Low Resource Sequence Tagging using Sentence Reconstruction

Tal Perl
Tel Aviv University
talperl@mail.tau.ac.il

Sriram Chaudhury
Wipro
sriramchaudhury@gmail.com

Raja Giryes
Tel Aviv University
raja@tauex.tau.ac.il

Abstract

This work revisits the task of training sequence tagging models with limited resources using transfer learning. We investigate several proposed approaches introduced in recent works and suggest a new loss that relies on sentence reconstruction from normalized embeddings. Specifically, our method demonstrates how by adding a decoding layer for sentence reconstruction, we can improve the performance of various baselines. We show improved results on the CoNLL02 NER and UD 1.2 POS datasets and demonstrate the power of the method for transfer learning with low-resources achieving 0.6 F1 score in Dutch using only one sample from it. The code is publicly available at: https://github.com/tperl/Low-Resource-Sequence-Tagging-using-Sentence-Reconstruction.

1 Introduction

The increased popularity of deep learning led to a giant leap in natural language processing (NLP). Tasks such as neural machine translation (Lample et al., 2018a; Gu et al., 2018), sentiment analysis (Patro et al., 2018) and question answering (Ran et al., 2019) achieved impressive results.

A major limitation of deep learning is the need for huge amounts of training data. Thus, when dealing with low resource datasets, transfer learning is a common solution. A popular approach in NLP is training a language model for getting a good context-based word representation. Language models such as Bert (Devlin et al., 2019), Roberta (Liu et al., 2019b), ELMO (Peters et al., 2018), and XL-net (Yang et al., 2019) that are trained on very large corpora, are used by the community for different NLP tasks. This "transfer-learning" across tasks within the same language relies on fine-tuning a language model for a specific task (Sun et al., 2019).

This work focuses on transfer learning between *different* languages. Some approaches have been suggested for it. Yang et al. (2017) have proposed using joint training with a large dataset as a source and a small dataset as a target. Zou et al. (2018) have shown how by aligning sentence representations using an adversarial loss, they were able to transfer knowledge between two languages.

Contribution. This work analyzes the contribution of various techniques proposed for transfer learning between languages for the task of sequence tagging. In particular, we evaluate joint training and adversarial learning. Moreover, we propose a novel regularization technique, namely, we add a reconstruction loss with ℓ_2 normalization. We show that the addition of this loss improves the performance of various sequence tagging tasks when doing transfer learning.

Our strategy shows promising results for training models without being language-specific, which saves expensive labeling time. An important characteristic of our technique is its ability to provide good tagging in "few-shot learning" (Fei-Fei et al., 2006). We achieve this result by adding to the small dataset, a larger corpus corresponding to another language. Our proposed loss improves the transfer of information and thus the tagging accuracy. We demonstrate our approach on the ConLL02/03 and the Universal Dependency (UD) 1.2 datasets.

2 Related Work

Solving sequence tagging tasks, such as named entity recognition (NER) or part of speech (POS), using statistical methods has been studied for more than two decades. Early solutions used hidden markov models (HMMs) (Bikel et al., 1997), support-vector machines (SVMs) (Isozaki and Kazawa, 2002) and conditional random fields (CRF, Lafferty et al., 2001), we focus on a more

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2692–2698
July 5 - 10, 2020.

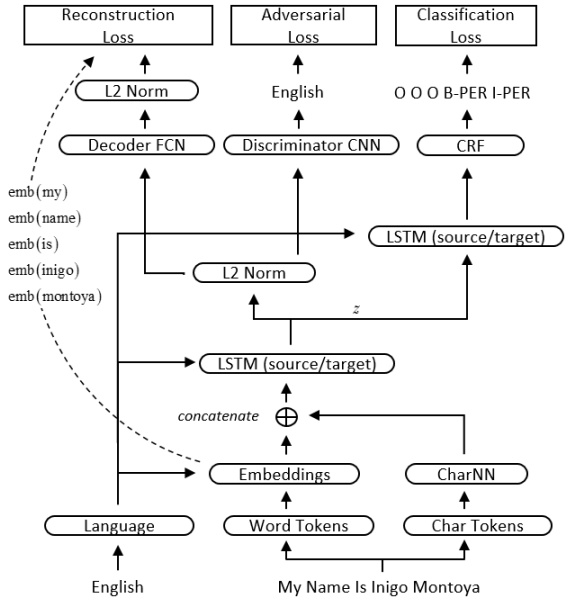

Figure 1: Proposed Method. Notice that the reconstruction loss labels are taken from the embeddings lookup table. This can be replaced by context-aware embeddings. The LSTMs are language-specific and are fed by the relevant embeddings per sample. We normalize the sentence representation for all sentences and the word representation as well.

modern approach using common deep learning-based approaches that significantly improve the performance.

