Alignment-free Cross-lingual Semantic Role Labeling

Rui Cai and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
Rui.Cai@ed.ac.uk mlap@inf.ed.ac.uk

Abstract

Cross-lingual semantic role labeling (SRL) aims at leveraging resources in a source language to minimize the effort required to construct annotations or models for a new target language. Recent approaches rely on word alignments, machine translation engines, or preprocessing tools such as parsers or taggers. We propose a cross-lingual SRL model which only requires annotations in a source language and access to raw text in the form of a parallel corpus. The backbone of our model is an LSTM-based semantic role labeler jointly trained with a semantic role compressor and multilingual word embeddings. The compressor collects useful information from the output of the semantic role labeler, filtering noisy and conflicting evidence. It lives in a multilingual embedding space and provides direct supervision for predicting semantic roles in the target language. Results on the Universal Proposition Bank and manually annotated datasets show that our method is highly effective, even against systems utilizing supervised features.[1]

1 Introduction

Semantic role labeling (SRL) is the task of identifying the arguments of semantic predicates in a sentence and labeling them with a set of predefined relations (e.g., "who" did "what" to "whom," "when," and "where"). It has emerged as an important technology for a wide spectrum of applications ranging from machine translation (Aziz et al., 2011; Marcheggiani et al., 2018) to information extraction (Christensen et al., 2011), and summarization (Khan et al., 2015).

There have been considerable efforts on developing annotated resources for semantic role labeling (Palmer et al., 2005; Zaghouani et al., 2010) which in turn have greatly facilitated the development of the various models designed to automatically predict semantic roles. Recent years have seen the successful application of neural network models to SRL (Zhou and Xu, 2015; He et al., 2017; Marcheggiani et al., 2017) which forego the need for extensive feature engineering. Despite recent advances in representational learning, a perennial problem with building SLR systems lies in the paucity of training data since semantic role annotations are available for only a handful of the world's languages. As a result, much previous work has focused on *cross-lingual* SRL which aims at leveraging existing resources in a source language to minimize the effort required to construct a model or annotations for a new target language.

Annotation projection is a popular approach which transfers annotations from a source to a target language via automatic word alignments (Padó and Lapata, 2005; van der Plas et al., 2011; Aminian et al., 2019). Although very intuitive, it is sensitive to the quality of the parallel data, the performance of the source-language SRL model, and the accuracy of alignment tools, all of which introduce noise. *Translation-based* approaches (Täckström et al., 2012; Fei et al., 2020; Rasooli and Collins, 2015) aim to alleviate the noise brought by the the source-side labeler by directly translating the gold-standard data into the target language. A third alternative is *model transfer* where a source-language model is modified in a way that it can be directly applied to a new language, e.g., by employing cross-lingual word representations (Täckström et al., 2012; Swayamdipta et al., 2016; Daza and Frank, 2019a) and universal POS tags (McDonald et al., 2013).

Word alignment noise poses serious problems for both annotation-projection and translation-based methods (the latter still rely on alignment tools to transfer word-level labels from source to

[1] Our code and data can be downloaded from https://github.com/RuiCaiNLP/SRL_CPS.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3883–3894,
November 16–20, 2020.

target). For example, there could be many-to-one alignments, leading to semantic role conflicts in the target language. Some form of filtering is often introduced to reduce the impact of this noise, e.g., parallel sentence pairs are discarded according to projection density (Aminian et al., 2019) or alignment confidence (Fei et al., 2020). In addition, translation-based approaches rely on high performance translation engines, which are often trained on large-scale parallel corpora. Unfortunately, neither adequate MT nor high-quality parallel data can be guaranteed when dealing with low-resource languages. Model transfer is an appealing alternative, however, it relies on accurate features based on lemmas, POS tags, and syntactic parse trees (Kozhevnikov and Titov, 2013; Fei et al., 2020) which are themselves obtained with access to additional annotation. It is not realistic to assume that treebank-style resources will be available for low-resource languages.

In this paper, we propose a novel method for cross-lingual SRL which does not rely on word alignments, machine translation or pre-processing tools such as parsers or taggers. Aside from semantic role annotations in the source language, we only assume access to raw text in the form of a parallel corpus. The backbone of our model is an LSTM-based semantic role labeler jointly trained with multi-lingual word embeddings and a semantic role compressor. The compressor distills useful information pertaining to arguments, predicates and their roles from the output of the semantic role labeler (e.g., by automatically filtering unrelated or conflicting information). Importantly, the compressor lives in a multilingual space and can provide direct supervision for predicting semantic roles in the target language, sidestepping intermediaries like word-level alignments and machine translation.

For evaluation, we make use of several multilingual benchmarks. These include the Universal Proposition Bank (UPB; Akbik et al. 2016), a recently released resource which contains semi-automatically created annotations under a unified labeling scheme for several languages, and a French corpus (van der Plas et al., 2010) which follows PropBank-style annotations (Palmer et al., 2005). We also release two additional manually labeled resources in Chinese and German, which we hope will be useful for future research.[2] Experimental results show that our method is highly effective across languages and annotation schemes, even compared against systems making use of supervised features.

