improvements with a simple method of stacking another task-specific classification layer and sharing the middle biaffine layers, this could be an interesting direction to explore with better parameter-sharing schemes.

Experiments	Target Languages	SRL Style	Same Frames?	Compatible Roles?	Main SRL Setting
EWT/UPB† (§3.2)	de,fr,it,es,pt,fi	Dependency-based	Yes	Yes	Zero-shot
EWT/FiPB (§3.3)	fi	Dependency-based	No	Yes	Semi-supervised
CoNLL-2009 (§3.4)	cs,zh,es,ca	Dependency-based	No	No	Semi-supervised
OntoNotes (§3.5)	zh,ar	Span-based	No	Yes	Semi-supervised

Table 1: An overview of our experiments. Here "Same Frames?" denotes whether different languages utilize the same semantic frames, and "Compatible Roles?" denotes whether the roles labels are the same. (†UPB is created semi-automatically, while other datasets use directly or are converted from manual annotations.)

encoded representations to judge whether it triggers a semantic frame.

2.3.2 Dependency-based SRL

For dependency-based SRL, the problem can be again formalized as a pairwise labeling task, and we treat it in a similar way as in the syntax decoder:

$$p(r_s|w_P, w_A) = \frac{e^{\text{score}_{r_s}(h_P,h_A)}}{\sum_{r' \in \mathcal{R}_s \cup \{\epsilon\}} e^{\text{score}_{r'}(h_P,h_A)}}$$

Here $p(r_s|w_P, w_A)$ denotes the probability that a predicate w_P takes w_A as an argument with the semantic role r_s (or ϵ, which denotes no semantic relation). Again we use biaffine modules for scoring and cross-entropy as the objective function.

2.3.3 Span-based SRL

Predicting argument spans is usually cast as a sequence labeling problem, with most recent neural SRL models adopting a simple BIO-tagging decoder (Zhou and Xu, 2015; He et al., 2017; Tan et al., 2018; Shi and Lin, 2019). In this work, we further consider a two-step syntax-aware approach (Zhang et al., 2021), where the first step identifies the argument head and a second step decides span boundaries given the head identified in the first step. Here, the first step is exactly the task of dependency-based SRL and we use the same decoder. For the second step, we adopt the span selection method from extractive question answering (Wang and Jiang, 2016; Devlin et al., 2019) and use two classifiers to decide the start and end of the span given the head word.

2.4 Training Scheme

To deal with the multi-task and multilingual scenarios, we adopt a simple training scheme. For each training step, we first sample a task (parsing or SRL), and then a language (source or target). Based on these, we sample a batch of instances from the corresponding dataset and train the model on the selected task. In our experiments, we apply fixed sampling rates for the selection of tasks and languages (1:2 for parsing vs. SRL and 1:1 for source vs. target). In preliminary experiments, we also tried varying sampling rates, but did not find obvious improvements. Exploration of more sophisticated training schemes is left to future work.

3 Experimental Results

3.1 General Settings

We conduct comprehensive experiments with three groups of datasets:[2] 1) English Web Treebank (EWT) (Silveira et al., 2014), Universal Proposition Banks (UPB v1.0) (Akbik et al., 2015, 2016b) and Finnish PropBank (FiPB) (Haverinen et al., 2015); 2) CoNLL-2009 (Hajič et al., 2009); and 3) OntoNotes v5.0 (Hovy et al., 2006; Weischedel et al., 2011). Table 1 gives an overview of the experimental settings for each dataset; please refer to Appendix A for more details.

We take English as the source language[3] and transfer to other target languages. For experiments on UPB and FiPB, we assemble the English SRL dataset with EWT and its SRL annotations from PropBank v3. For CoNLL-2009 and OntoNotes, we utilize the corresponding English sets. For evaluation, we calculate labeled F1 score for arguments. Conventionally, predicate senses are also evaluated for dependency-based SRL. However, cross-lingual transfer of sense disambiguation provides a non-trivial challenge (Akbik et al., 2016a), since it is lexicon-based and language-dependent. Moreover, argument labeling can be more related with dependency syntax, while sense disambiguation is more on the semantic side and semantic-oriented signals (like bilingual dictionaries or parallel corpora) may be more directly effective to enhance cross-lingual transfer. Therefore, in this work we

[2] We focus on PropBank-style SRL annotations since there are more annotations available across different languages than in other formalisms. The explored method can be extended to other formalisms which we leave to future exploration.

[3] Please refer to Appendix C.3 for experiments taking other languages as the source.

Method	DE	FR	IT	ES	PT	FI	AVG
NoSyn	$57.85_{\pm 0.34}$	$51.41_{\pm 0.18}$	$55.79_{\pm 0.42}$	$50.08_{\pm 0.16}$	$52.53_{\pm 0.40}$	$43.78_{\pm 0.57}$	$51.91_{\pm 0.17}$
EnSyn	$57.78_{\pm 0.43}$	$51.64_{\pm 0.16}$	$54.90_{\pm 0.68}$	$50.15_{\pm 0.36}$	$52.65_{\pm 0.22}$	$44.41_{\pm 0.76}$	$51.92_{\pm 0.31}$
TargetSyn	$56.84_{\pm 1.42}$	$57.55_{\pm 1.01}$	$55.78_{\pm 2.04}$	$52.40_{\pm 1.20}$	$56.32_{\pm 1.54}$	$52.32_{\pm 1.20}$	$55.20_{\pm 1.20}$
FullSyn	$59.70_{\pm 0.61}$	$59.38_{\pm 0.37}$	$60.75_{\pm 0.28}$	$55.57_{\pm 0.36}$	$59.78_{\pm 0.28}$	$55.94_{\pm 0.56}$	$58.52_{\pm 0.20}$
SEQ(MLM)	$57.52_{\pm 0.67}$	$52.09_{\pm 0.77}$	$56.56_{\pm 0.55}$	$50.60_{\pm 0.47}$	$53.34_{\pm 0.47}$	$44.56_{\pm 0.80}$	$52.45_{\pm 0.29}$
SEQ(Syn)	$59.73_{\pm 0.39}$	$56.05_{\pm 0.44}$	$61.11_{\pm 0.29}$	$55.32_{\pm 0.29}$	$58.15_{\pm 0.21}$	$55.23_{\pm 0.49}$	$57.60_{\pm 0.15}$
GCN(Gold)	$63.78_{\pm 0.50}$	$56.44_{\pm 0.40}$	$61.96_{\pm 0.73}$	$56.77_{\pm 0.36}$	$59.79_{\pm 0.28}$	$55.29_{\pm 0.33}$	$59.01_{\pm 0.30}$
GCN(Pred)	$61.15_{\pm 0.37}$	$55.44_{\pm 0.36}$	$60.69_{\pm 0.65}$	$54.49_{\pm 0.37}$	$58.09_{\pm 0.27}$	$53.75_{\pm 0.32}$	$57.27_{\pm 0.29}$

Table 2: UPB development Arg-F1(%) scores in the English-to-others zero-shot setting (with mBERT).

focus on arguments and do not perform or evaluate sense disambiguation, following the conventions of span-based SRL.

For syntactic resources, we use either UD treebanks or convert constituency trees to dependencies using Stanford CoreNLP (Manning et al., 2014). In most of our settings, we assume access to multilingual syntax annotations for both source and target languages. We regard this as a practical setting since UD treebanks are available for a wide range of languages and syntactic annotations may be easier to obtain than semantic ones.

