Originally published in: Inui, Kentaro/Jiang, Jing/Ng, Vincent/Wan, Xiaojun (eds.): Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3.–7. November 2019, Hong Kong, China. - Stroudsburg, PA, USA: The Association for Computational Linguistics, 2019. Pp. 603-615

Translate and Label! An Encoder-Decoder Approach for Cross-lingual Semantic Role Labeling

Angel Daza and Anette Frank
Leibniz ScienceCampus "Empirical Linguistics and Computational Language Modeling"
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{daza,frank}@cl.uni-heidelberg.de

Abstract

We propose a Cross-lingual Encoder-Decoder model that simultaneously translates and generates sentences with Semantic Role Labeling annotations in a resource-poor target language. Unlike annotation projection techniques, our model does not need parallel data during inference time. Our approach can be applied in monolingual, multilingual and cross-lingual settings and is able to produce dependency-based and span-based SRL annotations. We benchmark the labeling performance of our model in different monolingual and multilingual settings using well-known SRL datasets. We then train our model in a cross-lingual setting to generate new SRL labeled data. Finally, we measure the effectiveness of our method by using the generated data to augment the training basis for resource-poor languages and perform manual evaluation to show that it produces high-quality sentences and assigns accurate semantic role annotations. Our proposed architecture offers a flexible method for leveraging SRL data in multiple languages.

1 Introduction

Semantic Role Labeling (SRL) extracts semantic predicate-argument structure from sentences. This has proven to be useful in Neural Machine Translation (NMT) (Marcheggiani et al., 2018), Multi-document-summarization (Khan et al., 2015), AMR parsing (Wang et al., 2015) and Reading Comprehension (Mihaylov and Frank, 2019). SRL consists of three steps: i) predicate detection, ii) argument identification and iii) role classification. In this work we focus on PropBank SRL (Palmer et al., 2005), which has proven its validity across languages (van der Plas et al., 2010). While former SRL systems rely on syntactic features (Punyakanok et al., 2008; T¨ackstr¨om et al., 2015), recent neural approaches learn to model both ar-gument detection and role classification given a

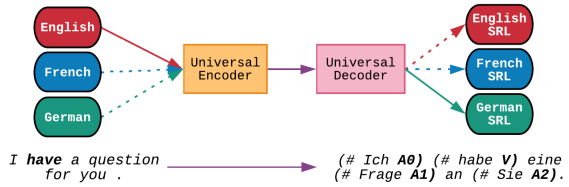

Figure 1: We propose an Encoder-Decoder model that translates a sentence into a target language and applies SRL labeling to the translated words. In this example we translate from English to German and label roles for the predicate *have*.

predicate (Marcheggiani et al., 2017; He et al., 2017), and even jointly predict predicates inside sentences (He et al., 2018; Cai et al., 2018). While these approaches alleviate the need for pipeline models, they require sufficient amounts of training data to perform adequately. To date, such models have been tested primarily for English, which offers a considerable amount of high-quality training data compared to other languages. The lack of sufficiently large SRL datasets makes it hard to straightforwardly apply the same architectures to other languages and calls for methods to augment the training data in lower-resource languages.

There is significant prior work on SRL data augmentation (Hartmann et al., 2017), annotation projection for monolingual (Fürstenau and Lapata, 2012; Hartmann et al., 2016), and cross-lingual SRL (Pado´ and Lapata, 2009; van der Plas et al., 2011; Akbik et al., 2015, 2016). A drawback of cross-lingual projection is that even at prediction time it requires parallel sentences, a semantic role labeler on the source side, as well as syntactic information for both language sides. Thus, it is desirable to design an architecture that can make use of existing annotations in more than one lan-

guage and that learns to translate input sentences to another language while transferring semantic role annotations from the source to the target.

Techniques for low-resource Neural Machine Translation (NMT) show the positive impact on target predictions by adding more than one language during training, such as Multi-source NMT (Zoph and Knight, 2016) and Multilingual NMT (Johnson et al., 2017; Firat et al., 2016a), whereas Mulcaire et al. (2018) show the advantages of training a single polyglot SRL system that improves over monolingual baselines in lower-resource settings. In this work, we propose a general Encoder-Decoder (Enc-Dec) architecture for SRL (see Figure 1). We extend our previous Enc-Dec approach for SRL (Daza and Frank, 2018) to a cross-lingual model that translates sentences from a source language to a (lower-resource) target language, and during decoding jointly labels it with SRL annotations.[1]

Our contributions are as follows:

- We propose the first cross-lingual multilingual Enc-Dec model for PropBank SRL.
- We show that our cross-lingual model can generate new labeled sentences in a target language without the need of explicit syntactic or semantic annotations at inference time.
- Cross-lingual evaluation against a labeled gold standard achieves good performance, comparable to monolingual SRL results.
- Augmenting the training set of a lower-resource language with sentences generated by the cross-lingual model achieves improved F1 scores on the benchmark dataset.
- Our universal Enc-Dec model lends itself to monolingual, multilingual and crosslingual SRL and yields competitive performance.

