# 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-


*Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2822–2828
Copenhagen, Denmark, September 7–11, 2017. 

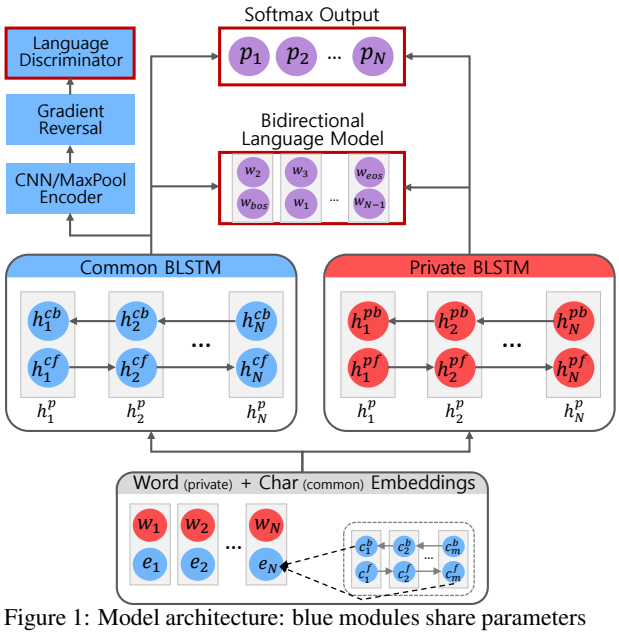


Figure 1: Model architecture: blue modules share parameters for all the languages and red modules have different parameters for different languages. $w_i$ and $e_i$ denote the $i$-th word vector and the $i$-th character vector composition, respectively. $h_i^{cf}$, $h_i^{cb}$, $h_i^{pf}$, and $h_i^{pb}$ denote the $i$-th hidden outputs of the forward common LSTM, the backward common LSTM, the forward private LSTM, and the backward private LSTM, respectively. $h_i^c$ and $h_i^p$ denote the concatenated output of the common BLSTM and the private BLSTM, respectively. Violet circles represent target labels that are predicted with different parameters for different languages, where the inputs are output summation of the common BLSTM and the private BLSTM. The model is trained with three objectives denoted with red boxes.

lows 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),[1] 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 formulate as:

$$\mathcal{L}_p = -\sum_{i=1}^{S}\sum_{j=1}^{N} p_{i,j} \log\left(\hat{p}_{i,j}\right), \qquad (1)$$

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

[1] We 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 language classification errors are passed back with opposed sign to the sentence encoder, which adversarially encourages the sentence encoder to be language-agnostic. The loss function of the language classifier is formulated as:

$$\mathcal{L}_a = -\sum_{i=1}^{S} l_i \log \hat{l}_i, \tag{2}$$

where $S$ is the number of sentences, $l_i$ is the language of the $i$-th sentence, and $\hat{l}_i$ is the softmax output of the tagging. Note that though the language classifier is optimized to minimize the language 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 language 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 <eos>" and the backward LSTM predicts "<bos> I am". This objective encourages the BLSTM layers and the embedding 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 modeling module. It can be formulated as:

$$\mathcal{L}_l = -\sum_{i=1}^{S}\sum_{j=1}^{N} \log\left(P\left(w_{j+1}|f_j\right)\right) + \log\left(P\left(w_{j-1}|b_j\right)\right), \tag{3}$$

where $f_j$ and $b_j$ represent the $j$-th outputs of the forward direction and the backward direction, respectively, given the output sum of the common BLSTM and the private BLSTM.