We adopt pre-trained multilingual language models (multilingual BERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020)) to initialize our encoders and fine-tune the full models. We use the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 2e-5. We train the models for 100K steps with a batch size around 1024 tokens for each step. All models are trained and evaluated on one GTX 1080 Ti GPU, and training one model usually takes around half a day.

3.2 UPB

UPB annotates[4] target languages with English PropBank frames, which allows us to explore zero-shot experiments without any target SRL training resources. We follow the setting of (Fei et al., 2020a): training the models with English SRL annotations (EWT) and directly applying them to the target languages. In this experiment only we assume predicates are given since UPB is limited to verbal predicates, which leads to discrepancies between source and target predicate annotations. For the syntactic resources, we take the corresponding treebanks (upon which UPB is annotated) from UD v1.4 (Nivre et al., 2016a) and simply include them as additional training data for syntactic supervision.

3.2.1 Comparisons

We first compare several strategies on the usage of syntax, and results on the development set are shown in Table 2. Here we utilize multilingual BERT (mBERT) for the basic encoder. The table is split into three groups:

- **Syn** varies which syntactic resources are used. The four rows denote no syntax (NoSyn), only source (English; EnSyn), only targets (other six languages; TargetSyn) and full syntactic resources (English plus other six; FullSyn). Here, only adding source syntax is not helpful, but target syntax information is generally beneficial. Furthermore, combining both source and target syntax leads to the best results.
- **SEQ** explores a sequential two-stage fine-tuning scheme (Phang et al., 2018; Wang et al., 2019): first training the model with an auxiliary task (syntax or others) and then with the target task (SRL). Using syntactic parsing as the intermediate task can bring clear improvements, but it is slightly worse than the MTL scheme. Here, we also explore a masked language model (MLM) intermediate objective (Devlin et al., 2019) as a baseline, using the raw texts of the UD treebanks. Though it can slightly improve the results, the gains are much smaller than those due to syntax.
- **GCN** uses syntax as inputs. We stack a graph convolutional network (GCN) (Kipf and Welling, 2017) between the encoder and decoders to encode input dependency trees. Specifically, we adopt the architecture of (Marcheggiani and Titov, 2017). Using gold trees in this setting out-performs the MTL strategy. However, when using predicted syntax,[5] error propagation re-

[4] Notice that UPB is created in a semi-automatic way without fully manually validated test sets, but it provides a test-bed for evaluating zero-shot cross-lingual transfer.

[5] We obtain predicted syntax trees with our own BERT-based parsers, which achieve strong results (dev-LAS%): 89.6(de), 91.6(fr), 93.7(it), 89.7(es), 92.1(pt) and 93.2(fi).

EnSRL	Syntax	0.1K		1K		10K	
		Dev	Test	Dev	Test	Dev	Test
No	No	$43.25_{\pm0.50}$	$44.76_{\pm0.82}$	$69.02_{\pm0.32}$	$70.29_{\pm0.54}$	$82.51_{\pm0.40}$	$82.91_{\pm0.35}$
No	Yes	$58.75_{\pm0.49}$	$58.32_{\pm0.80}$	$73.91_{\pm0.33}$	$74.06_{\pm0.30}$	$82.71_{\pm0.13}$	$83.24_{\pm0.17}$
Yes	No	$60.76_{\pm0.56}$	$60.42_{\pm0.89}$	$73.23_{\pm0.37}$	$73.67_{\pm0.68}$	$\mathbf{82.92}_{\pm0.30}$	$\mathbf{83.35}_{\pm0.27}$
Yes	Yes	$\mathbf{68.36}_{\pm0.17}$	$\mathbf{67.22}_{\pm0.39}$	$\mathbf{75.98}_{\pm0.14}$	$\mathbf{76.13}_{\pm0.18}$	$82.73_{\pm0.18}$	$83.34_{\pm0.22}$

Table 3: FiPB Arg-F1(%) scores in English/Finnish settings (with different numbers of Finnish SRL sentences). "EnSRL" indicates whether using English SRL, and "Syntax" denotes whether using syntactic annotations.

Method	DE	FR	IT	ES	PT	FI
mBERT/NoSyn	55.0	49.9	53.1	49.7	51.0	44.7
mBERT/FullSyn	57.5	56.8	58.3	56.2	58.9	54.4
XLM-R/NoSyn	57.5	50.8	54.3	51.5	53.1	51.8
XLM-R/FullSyn	60.2	56.6	**60.6**	57.3	**59.5**	**59.9**
Fei et al. (2020a)	**65.0**	**64.8**	58.7	**62.5**	56.0	54.5

Table 4: UPB test Arg-F1(%) scores in the English-to-others zero-shot setting (averaged over five runs).

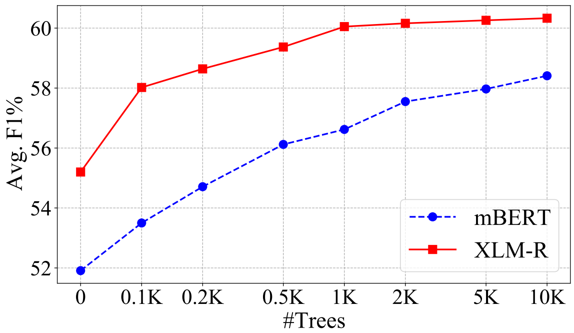

Figure 1: Averaged UPB development results versus number of trees (in log scale) used per language.

duces the observed benefit. The MTL scheme is an attractive alternative strategy considering its competitive performance and model simplicity.

3.2.2 Main Results

The test results are listed in Table 4. Similar to the trends in the development sets, including syntactic signals brings clear improvements, especially for the more distant Finnish language. Using XLM-R, which is pre-trained on more data than mBERT, is also helpful,[6] upon which syntax can still bring further benefits. We also compare with the results from (Fei et al., 2020a), which translates and projects source SRL instances to target languages for training. The translation-based method performs strongly for German, French and Spanish. Considering that German and French are commonly used languages in machine translation research, availability of high-quality translation systems may be one of the contributing factors. Our syntax-enhanced models are generally competitive for other languages. It would be interesting to further explore the combination of translation and syntax in future work.

3.2.3 Varying Treebank Sizes

We further vary the number of available syntax trees for the auxiliary parsing task, for which Figure 1 shows the results. We randomly sample a fixed number of trees for each of the languages (both source and target) and again include them in training. The results indicate that we do not need the full treebanks to obtain good results. Especially with XLM-R, 1K trees from each language can already lead to gains comparable to the 10K case.

3.3 FiPB[7]

Similar to the experiments on UPB, we take English SRL annotations from EWT as the source. FiPB adopts (almost) the same argument role set[8] as the English ones and we use a shared SRL decoder for both languages. In preliminary experiments, we find that this sharing strategy performs better than using separate, language-specific decoders. For syntax, we again take corresponding English and Finnish treebanks from UD v1.4.

3.3.1 Results

The main results on FiPB are listed in Table 3. In the lowest-resource scenario (0.1K Finnish SRL sentences), both English SRL and syntax are quite helpful, and combining them leads to further improvements. The trend is similar if given 1K target SRL annotations, but the gaps decrease. Finally,

[6] Due to better performance, XLM-R is used in the remaining experiments.