2 An Extensible Model for SRL

2.1 One Model to Treat Them All

We define the SRL task as a sequence transduction problem: given an input sequence of tokens $X = x_1, ..., x_i$, the system is tasked to generate a sequence $Y = y_1, ..., y_j$ consisting of words interleaved with SRL annotations. Defining the task in this fashion allows X and Y to be of different lengths and therefore target sequences may also

[1]Code is available at: https://github.com/Heidelberg-NLP/SRL-S2S.

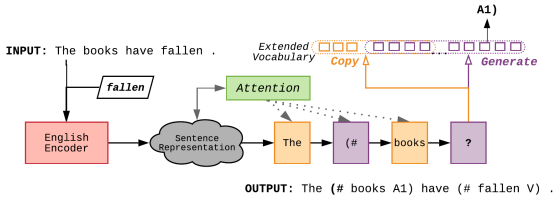

Figure 2: Monolingual Enc-Dec model for SRL with copying (Daza and Frank, 2018). We generalize this architecture to multilingual and cross-lingual SRL.

contain word tokens of different languages if desired. This means that we could train an Enc-Dec model that learns not only to label a sentence, but to jointly translate it while applying SRL annotations directly to the target language. Moreover, following conceptually the multilingual Enc-Dec model proposed by Johnson et al. (2017), we can train a single model that allows for joint training with multiple language pairs while sharing parameters among them. We apply a similar joint multilingual learning method to produce structured output sequences in the form of translations enriched with SRL annotations on the (lower-resource) target language (cf. Figure 3). We will apply this universal structure-inducing Enc-Dec model to the Semantic Role Labeling task, and show that it can be deployed in three different settings:

i) **monolingual**: encode a sentence in a given language and learn to decode a labeled sequence by reproducing the source words and inserting the appropriate structure-indicating labels in the output (cf. Figure 2). A copying mechanism (Gu et al., 2016) allows this model to reproduce the input sentence as faithfully as possible.

ii) **one-to-one multilingual**: train a single, joint model to generate n different structure-enriched target languages given inputs in the same language. For example: Labeled English (*EN-SRL*) given an *EN* sentence or Labeled German (*DE-SRL*) given a *DE* sentence. This multilingual model still relies on copying to relate each labeled output sentence to its corresponding input counterpart. However, unlike (i), it has the advantage of sharing parameters among languages.

iii) **cross-lingual**: generate outputs in n different target languages given inputs in m different source languages, for example: Labeled German (*DE-SRL*) and Labeled French (*FR-SRL*) given an *EN* sentence (see Figure 3). In this setting, we do not restrict the model to *copy* words from the

source sentence but train it to *translate* them.
In Section 2.2 we describe how the basic Enc-Dec model for SRL is constructed and in Section 2.3 we describe the additional components that allow us to generalize this architecture to the one-to-one multilingual and cross-lingual scenarios.

2.2 Encoder-Decoder Architecture

We reimplement and extend the Enc-Dec model with attention (Bahdanau et al., 2015) and copying (Gu et al., 2016) mechanisms for SRL proposed by Daza and Frank (2018). This model encodes the source sentence and decodes the input sequence of words (in the same language) interleaved with SRL labels.

Data Representation. Similar to other prior work (Liu et al., 2018) and our own (Daza and Frank, 2018), we linearize the SRL structure in order to process it as a sequence of symbols suitable for the Enc-Dec architecture. We restrict ourselves to argument identification and labeling of one predicate at a time. We feed the gold predicate in training and inference, and process each sentence as many times as it has predicates. An opening bracket *(#* indicates the start of a labeled-argument region; a closing labeled bracket, e.g. *A0)*, indicates the ending and the tag of the labeled region (see Figure 2).

Vocabulary. We define a shared vocabulary consisting of all source and target words $\mathcal{V} = \{v_1, ..., v_N\} \cup \{UNK\}$ and the SRL labels $\mathcal{L} = \{l_1, ..., l_M\}$. In addition, we employ a per-instance extension set $\mathcal{X} = \{x_1 ..., x_{T_x}\}$ containing all words from the source sequence. Our final vocabulary is $\mathcal{V} \cup \mathcal{L} \cup \mathcal{X}$.

Encoder. In our prior work (Daza and Frank, 2018) we used a 2-layer BiLSTM as encoder. In this paper, we adopt the Deep BiLSTM Encoder from He et al. (2017) which has been shown to work well for SRL models. Again following He et al. (2017), we define the encoder input vector x_i as the concatenation of a word embedding w_i and a binary predicate-feature embedding p_i indicating at each time-step whether the current word is a predicate or not [2]. The encoder outputs a series of hidden states $h_1, ..., h_{T_x}$ representing each token. We refer to this series of states as $\mathbf{H}$.

Attention. To improve the access to the source sentence representation, we include the attention mechanism proposed by Bahdanau et al. (2015), which computes a context vector at each time step *t* based on $\mathbf{H}$ and the current decoder state.