All the three loss functions are added to be optimized altogether as:

$$\mathcal{L} = w_s\left(\mathcal{L}_p + \lambda\mathcal{L}_a + \lambda\mathcal{L}_l\right), \tag{4}$$

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 language and the target language. Since the source language has a larger train set and we are focusing on improving the performance of the target language, $w_s$ is set to 1 when training the target language. 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 training sentences.[2] We chose datasets with 1k to 14k training sentences. The number of tag labels differs 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 languages to be the same as 1,280 for fair comparison among different languages. The remainder training examples of the target languages are still used for both language-adversarial training and bidirectional 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 significantly 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 evaluated. 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 modeling and with it, respectively.

| | | Target only | | Source (English) → Target | | | | |
|---|---|---|---|---|---|---|---|---|
| Language Family | Language | c | c,l | c,l | p,l | c,p,l | c,l+a | c,p,l+a |
| Germanic | Swedish | 93.26 | 94.31 | 94.36 | 94.39 | 94.51 | 94.38 | **94.63** |
| | Danish | 92.13 | 93.41 | 93.34 | 93.76 | 94.05 | 93.74 | **94.26** |
| | Dutch | 83.24 | 84.73 | 85.20 | 84.92 | 84.85 | 84.99 | **85.83** |
| | German | 89.27 | 90.69 | 90.06 | 90.40 | 90.01 | 90.14 | **90.71** |
| | Avg | 89.47 | 90.78 | 90.74 | 90.87 | 90.86 | 90.82 | **91.36** |
| Slavic | Slovenian | 93.06 | 93.79 | 93.83 | 94.06 | **94.20** | 93.93 | 94.06 |
| | Polish | 91.30 | 91.30 | 91.69 | **92.11** | 91.86 | 91.77 | **92.11** |
| | Slovak | 86.53 | 89.56 | 90.11 | 89.88 | 89.98 | **90.40** | 90.01 |
| | Bulgarian | 93.45 | 95.27 | 95.33 | 95.50 | 95.52 | 95.25 | **95.65** |
| | Avg | 91.09 | 92.48 | 92.74 | 92.89 | 92.89 | 92.84 | **92.95** |
| Romance | Romanian | 93.20 | 94.09 | **94.22** | 94.17 | 94.05 | 93.91 | 94.20 |
| | Portuguese | 94.23 | 95.18 | 95.42 | 95.15 | **95.55** | 95.36 | 95.51 |
| | Italian | 93.80 | **95.95** | 95.79 | 95.61 | 95.84 | 95.70 | 95.92 |
| | Spanish | 91.94 | 93.34 | 93.34 | 93.31 | 93.29 | 92.94 | **93.44** |
| | Avg | 93.29 | 94.64 | 94.69 | 94.56 | 94.68 | 94.48 | **94.77** |
| Indo-Iranian | Persian | 93.91 | 94.63 | 94.68 | 94.79 | 94.78 | 94.49 | **94.83** |
| Uralic | Hungarian | 93.20 | 93.27 | 94.40 | 94.66 | **94.69** | 94.29 | 94.45 |
| | Total Avg | 91.61 | 92.82 | 92.98 | 93.05 | 93.08 | 92.95 | **93.26** |

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.) $c$: using common BLSTM, $p$: using private BLSTMs, $l$: bidirectional language modeling objectives, $a$: language-adversarial training. (Underlined scores denote that the differences between the highest score of the other models and those scores are statistically insignificant with McNemar's $\chi^2$ test with $p$-value $< 0.05$.)

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.[3] 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 per-

[3]The results in detail are shown in the first authors dissertation Kim (2017).