[7] Starting from this experiment, we focus on the semi-supervised setting where varying amounts of target SRL annotations are used during training.

[8] Except for two Finnish specific roles (ArgM-CSQ and ArgM-PRT) which only account for around 2% of labels.

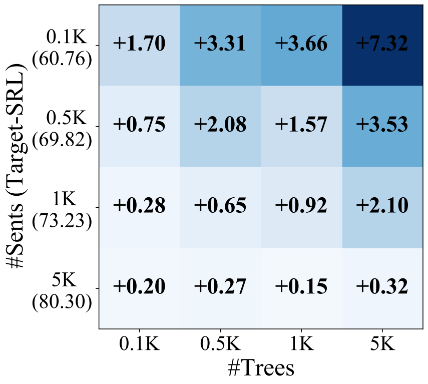

Figure 2: Improvements (F1 scores on FiPB development set) over no-syntax baselines (shown in parentheses at the y-axis) with various training sizes.

when given enough target training instances as in the 10K scenario, the gains due to extra resources (either English SRL or syntax) are negligible. In this case, the model may have already learned most of the patterns from rich target SRL annotations.

3.3.2 Varying Training Sizes

We further vary both syntax and target-SRL training sizes, and the influence on model performance is shown in Figure 2. Here, all the models are trained using all English SRL and varying amounts of Finnish SRL sentences. The numbers in parentheses at y-axis show the F1 scores of baseline models without syntax. As expected, syntax is more helpful when we have less target-SRL and more syntactic resources (towards the right corner in the figure). When we have more target SRL annotations, syntactic resources become less helpful. Nevertheless, in low-resource scenarios, even small quantities of syntactic annotation can bring clear improvements.

3.3.3 Analysis

We further perform analysis on the development results in the 1K case, as shown in Table 5. In the first group of role label breakdowns, adding syntax particularly helps core arguments while adding English SRL helps more on non-core arguments. Finally, combining both leads to the best results overall. In the second group, we break down arguments by their syntactic distance to the predicates. The results show that syntactic supervision is still beneficial when the predicate and the argument are two edges away (d=2). However, when syntax distance is larger, direct syntactic supervision becomes less helpful.

	Base	+Syntax	+EnSRL	+Both
ARG0 (12%)	75.72	$\underline{81.08}$	80.11	**82.94**
ARG1 (36%)	77.21	$\underline{83.26}$	81.29	**84.24**
ARG2 (14%)	65.85	$\underline{70.41}$	69.93	**71.84**
ARGM (35%)	60.76	64.73	$\underline{65.46}$	**68.33**
d=1 (84%)	75.62	$\underline{78.45}$	78.05	**80.35**
d=2 (15%)	58.41	$\underline{63.26}$	61.31	**64.22**
d>2 (1%)	24.07	$\underline{25.84}$	**28.61**	25.39
$\xleftarrow{\text{nmod}}$ (23%)	61.01	62.95	$\underline{64.69}$	**66.80**
$\xleftarrow{\text{nsubj}}$ (14%)	85.72	**88.15**	85.35	$\underline{87.89}$
$\xleftarrow{\text{dobj}}$ (13%)	90.10	**93.37**	91.20	$\underline{93.25}$
$\xleftarrow{\text{advmod}}$ (9%)	64.40	65.17	$\underline{68.18}$	**69.04**
$\xrightarrow{\text{acl}}$ (4%)	83.44	85.17	$\underline{86.20}$	**87.13**
$\xrightarrow{\text{cop}}$ (4%)	91.63	**97.79**	94.24	$\underline{96.63}$
$\xleftarrow{\text{xcomp}}$ (3%)	70.83	$\underline{75.21}$	73.01	**77.29**
$\xleftarrow{\text{aux}}$ (3%)	95.21	$\underline{97.53}$	95.67	**97.97**
$\xleftarrow{\text{nsubj}}\xrightarrow{\text{cop}}$ (3%)	95.75	$\underline{98.37}$	94.94	**98.60**
$\xleftarrow{\text{advcl}}$ (3%)	51.80	57.83	$\underline{68.64}$	**70.31**

Table 5: Analysis (F1% breakdown) on the FiPB development set (1K setting). The first block denotes breakdowns on argument roles, the second denotes syntactic distance between predicate and argument words, and the third denotes the syntactic path between them. The numbers in parentheses denote percentages. **Bold** and $\underline{\text{underlined}}$ numbers indicate the best and second-best results respectively.

In the third group, we look at the labeled syntactic paths between the arguments and the predicates. For example, "$\xleftarrow{\text{nmod}}$" denotes that the argument is a syntactic modifier of the predicate and the dependency relation is "nmod", while "$\xrightarrow{\text{acl}}$" denotes the argument is the syntactic head of the predicate with the dependency relation of "acl". We show the results on top-ten frequent paths, which cover around 80% of all the arguments. According to the breakdown results, the syntactic supervision helps more on the edges of subject, direct object and some functional relations (like copula), while English SRL is more beneficial on the more semantic links, such as adverbial words and clauses. This agrees with our analysis on the argument roles: the syntax helps more on the core arguments, which are usually directly connected as subjects or objects, while English SRL helps more on "ArgM"s, which tends to be adverbial.

3.4 CoNLL-2009

The original SRL annotations of CoNLL-2009 are based on language-specific syntax, causing the argument head words to disagree with UD conventions. We thus follow Pražák and Konopík

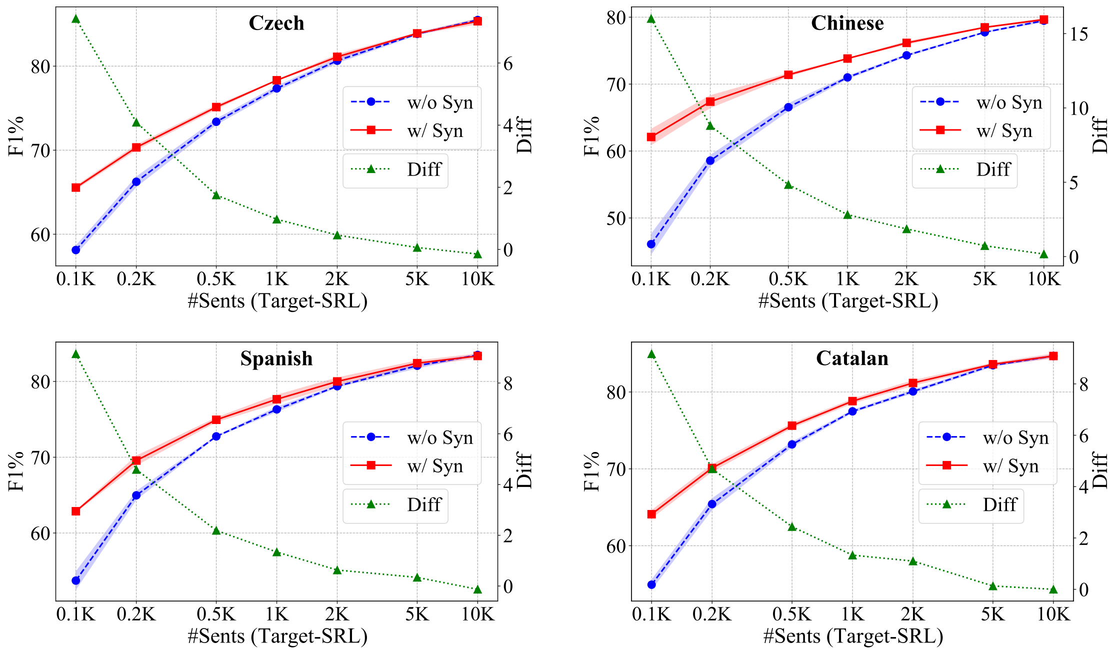

Figure 3: Test results of CoNLL-2009 semi-supervised experiments. Here the x-axis denotes the number (in log scale) of target-SRL annotated sentences available for training. The results are averaged over three runs, and the shaded areas indicate the ranges of standard deviations.