Decoder. We use a single-layer Decoder with LSTM cells (Hochreiter and Schmidhuber, 1997) and a copying mechanism. It emits an output token y_t from a learned score ψ_g over the vocabulary at each time step *t* given its state s_t, the previous output token y_{t-1}, and the attention context vector c_t. In addition, a copying score ψ_c is calculated. The decoder learns from these scores when to generate a new token and when to copy from the encoded hidden states $\mathbf{H}$. Formally we compute the scores as:

$$\begin{aligned} \psi_g(y_t = v_i) &= W_o[s_t; c_t], \;\; v_i \, \epsilon \, \mathcal{V} \cup \mathcal{L} \\ \psi_c(y_t = x_j) &= \sigma(h_j^T W_c) s_t, \;\; x_j \, \epsilon \, \mathcal{X} \end{aligned} \tag{1}$$

where $W_o \epsilon \mathbb{R}^{N \times 2d_s}$ and $W_c \epsilon \mathbb{R}^{d_h \times d_s}$ are learnable parameters and s_t, c_t are the current decoder state and context vector, respectively. These scores are used to compute two distributions: one for the likelihood of copying ($\mathbf{c}$) y_t and another for the likelihood of generating ($\mathbf{g}$) y_t. Formally:

$$\begin{aligned} p(y_t | s_t, y_{t-1}, c_t, \mathbf{H}) = p(y_t, \mathbf{g} | s_t, y_{t-1}, c_t) + \\ p(y_t, \mathbf{c} | s_t, y_{t-1}, \mathbf{H}) \end{aligned} \tag{2}$$

The two distributions are then normalized by a final *softmax* layer from which we compute a joint likelihood of y_t and choose the token with the highest score within this joint likelihood.

2.3 Multilingual Extensions

We generalize the monolingual Enc-Dec model for SRL to a multilingual SRL system by adding two main components:

Translation Token. Like Johnson et al. (2017), we prefix the source sequence with a special token that indicates the expected language of the target sequence. If the source is in *EN* and the target is a German sentence with SRL labels, the source sentence will be preceded by the token *<2DE-SRL>*.

Language Indicator Embeddings. We want the model to profit from the common role label inventory used across languages, yet at the same time there are subtle differences in role labeling and how roles are linguistically marked in the different languages[3]. Hence, we define *N* different language indicators (e.g., *FR, DE*) and represent

[2] These two additions already show improvements compared to the reported results in Daza and Frank (2018).

[3] e.g. the role A2 (Beneficiary) can be PP in *EN* and *FR*, but dative NP in *DE* (DativeNP)

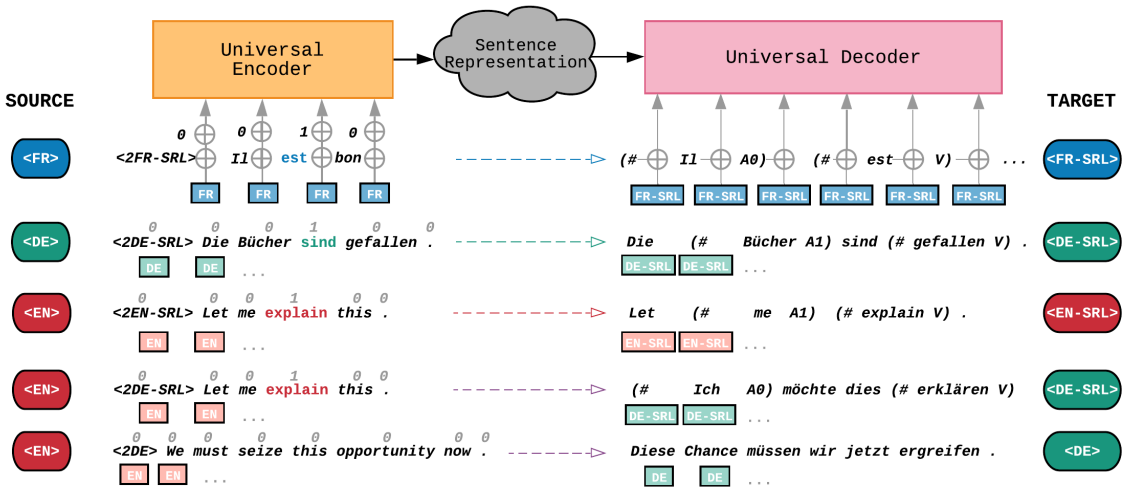

Figure 3: A universal structure-inducing Enc-Dec model with copying and sharing of parameters across languages that jointly translates and labels sentences. It can also be trained cross-lingually and be augmented with classic machine translation data (as shown in the two bottom rows of the figure).

each of them with a randomly initialized language indicator vector that we tune during training. The model can use these language indicator embedding vectors to leverage language-specific properties when generating SRL annotations. Also, by using these embeddings in the decoder, we can help it to stay consistent regarding the language it generates[4].

Thus, in all multilingual settings, at each time step t we feed the **Encoder** with a concatenation of the previous encoder state h_{t-1}, the word embedding w_t of the current token, the embedded predicate indicator p_t and the language indicator embedding l_t. The Encoder state update is defined as:

$$h_t = LSTM([h_{t-1}; w_t; p_t; l_t]) \quad (3)$$

Likewise, on the **Decoder** side we concatenate the representations for both word tokens and label tokens with the language indicator vector to produce tokens in a specific l anguage. For SRL-labeled output sentences the indicator token for the language embedding is *DE-SRL, FR-SRL, ...* depending on the target language. Formally, at each time step the decoder updates its state by taking into account the previous decoder state s_{t-1}, the previous generated token[5] y_{t-1}, the language indicator embedding l_{t-1} and the attention context vector c_t:

$$s_t = LSTM([s_{t-1}; y_{t-1}; l_{t-1}; c_t]) \quad (4)$$

3 Data

3.1 SRL Monolingual Datasets

Two labeling schemes have been established for PropBank SRL: span-based and dependency-based. In the former, arguments are characterized as word-spans. This scheme was introduced in the CoNLL-05 Shared-task (Carreras and M`arquez, 2005) and is available only for English. In the dependency-based SRL format, only the syntactic heads of arguments are labeled. This format was defined in the CoNLL-09 Shared task (Haji˘c et al., 2009), which includes SRL labeled data for seven languages.[6]