Our contributions can be summarizes as follows: (a) we propose a knowledge-lean model which does not rely on alignments, machine translation or sophisticated linguistic preprocessing; (b) we introduce the concept semantic role compressor which is important at filtering noisy information and can be potentially useful for other crosslingual tasks (e.g., dependency parsing); (3) we release two manually annotated datasets which will further advance cross-lingual semantic role labeling complementing previous work (Aminian et al., 2019; Fei et al., 2020) which reports result on semi-automatically created annotations).

2 Model

Figure 1 provides a schematic overview of our model. We assume we have access to semantic role annotations in a source language (e.g., English) and a parallel corpus of source-target sentences (e.g., English-French). Our model is jointly trained to predict semantic roles in the source *and* target languages. It has two main components, namely a semantic role *labeler*, and a semantic role *compressor*. The role labeler consists of:

- an input layer which takes *multilingual* word embeddings and predicate indicator embeddings as input;
- a bidirectional LSTM (BiLSTM) encoder which takes as input the representation of each word in a sentence and produces context-dependent representations;
- a biaffine scorer to calculate the score of each semantic role for each word.

While the semantic role compressor consists of:

- an input layer which again combines multilingual word embeddings and semantic role distributions for each word in the sentence;
- a bidirectional LSTM (BiLSTM) encoder which produces compressed semantic role information for an input sentence;
- a biaffine scorer which calculates the similarity between compressed representations of semantic roles and input words.

In the following sections we describe these two components more formally.

[2] Our annotations are available from https://github.com/RuiCaiNLP/ZH_DE_Datasets.

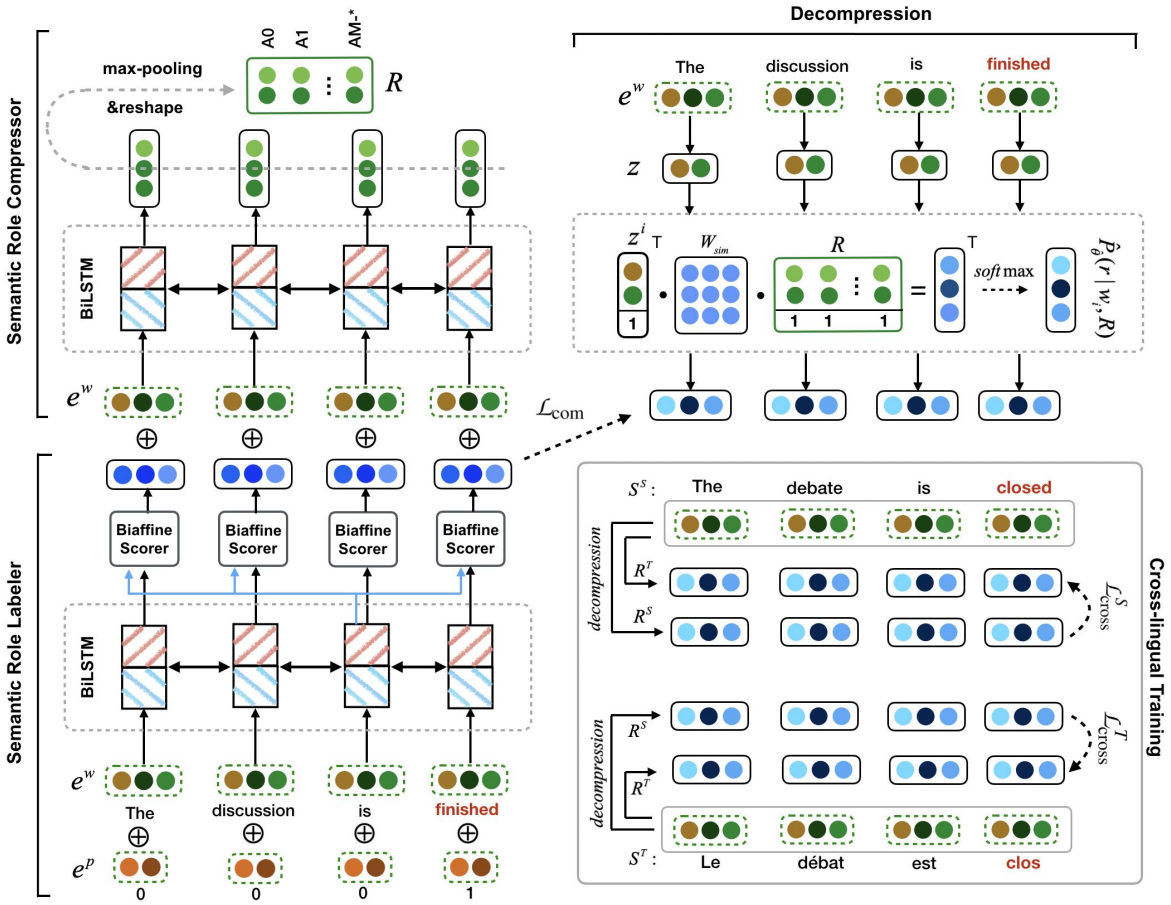

Figure 1: Model overview: semantic role labeler (left-bottom) and semantic role compressor (left-top). The right-top part presents the process of decompression after obtaining R. The right-bottom part illustrates cross-lingual training given an English-French sentence pair (S^S and S^T) from the Europarl parallel Corpus, where R^S and R^T are the output of the compressor taking S^S and S^T as input, respectively. Best viewed in color.