(2017) and convert[9] them to UD-based argument heads. We take five languages from CoNLL-2009 where we can obtain corresponding UD trees for the source sentences. For English and Chinese, we use Stanford CoreNLP to convert the constituency trees to dependencies, while for Czech, Spanish and Catalan, we assign dependency trees from corresponding UD v2.7 (Zeman et al., 2020) treebanks.[10] Moreover, since role labels are not compatible across languages in CoNLL-2009, we utilize separate SRL decoders for source and target languages. In preliminary experiments, we tried several parameter-sharing strategies but did not find obvious improvements. Explorations on more complex sharing and regularization methods (Jindal et al., 2020) are left to future work.

3.4.1 Results

We again take English as the resource-rich source language and the other four as lower-resource targets. We run experiments separately for each target language, which means all experiments are bilingual (with the exception of XLM-R pretraining). For syntax, since different languages have different treebank sizes, we randomly sample 10K trees for both source and target languages. The results are shown in Figure 3. The patterns are consistent among all languages and similar to previous experiments on FiPB: syntax is clearly helpful in low-resource scenarios, but as we have access to more target SRL annotations, the gaps decrease and finally diminish in the high-resource scenarios.

Syntax	Spanish		Catalan	
	LAS%	ArgF1%	LAS%	ArgF1%
NoSyntax	-	54.6$_{\pm 1.2}$	-	54.0$_{\pm 0.9}$
Spanish	**86.9**$_{\pm 0.1}$	**63.6**$_{\pm 0.7}$	67.9$_{\pm 7.1}$	59.0$_{\pm 0.9}$
Catalan	77.0$_{\pm 1.0}$	61.0$_{\pm 0.9}$	**85.7**$_{\pm 0.4}$	**63.9**$_{\pm 0.2}$
French	64.2$_{\pm 9.1}$	57.9$_{\pm 0.8}$	58.7$_{\pm 2.0}$	55.4$_{\pm 0.8}$
Italian	66.1$_{\pm 3.6}$	58.1$_{\pm 0.5}$	56.4$_{\pm 6.4}$	57.0$_{\pm 0.9}$
Portuguese	69.5$_{\pm 3.0}$	57.6$_{\pm 1.7}$	58.6$_{\pm 5.7}$	56.8$_{\pm 0.8}$

Table 6: Development results of Spanish and Catalan CoNLL-2009 semi-supervised experiments (0.1K target-SRL) using syntax from different languages.

3.4.2 Using Other Treebanks

We further explore the scenarios where we do not directly have syntactic annotations for the target language. Considering that the parsing task can also benefit from cross-lingual transfer, we can utilize treebanks from nearby languages for syntactic supervision. We take Spanish and Catalan (the 0.1K target SRL case) for this analysis and the results are shown in Table 6. We further ex-

[9] Please refer to Appendix B for the conversion details.

[10] PDT for Czech and AnCora for Spanish and Catalan.

Method	Syntax	0.1K		1K		10K	
		Dev	Test	Dev	Test	Dev	Test
Chinese							
BIO	No	$50.59_{\pm 0.38}$	$49.67_{\pm 0.28}$	$62.81_{\pm 0.15}$	$62.58_{\pm 0.24}$	$70.04_{\pm 0.18}$	$70.37_{\pm 0.08}$
TwoStep	No	$49.26_{\pm 0.44}$	$48.53_{\pm 0.67}$	$63.08_{\pm 0.05}$	$63.22_{\pm 0.12}$	$70.43_{\pm 0.12}$	$70.80_{\pm 0.12}$
BIO	Yes	$53.47_{\pm 0.24}$	$52.81_{\pm 0.20}$	$64.55_{\pm 0.21}$	$64.49_{\pm 0.11}$	$70.26_{\pm 0.18}$	$70.58_{\pm 0.22}$
TwoStep	Yes	$\mathbf{56.16}_{\pm 0.14}$	$\mathbf{55.52}_{\pm 0.23}$	$\mathbf{65.36}_{\pm 0.22}$	$\mathbf{65.65}_{\pm 0.15}$	$\mathbf{70.66}_{\pm 0.13}$	$\mathbf{71.04}_{\pm 0.12}$
Arabic							
BIO	No	$46.14_{\pm 0.73}$	$44.87_{\pm 1.10}$	$59.72_{\pm 0.51}$	$58.80_{\pm 0.26}$	$69.67_{\pm 0.12}$	$67.87_{\pm 0.42}$
TwoStep	No	$46.33_{\pm 0.26}$	$45.28_{\pm 0.48}$	$59.91_{\pm 0.39}$	$59.53_{\pm 0.62}$	$70.17_{\pm 0.25}$	$\mathbf{68.46}_{\pm 0.30}$
BIO	Yes	$49.23_{\pm 0.27}$	$49.13_{\pm 0.31}$	$61.36_{\pm 0.23}$	$60.89_{\pm 0.21}$	$70.02_{\pm 0.25}$	$67.92_{\pm 0.22}$
TwoStep	Yes	$\mathbf{51.68}_{\pm 0.33}$	$\mathbf{51.50}_{\pm 0.50}$	$\mathbf{61.81}_{\pm 0.45}$	$\mathbf{61.70}_{\pm 0.61}$	$\mathbf{70.19}_{\pm 0.16}$	$68.28_{\pm 0.31}$

Table 7: OntoNotes Arg-F1(%) scores in English-sourced semi-supervised settings (with different numbers of target SRL training sentences). "BIO" indicates using a BIO-based sequence labeling decoder and "TwoStep" denotes the syntactically-aware decoding method which first extracts head words then decides span boundaries.

plore three Romance languages: French, Italian and Portuguese. As expected, directly using target-language syntax obtains the best results. Spanish and Catalan, which are closely related languages, benefit each other the most. Nevertheless, compared with the NoSyntax baseline, syntactic information from all these languages are helpful. This result is of practical interest when transferring to a truly low-resource language where syntactic annotations may also be limited. Finding a related language with rich syntactic resources for auxiliary training signals is a promising way to improve performance.

3.5 OntoNotes

Finally, we turn to span-based SRL where the extraction of full argument spans is required. Utilizing OntoNotes annotations, we still take English as the source and Chinese or Arabic as the target. Similar to FiPB, the argument roles are compatible with PropBank-style English roles and we use a shared SRL decoder for both the source and target languages. We adopt data splits from the CoNLL12 shared task (Pradhan et al., 2012). Similar to those of CoNLL-2009, for English and Chinese, we convert constituencies to dependencies with Stanford CoreNLP. For Arabic, we assign dependency trees from Arabic-NYUAD (Taji et al., 2017) treebank of UD v2.7.