We will use span-based and dependency-based data for English to benchmark the monolingual system. For the multilingual experiments, we use the dependency-based annotations, given that there is labeled data available in different languages on this format. Specifically, w e u se the English and German portions of CoNLL-09 and the automatically annotated French SRL corpus of van der Plas et al. (2011) for training and the human-labeled sentences from van der Plas et al. (2010) for testing. Both corpora contain a similar

[4]Johnson et al. (2017) only use a translation token, but our training data is significantly smaller than theirs.

[5]During training we use teacher forcing, feeding the gold target token instead of the previously generated token

[6]Note that CoNLL-09 is not a parallel corpus. All data was annotated independently and later ported to CoNLL.

Mono-lingual	Language	Train		Test
		# Sents	w/ 1-Pred	w/ 1-Pred
CoNLL-05	EN [Span]	75,187	94,497	5,476
CoNLL-09	EN [Dep]	39,279	180,446	10,626
CoNLL-09	DE [Dep]	36,020	39,138	2,044
v.d. Plas	FR [Dep]	40,075	73,094	2,036

Table 1: Train and Test Data for Monolingual Models. We show the original number of sentences and the size of the "expanded" data with one copy per predicate.

Cross-lingual Model	# Sentences
EN - DE-SRL (Akbik, 2015)	63,397
EN - FR-SRL (Akbik, 2015)	40,827
EN - FR (UN)	100,000
EN - DE (Europarl)	100,000

Table 2: Data used for Cross-lingual Models: From the SRL parallel data available we take 90% for training and use the rest as a *Dev* set for our experiments. We add the non-labeled data (from UN and Europarl) during training to enforce translation knowledge.

label set as the English PropBank[7]. For statistics on the size of the datasets see Table 1.

3.2 Datasets for Cross-lingual SRL

We use the dependency-based labeled German and French SRL corpus from Akbik et al. (2015) which was produced via annotation projection. These sentences are already pre-filtered to ensure that the predicate sense of the source predicate is preserved in the target sentence. Since the role labels are projected from automatically PropBank-parsed English sentences, all languages share the same label set. The underlying corpus for this dataset is composed of Machine Translation (MT) parallel corpora: Europarl (Koehn, 2005) for *EN-DE* (about 63K sents), and UN (Ziemski et al., 2016) for *EN-FR* (about 40K sents).

Since we only had access to the labeled sentences (target-side), we constructed our parallel training pairs *EN* to *FR-SRL* and *EN* to *DE-SRL* by finding the original source English counterparts. We use Flair (Akbik et al., 2018) to predict PropBank frames on the English source sentences and find the alignment to the labeled predicate on the target side using fast-align (Dyer et al., 2013).

In addition to the parallel SRL-labeled data, we choose a subset of 100K parallel (non-labeled) sentences for each language pair from the mentioned MT datasets (Europarl and UN corpora) to improve the translation quality of the model, we use 90% for training and the rest as a development set. The data is summarized in Table 2.

[7]French data was directly annotated using the English labelset but German CoNLL-09 contains additional core labels A5-A9 and does not contain *AM-* modifier labels

4 Experiments and Results

4.1 General Settings

We use the AllenNLP (Gardner et al., 2018) Enc-Dec model as a basis for our implementation. Our model is trained to minimize the negative log-likelihood of the next token. Hyperparameters and model sizes are provided in Supplement A.1. We use pre-trained word embeddings (fine-tuned during training) for the 3 languages: GloVe (Pennington et al., 2014) for *EN* and the pre-trained vectors from Grave et al. (2018) for *FR* and *DE*. We also train versions with contextual word representations: pre-trained English 1024-dimensional ELMo (Peters et al., 2018) and multilingual 768-dimensional BERT-small (Devlin et al., 2019) representations.

4.2 Monolingual Experiments and Results

We train three separate monolingual versions for *EN*, *DE* and *FR*. We first benchmark our system against a wide variety of English models (span- and dependency-based) that perform the role classification task with gold predicates to show that our labeling performance is competitive with the existent SOTA neural models for English. This is shown in Table 3. The performance of *DE* and *FR* is shown in Table 4 where we compare all monolingual systems for the three languages (top half), against the one-to-one multilingual versions (bottom half). Results for *EN* show that the Enc-Dec architecture is competitive with the GloVe-based models (although still 4 F1 points below SOTA in most cases), however it benefits more from ELMo, achieving SOTA results for span-based and dependency-based SRL.