2.1 Semantic Role Labeler

Input Layer and Encoder For each sentence, the representation of i-th word w_i is the concatenation of *multilingual* contextualized word embeddings $e^w_{w_i}$ and *predicate* indicator embedding $e^p_{w_i}$. The former are pretrained on a large-scale unlabeled corpus and their parameters stay frozen during the training of our model. Predicate embeddings are randomly initialized and updated constantly during model training. Unlike previous supervised SRL approaches (Roth and Lapata, 2016; Cai and Lapata, 2019; He et al., 2019), our model does not make use of any syntactic information (e.g., POS-tags, dependency relations) since we cannot assume it will be available for low-resource languages.

Following Marcheggiani et al. (2017), sentences are represented using a multi-layer bi-directional LSTM (Hochreiter and Schmidhuber, 1997); the BiLSTM receives at time step t representation x for each word and recursively computes two hidden states, one for the forward pass ($\overrightarrow{h}_t$), and another one for the backward pass ($\overleftarrow{h}_t$). Each word is the concatenation of its forward and backward LSTM state vectors $h_t = \overrightarrow{h}_t \circ \overleftarrow{h}_t$.

Biaffine Role Scorer Once the high-level BiLSTM encoder produces representations h for each word, two distinct non-linear transformations are applied to predicate w_p (being considered at the time) and word w_i, respectively:

$$\begin{aligned} h'_{w_p} &= f(W_p h_{w_p} + b_p) \\ h'_{w_i} &= f(W_w h_{w_i} + b_w) \end{aligned} \tag{1}$$

where f is a non-linear activation function (we use Leaky ReLu). The score $s(r_j, h'_{w_i}, h'_{w_p})$ of semantic role r_j between current predicate w_p and word w_i is calculated as:

$$\begin{aligned} s(r_j, h'_{w_i}, h'_{w_p}) &= h'^{\top}_{w_i} W_{r_j} h'_{w_p} \\ &\quad + U_{r_j}(h'_{w_i} \circ h'_{w_p}) + b_{r_j} \end{aligned} \tag{2}$$

where W_{r_j}, U_{r_j}, and b_{r_j} are parameters specific to role r_j, and are updated during training.

Both the binaffine role scorer and SRL encoder are illustrated in Figure 1 (bottom left part).

Predicate Identification and Disambiguation The SRL labeler presented thus far assumes that predicates are known. Although in most SRL datasets predicates are explicitly annotated, such annotations are absent from unlabeled parallel data, and our model would need to automatically identify predicates if it were to be useful in practice. To this end, we run two modules on top of the sentence encoder in order to identify the predicate and disambiguate its senses. Each module is a multilayer perceptron (MLP) with a softmax layer, and is trained jointly with semantic role labeler.

2.2 Semantic Role Compressor

The semantic role compressor operates over the output of the semantic role labeler; it aims to relate each semantic role to specific words and compress this information into a fixed-size matrix.

Semantic Information Compression Although the semantic role labeler produces a label for each word in the sentence, most words will bear the label "NULL", which indicates that they are not arguments of the predicate of interest. In order to provide useful supervision to the target language, we filter out information about non-argument words. Specifically, we compress the output of the semantic role labeler into a hidden representation which only records information about arguments. In theory, each semantic role appears no more than once in a sentence, so we propose to use a fixed-size matrix $R \in \mathbb{R}^{n_r \times d_r}$ to represent compressed information, where n_r is the size of semantic role set, and d_r denotes the length of hidden representation for each semantic role.

The semantic role compressor will bind word w_i to its corresponding role. Like the semantic role labeler, the compressor also operates over word embeddings (see upper left part in Figure 1); for sentence S, word w_i is represented by $P_\theta(r|w_i, w_p, S) \circ e^w_{w_i}$, where $e^w_{w_i}$ is the multilingual embedding of w_i, and $P_\theta(r|w_i, w_p, S)$ is the probability distribution over roles produced by the semantic role labeler:

$$P_\theta(r|w_i, w_p, S) = \text{softmax}\{s(r_1, h'_{w_i}, h'_{w_p}), \ldots, s(r_{n_r}, h'_{w_i}, h'_{w_p})\} \quad (3)$$

where θ are the parameters of the semantic role labeler. Analogously to the semantic role labeler, a multi-layer BiLSTM yields sentence representations (see upper block in Figure 1). At time step t, forward and backward hidden states $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are concatenated and then fed to a non-linear layer. A max-pooling layer thereafter gathers global information from hidden features at each time step, and compresses them into a fixed-size vector:

$$R = \max_{t=1}^{n} f(W_1[\overrightarrow{h}_t \circ \overleftarrow{h}_t] + b_1) \quad (4)$$

where W_1 is a weight matrix, b_1 is a bias term for the hidden state vector, and n is the length of sentence. For the sake of decompression (see next section), R is reshaped from a vector into a matrix with n_r rows and d_r columns (see very top in Figure 1, left side).