Collobert et al. (2011) demonstrated the great potential of using neural networks for various NER tasks. Huang et al. (2015) proposed the Bidirectional-LSTM (Bi-LSTM) CRF and Lample et al. (2016) presented a promising architecture for NER by adding character embeddings to its input. Peng and Dredze (2016) used recurrent neural networks (RNN) for NER and word segmentation in Chinese. In the context of transfer learning for sequence tagging, Yang et al. (2017) showed that by using hierarchical RNNs and joint training, it is possible to transfer knowledge between domains of different corpora and different languages.

Cao et al. (2018) exhibited that using self-attention and an adversarial loss, they were able to perform transfer learning between two different domains in Chinese. Yadav et al. (2018) showed that Deep Affix Features is beneficial to NER. Jiang et al. (2019) used DARTS neural architecture search (Liu et al., 2019a) to improve NER. Lin et al. (2018) showed that by using multi-lingual multi-task architecture they were able to get interesting results. Devlin et al. (2019) introduced a new representation scheme for NLP tasks achieving impressive NER results. Clark et al. (2018) proposed a new method for getting improved representations of Bi-LSTM of sentence encoders using labeled and unlabeled data.

Barone and Valerio (2016) showed that using an adversarial loss (Goodfellow et al., 2014) may lead to a better word representation. In addition, Adel et al. (2018) used an adversarial loss for getting better sentence representation. Tzeng et al. (2017) demonstrated how by aligning deep representations using an adversarial loss, they transfer knowledge from one domain to another. Lample et al. (2018a) exhibited this approach for unsupervised machine translation. Inspired by these strategies, we propose a method for transfer learning between different languages for sequence tagging. Specifically, we focus on sentence representation alignment.

3 Our Approach

This section describes our sentence reconstruction approach for improving low resource sequence tagging tasks. Many successful sequence tagging network models are composed of an encoder-decoder structure. We suggest adding to them a new decoder branch comprised of a fully convolutional network (FCN) and an ℓ_2 loss term for reconstructing the word embeddings of the input sentence.

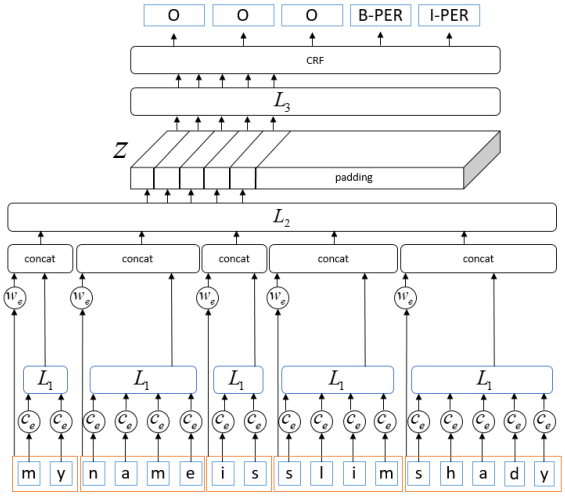

Figure 2: Baseline similar to Lample et al. (2016).

To analyze the effectiveness of our proposed technique, we evaluate its contribution compared to other recently proposed strategies for transfer learning across languages: weight sharing and adversarial alignment. For completeness, we briefly

	Baseline	L2	TL	(TL)+(L2)	(TL) + Adversarial	(TL) + (L2)+ Adversarial	(Yang et al., 2017)
English	89.1	89.3	89.6	89.9	89.5	90.1	**91.26**
Spanish	85.84	86	86.1	86.2	84.8	**86.3**	85.77
Dutch	86.67	87.18	87.1	87.62	85.7	**87.64**	85.19
English (0.1)	83.1	82.7	85.5	86.1	85.8	**86.5**	86.5
Spanish (0.1)	76.4	76.47	78.7	**78.5**	77.8	77.8	76.5
Dutch (0.1)	74.8	75.8	79	**80**	77.9	79.5	-
English (0.01)	44.75	44.8	73.8	74.17	73.8	**74.3**	72.6
Spanish (0.01)	33.3	43.6	63.3	64.98	65.8	**67.87**	60.4
Dutch (0.01)	40.7	42.9	62.5	64.75	68.56	**68.93**	-

Table 1: Ablation results on NER ConLL02/03 compared to (Yang et al., 2017), using sentence reconstruction (L2), using weight sharing based transfer learning (TL), using the adversarial loss and combining them all together.

describe the baseline we are using and each of these methods. Then, we present our new auxiliary loss.

3.1 Baseline

Our base model follows Lample et al. (2016). Specifically, we run an LSTM (Hochreiter and Schmidhuber, 1997) on the character tokens, concatenate the output to the word embeddings and run an additional LSTM. We then feed its output, denoted z, to another LSTM with a CRF at its end, which produces the sequence tagging, whether it is POS or NER. See Fig. 4 for our baseline.