4.3 Multilingual Experiments and Results

We train a single multilingual model with the concatenation of the training data for the three languages *EN*, *DE* and *FR* that we previously used on the monolingual experiments. We use a common vocabulary for the three languages and keep all tokens that occur more than 5 times in the combined dataset. We train the model with batches containing instances randomly chosen from the individual languages (this means that each batch might contain examples from different language pairs).

Type	Model	Word Repres.	CoNLL-05 WSJ	CoNLL-05 OOD	CoNLL-09 WSJ	CoNLL-09 OOD
Span SRL	He 2017	GloVe	84.6	73.6	-	-
	Daza 2018	GloVe	79.2	68.4	-	-
	He, 2018	ELMo	83.9	73.7	-	-
	Tan, 2018	GloVe	84.8	74.1	-	-
	Strubell 18 [LISA]	GloVe	84.6	74.5	-	-
	Strubell 18 [LISA*]	ELMo	86.5	78	-	-
	Ouchi 2018	ELMo	**88.5**	79.6	-	-
Dep SRL	Roth 2016	DPE*	-	-	87.7	76.1
	Marcheggiani 2017	Dyer*	-	-	87.7	77.7
	Cai et al 2018	GloVe	-	-	89.6	79
Dep and Span SRL	FitzGerald 2015	GloVe	80.3	72.2	87.8	75.5
	Li 2019	ELMo	87.7	-	90.4	-
	Ours [Mono]	GloVe	80.4	70.5	85.5	75.7
	Ours [Mono]	ELMo	88.3	**80.9**	**90.8**	**84.1**

Table 3: CoNLL-09 and CoNLL-05 Test Sets for English. Our model with ELMo shows SOTA performance on both types of SRL. LISA* only reports ELMo with predicted predicates; DPE*: dependency path embeddings; Dyer*: Dyer et al. 2015.

Model	EN-Test	DE-Test	FR-Test
SOTA models*	90.4	80.1	73
Ours-EN [Mono + GloVe]	85.5	-	-
Ours-DE [Mono + GloVe]	-	61.9	-
Ours-FR [Mono + GloVe]	-	-	70.3
Mulcaire 2018 [Multi + GloVe]	86.5	69.9	-
Ours [Multi + GloVe]	87	68.2	70.5
Ours [Multi + ELMo]	91.1	75.7	70.7
Ours [Multi + BERT]	89.7	77.2	72.4

Table 4: F1 scores for role labeling on dependency-based SRL data. EN and DE Tests: CoNLL-09; FR-Test: van der Plas et al. (2011). State of the art (SOTA) models* are: Cai et al. (2018) [GloVe] for EN, Roth and Lapata (2016) [Dependency-path Embeddings] for DE and van der Plas et al. (2014) [Non-neural] for FR, respectively.

Multilingual training yields improvement on the three languages studied in this paper when compared to our monolingual baselines, particularly for German, which shows more than 6 points (F1) of improvement. In addition, we compare with the polyglot SRL system of Mulcaire et al. (2018) (which also leverages data from multiple languages during training), obtaining better results for English using GloVe. We then show that adding contextual representations to our model results in bigger improvements across the board.

4.4 Cross-Lingual Experiments and Results

Training. After validating the robustness of our architecture when handling different languages at the same time, we now train a cross-lingual SRL version. This setting differs from the previous two because the model needs to learn two tasks: besides generating appropriate SRL labels, it needs to translate from source into a target language. To do so, we train a single model using the concatenation of the parallel datasets listed in Table 2 and described in Section 3.2. We further include Machine Translation (MT) data to reinforce the translation knowledge of the model, so that it can generate fluent (labeled) target sentences. As in the multilingual experiments, we train the model with alternating batches of instances randomly chosen from the individual language pairs. Note that the amount of MT data that we can add is restricted: the labeled multilingual data is relatively small and labeling performance suffers when the MT data gets too dominant.

Evaluating Cross-lingual SRL. As in classical MT, evaluation is difficult, since the system outputs will approximate a target reference but will never be guaranteed to match it. Hence in this setting we do not have a proper gold standard to evaluate the labeled outputs, since we are generating labeled target sentences from scratch. Similar to MT research, we apply BLEU score (Papineni et al., 2002) to measure the closeness of the outputs against our *Dev* Set.

The upper part of Table 5 compares the scores of two versions of the Enc-Dec model trained on the cross-lingual data from Table 2 systems, one using GloVe embeddings and the second using BERT, respectively. To better distinguish translation vs. labeling quality, we compute BLEU scores for the system outputs against labeled reference sentences in three different ways: on *words only*, *labels only*, and on *full labeled sequences* (both word and label outputs). We see that the prediction of words is similar in the two languages, but labeling is more difficult for *DE* than for *FR* for both systems. Also we observe that adding multilingual BERT is very helpful to obtain even more fluent and correct labeled outputs (according to BLEU) resulting in ca. +9 points in German and +5 in French on the full sequences. This is very important given that we have a small training set compared to classic NMT scenarios.

The bottom part of Table 5 shows the scores when restricting the evaluation to sentences with score ≥ 10. We observed that this threshold[8] is a good trade-off in both the amount of kept sentences (above the threshold) and average BLEU

[8]We tried with thresholds of 5, 10, 20 and 30.

