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), Multidocument-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; Tackström et al., 2015), recent neural approaches learn to model both argument detection and role classification given a 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-
language 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.

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 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

---

Figure 2: Monolingual Enc-Dec model for SRL with copying (Daza and Frank, 2018). We generalize this architecture to multilingual and cross-lingual SRL.
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 \( \mathcal{X} = \{x_1, ..., x_T\} \) 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 \( h_1, ..., h_T \), 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 \( \psi_g \) 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 \( \psi_c \) 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:

\[
\begin{align*}
\psi_g(y_t = v_i) &= W_0[s_t; c_t], \quad v_i \in \mathcal{V} \cup \mathcal{L} \\
\psi_c(y_t = x_j) &= \sigma(h_j^T W_c s_t), \quad x_j \in \mathcal{X}
\end{align*}
\]

where \( W_0 \in \mathbb{R}^{N \times 2d_s} \) and \( W_c \in \mathbb{R}^{d_s \times d_s} \) 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 \( \mathcal{C} \) \( y_t \) and another for the likelihood of generating \( \mathcal{G} \) \( y_t \). Formally:

\[
p(y_t | s_t, y_{t-1}, c_t, \mathbf{H}) = p(y_t, \mathbf{g} | s_t, y_{t-1}, c_t) + p(y_t, \mathbf{c} | s_t, y_{t-1}, \mathbf{H})
\]

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 \( \text{EN} \) and the target is a German sentence with SRL labels, the source sentence will be preceded by the token \( \langle 2DE-SRL \rangle \).

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., \( \text{FR} \), \( \text{DE} \)) 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 \( \text{EN} \) and \( \text{FR} \), but dative NP in \( \text{DE} \) (DativeNP)
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:

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

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 language. 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 \(y_{t-1}\), the language indicator embedding \(l_{t-1}\) and the attention context vector \(c_t\):  

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

### 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 Marquez, 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 (Hajič 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, we use 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.
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.

<table>
<thead>
<tr>
<th>Language</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL-05</td>
<td>75,187</td>
<td>4,076</td>
</tr>
<tr>
<td>CoNLL-09</td>
<td>39,279</td>
<td>10,626</td>
</tr>
<tr>
<td>CoNLL-09</td>
<td>36,020</td>
<td>2,044</td>
</tr>
<tr>
<td>v.d. Pas</td>
<td>40,075</td>
<td>2,036</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Cross-lingual Model</th>
<th># Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN - DE-SRL (Akbik, 2015)</td>
<td>63,397</td>
</tr>
<tr>
<td>EN - FR-SRL (Akbik, 2015)</td>
<td>40,827</td>
</tr>
<tr>
<td>EN - FR (UN)</td>
<td>100,000</td>
</tr>
<tr>
<td>EN - DE (Europarl)</td>
<td>100,000</td>
</tr>
</tbody>
</table>

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.

### 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).
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 \( \geq 10 \). We observed that this threshold is a good trade-off in both the amount of kept sentences (above the threshold) and average BLEU.

---

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

---

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.


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.
<table>
<thead>
<tr>
<th>Model [Filter]</th>
<th>German F-Seq Word Label</th>
<th>French F-Seq Word Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>XL-GloVe [All]</td>
<td>18.86 17.17 25.52</td>
<td>28.99 17.36 32.76</td>
</tr>
<tr>
<td>XL-BERT [All]</td>
<td>27.22 27.36 29.59</td>
<td>33.59 22.48 37.17</td>
</tr>
<tr>
<td>XL-GloVe ≥ 10</td>
<td>30.58 36.71 51.68</td>
<td>38.99 43.79 61.73</td>
</tr>
<tr>
<td>XL-BERT ≥ 10</td>
<td>36.95 41.36 55.73</td>
<td>42.66 46.52 65.32</td>
</tr>
</tbody>
</table>

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 ≥ 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) 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. 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.

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, 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.

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

---

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

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.

---

11i.e., the generated sentences for which we measured a BLEU score ≥ 10 against the source using back-translation.
12We 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.

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.

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.

---

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).
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. Kozhevennikov 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 labels 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 interactive word alignments between English and French with additional 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) propose 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


Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daihue Kawahara, Maria Antònia Martí, Lluís Márquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu,


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


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


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


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