Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources

Joo-Kyung Kim†, Young-Bum Kim‡, Ruhi Sarikaya†, Eric Fosler-Lussier†
†The Ohio State University, Columbus, OH 43210, USA
‡Amazon Alexa, Seattle, WA 98121, USA

Abstract

Training a POS tagging model with cross-lingual transfer learning usually requires linguistic knowledge and resources about the relation between the source language and the target language. In this paper, we introduce a cross-lingual transfer learning model for POS tagging without ancillary resources such as parallel corpora. The proposed cross-lingual model utilizes a common BLSTM that enables knowledge transfer from other languages, and private BLSTMs for language-specific representations. The cross-lingual model is trained with language-adversarial training and bidirectional language modeling as auxiliary objectives to better represent language-general information while not losing the information about a specific target language. Evaluating on POS datasets from 14 languages in the Universal Dependencies corpus, we show that the proposed transfer learning model improves the POS tagging performance of the target languages without exploiting any linguistic knowledge between the source language and the target language.

1 Introduction

Bidirectional Long Short-Term Memory (BLSTM) based models (Graves and Schmidhuber, 2005), along with word embeddings and character embeddings, have shown competitive performance on Part-of-Speech (POS) tagging given sufficient amount of training examples (Ling et al., 2015; Lample et al., 2016; Plank et al., 2016; Yang et al., 2017).

Given insufficient training examples, we can improve the POS tagging performance by cross-lingual POS tagging, which exploits affluent POS tagging corpora from other source languages. This approach usually requires linguistic knowledge or resources about the relation between the source language and the target language such as parallel corpora (Täckström et al., 2013; Duong et al., 2013; Kim et al., 2015a; Zhang et al., 2016), morphological analyses (Hana et al., 2004), dictionaries (Wisniewski et al., 2014), and gaze features (Barrett et al., 2016).

Given no linguistic resources between the source language and the target language, transfer learning methods can be utilized instead. Transfer learning for cross-lingual cases is a type of transductive transfer learning, where the input domains of the source and the target are different (Pan and Yang, 2010) since each language has its own vocabulary space. When the input space is the same, lower layers of hierarchical models can be shared for knowledge transfer (Collobert et al., 2011; Kim et al., 2015b; Yang et al., 2017), but that approach is not directly applicable when the input spaces differ.

Yang et al. (2017) used shared character embeddings for different languages as a cross-lingual transfer method while using different word embeddings for different languages. Although the approach showed improved performance on Named Entity Recognition, it is limited to character-level representation transfer and it is not applicable for knowledge transfer between languages without overlapped alphabets.

In this work, we introduce a cross-lingual transfer learning model for POS tagging requiring no cross-lingual resources, where knowledge transfer is made in the BLSTM layers on top of word embeddings and character embeddings. Inspired by Kim et al. (2016)’s multi-task slot-filling model, our model utilizes a common BLSTM for representing language-generic information, which al-
allows knowledge transfer from other languages, and private BLSTMs for representing language-specific information. The common BLSTM is additionally encouraged to be language-agnostic with language-adversarial training (Chen et al., 2016) so that the language-general representations to be more compatible among different languages.

Evaluating on POS datasets from 14 different target languages with English as the source language in the Universal Dependencies corpus 1.4 (Nivre et al., 2016), the proposed model showed significantly better performance when the source language and the target language are in the same language family, and competitive performance when the language families are different.

2 Model

Cross-Lingual Training Figure 1 shows the overall architecture of the proposed model. The baseline POS tagging model is similar to Plank et al. (2016)’s model, and it corresponds to having only word+char embeddings, common BLSTM, and Softmax Output in Figure 1. Given an input word sequence, a BLSTM is used for the character sequence of each word, where the outputs of the ends of the character sequences from the forward LSTM and the backward LSTM are concatenated to the word vector of the current word to supplement the word representation. These serve as an input to a BLSTM, and an output layer are used for POS tag prediction.

For the cross-lingual transfer learning, the character embedding, the BLSTM with the character embedding (Yang et al., 2017), and the common BLSTM are shared for all the given languages while word embeddings and private BLSTMs have different parameters for different languages.