	German			French		
Model [Filter]	**F-Seq**	**Word**	**Label**	**F-Seq**	**Word**	**Label**
XL-GloVe [All]	18.86	17.17	25.52	28.99	17.36	32.76
XL-BERT [All]	27.22	27.36	29.59	33.59	22.48	37.17
XL-GloVe [$\geq$ 10]	30.58	36.71	51.68	38.99	43.79	61.73
XL-BERT [$\geq$ 10]	36.95	41.36	55.73	42.66	46.52	65.32

Table 5: Cross-lingual (XL) system results using BLEU score on individual languages inside the *Dev* set. We compute BLEU on labeled sequences (F-Seq), and separately for words and only labels. We also show scores when pre-filtering on F-Seq with BLEU $\geq$ 10.

score increase (presumably sentence quality). By keeping only the filtered subset of sentences we get an improvement on average of approx. 10 BLEU points on the full sequences (*F-Seq*), and almost double the score for *labels only*. This holds for GloVe and BERT versions on both languages.

Output Filtering and Data Generation. We use our cross-lingual model as a labeled data generator by applying it on *EN* sentences from Europarl (100K) and UN corpora (100K)[9] and let the model predict *DE-SRL* and *FR-SRL* as target languages. This results in *unseen* German and French labeled sentences. Since we cannot guarantee that the generated sentences preserve the source predicate meaning, we filter all outputs by keeping only those that come close to the original sentence meaning. We approximate this by back-translating the generated outputs (stripping the labels and keeping only the words) using the pre-trained *DE-EN* model from OpenNMT (Klein et al., 2017).

We compare the back-translations to the sentences that we originally presented to the system and, using the previously described filtering heuristic, we keep only those whose BLEU score is equal or greater than 10. The logic behind this is that if the back-translation is close enough to the source, the generated sentence preserves a fair amount of the original sentence meaning[10]. With this strategy, after applying the *BLEU filter*, we end up with a parallel dataset of 44K generated sentences for *(EN, DE-SRL)* and 32K for (*EN*, *FR-SRL*). In the next section we show more detailed evaluation measures of the system outputs, focusing on the filtered dataset that we just described.

[9]Note that these are taken from a different subset than the parallel sentences used during training.

[10]BLEU score is used as a naive approach to avoid excessively noisy data but we could also develop, for example, a semantic similarity metric to also keep sentences that are close enough to the original predicate sense meaning.

4.5 Cross-Lingual Detailed Evaluation

We are aware that BLEU score gives only a rough estimate of the actual quality of the outputs, therefore we propose to measure the performance of our system in two more detailed evaluation settings: (i) a small-scale **human evaluation** where we evaluate the assigned SRL labels against 226 sentences that were manually judged and annotated to give an estimation of the quality of the generated data, (ii) an **extrinsic evaluation** using labeled sentences generated by our system to augment the training set for a resource-poor language. We conduct the extrinsic evaluation on German and French and the manual evaluation only on the German data, which proved to be the more challenging language compared to French.

4.5.1 Human Evaluation

To provide an in-depth quality assessment of the generated sentences, we create a small-scale gold standard consisting of 226 sentences. To select a representative sample from our newly generated labeled sentences,[11] we analyze the distribution of labels in the data and apply stratified sampling to cover as many predicates as possible and as many role label variants as possible. We judge these sentences on the quality of the generated language and annotate them with PropBank roles.

SRL Gold Standard. As we are lacking trained PropBank annotators, we mimic the question-based role annotation method of He et al. (2015), who constructed QA pairs in order to label the predicate-argument structure of verbs. The annotation involves several subtasks: The first is to generate questions targeting a specific verb in a sentence and to mark as answers a subset of words from the same sentence. The next subtask is to choose the head word of each selected subset and to assign a PropBank label to this head according to a table that correlates WH-phrases with the most likely label.[12]

We ask two linguistically trained annotators to perform the whole task independently and compute Krippendorff's Alpha (Krippendorff, 1980) on the role labels, which results in an inter-annotator agreement score of 82.83. We resolved conflicting annotations through discussion among

[11]i.e., the generated sentences for which we measured a BLEU score $\geq$ 10 against the source using back-translation.

[12]We provide this correlation table and the full annotation guidelines in appendix A.2 in the Supplement.

the annotators. The resulting gold standard contains 737 annotated roles. Notably, the most prominent roles (as in the CoNLL datasets) are A0 and A1 which are normally related to the agent and the patient in sentences, but the annotated data also includes modifier roles such as temporal, modal, discourse markers, among others[13].

Translation Quality. We ask two different annotators to score each output sentence (they see only the words, not the labels) on a scale of 1-5 for *Quality* (1: 'is completely ungrammatical'; 5: 'is perfectly grammatical') and for *Naturalness* (1: 'The sentence is not what a native speaker would write'; 5: 'The sentence could have been written by a native speaker'). We obtain a high average score of 4.4 for *Quality* and 4.2 for *Naturalness*.