Decompression Semantic roles in a sentence can be obtained by combining compressed information in R with the multilingual embedding of each word, and this process is referred to as decompression. Concretely, for i-th word and j-th role, we use a biaffine scorer[3] to calculate the similarity between $e^w_{w_i}$ and R_j. We first perform a non-linear transformation for word embedding $e^w_{w_i}$:

$$z_i = f(W_2 e^w_{w_i} + b_2) \quad (5)$$

where z_i contains hidden features for word w_i. And then, use a biaffine scorer to calculate the similarity score between z_i and R_j:

$$\hat{s}(z_i, R_j) = z_i \top W_{sim} R_j + U_{sim}(z_i \circ R_j) + b_{sim} \quad (6)$$

where W_{sim}, U_{sim}, and b_{sim} are parameters updated during training. For word w_i, the final probability distribution over semantic roles is obtained by applying a softmax operation on the scores of all semantic roles:

$$\hat{P}_{\hat{\theta}}(r|w_i, R) = \text{softmax}\{\hat{s}(z_i, R_1), \ldots, \hat{s}(z_i, R_{n_r})\} \quad (7)$$

where $\hat{\theta}$ are the parameters of the compressor. Figure 1 (right upper part) illustrates decompression.

Gaussian Noise In order to improve the robustness of the compressor, we inject Gaussian noise to word embeddings. This is an effective regularization method (Liu et al., 2019) which improves

[3]The score for the label "NULL" is fixed to 0, as R does not record information for non-argument words.

the model's ability to generalize to unseen inputs from different languages. The final embeddings are: $e^w = [e^w_{w_1} + N_1, ..., e^w_{w_n} + N_n]$, where $\mathbf{N} \sim \mathcal{N}(0, 0.1\mathbf{I})$ and n is the length of the sentence.

2.3 Training

In our learning setting, semantic role annotations are only available in the source-language. We therefore rely on (unlabeled) parallel data to provide cross-lingual supervision for the target-language. During each iteration, we randomly select a batch from the annotated source-language for supervised training and a batch from the parallel data for cross-lingual training.

Supervised Training We train the semantic role labeler in the source language in a supervised fashion, using a cross-entropy loss objective:

$$\mathcal{L}_{\text{ce}} = \frac{1}{n} \sum_{i=1}^{n} t_i \log P_\theta(r|w_i, w_p, S) \quad (8)$$

where n is the length of sentence and $t_i \in \mathbb{R}^{n_r}$ are one-hot ground truth representations. When training the compressor network, the objective is defined as the KL-divergence between the input distribution (produced by semantic role labeler) and the output distribution of the compressor:

$$\mathcal{L}_{\text{com}} = \frac{1}{n} \sum_{i=1}^{n} D(P_\theta(r|w_i, w_p, S), \hat{P}_{\hat{\theta}}(r|w_i, R)) \quad (9)$$

where D is a distance function between probability distributions (we use the Kullback-Leibler divergence). The final objective $\mathcal{L}_{\text{sup}}$ for supervised learning is the sum of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{com}}$.

Cross-lingual Training Given an *unlabeled* parallel source-target sentence pair (S^S and S^T), we first perform predicate identification on both sentences and randomly choose a predicate w^S_p in S^S as the current predicate of interest. We then find, amongst all words identified as predicates in S^T, predicate w^T_p which has the highest word embedding similarity with w^S_p.

By feeding word embeddings and predicate information into our model, we obtain compressed role representations R^S and R^T for source and target sentences S^S and S^T. Recall that we must apply decompression in order to obtain role specific information for S^S and S^T. Since decompression operates over multilingual representations, it is relatively straightforward to obtain semantic roles for source and target sentences. In fact, we apply R^S and R^T on both S^S and S^T and compare the outcome (see Figure 1, bottom part, right side). The training objectives are defined as:

PropBank v3	UPB					
EN	DE	FR	IT	ES	PT	FI
272,380	997	298	489	1,995	936	716

CoNLL-09	van der Plas et al. (2010)
EN	FR
39,279	1,000

ProBank v3	UPB (manually re-labeled)	
EN	ZH	DE
272,380	304	258

Table 1: Annotated data used in our experiments. We show the English source annotations (left column) used for *training* and corresponding target annotations used for *testing* in various languages.

$$\mathcal{L}^S_{\text{cross}} = \frac{1}{n_S} \sum_{i=1}^{n} D(\hat{P}_{\hat{\theta}}(r|w^S_i, R^S), \hat{P}_{\hat{\theta}}(r|w^S_i, R^T)) \quad (10)$$

$$\mathcal{L}^T_{\text{cross}} = \frac{1}{n_T} \sum_{i=1}^{n} D(\hat{P}_{\hat{\theta}}(r|w^T_i, R^S), \hat{P}_{\hat{\theta}}(r|w^T_i, R^T)) \quad (11)$$

where n_S and n_T are the length of S^S and S^T, respectively.