3.5.1 Results

In this experiment, we specifically compare two SRL decoders. The first one casts the task as a BIO-based sequence labeling problem. We further add a standard linear-chain conditional random field (CRF) (Lafferty et al., 2001), which we found consistently helpful in preliminary experiments. The other one is the two-step decoder described in §2.3. As shown in Table 7, the trends are similar for both Chinese and Arabic. With regard to auxiliary syntactic supervision, we find similar trends to previous experiments: in low-resource scenarios, syntactic supervision is beneficial for both decoders, but as the availability of target SRL resources increases, the gaps become smaller until diminished. The more interesting comparisons are between the two decoders: when not using syntactic supervision, their performances are comparable; but when trained with auxiliary signals from syntax, the syntax-aware two-step decoder performs better than the BIO tagger, especially in low-resource cases. Please refer to Appendix C.4 for more detailed error analysis.

3.5.2 Syntax with Genre Mismatches

Since English and Chinese OntoNotes also annotate six different genres of text, we further explore scenarios where the syntax and SRL datasets have genre mismatches. We still take all English instances for multilingual training, but split the Chinese corpus according to genres, including broadcast conversation (bc), broadcast news (bn), magazine (mz), newswire (nw), telephone conversation (tc) and web (wb). We focus on the low-resource scenario where 0.1K Chinese SRL sentences on the target genre are available. The development results are shown in Figure 4. When the genre of syntactic supervision matches the target SRL, the improvements are the largest. Nevertheless, even in the case of genre mismatches, syntax can still be beneficial, especially within similar genres. We further find a positive correlation (Pearson corre-

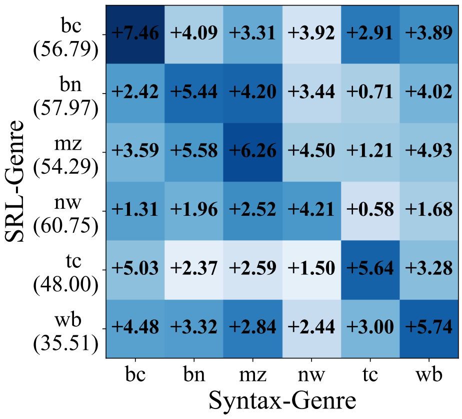

Figure 4: Improvements (F1 scores on Chinese OntoNotes development set) over no-syntax baselines (shown in parentheses at the y-axis) with syntactic supervision of different genres.

lation is 0.73; Spearman is 0.78) between these improvements and genre similarities calculated by the centroids of mBERT representations (Aharoni and Goldberg, 2020). This may provide a mechanism for selecting the most beneficial syntactically annotated instances.

4 Related Work

4.1 Cross-lingual SRL

Recently there have been increasing interests in cross-lingual SRL, where SRL annotations from high-resource languages are utilized to help low-resource ones. One straightforward method is data transfer, using either annotation projection (Yarowsky and Ngai, 2001) or translation (Tiedemann and Agić, 2016) to create SRL instances for target languages (Padó and Lapata, 2009; Akbik et al., 2015; Aminian et al., 2019; Fei et al., 2020a). A related idea is to utilize parallel corpus to introduce cross-lingual signals (Daza and Frank, 2019, 2020; Cai and Lapata, 2020). Another method is model transfer which we focus in this work: directly applying the model trained with source languages to target ones (Kozhevnikov and Titov, 2013, 2014; Fei et al., 2020b). This method requires shared representations for different languages, which recent multilingual pre-trained encoders (Devlin et al., 2019; Conneau et al., 2020) are good at (Wu and Dredze, 2019; Pires et al., 2019). We take these multilingual encoders as the backbone of our models, since they have been shown effective for SRL across multiple languages (He et al., 2019; Conia and Navigli, 2020). In the semi-supervised settings where some target SRL annotations are available, the models are actually trained in a polyglot way (Mulcaire et al., 2018, 2019). Our work differs in the focus on low-resource scenarios.

Cross-lingual SRL still remains challenging due to data scarcity and annotation heterogeneity. To create multilingual SRL data, Akbik et al. (2015) utilize parallel corpus to create target SRL annotations with filtered projection, and Daza and Frank (2020) create the X-SRL dataset through translation and projection with multilingual contextualized representations, offering a multilingual parallel SRL corpus. To deal with heterogeneous SRL formalism, Jindal et al. (2020) adopt an argument regularizer to encourage cross-lingual argument matching, and Conia et al. (2021) introduce a unified model, which may implicitly learn to align heterogeneous linguistic resources.

4.2 Syntax and SRL

Even with recent developments in neural network modeling, with which syntax-agnostic models have been shown to match linguistically-informed counterparts (Marcheggiani et al., 2017; Cai et al., 2018), syntax has also been found helpful for SRL (Marcheggiani and Titov, 2017; Swayamdipta et al., 2018; Strubell et al., 2018; He et al., 2018; Cai and Lapata, 2019; Shi et al., 2020; Fei et al., 2021). In this work, we further explore the helpfulness of syntax for cross-lingual SRL. While previous work on this topic mainly uses syntax as input features (Kozhevnikov and Titov, 2013; Pražák and Konopík, 2017; Fei et al., 2020b), we adopt a simpler strategy utilizing it as an auxiliary training signal via multitask learning (Caruana, 1997; Ruder, 2017), which has also been found beneficial for monolingual SRL (Swayamdipta et al., 2018; Strubell et al., 2018; Cai and Lapata, 2019).

5 Conclusion

In this work, we provide a comprehensive empirical exploration of the helpfulness of syntactic supervision for cross-lingual SRL. With extensive evaluations across a variety of datasets and settings, we show that auxiliary syntactic signals are generally beneficial, especially in low-resource SRL cases. We hope that this work can shed some light on the relations between syntax and SRL in the cross-lingual scenarios.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

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, Xinyu Guan, and Yunyao Li. 2016a. Multilingual aliasing for auto-generating proposition Banks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3466–3474, Osaka, Japan. The COLING 2016 Organizing Committee.

Alan Akbik, Vishwajeet Kumar, and Yunyao Li. 2016b. Towards semi-automatic generation of proposition Banks for low-resource languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 993–998, Austin, Texas. 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.

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*, pages 2753–2765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rui Cai and Mirella Lapata. 2019. Syntax-aware semantic role labeling without parsing. *Transactions of the Association for Computational Linguistics*, 7:343–356.

Rui Cai and Mirella Lapata. 2020. Alignment-free cross-lingual semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3883–3894, Online. Association for Computational Linguistics.

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

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. 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. Association for Computational Linguistics.

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.

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. 2019. Translate and label! an encoder-decoder approach for cross-lingual 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 603–615, Hong Kong, China. 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.

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.

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

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 (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. 2021. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, Online. Association for Computational Linguistics.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020a. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Hao Fei, Meishan Zhang, Fei Li, and Donghong Ji. 2020b. Cross-lingual semantic role labeling with model transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2427–2437.

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

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Katri Haverinen, Jenna Kanerva, Samuel Kohonen, Anna Missilä, Stina Ojala, Timo Viljanen, Veronika Laippala, and Filip Ginter. 2015. The finnish proposition bank. *Language Resources and Evaluation*, 49(4):907–926.