SRL Performance on Gold Standard. We use our human-annotated sentences to measure the automatic labeling performance of our cross-lingual SRL model which we call XL-BERT). We obtain 73.21 F1 score (73.33 precision, 73.1 recall). We also measure the performance of the ZAP label projection system of Akbik and Vollgraf (2018) on this data (we only consider arguments of the predicates that were annotated). ZAP obtains a low F1 score of 56.03 (42.65 precision, 81.7 recall). Thus, XL-BERT shows much better, and more precise results compared to this baseline and achieves overall very acceptable and stable labeling quality. This shows that the joint translation-labeling task is successful. ZAP, by contrast, shows more unstable results, which might be due to word alignment noise. Although we train on such data, our model can also loose some of this noise, given that the same model is trained to produce more than one labeled language, namely *FR-SRL* and *DE-SRL*.

4.5.2 Extrinsic Task: Data Augmentation

Finally, we augment the training sets of our two resource-poor languages *DE* and *FR*, in portions of 10K until we cover the complete generated data. We compare the increase in F1 score when training models with different amounts of additional data. We also add a comparison of the improvement achieved when adding the same amount of sentences produced by the labeled projection method of Akbik et al. (2015). We see in Table 6 that adding our German data shows improvement in F1 score, despite the fact that the CoNLL-09 label scheme has arguments not seen in our training data (namely A5-A9). Presumably we see this improvement because the frequency of the major roles is more prominent. In the case of French, we don't see significant improvement, however also here the addition of projected data shows a similar trend.

[13]The label distribution is given in the Supplement, A.3.

Model + Training Data	Data Size	F1 Test
DE [Mono] *(Original)*	39K	61.9
DE [Mono] + *LabelProj*	83K	62.37
DE [Mono] + OurGen10K	49K	62.4
DE [Mono] + OurGen20K	59K	62.46
DE [Mono] + OurGen30K	69K	62.81
DE [Mono] + OurGenALL	**83K**	**63.57**
FR [Mono] *(Original)*	73K	70.3
FR [Mono] + *LabelProj*	105K	70.45
FR [Mono] + OurGen10K	83K	70.33
FR [Mono] + OurGen20K	**93K**	**70.52**
FR [Mono] + OurGenALL	105K	70.39

Table 6: We retrain the monolingual systems *DE, FR* using the original training sets (BL: *Original*) shown in Table 1 and inject our generated data in different sizes. We also compare to the stronger baseline *LabelProj* where we add data created by label projection (Akbik et al., 2015).

5 Related Work

Encoder-Decoder Models. A wide range of NMT models are based on the Encoder-Decoder approach (Sutskever et al., 2014) with attention mechanism (Bahdanau et al., 2015; Luong et al., 2015). More recent architectures (Zoph and Knight, 2016; Firat et al., 2016a) show that training with multiple languages performs better than one-to-one NMT. Multilingual models have also been trained to perform Zero-shot translation (Johnson et al., 2017; Firat et al., 2016b). The Enc-Dec approach has been tested in many tasks that can be formulated as a sequence transduction problem: syntactic parsing (Vinyals et al., 2015), AMR and Semantic Parsing (Konstas et al., 2017; Dong and Lapata, 2016) and SRL (Daza and Frank, 2018). The most similar approach to ours is Zhang et al. (2017), who propose a cross-lingual Enc-Dec that produces OpenIE-annotated English given a Chinese sentence. However, their setup is easier than ours since they have a reliable labeler on the target side, facilitating the generation of more training data unlike us who are interested in labeling the resource-poor language.

Cross-lingual Annotation Projection. A common approach to address the lack of annotations is projecting labels from English to a lower-resource language of interest. This has shown good results in the transfer of semantic information to target languages. Kozhevnikov and Titov (2013) propose an unsupervised method to transfer SRL labels to another language by training on the source side and using shared feature representations for predicting on the target side. Pado´ and Lapata (2009) project FrameNet (Baker et al., 1998) SRL la-bels by searching for the best alignment in source and target constituent trees, defining label transfer as an optimization problem in a bipartite graph. van der Plas et al. (2011) use intersective word alignments between English and French with ad-ditional filtering heuristics to determine whether a PropBank label should be transferred and then use this to train a joint syntactic-semantic parser for both languages. Akbik et al. (2015) proposes a higher-confidence projection by first creating a system with high precision and low recall and then using a bootstrap approach to augment the labeled data.

Separately, Minard et al. (2016) generated a multilingual event and time parallel corpus including SRL annotations. Their corpus was manually annotated on the English side and automatically projected to Italian, Spanish, and Dutch based on the manual alignment of the annotated elements. Unfortunately, the authors do not report the performance of the SRL task, making it difficult for us to use their data for benchmarking.

Semantic Role Labeling. Span-based SRL only exists on English data (Zhou and Xu, 2015; He et al., 2018; Strubell et al., 2018; Ouchi et al., 2018). Dependency-based SRL models such as (Marcheggiani and Titov, 2017; Cai et al., 2018; Li et al., 2019) are the state-of-the-art for English. For French, we compare against van der Plas et al. (2014) since we did not find more recent work for that language. Roth and Lapata (2016) show a model based on dependency path embeddings that achieved SOTA in English and German. The Polyglot SRL model of Mulcaire et al. (2018) shows some improvement over monolingual baselines when aggregating all multilingual data available from CoNLL-09, while more refined integration did not show further improvement. Their system does not perform better than our multilingual models for English and German.