| | | Target only | | Source (English) → Target | | | | |
|---|---|---|---|---|---|---|---|---|
| Language Family | Language | p | p,l | p,l | c,l | p,c,l | c,l+a | p,c,l+a |
| Germanic | Swedish | 87.43 | 90.49 | **91.02** | 90.45 | 90.48 | 90.72 | 90.70 |
| | Danish | 86.42 | 90.00 | 90.74 | 90.69 | 90.02 | 90.16 | **90.79** |
| | Dutch | 76.76 | 82.24 | **82.61** | 82.46 | 82.10 | 82.58 | 82.15 |
| | German | 86.25 | 88.95 | 89.10 | 88.69 | 88.93 | 88.08 | **89.68** |
| | Avg | 84.22 | 87.92 | **88.37** | 88.07 | 87.88 | 87.88 | 88.33 |
| Slavic | Slovenian | 87.02 | 89.97 | 90.29 | 90.00 | 90.32 | 89.58 | **90.59** |
| | Polish | 82.10 | 84.13 | 85.21 | 85.41 | 85.30 | 85.46 | **85.50** |
| | Slovak | 76.22 | 81.03 | 82.95 | **83.40** | 82.68 | 82.70 | 83.17 |
| | Bulgarian | 87.32 | **92.81** | 92.68 | 92.07 | 92.30 | 92.20 | 92.39 |
| | Avg | 83.16 | 86.98 | 87.78 | 87.72 | 87.65 | 87.48 | **87.91** |
| Romance | Romanian | 88.67 | **91.44** | **91.44** | 90.87 | 91.22 | 90.85 | 91.37 |
| | Portuguese | 90.66 | 93.73 | 93.55 | 93.90 | 93.81 | 93.58 | **94.20** |
| | Italian | 89.78 | 93.99 | 93.82 | 93.27 | 93.46 | 93.51 | **94.00** |
| | Spanish | 85.91 | **91.07** | 90.59 | 90.59 | **91.07** | 90.17 | 90.88 |
| | Avg | 88.76 | 92.56 | 92.35 | 92.16 | 92.39 | 92.03 | **92.61** |
| Indo-Iranian | Persian | 90.64 | **92.40** | 91.98 | 91.97 | 92.12 | 92.18 | 91.83 |
| Uralic | Hungarian | 89.14 | 90.65 | 91.45 | 91.48 | 90.91 | **91.52** | 90.72 |
| | Total Avg | 86.02 | 89.49 | 89.82 | 89.66 | 89.62 | 89.52 | **89.86** |

Table 2: POS tagging accuracies (%) with 320 tag-labeled training examples for each target language. All the training examples are still used for the other objectives.

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 to 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]The extended work in detail are shown in Kim (2017).

[5]learning rate=0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$.

## 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).

## References


Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. In *International Conference on Machine Learning (ICML)*.

Maria Barrett, Frank Keller, and Anders Søgaard. 2016. Cross-lingual transfer of correlations between parts of speech and gaze features. In *Proceedings of COLING*, pages 1330–1339.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 343–351.

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Weinberger, and Claire Cardie. 2016. Adversarial deep averaging networks for cross-lingual sentiment classification. In *arXiv:1606.01614*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learing Research (JMLR)*, 12:2493–2537.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised POS tagging with bilingual projections. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 634–639.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 1019–1027.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17:1–35.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.

Jiri Hana, Anna Feldman, and Chris Brew. 2004. A resource-light approach to Russian morphology: Tagging Russian using Czech resources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 222–229.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2342–2350.

Joo-Kyung Kim. 2017. *Linguistic Knowledge Transfer for Enriching Vector Representations*. Ph.D. thesis, The Ohio State University.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Young-Bum Kim, Benjamin Snyder, and Ruhi Sarikaya. 2015a. Part-of-speech taggers for low-resource languages using CCA features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1292–1302.

Young-Bum Kim, Karl Stratos, and Ruhi Sarikaya. 2016. Frustratingly easy neural domain adaptation. In *Proceedings of COLING*, pages 387–396.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 2015b. New transfer learning techniques for disparate label sets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 473–482.

Diederik P. Kingma and Jimmy Lei Ba. 2015. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 260–270.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernández Astudillo, Silvio Amir, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1520–1530.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML 2013 Workshop on Deep Learning for Audio, Speech, and Language Processing*.

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

Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Fabricio Chalub, Çar Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökrmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Linh Hà M, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phng Lê Hng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyn Th, Huyn Nguyn Th Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2016. Universal dependencies 1.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Ohio Supercomputer Center. 1987. Ohio supercomputer center. http://osc.edu/ark:/19495/f5s1ph73.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359.

Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout improves recurrent neural networks for handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 285–290.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 412–418.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics (TACL)*, 1:1–12.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1779–1785.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *International Conference on Learning Representations (ICLR)*.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag – multilingual POS tagging via coarse mapping between embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1307–1317.