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 (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

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)*, pages 5350–5359, Hong Kong, China. Association for Computational Linguistics.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2061–2071, Melbourne, Australia. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Ishan Jindal, Yunyao Li, Siddhartha Brahma, and Huaiyu Zhu. 2020. CLAR: A cross-lingual argument regularizer for semantic role labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3113–3125, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1200, Sofia, Bulgaria. Association for Computational Linguistics.

Mikhail Kozhevnikov and Ivan Titov. 2014. Cross-lingual model transfer using feature representation projection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–585, Baltimore, Maryland. Association for Computational Linguistics.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky.

2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

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)*, pages 411–420, Vancouver, Canada. Association for Computational Linguistics.

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

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. 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 3912–3918, Minneapolis, Minnesota. Association for Computational Linguistics.

Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith. 2018. Polyglot semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 667–672, Melbourne, Australia. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, and et al. 2016a. Universal dependencies 1.4. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016b. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

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*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Ondřej Pražák and Miloslav Konopík. 2017. Cross-lingual SRL based upon Universal Dependencies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 592–600, Varna, Bulgaria. INCOMA Ltd.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Tianze Shi, Igor Malioutov, and Ozan Irsoy. 2020. Semantic role labeling as syntactic dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7551–7571, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

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*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep semantic role labeling with self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jörg Tiedemann and Zeljko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research*, 55:209–248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*, pages 54–63.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, and et al. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2021. Comparing span extraction methods for semantic role labeling. In *Proceedings of the 5th Workshop on Structured Prediction for NLP (SPNLP 2021)*, pages 67–77, Online. Association for Computational Linguistics.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. 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 1127–1137, Beijing, China. Association for Computational Linguistics.

Appendices

A Detailed Experiment Settings

A.1 Datasets

Table 8 presents the statistics of the datasets that we utilize. The details of each (group of) dataset are described in the following.

EWT/UPB/FiPB is the group where we assemble the English SRL dataset with English Web Treebank[11] (EWT) and PropBank v3[12], and utilize it as the source annotations. We take Universal Proposition Banks[13] (UPB v1.0) (Akbik et al., 2015, 2016b) and Finnish PropBank[14] (FiPB) (Haverinen et al., 2015) for the targets. UPB annotates target langauges with English PropBank frames and role labels. This allows zero-shot cross-lingual learning, which is our main setting for experiments with UPB. In the UPB experiment only we assume predicates are given since there are discrepancies between source and target predicate annotations. In experiments with FiPB (as well as CoNLL-2009 and OntoNotes), we focus on semi-supervised multilingual scenarios with end-to-end models that perform both predicate identification and argument labeling. FiPB is a collection of semantic frames built on top of the Turku Dependency Treebank (TDT). The frames are Finnish specific, but the role labels are (almost) the same as the PropBank ones (Arg0, Arg1, ..., ArgM-*). FiPB defines only two additional ArgMs: CSQ (consequence) and PRT (phrasal marker).

CoNLL-2009 shared task[15] (Hajič et al., 2009) provides SRL resources for a variety of languages[16,17]. Since they are built from different language-specific datasets, there are no consistent predicate and argument role labels across all languages (though there are shared ones between certain language pairs, like English-Chinese and Spanish-Catalan). Moreover, the dependency-based SRL annotations are based on language-specific dependencies. To further encourage shared structures, we convert them to the ones based on UD. Details of the conversion are described in the next section.

OntoNotes annotates a large corpus[18] in three languages (English, Chinese and Arabic) with various layers of structural information. We take the SRL annotations from it for our experiments. For English, we utilize the data[19] from (Pradhan et al., 2013), while for Chinese and Arabic, we directly use those provided by CoNLL12[20]. For all the languages, we also follow the data splittings of CoNLL12. Similar to FiPB, the SRL annotations in OntoNotes utilize language-specific frames but compatible argument role sets.

A.2 Hyper-parameters

Without specifications, we use pre-trained multilingual language models (mBERT or XLM-R) to initialize the encoders and fine-tune the full models in our experiments. The parameter numbers of the full models are 185M and 285M, for those with mBERT and XLM-R respectively. For the hyper-parameter settings, we mainly follow common practice, and only slightly tune them in preliminary experiments[21]. We apply dropout rates of 0.1 to the encoder and 0.2 to the decoders. We use Adam as the optimizer with an initial learning rate of 2e-5. The learning rate is linearly decayed towards 2e-6 through the training process. The models are trained for 100K steps, where each step contains a batch of around 1024 tokens. We evaluate the model on the development set every 1K steps and the best model is selected by validation results. For zero-shot experiments, we simply validate with the source development set. For semi-supervised experiments, we use the target language, but down-sample the original development set to 10% of the target training size. All the training and evaluations are performed on one GTX 1080 Ti GPU. Training one model takes around half a

[11] https://catalog.ldc.upenn.edu/LDC2012T13

[12] https://github.com/propbank/propbank-release

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

[14] https://github.com/TurkuNLP/Finnish_PropBank/tree/data

[15] https://ufal.mff.cuni.cz/conll2009-st/

[16] https://catalog.ldc.upenn.edu/LDC2012T03

[17] https://catalog.ldc.upenn.edu/LDC2012T04

[18] https://catalog.ldc.upenn.edu/LDC2013T19

[19] https://cemantix.org/data/ontonotes.html

[20] https://conll.cemantix.org/2012/data.html

[21] The ranges are: learning rate ∈ {1e-5, 2e-5, 4e-5}, dropout ∈ {0.1, 0.2}, batch-size ∈ {512, 1024, 2048}, steps ∈ {100K, 150K, 200K}. We do not try all combinations but decide by random sampling as well as heuristics, due to computation limitations. We find that the hyper-parameters do not influence the results much if set in reasonable ranges.

	Train			Development			Test		
	Sent.	Pred.	Arg.	Sent.	Pred.	Arg.	Sent.	Pred.	Arg.
EWT									
English (EN)	12.5K	40.5K	122.0K	2.0K	5.0K	14.6K	2.1K	4.8K	14.1K
UPB									
German (DE)	14.1K	21.3K	52.2K	0.8K	1.2K	2.8K	1.0K	1.3K	3.3K
French (FR)	14.6K	29.3K	39.0K	1.6K	3.0K	4.1K	0.3K	0.6K	0.9K
Italian (IT)	12.8K	25.6K	52.7K	0.5K	1.0K	2.1K	0.5K	1.0K	2.2K
Spanish (ES)	28.5K	73.3K	149.6K	3.2K	8.3K	16.9K	2.0K	5.4K	11.4K
Portuguese (PT)	7.5K	16.8K	33.1K	0.9K	2.1K	4.2K	0.9K	2.1K	4.2K
Finnish (FI)	12.2K	25.6K	54.3K	0.7K	1.5K	3.1K	0.6K	1.5K	3.1K
FiPB									
Finnish (FI)	12.2K	27.4K	72.1K	0.7K	1.6K	4.1K	0.6K	1.5K	4.1K
CoNLL-2009									
English (EN)	39.3K	179.0K	393.7K	1.3K	6.4K	13.9K	2.4K	10.5K	23.3K
Czech (CS)	38.7K	414.2K	365.9K	5.2K	55.5K	49.2K	4.2K	44.6K	39.3K
Chinese (ZH)	22.3K	102.8K	231.9K	1.8K	8.1K	18.6K	2.6K	12.3K	27.7K
Spanish (ES)	14.3K	43.8K	99.1K	1.7K	5.1K	11.6K	1.7K	5.2K	11.8K
Catalan (CA)	13.2K	37.4K	84.4K	1.7K	5.1K	11.5K	1.9K	5.0K	11.3K
OntoNotes									
English (EN)	75.2K	188.9K	622.5K	9.6K	23.9K	78.1K	9.5K	24.5K	80.2K
Chinese (ZH)	36.5K	117.1K	365.3K	6.1K	16.6K	51.0K	4.5K	15.0K	46.6K
Arabic (AR)	7.4K	20.0K	65.8K	0.9K	2.5K	8.0K	1.0K	2.3K	7.6K

Table 8: Statistics of the datasets. The entries denote numbers of sentences (Sent.), predicates (Pred.) and arguments (Arg.).

day while decoding is fast with several hundreds of sentences processed per second. All the results reported in this work are averaged over three (for most ablation studies) or five (for most test results) runs. The evaluation of arguments follows the standard evaluation script of `srl-eval.pl` [22].