In order to improve the performance of the semantic role compressor on the source and target language, we train it using parallel sentence pairs by minimizing:

$$\mathcal{L}^S_{\text{com}} = \frac{1}{n_S} \sum_{i=1}^{n} D(P_\theta(r|w^S_i, w^S_p, S^S), \hat{P}_{\hat{\theta}}(r|w^S_i, R^S)) \quad (12)$$

$$\mathcal{L}^T_{\text{com}} = \frac{1}{n_T} \sum_{i=1}^{n} D(P_\theta(r|w^T_i, w^T_p, S^T), \hat{P}_{\hat{\theta}}(r|w^T_i, R^T)) \quad (13)$$

The final training loss during cross-lingual training $\mathcal{L}_{\text{cross}}$ is the sum of above losses:

$$\mathcal{L}_{\text{cross}} = \mathcal{L}^S_{\text{cross}} + \mathcal{L}^T_{\text{cross}} + \mathcal{L}^S_{\text{com}} + \mathcal{L}^T_{\text{com}} \quad (14)$$

3 Experiments

3.1 Datasets

We trained our model using English as the source language and obtained semantic role labelers in German (DE), Spanish (ES), Finish (FI), French (FR), Italian (IT), Portuguese (PT), and Chinese (ZH). For English, we used the Proposition Bank (v3; Palmer et al. 2005) and the annotations provided as part of the CoNLL-09 shared task (Hajič

et al., 2009). We used the Europarl parallel corpus (Koehn, 2005) for the European languages and a large-scale EN-ZH parallel corpus (Xu, 2019) for Chinese. We provide details regarding the size of the parallel corpora in the Appendix. We compared our model against previous methods on the Universal Proposition Bank (UPB, v1.0; Akbik et al. 2016), which is built upon the Universal Dependency Treebank (UDT, v1.4) and the Proposition Bank (PB, v3.0). All languages in the UBP follow a unified dependency-based SRL annotation scheme. In order to comply with this scheme, we converted argument spans in the English Proposition Bank to dependency-based arguments by labeling the syntactic head of each span.

As UPB adopts a semi-automatic annotation procedure, it unavoidably contains a certain amount of errors. We therefore also tested our model on manually annotated datasets which are few and far between, presumably due to the labeling effort involved. An existing dataset (van der Plas et al., 2010) provides SRL labels for French following an annotation scheme similar to CoNLL-09 for English (Hajič et al., 2009). The CoNLL-09 shared task provides semantic role annotations for seven languages, but the role sets differ across languages, and it is far from trivial to unify them. To this end, we created two manual resources, by randomly sampling 258 German and 304 Chinese sentences from UPB. The manual annotation was performed by native speakers following the annotation guidelines of UPB which in turn follows the English Proposition Bank. Table 1 provides a breakdown of labeled data used in our experiments.

3.2 Model Configuration

Our model was implemented in PyTorch and optimized using the Adam optimizer (Kingma and Ba, 2014). Word embeddings were initialized using the officially released multilingual BERT (base; cased version; Devlin et al. 2019). The parameters of BERT are fixed during training in order to preserve the cross-lingual nature of the embeddings. Hyperparameter values (for all languages) are shown in Table 2.

Hyperparameters	value
multilingual BERT embeddings size	768
predicate indicator embeddings size	16
batch size	30
learning rate	0.001
Bi-LSTM hidden states size	400
BiLSTM depth	3
hidden feature size in biaffine scorer	300
Bi-LSTM hidden states size	256
BiLSTM depth	2
compressed role representation size	30
hidden feature size in biaffine scorer	30

Table 2: Hyperparameter settings for input and training (first block), semantic role labeler (second block) and semantic role compressor (third block).

3.3 Results on Universal Proposition Bank

We compared our model against several baselines on the UPB test set. These include two transfer methods: *Bootstrap* (Aminian et al., 2017) and *CModel* (Aminian et al., 2019), which perform annotation projection through parallel data and filter word alignments empirically. We also report the results of two strong mixture-of-experts models which focus on combining language specific features automatically (*MOE*; Guo et al. 2018), and also on learning language-invariant features with a multinomial adversarial network as a shared feature extractor (*MAN-MOE*; Chen et al. 2019). We also include a recently proposed translation-based model (*PGN*; Fei et al. 2020) which performs competitively on UPB; this system directly translates the source annotated corpus into the target language, and then performs annotation projection and filtering similar to *Bootstrap* and *CModel*.