The outputs of the common BLSTM and the private BLSTM of the current language are summed to be used as the input to the softmax layer to predict the POS tags of given word sequences. The loss function of the POS tagging can be formulized as:

\[ \mathcal{L}_p = - \sum_{i=1}^{S} \sum_{j=1}^{N} p_{i,j} \log (\hat{p}_{i,j}) , \]  

(1)

where \( S \) is the number of sentences in the current minibatch, \( N \) is the number of words in the current sentence, \( p_{i,j} \) is the label of the \( j \)-th tag of the \( i \)-th sentence in the minibatch, and \( \hat{p}_{i,j} \) is the predicted tag. In addition to this main objective, two more objectives for improving the transfer learning are described in the following subsections.

Language-Adversarial Training We encourage the outputs of the common BLSTM to be language-agnostic by using language-adversarial training (Chen et al., 2016) inspired by domain-adversarial training (Ganin et al., 2016; Bousmalis et al., 2016). First, we encode a BLSTM output sequence as a single vector using a CNN/MaxPool encoder, which is implemented the same as a CNN for text classification (Kim, 2014). The encoder is with three convolution filters whose sizes are 3, 4, and 5. For each filter, we pass the BLSTM output sequence as the input sequence and obtain a single vector from the filter output by using max pooling, and then \( \tanh \) activation function is used for transforming the vector. Then, the vector outputs of the three filters are concatenated and forwarded to the language discriminator through the gradient reversal layer. The discriminator is implemented

---

1We also tried isolated character-level modules but the overall performance was worse.
as a fully-connected neural network with a single
hidden layer, whose activation function is Leaky
ReLU (Maas et al., 2013), where we multiply 0.2
to negative input values as the outputs.

Since the gradient reversal layer is below the
language classifier, the gradients minimizing lan-
guage classification errors are passed back with
opposed sign to the sentence encoder, which ad-
versarially encourages the sentence encoder to be
language-agnostic. The loss function of the lan-
guage classifier is formulated as:

$$L_a = - \sum_{i=1}^{S} l_i \log \hat{f}_i,$$

where \( S \) is the number of sentences, \( l_i \) is the lan-
guage of the \( i \)-th sentence, and \( \hat{f}_i \) is the softmax
output of the tagging. Note that though the lan-
guage classifier is optimized to minimize the lan-
guage classification error, the gradient from the
language classifier is negated so that the bottom
layers are trained to be language-agnostic.

**Bidirectional Language Modeling** Rei (2017)
showed the effectiveness of the bidirectional lan-
guage modeling objective, where each time step of
the forward LSTM outputs predicts the word of the
next time step, and each of the backward LSTM
outputs predicts the previous word. For example,
if the current sentence is “I am happy”, the forward
LSTM predicts “am happy” and the back-
ward LSTM predicts “<bos> I am”. This objective
encourages the BLSTM layers and the embed-
ing layers to learn linguistically general-purpose
representations, which are also useful for specific
downstream tasks (Rei, 2017). We adopted the
bidirectional language modeling objective, where
the sum of the common BLSTM and the private
BLSTM is used as the input to the language model-
ing module. It can be formulated as:

$$L_t = - \sum_{i=1}^{S} \sum_{j=1}^{N} \log (P(w_{j+1}|f_j)) +$$
$$\log (P(w_{j-1}|b_j)),$$

where \( f_j \) and \( b_j \) represent the \( j \)-th outputs of the
forward direction and the backward direction, re-
spectively, given the output sum of the common
BLSTM and the private BLSTM.

All the three loss functions are added to be op-
timized altogether as:

$$L = w_s (L_p + \lambda L_a + \lambda L_t),$$

where \( \lambda \) is gradually increased from 0 to 1 as
epoch increases so that the model is stably trained
with auxiliary objectives (Ganin et al., 2016). \( w_s \)
is used to give different weights to the source lan-
guage and the target language. Since the source
language has a larger train set and we are focusing
on improving the performance of the target lan-
guage, \( w_s \) is set to 1 when training the target lan-
guage. For the source language, instead, it is set as
the size of the target train set divided by the size
of the source train set.

### 3 Experiments

For the evaluation, we used the POS datasets from
14 different languages in Universal Dependencies
corpus 1.4 (Nivre et al., 2016). We used English
as the source language, which is with 12,543 train-
ing sentences.\(^2\) We chose datasets with 1k to 14k
training sentences. The number of tag labels dif-
fers for each language from 15 to 18 though most
of them are overlapped within the languages.