B UD-based Conversion for CoNLL-2009

The SRL annotations of argument heads in CoNLL-2009 are based on Language-Specific Dependency (LSD) trees rather than Universal Dependencies (UD). To convert argument heads between different syntactic formalism, we adopt a simple path-based method. Assuming that for a predicate p, it has an argument whose head is a according to the original tree, the conversion aims to find a new head according to the new tree:

1. In the new tree, find the lowest common ancestor c of the predicate p and the original argument head a.
2. Go down from c to a in the new tree, locate the first word (except for the predicate p) that is a descendant of a (or a itself) in the original tree and make it the new head.

We will illustrate this procedure with the example in Figure 5. Here, the predicate is the verb "ran" and it has an "ArgM-LOC" argument, whose full span is "in the park". According to the language-specific dependency tree, the word "in" is the direct child of the verb and thus becomes the argument head. Nevertheless, according to UD, the content word "park" is the direct child and we want to convert the argument head to it. Firstly, we find the lowest common ancestor of "ran" (the predicate) and "in" (the old argument head) in the UD tree, which is "ran" itself. Then we go down from this ancestor ("ran") towards the old argument head ("in"): the visiting path should be $ran \rightarrow park \rightarrow in$. We find that "park" is the first word that is a descendant of "in" in the original tree and therefore "park" is assigned as the new argument head.

In this work, we take five languages from CoNLL-2009: English, Czech, Chinese, Spanish, Catalan, for which we can obtain or convert to gold UD trees[23]. Table 9 gives some results on the

[22] https://www.cs.upc.edu/~srlconll/soft.html

[23] For Chinese and English, we use CoreNLP to convert from constituencies to UD. For Czech, we use UD_PDT, while

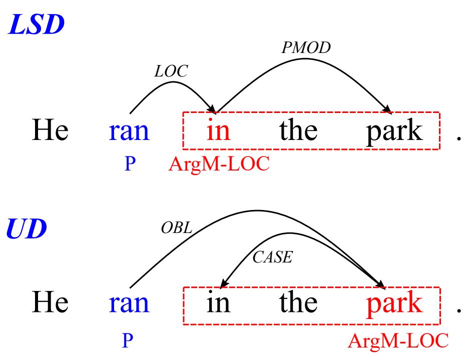

Figure 5: An example for the conversion between language-specific dependencies (LSD) and universal dependencies (UD). For brevity, we only show the important dependency edges. Here, "ran" is the predicate (P) and the argument head is "in" with LSD and "park" with UD. The conversion between these two can be done by comparing the syntactic paths and descendants with the old and new trees.

Language	UAS%	Arg-Agree%	Roundtrip-Agree%
English	50.88	72.27	99.05
Czech*	46.36	97.00	73.38
Chinese	60.02	81.94	99.87
Spanish	58.01	69.92	100.00
Catalan	58.99	72.61	100.00

Table 9: Agreements between Language-Specific Dependencies (LSD) and Universal Dependencies (UD). Here, "UAS" denotes the unlabeled attachment scores when comparing LSD and UD trees, "Arg-Agree" denotes the agreement rates on argument heads between original argument heads and those converted to UD, while "Roundtrip-Agree" denotes the agreement rates with round-trip styled conversions: first converting from LSD to UD and then converting back to LSD. (* Czech is a special case where the original argument heads seem to mostly agree with UD.)

agreements between different syntactic formalism. Although on overall syntactic attachments, LSD disagrees much with UD (the highest UAS is 60% for Chinese), the argument head agreement rates are much higher (the lowest argument agreement rate is around 70% for Spanish). If adopting our converting method, a round-trip styled conversion (converting from LSD to UD and then back to LSD) can almost recover all the arguments, showing the effectiveness of our method. Notice that Czech is an exception where the original argument heads seem to already mostly follow the UD trees.

We further perform SRL experiments with different syntactic formalism. The settings are the same as our main experiments on CoNLL-2009. Here, we take full English SRL and 1K target SRL sentences. Table 10 lists the target development SRL results. Similar to the results in our main experiments, syntactic supervision is beneficial for all languages, and this holds true for both the original language-specific dependencies and the universal dependencies. Interestingly, using original syntax trees and argument heads performs the best, especially for Spanish and Catalan. Through error analysis, we find that for these two languages, the "LSD+Orig." model is much better than the "UD+UD" model mainly on arguments whose original head word is preposition (5 F1 points better for Catalan and 3 points better for Spanish). The reason might be that prepositional word types appear more frequently than content words like nouns and proper nouns, and might be easier to extract if adopting LSD and using prepositions as argument heads, especially at low-resource scenarios.

Syntax	SRL	Czech	Chinese	Spanish	Catalan
No	Orig.	77.51	72.05	77.42	77.57
No	UD	-	71.79	76.83	76.88
UD	Orig.	78.48	74.75	78.22	78.89
UD	UD	-	75.02	78.47	78.64
LSD	Orig.	**78.94**	**75.11**	**79.52**	**80.02**
LSD	UD	-	74.45	78.13	78.78

Table 10: CoNLL-2009 development Arg-F1(%) scores (on original argument heads) with different training resources. For syntax, we have the options of "No" (no auxiliary syntactic supervision), "LSD" (original language-specific dependencies) and "UD" (universal dependencies). For SRL, we have the options of "Orig." (original argument heads) and "UD" (argument heads converted according to UD trees). If training with UD-SRL, we adopt a post-processing step and convert the argument heads back to original ones with LSD for fair comparisons. Notice that for Czech, we do not have results for UD-SRL since there are no easy ways to convert arguments back to the original ones (which disagree much with LSD and slightly disagree with UD).

Though UD seems slightly less effective than the original LSD in this experiment, we still utilize UD-based ones (both syntax and SRL) in our main experiments, considering the potential to extend to more languages. It would also be interesting to explore the combination of different syntactic formalism, which we leave to future work.

for Spanish and Catalan, we use UD_AnCora.