Table 3 shows labeled F-scores (using automatically predicted predicate senses) on the test portion of the Universal Proposition Bank. The various languages are ordered according to their typological distance to English based on word order (Ahmad et al., 2019a) with Portuguese being closest and Finnish farthest. As can be seen, our model outperforms previous systems on DE, FR and PT, and is on average better. It is worth noting that, in addition to pretrained word-alignment tools, both *Bootstrap* and *PGN* utilize supervised part-of-speech (POS) tags for the target language. However, our model still achieves the best average F-score (61.1%) without employing any additional features. Pairwise differences in F_1 between our model *MAN-MOE*, *CModel*, and *PGN*) are all statistically significant ($p < 0.05$) using stratified shuffling (Noreen, 1989).

3.4 Results on Human-labeled Data

As UPB annotations are semi-automatic and possibly contain projection errors, we further compared

Models	PT	FR	ES	IT	DE	FI	avg
Dist. to EN	0.09	0.09	0.12	0.12	0.14	0.20	0.13
Bootstrap	53.9	63.4	52.2	52.3	55.0	53.1	55.0
CModel	56.5	58.5	56.0	55.5	57.0	**58.9**	57.1
MAN-MOE	55.2	65.3	**62.8**	57.1	64.3	52.3	59.4
MoE	55.5	63.3	60.3	56.7	63.2	50.6	58.2
PGN	56.0	64.8	62.5	**58.7**	65.0	54.5	60.3
Ours	**57.8**	**66.2**	61.5	57.6	**65.7**	57.6	**61.1**

Table 3: Results (F_1) on UPB test sets for six languages. Results for comparison systems are taken from previous papers (Aminian et al., 2019; Fei et al., 2020).

Models	FR	DE	ZH	avg
CModel	68.5	66.9	62.3	65.9
MAN-MOE	72.8	69.2	64.7	68.9
PGN	73.2	70.1	65.4	69.5
Ours	**75.3**	**71.4**	**68.5**	**71.7**

Table 4: Results (F_1) on manually annotated test sets for German, French, and Chinese. Pairwise differences between our model and previous systems are all statistically significant ($p < 0.05$) using stratified shuffling (Noreen, 1989).

our model against manual annotations on French, German, and Chinese (see Table 1). Since previous models have not provided results on these datasets, we re-implemented three strong comparison systems, i.e., *CModel*, *MAN-MOE*, and *PGN*. Details on our implementation are in the Appendix.

Our results are summarized in Table 4, where languages are ordered in terms of their word order distance to English (Ahmad et al., 2019a). We note that our approach significantly outperforms previously published models on these three languages. All systems perform best on French which is perhaps unsurprising given that it is closest to English and worst on Chinese which is least related to English. This suggests that transferring SRL annotations between languages with similar word orders could be an easier task.

3.5 Ablation Study and Analysis

To investigate the contribution of the semantic role compressor and cross-lingual training, we conducted a series of ablation studies on the manually annotated DE, FR, and ZH datasets. Evaluation in these experiments excludes the accuracy of predicate disambiguation, since we wish to focus on the SRL model per se.

Our experiments are summarized in Table 5. The first block shows the performance of the full model.

Models	DE	FR	ZH
Ours	63.4	68.8	60.4
w/o BERT	47.7	52.6	44.5
w/o BERT (+position)	55.3	60.5	53.0
w/o Gaussian noise	61.7	66.2	57.7
w/o cross-lingual training	52.5	59.8	49.5
w/o compressor (+attention)	51.7	59.5	47.1

Table 5: Ablations on manually annotated datasets.

In the second block, we assess the effect of different kinds of word representations. First, we substitute multilingual BERT embeddings with MUSE embeddings (Lample et al., 2018), which were obtained by aligning (monolingual) fastText embeddings for various languages onto a universal space. We can see that the performance of our model drops significantly. One important reason is that MUSE embeddings are not contextualized; as a result, a word appearing multiple times in the same sentence will receive the same embedding, even when it occupies different semantic roles, which in turn leads to conflicts during decompression. One solution is concatenating MUSE with word position embeddings during compression and decompression (see Appendix for details). This improves SRL performance from 47.7% (DE), 52.6% (FR), and 44.5% (ZH) to 55.3%, 60.5% and 53.0%, but is still inferior to the original model. Next, we remove Gaussian noise from the model and as can be seen there is a drop in performance indicating that it further boosts SRL accuracy.

In the third block, we remove cross-lingual training, and observe a significant drop in F-score over the full model. In order to verify the need for semantic role compression, we substitute the compressor with an attention-based module (Bahdanau et al., 2015) and proceed to train our model as described in Section 2.3. Specifically, we obtain soft alignments and use these to weight all annotations $P_\theta(r|w_i, w_p, S)$, thereby obtaining an expectation over role assignments. The alignment module and the basic semantic role labeler are trained jointly during cross-lingual training. We can see that performance drops substantially for all three languages compared to the full model. The reason might be that the output of the semantic role labeler is noisy and attention often creates labeling conflicts (e. two words show high confidence for the same semantic role). However, our compressor can filter out this noise and resolve conflicts more effectively.