Table 1 shows the POS tagging accuracies of
different transfer learning models when we limited
the number of training sentences of the target lan-
guages to be the same as 1,280 for fair comparison
among different languages. The remainder train-
ing examples of the target languages are still used
for both language-adversarial training and bidirec-
tional language modeling since the objectives do
not require tag labels. Training with only the train
sets in the target languages (\( c \)) showed 91.61% on
average. When bidirectional language modeling
objective is used (\( c, l \)), the accuracies were signifi-
cantly increased to 92.82% on average. Therefore,
we used the bidirectional language modeling for
all the transfer learning evaluations.

With transfer learning, the three cases of using
only the common BLSTM (\( c \)), using only the private
BLSTMs (\( p \)), and using both (\( c, p \)) were evalu-
ated. They showed better average accuracies than target only cases, but they showed
mixed results. However, our proposed model
\((c, p, l + a)\), which utilizes both the common
BLSTM with language-adversarial training and
the private BLSTMs, showed the highest average
score, 93.26%. For all the Germanic languages,
where the source language also belongs to, the
accuracies are significantly higher than those of

\(^2\)The accuracies of English POS tagging are 94.01 and
94.33 for models without the bidirectional language model-
ing and with it, respectively.
other transfer learning models. For the languages belonging to Slavic, Romance, or Indo-Iranian, our model shows competitive performance with the highest average accuracies among the compared models. Since languages in the same family are more likely to be similar and compatible, it is expected that the gain from the knowledge transfer to the languages in the same family to be higher than transferring to the languages in different families, which was shown in the results. This shows that utilizing both language-general representations that are encouraged to be more language-agnostic and language-specific representations effectively helps improve the POS tagging performance with transfer learning.

Table 2 shows the results when using 320 tag-labeled training sentences. In this case, transfer learning methods still show better accuracies than target-only approaches on average. However, the performance gain is weakened compared to using 1,280 labeled training sentences and there are some mixed results. In several cases, just utilizing private BLSTMs without the common BLSTM showed better accuracies than utilizing the common BLSTM.

When training with only 32 tag-labeled sentences, which is an extremely low-resourced setting, transfer learning methods still showed better accuracies than target-only methods on average. However, not using the common BLSTM in transfer learning models showed better performance than using it on average. The main reason would be that we are not given a sufficient number of labeled training sentences to train both the common BLSTM and the private BLSTMs. In this case, just having private BLSTMs without the common BLSTM can show better performance. We also evaluated the opposite cases, which use all the tag-labeled training sentences in the target languages, and they showed mixed results. For example, the accuracy of German with the target only model is 93.31% while that of the proposed model is 93.04%. This is expected since transfer learning is effective when the target train set is small.

An extension of this work is utilizing multiple languages as the source languages. Since we have four languages for each of Germanic, Slavic, and Romance language families, we evaluated the performance of those languages using the other languages in the same families as the source languages expecting that languages in the same language family are more likely to be helpful each other. For the efficiency, we performed multi-task learning for multiple languages rather than differentiating the targets from sources. When we tried to use 1,280, 320, and 32 tag-labeled training sentences for each language in the multi-source settings, the results showed noticeably better performance than using it on average. The results in detail are shown in the first authors dissertation Kim (2017).

Table 1: POS tagging accuracies (%) when setting the numbers of the tag-labeled training examples of the target languages to be the same as 1,280 (The remaining training examples are still used for the language modeling and the adversarial training.)