3.2 Weight sharing

Yang et al. (2017) have shown that sharing weights between architectures that correspond to different languages leads to transferring knowledge between them. Our joint training model is inspired by their "Cross Lingual Transfer" with the difference that we use a single CRF that is applied to the output of both LSTMs. See Fig. 3 for a schematic of the our modified version.

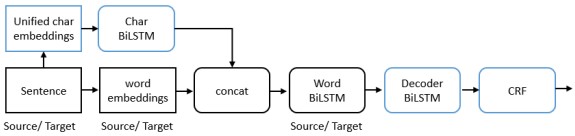

Figure 3: Our modified version of Yang et al. (2017)'s weight sharing. In blue are modules shared between source and language sentences.

3.3 Adversarial loss

The baseline described above essentially learns a sentence hidden representation, z. For aligning representations from different languages, we feed this feature vector to a 1D CNN which encodes it and outputs a softmax class and acts as a discriminator. We add a switch layer in the input

	ES	NL	EN
(Gillick et al., 2015)	82.95	82.84	86.50
(Luo et al., 2015)	-	-	91.20
(Lample et al., 2016)	85.75	81.74	90.94
(Yang et al., 2017)	85.77	85.19	91.26
(Lin et al., 2018)	85.88	86.55	-
(Yadav et al., 2018)	**87.26**	87.54	90.86
(Baevski et al., 2019)	-	-	93.5
(Jiang et al., 2019)	-	-	**93.47**
(Straková et al., 2019)	-	-	93.38
Our baseline	85.84	86.67	89
Our transfer	86.3	**87.64**	90.1

Table 2: Method results F1 score on CoNLL 2002/2003 compared to state of the art.

that arbitrates between feeding sentences from the source and target language (each uses its respective word embedding). We train the discriminator on the normalized hidden representations generated by each sentence $Z = z/||z||_2$. Thus, given the possible labels l_i, l_j of the predicted language, for an input with label l_i/l_j, the discriminator will try to predict l_i/l_j. The generator will try to fool the discriminator and cause it to predict the opposite (l_j/l_i). The adversarial loss $\mathcal{L}_{adv}$ is the sum of the discriminator loss $\mathcal{L}_D$ and the generator loss $\mathcal{L}_G$ as follows (Lample et al., 2018a):

$$\mathcal{L}_D(\theta_D, Z|\theta_D) = -E_{(s_i,l_i)}[\log p_D(l_i|e(s_i,l_i)],$$
$$\mathcal{L}_G(\theta_{enc}, Z|\theta_D) = -E_{(s_i,l_i)}[\log p_D(l_j|e(s_i,l_i)],$$
$$\mathcal{L}_{adv} = \mathcal{L}_G + \mathcal{L}_D, \quad (1)$$

where s_i is the input sentence, $e(\cdot)$ the encoder function, and θ_D and θ_{enc} are the discriminator's and the encoder's parameters, respectively.

3.4 Reconstruction loss

An adversarial training scheme can still reach trivial representations, meaning the generator produces sentence representations that do not contain meaningful information of the original sentences. There-

	ES	NL	RO
(Heinzerling and Strube, 2019)	**96.5**	93.8	89.7
(Plank et al., 2016)	95.74	93.3	-
(Yasunaga et al., 2018)	96.44	93.09	91.46
Ours baseline	96	93.1	91.45
Ours transfer	96.4	**93.8**	**93.04**

Table 3: Method results accuracy on UD 1.2 Part of speech (POS) compared to the state-of-the-art.

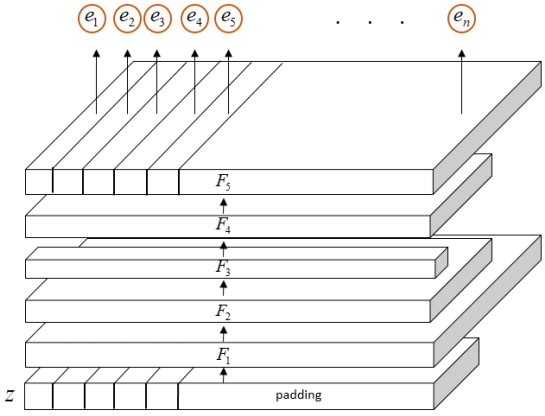

Figure 4: Our proposed fully convolutional network for learning the input sentence embeddings