6 Conclusions

We presented the first cross-lingual SRL system that translates a sentence and concurrently labels it with PropBank roles. The proposed Enc-Dec architecture is flexible: as a *monolingual* system the model achieves SOTA for English PropBank role labeling, the *multilingual* SRL system shows that joining multiple languages improves SRL performance over the monolingual baselines, and a *cross-lingual* system can be used to generate SRL-labeled data for lower-resource languages. Evaluation of the cross-lingual system shows that the quality-filtered sentences are highly grammatical and natural, and that the generated PropBank labels can be more precise than label projection. Using our labeled data beats a label projection baseline when using it to augment the training set of a lower-resource language.

An advantage of our proposed model is that it does not need parallel data at inference time. Our current model can possibly be further improved by adding more automatically generated data in the data augmentation scenario, or by targeted selection in an active learning setting. Current limitations of the system may be alleviated by pre-training the model to acquire better translation knowledge from larger training data, and by developing more refined filtering methods.

In future work we also aim to make the system more flexible, by extending it to few-shot or zero-shot learning, to alleviate the need for an initial big annotated set, and thus to be able to generate SRL data for truly resource-poor languages. Further challenges for this novel architecture are to extend it to joint predicate and role labeling for more than one predicate at a time.

Acknowledgements

We thank the reviewers for their insightful comments. This research is funded by the Leibniz ScienceCampus Empirical Linguistics & Computational Language Modeling, supported by Leibniz Association grant no. SAS2015-IDS-LWC and by the Ministry of Science, Research, and Art of Baden-Wurttemberg. We thank NVIDIA Corporation for donating GPUs used in this research. We are grateful to our annotators and to Éva Mújdricza-Maydt for her assistance with the human evaluation setup.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alan Akbik, L.b Chiticariu, M.b Danilevsky, Y.bLi, S.b Vaithyanathan, and H.b Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. *ACL-IJCNLP 2015*, 1:397–418.

Alan Akbik, Vishwajeet Kumar, and Yunyao Li. 2016. Towards Semi-Automatic Generation of Proposition Banks for Low-Resource Languages. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 993–998.

Alan Akbik and Roland Vollgraf. 2018. ZAP: An open-source multilingual annotation projection framework. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Collin F. Baker, Charles J. Fillmore, John B. Lowe, Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 17th international conference on Computational linguistics*, volume 1, page 86, Morristown, NJ, USA. 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 *COLING*, pages 2753–2765. Association for Computational Linguistics.

Xavier Carreras and Llu´ıs M`arquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 152–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2018. A sequence-to-sequence model for semantic role labeling. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 207–216. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, andKristina 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.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33–43. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *HLT-NAACL*, pages 866–875. The Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Lan-guage Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Hagen Fürstenau and Mirella Lapata. 2012. Semi-Supervised Semantic Role Labeling via Structural Alignment. *Journal of Computational Linguistics*, 38(1):135–171.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. 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: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Silvana Hartmann, Judith Eckle-Kohler, and Iryna Gurevych. 2016. Generating training data for semantic role labeling based on label transfer from linked lexical resources. *Transactions of the Association for Computational Linguistics*, 4:197–213.

Silvana Hartmann, Éva Mújdricza-Maydt, Ilia Kuznetsov, Iryna Gurevych, and Anette Frank. 2017. Assessing SRL frameworks with automatic training data expansion. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 115–121, Valencia, Spain. Association for Computational Linguistics.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory.*Neural Comput.*, 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Atif Khan, Naomie Salim, and Yogan Jaya Kumar. 2015. A framework for multi-document abstractive summarization based on semantic role labelling. *Appl. Soft Comput.*, 30(C):737–747.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-Sequence Models for Parsing and Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada.

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. Association for Computational Linguistics.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage.

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

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439, Melbourne, Australia. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

Diego Marcheggiani, Joost 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.

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. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2019. Discourse-Aware Semantic Self-Attention For Narrative Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Bego˜na Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of the Tenth International Confer-ence on Language Resources and Evaluation (LREC 2016)*, pages 4417–4422, Portoroz, Slovenia. Euro-pean Language Resources Association (ELRA).

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.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1630–1642.

Sebastian Pado´ and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):307–340.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*.

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

Lonneke van der Plas, Marianna Apidianaki, and Chenhua Chen. 2014. Global methods for cross-lingual semantic role and predicate labelling. In *COLING*.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *ACL (Short Papers)*, pages 299–304. The Association for Computer Linguistics.

Lonneke van der Plas, Tanja Samardzic, and Paola Merlo. 2010. Cross-lingual validity of propbank in the manual annotation of french. In *Linguistic Annotation Workshop*, pages 113–117. Association for Computational Linguistics.

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

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202, Berlin, Germany. Association for Computational Linguistics.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41.

Oriol Vinyals, Ł ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2773–2781. Curran Associates, Inc.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. A Transition-based Algorithm for AMR Parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375. Association for Computational Linguistics.

Sheng Zhang, Kevin Duh, and Benjamin Van Durme. 2017. MT/IE: Cross-lingual open information extraction with neural sequence-to-sequence models.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Valencia, Spain. 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.

Micha 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 2016)*, Paris, France. European Language Resources Association (ELRA).

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Con-ference of the North American Chapter of the Asso-ciation for Computational Linguistics: Human Lan-guage Technologies*, pages 30–34, San Diego, Cali-fornia. Association for Computational Linguistics.