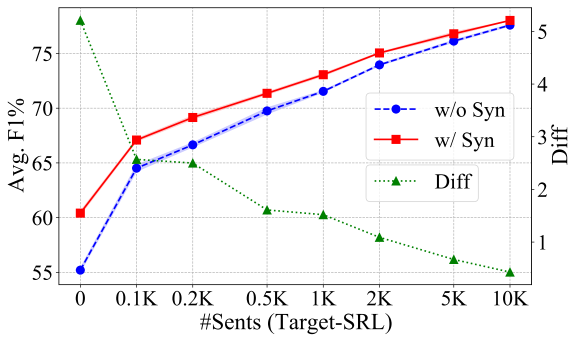

Figure 6: Averaged UPB development results versus number of SRL target sentences (in log scale) utilized per target language (using XLM-R for encoder).

EnSRL	Syntax	0.1K	1K	10K
No	No	$3.03_{\pm 0.82}$	$13.58_{\pm 0.59}$	$44.11_{\pm 0.40}$
No	Yes	$28.54_{\pm 1.35}$	$41.50_{\pm 0.68}$	$56.14_{\pm 0.29}$
Yes	No	$5.81_{\pm 0.18}$	$20.29_{\pm 0.42}$	$48.57_{\pm 0.21}$
Yes	Yes	$\mathbf{39.06}_{\pm 0.40}$	$\mathbf{46.05}_{\pm 0.75}$	$\mathbf{58.05}_{\pm 0.34}$

Table 11: FiPB development Arg-F1(%) scores in English/Finnish settings (with different number of Finnish SRL sentences) with randomly initialized encoders. "EnSRL" indicates whether using English SRL, and "Syntax" denotes whether using syntax.

FiSRL	Syntax	0.1K	1K	10K
No	No	$48.30_{\pm 0.57}$	$73.23_{\pm 0.21}$	$83.61_{\pm 0.14}$
No	Yes	$58.63_{\pm 0.71}$	$74.10_{\pm 0.28}$	$83.51_{\pm 0.28}$
Yes	No	$62.35_{\pm 0.29}$	$74.92_{\pm 0.32}$	$83.38_{\pm 0.06}$
Yes	Yes	$\mathbf{67.41}_{\pm 0.32}$	$\mathbf{75.91}_{\pm 0.22}$	$\mathbf{83.98}_{\pm 0.09}$

Table 12: EWT development Arg-F1(%) scores in Finnish(FiPB)/English settings (with different number of English SRL sentences). "FiSRL" indicates whether using Finnish SRL, and "Syntax" denotes whether using syntax.

Method	Syntax	0.1K	1K	10K
Chinese → English				
BIO	No	$57.29_{\pm 0.70}$	$71.41_{\pm 0.53}$	$79.43_{\pm 0.05}$
TwoStep	No	$57.17_{\pm 1.17}$	$72.64_{\pm 0.09}$	$\mathbf{79.95}_{\pm 0.09}$
BIO	Yes	$56.67_{\pm 0.52}$	$72.03_{\pm 0.28}$	$79.41_{\pm 0.10}$
TwoStep	Yes	$\mathbf{60.42}_{\pm 0.27}$	$\mathbf{73.53}_{\pm 0.14}$	$79.77_{\pm 0.06}$
Arabic → English				
BIO	No	$59.78_{\pm 0.37}$	$72.07_{\pm 0.20}$	$79.34_{\pm 0.05}$
TwoStep	No	$59.71_{\pm 0.71}$	$72.58_{\pm 0.22}$	$79.61_{\pm 0.21}$
BIO	Yes	$58.03_{\pm 0.73}$	$72.48_{\pm 0.09}$	$79.28_{\pm 0.08}$
TwoStep	Yes	$\mathbf{60.92}_{\pm 0.14}$	$\mathbf{73.23}_{\pm 0.10}$	$\mathbf{79.66}_{\pm 0.24}$

Table 13: OntoNotes English development Arg-F1(%) scores in semi-supervised settings (with different number of target SRL training sentences), using Chinese or Arabic as the source language.

C Extra Results

C.1 Semi-supervised Results on UPB

We also experiment with semi-supervised settings on the UPB datasets. We still take English as the source and randomly sample SRL training instances for target languages and train the models alongside all these source examples. The results are shown in Figure 6, where adding target SRL annotations can bring obvious improvements. Nevertheless, including syntactic supervision is still helpful, particularly in low-resource scenarios.

C.2 No Pre-trained Initialization

In the main experiments, we utilize pre-trained multilingual language models to initialize the encoders. Here, we explore the case where no such initialization is performed (taking FiPB as an example). All other settings are the same as previous, except that the models are all randomly initialized. The training scheme is slightly modified: we perform learning rate warmup for the first 10K steps and increase the maximum learning rate to 1e-4. The results on the development sets are shown in Table 11. There are no surprises that the scores are much lower than those with pre-trained models. Interestingly, though both English SRL and syntax can provide improvements in both low-resource and high-resource cases, syntax is much more helpful. A possible reason is that the multilingual pretraining provides shared representations across languages, without which the extra supervision from other languages may be much less effective.

C.3 Other Languages as Source

In our main experiments, we take English as the source language since it is usually the language that has the most abundant resources. Here, we take some other languages as the source and English as the target. Specifically, we use FiPB/EWT and OntoNotes for these experiments, where other settings exactly follow those of the main experiments. The development results are shown in Table 12 and 13. The general trends are very similar to those in the English-as-source experiments, where syntax supervision is generally helpful, especially in low-resource scenarios. There are many other interesting settings that are not covered in this work, such as multi-source transfer and direct transfer among non-English languages. We leave the explorations of these to future work.

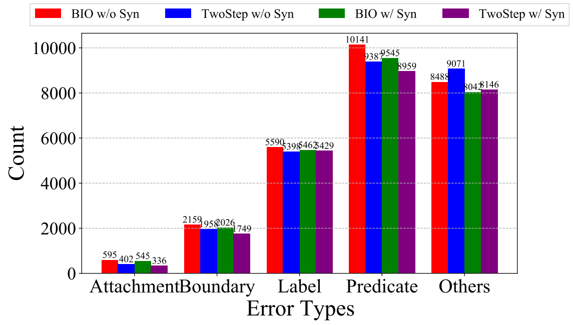

Figure 7: Error breakdowns of arguments on the OntoNotes Chinese development set, 1K setting.

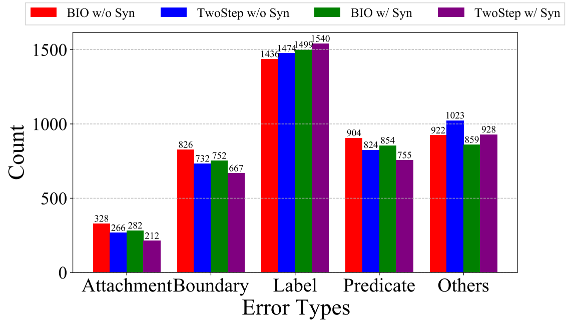

Figure 8: Error breakdowns of arguments on the OntoNotes Arabic development set, 1K setting.

C.4 Error Analysis on OntoNotes

We further perform error analysis on the Chinese and Arabic development set in the 1K setting. As shown in Figure 7 and 8, syntactic supervision and the syntax-aware TwoStep decoder make fewer errors related to phrasal attachments, span boundaries and predicate identification. Notice that the first two categories are closely related to syntax, which may explain why syntax-informed models make fewer such errors. In particular, the two-step model trained with syntactic supervision makes the fewest syntax-related errors. Together with its generally better overall F1 scores, these demonstrate the benefits of utilizing syntactic information alongside a suitable syntactically-aware model.