French	*SRL only*	Ours	Frequency(%)
A0	71.9	83.6	26%
A1	65.7	78.8	37%
A2	37.8	43.6	7%
AM-*	46.7	48.5	30%

Chinese	*SRL only*	Ours	Frequency(%)
A0	59.2	63.7	18%
A1	59.9	74.4	38%
A2	38.6	65.6	15%
AM-*	36.0	37.3	29%

Table 6: Results (F_1) on French and Chinese test sets grouped by gold role labels.

In Table 6, we present model performance for French and Chinese for different (gold) role labels. We compare the full model against an *SRL only* model without cross-lingual training. As shown in Table 6, cross-lingual training improves SRL performance in French and Chinese on all semantic roles.[4] For French, the most significant improvement comes from A1; for Chinese, cross-lingual training benefits labeling A1 and A2 significantly. Compared with A0, A1, and A2, the improvements on AM-* (modifiers for current predicate) are modest for both French and Chinese. One possible reason is that the head words of A0, A1 and A2 are usually nouns or adjectives, which tend to have fixed positions in parallel sentence pairs. However, modifiers can be optional and have more varied positions within and across languages, which increases the difficulty for cross-lingual learning.

4 Related Work

There has been a great deal of interest in cross-lingual transfer learning for SRL (Padó and Lapata, 2009; van der Plas et al., 2011; Kozhevnikov and Titov, 2013; Tiedemann, 2015; Zhao et al., 2018; Chen et al., 2019; Aminian et al., 2019; Fei et al., 2020). The majority of previous work has focused on two types of approaches, namely annotation projection and model transfer.

A variety of methods have been proposed to improve the quality of annotation projections due to alignment noise. These range from word and argument filtering techniques (Padó and Lapata, 2005, 2009), to learning syntax and semantics jointly (van der Plas et al., 2011), and iterative bootstrapping (Akbik et al., 2015; Aminian et al., 2017). In an attempt to reduce the reliance on supervised lexico-syntactic features for the target language, Aminian et al. (2019) make use of word and character features, and filter projected annotations according to projection density. Model transfer does not require parallel corpora or word alignment tools; nevertheless, it relies on accurate features such as POS tags (McDonald et al., 2013) or syntactic parse trees (Kozhevnikov and Titov, 2013) to enhance the ability to generalize across languages. Adversarial training is commonly used to extract language-agnostic features thereby improving the performance of cross-lingual systems (Chen et al., 2019; Ahmad et al., 2019b).

Translation-based approaches have been gaining popularity in cross-lingual dependency parsing (Rasooli and Collins, 2015; Tiedemann, 2015; Conneau et al., 2018) and have recently been applied to SRL (Fei et al., 2020). Daza and Frank (2019b) propose a cross-lingual encoder-decoder model that simultaneously translates *and* generates sentences with semantic role annotations in a resource-poor target language. Rather than *creating* annotations or models for a target language, other work aims to exploit the similarities between languages. Mulcaire et al. (2018) combine resources for multiple languages to create *polyglot* semantic role labelers and show that polyglot training can result in better labeling accuracy than a monolingual labeler.

An obstacle for developing cross-lingual SRL models is the absence of a unified annotation scheme for all languages. Although the CoNLL-09 shared task (Hajič et al., 2009) provides annotations for seven languages, the labeling schemes and role sets are not shared. To this end, van der Plas et al. (2010) build a French SRL dataset, following an annotation scheme similar to CoNLL-09 for English. Some recent cross-lingual SRL models (Aminian et al., 2017, 2019; Fei et al., 2020) make use of the publicly available Universal Proposition Bank (UPB; Akbik et al. 2015; Akbik and Li 2016), which annotates predicates and semantic roles following the English Proposition Bank 3.0 (Palmer et al., 2005). Since annotation projection is involved in the construction of UPB, the quality of UPB is also influenced by the quality of the parallel data, the performance of the source-language SRL model, and the accuracy of alignment tools.

[4]The proportion of A2 in Chinese is higher than in French, as the two languages follow different annotation schemes.

5 Conclusions

In this paper we developed a cross-lingual SRL model and demonstrated it can effectively leverage unlabeled parallel data without relying on word alignments or any other external tools. We have also contributed two quality controlled datasets (compatible with PropBank-style guidelines) which we hope will be useful for the development of cross-lingual models. Directions for future work are many and varied. Although our focus has been on dependency-based SRL, our model can be easily adapted to span-based annotations (Carreras and Màrquez, 2005; Pradhan et al., 2013). In this case, the semantic role compressor could be modified to represent entire spans rather than just head words while decompression would remain unchanged (it would still output a probability distribution for each word over all semantic roles). We also plan to extend our framework to semi-supervised learning, where a small number of annotations might also be available in the target language.