<table>
<thead>
<tr>
<th>Language Family</th>
<th>Language</th>
<th>Target only</th>
<th>Source (English)</th>
<th>Target only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germanic</td>
<td>Swedish</td>
<td>93.26</td>
<td>94.36</td>
<td>93.49</td>
</tr>
<tr>
<td></td>
<td>Danish</td>
<td>92.13</td>
<td>93.34</td>
<td>93.76</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td>83.24</td>
<td>85.20</td>
<td>84.92</td>
</tr>
<tr>
<td></td>
<td>German</td>
<td>89.27</td>
<td>90.06</td>
<td>90.30</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>89.47</td>
<td>90.78</td>
<td>90.74</td>
</tr>
<tr>
<td>Slavic</td>
<td>Slovenian</td>
<td>93.06</td>
<td>93.83</td>
<td>94.06</td>
</tr>
<tr>
<td></td>
<td>Polish</td>
<td>91.30</td>
<td>91.69</td>
<td>92.11</td>
</tr>
<tr>
<td></td>
<td>Slovak</td>
<td>86.53</td>
<td>90.11</td>
<td>89.88</td>
</tr>
<tr>
<td></td>
<td>Bulgarian</td>
<td>93.45</td>
<td>95.33</td>
<td>95.50</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>91.09</td>
<td>92.48</td>
<td>92.74</td>
</tr>
<tr>
<td>Romance</td>
<td>Romanian</td>
<td>93.20</td>
<td>94.17</td>
<td>94.05</td>
</tr>
<tr>
<td></td>
<td>Portuguese</td>
<td>94.23</td>
<td>95.42</td>
<td>95.15</td>
</tr>
<tr>
<td></td>
<td>Italian</td>
<td>93.80</td>
<td>95.79</td>
<td>95.61</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>91.94</td>
<td>93.34</td>
<td>93.31</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>93.29</td>
<td>94.54</td>
<td>94.09</td>
</tr>
<tr>
<td>Indo-Iranian</td>
<td>Persian</td>
<td>93.91</td>
<td>94.68</td>
<td>94.79</td>
</tr>
<tr>
<td></td>
<td>Hungarian</td>
<td>93.20</td>
<td>94.40</td>
<td>94.66</td>
</tr>
<tr>
<td></td>
<td>Avg</td>
<td>91.61</td>
<td>92.82</td>
<td>92.98</td>
</tr>
</tbody>
</table>
formance than the results of using English as a single source language. Considering that utilizing 1,280*3=3,840, 320*3=960, or 32*3=96 tag labels from three other languages showed better results than using 12,543 English tag labels as the source, we can see that the knowledge transfer from multiple languages can be more helpful than that from single resource-rich source language. We also tried to use Wasserstein distance (Arjovsky et al., 2017) for the adversarial training in the multi-source settings, but there were no significant differences on average.⁴

### Implementation Details
All the models were optimized using ADAM (Kingma and Ba, 2015)⁵ with minibatch size 32 for total 100 epochs and we picked the parameters showing the best accuracy on the development set to report the score on the test set. The dimensionalites of all the BLSTM related layers follow Plank et al. (2016)’s model. Each word vector is 128 dimensional and each character vector is 100 dimensional. They are randomly initialized with Xavier initialization (Glorot and Bengio, 2010). For stable training, we use gradient clipping, where the threshold is set to 5. The dimensionality of each hidden output of LSTMs is 100, and the hidden outputs of both forward LSTM and backward LSTM are concatenated, thereby the output of each BLSTM for each time step is 200. Therefore, the input to the common BLSTM and the private BLSTM is 128+200=328 dimensional. The inputs and the outputs of the BLSTMs are regularized with dropout rate 0.5 (Pham et al., 2014). For the consistent dropout usages, we let the dropout masks be identical for all the time steps of each sentence (Gal and Ghahramani, 2016). For all the BLSTMs, forget biases are initialized with 1 (Jozefowicz et al., 2015) and the other biases are initialized with 0. Each convolution filter output for the sentence encoding is 64 dimensional, and the three filter outputs are concatenated to represent each sentence with a 192 dimensional vector.

### 4 Conclusion
We introduced a cross-lingual transfer learning model for POS tagging which uses separate BLSTMs for language-general and language-specific representations. Evaluating on 14 different languages, including the source language, improved tagging accuracies in almost all the cases. Specifically, our model showed noticeably better performance when the source language and the target languages belong to the same language family, and competitively performed with the highest average accuracies for target languages in different families.

### Acknowledgments
We thank the anonymous reviewers for their helpful comments. All the experiments in this work were conducted with machines at Ohio Supercomputer Center (1987).

---

⁴The extended work in detail are shown in Kim (2017).
⁵learning rate=0.001, β₁ = 0.9, β₂ = 0.999, ϵ = 1e⁻⁸.
References


Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa,


Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout improves recurrent neural networks for handwriting recognition.