Acknowledgments

This work was supported by the European Research Council (award number 681760, "Translating Multiple Modalities into Text"). We thank the anonymous reviewers for their helpful feedback; we are grateful to Ling Jiang and Sabine Webber for their meticulous annotation efforts in the creation of our SRL datasets.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019a. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019b. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 372–382, Hong Kong, China. Association for Computational Linguistics.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

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

Alan Akbik and Yunyao Li. 2016. K-SRL: Instance-based learning for semantic role labeling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 599–608, Osaka, Japan. The COLING 2016 Organizing Committee.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2017. Transferring semantic roles using translation and syntactic information. *arXiv preprint arXiv:1710.01411*.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2019. Cross-lingual transfer of semantic roles: From raw text to semantic roles. *arXiv preprint arXiv:1904.03256*.

Wilker Aziz, Miguel Rios, and Lucia Specia. 2011. Shallow semantic trees for smt. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 316–322, Edinburgh, Scotland.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, California.

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

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In

Proceedings of the 6th International Conference on Konwledge Capture, pages 113–119, Banff, Canada.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2019a. Translate and label! an encoder-decoder approach for cross-lingual semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 603–615, Hong Kong, China. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2019b. Translate and label! an encoder-decoder approach for cross-lingual semantic role labeling. *arXiv preprint arXiv:1908.11326*.

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

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. *arXiv preprint arXiv:2004.06295*.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5350–5359, Hong Kong, China. Association for Computational Linguistics.

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

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

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

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.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1297–1303, Hong Kong, China. Association for Computational Linguistics.

Diego Marcheggiani, Joost Bastings, and Ivan Titov. 2018. Exploiting semantics in neural machine translation with graph convolutional networks. In *Proceedings of the the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, New Orleans, US.

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.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

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

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Sebastian Padó and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 859–866. Association for Computational Linguistics.

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

Marth Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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

Lonneke van der Plas, Tanja Samardzić, and Paola Merlo. 2010. Cross-lingual validity of PropBank in the manual annotation of French. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 113–117, Uppsala, Sweden. Association for Computational Linguistics.

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

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338, Lisbon, Portugal. Association for Computational Linguistics.

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.

Swabha Swayamdipta, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Greedy, joint syntactic-semantic parsing with stack LSTMs. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 187–197, Berlin, Germany. Association for Computational Linguistics.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Jörg Tiedemann. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, 109, pages 191–199. Linköping University Electronic Press.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. 2010. The revised Arabic PropBank. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 222–226, Uppsala, Sweden. Association for Computational Linguistics.

Han Zhao, Shanghang Zhang, Guanhang Wu, Geoffrey J Gordon, et al. 2018. Multiple source domain adaptation with adversarial learning.

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.

A Positional Features

When using non-contextualized MUSE embeddings (see the ablation study in Section 3.5), we resort to position embeddings to distinguish words appearing multiple times in the same sentence. Unlike standard transformers where positional features are bound to word indices, the positional features we used for word w_i just record the number of words which are same as w_i and appeared before w_i (shown in Figure 2).

Figure 2: Positional features for English-French parallel sentences.

We adopt these new positional features for two reasons. Firstly, during cross-lingual training, the length of parallel sentences S^S and S^T is usually different. More importantly, for i-th word w_i in S^S, its correspondence $w_j^{'}$ in S^T is not the i-th word in S^T in most cases. When performing cross-lingual training, it is important that w_i and $w_j^{'}$ have the same position embeddings, so that they can obtain similar result after decompression. As shown in Figure 2, *he* ("il" in French) appears twice in the English sentence, and its French counterpart shares the same positional features. Experimental results show that positional features can effectively improve cross-lingual training. However, there are still cases when the word order changes dramatically after translation and our position features do not work. The only solution seems to be to use contextualized embeddings like multilingual BERT or multilingual ELMo, where every word in a sentence will be assigned unique embeddings.

B External Tools

When implementing previous models, we used Google Translate[5] as our translation engine, and giza++[6] to obtain word alignments. Besides source-language corpus, translated corpus is also used for the training of *PGN* and *MAN-MOE*. When prepossessing the Chinese part in EN-ZH parallel corpus (containing about 5 million sentence pairs), we use Jieba[7] for tokenization. The Chinese testset in UPB is in traditional Chinese, and we use Zhtools[8] to convert it to simplified Chinese to be compatible with our EN-ZH parallel corpus which is also in simplified Chinese.

C Parallel Corpus Size

Europarl provides parallel data between English and 21 European languages. We evaluated our model on six European languages. Table 7 give the size of the various parallel corpora used in our experiments.

Language	size
German	1,920,209
Spanish	1,965,734
Finnish	1,924,942
Italian	1,909,115
Portuguese	1,960,407
French	2,007,723

Table 7: Number of sentence pairs in Europarl for six languages.

[5] ttps://translate.google.com/

[6] https://github.com/moses-smt/giza-pp

[7] https://github.com/fxsjy/jieba

[8] https://github.com/skydark/nstools/tree/master/zhtools