fore, we propose using the ℓ_2 loss for reconstructing the input sentence (word embeddings). We do so by applying on the hidden representation z a 1D FCN with 5 layers, convolution kernels of size 3 and the ReLU non-linearity. Notice that z is a sequence of embedding vectors. Thus, the output of the FCN is also a sequence of vectors, where each of them tries to estimate the embedding of the corresponding word in the input sentence. If the generated sentence is of a different length than the input, we use the padding embedding vector to make them even. We train this decoder together with the encoder in the network using the following reconstruction loss

$$\mathcal{L}_{auto}(\theta_{enc}, \theta_{dec}) = \sum_i \|\tilde{e}_i - e_i\|_2^2, \qquad (2)$$

where θ_{dec} are the FCN parameters, e_i is the embedding of the ith word in the input sentence and $\tilde{e}_i$ is the corresponding reconstructed embedding, which we normalize. The reconstruction loss acts as a regularization term, which improves results also when used by itself (see the ablation study).

We would like to emphasize the importance of normalizing the representing vectors. Its motivation is in the fact that transforming the vectors onto a unit sphere causes the model to learn to maximize the similarity between sentences and words.

	Baseline	**Our method**
Arabic	66.05 $\pm$ 1.29	**76.82** $\pm$ **0.24**
Bulgarian	52.41 $\pm$ 1.46	**84.86** $\pm$ **0.30**
Estonian	47.22 $\pm$ 0.48	**56.10** $\pm$ **0.16**
Finnish	49.00 $\pm$ 1.45	**79.91** $\pm$ **0.39**
French	63.34 $\pm$ 3.10	**87.19** $\pm$ **0.37**
German	77.10 $\pm$ 1.36	**87.66** $\pm$ **0.30**
Greek	60.43 $\pm$ 0.80	**87.66** $\pm$ **0.30**
Hebrew	65.13 $\pm$ 2.11	**85.50** $\pm$ **0.75**
Italian	63.46 $\pm$ 1.31	**88.88** $\pm$ **0.71**
Norwegian	78.55 $\pm$ 0.62	**91.06** $\pm$ **0.31**
Polish	52.05 $\pm$ 0.61	**80.84** $\pm$ **0.47**
Slovenian	53.50 $\pm$ 0.37	**83.93** $\pm$ **0.77**
Spanish	83.65 $\pm$ 0.16	**90.60** $\pm$ **0.04**

Table 4: Low resource testing for part of speech on UD 1.2 dataset. For each language we ran 3 random seeds and report the mean and std for the baseline and the proposed method.

Figure 1 presents a model with all the discussed regularization techniques. Notice that each component in this model can be applied separately. For example, we may apply our new reconstruction loss alone, or as an additional branch to the adversarial branch with or without weight sharing.

4 Experiments

We follow the experiments of Yang et al. (2017) to evaluate our approach for transfer learning between languages. We compare our proposed regularization to joint training and the adversarial loss. We start by evaluating the impact of each strategy alone, and then gradually combine the losses to each other. Our source-target pairs are built of English and a selected target language (Spanish, Dutch or Romanian). In NER, we test both directions of transfer learning, i.e English to Spanish and Spanish to English. In POS, English is always the source language. We focus on using word embeddings that are aligned across different languages, specifically "MUSE" (Lample et al., 2018b). Our motivation for choosing it is to leverage the word alignment, which makes the impact of the sentence alignment clearer.

Loss analysis. For understanding the impact of our approach, we test it with and without the other techniques for transfer learning between languages. We also compare to each of them being applied separately. Table 1 summarizes our results. Notice that our proposed loss improves the performance when combined with other methods and even when being applied alone. Also, we have found that the improvement gained by the adversarial loss is

	ES	NL	EN
(Yang et al., 2017)	16	-	40.1
Lin et al. (2018)	**60**	50	-
Our baseline	22	33	7.6
Our transfer	59.5	**61**	**43.1**

Table 5: F1 scores on CoNLL 2002/2003 for few shot training (0.001 of the data) compared to (Yang et al., 2017).

Language	Baseline	Method	Lin et al. (2018)
English	7.6	**34.6**	-
Spanish	7.6	**53**	50
Dutch	7.6	**60**	50

Table 6: F1 scores on CoNLL 2002/2003 for one shot training, compared to Lin et al. (2018).

marginal and therefore, we do not use it in the final model used in the next experiments, which consist of only weight sharing and our proposed ℓ_2 reconstruction loss.

Results. We evaluate our model on three tasks: (i) NER transfer learning compared to leading methods; (ii) NER transfer learning on a subset of the target data; and (iii) POS transfer. We achieve competitive results on Conll2002 Dutch/Spanish. For testing how competitive our approach is, we also compare to state-of-the-art methods. Moreover, we perform experiments on subsets of the data similar to Yang et al. (2017). These experiments exhibit the advantage of our model, especially when training on scarce data. For example, we show that using only nine samples in Spanish (0.001 of the data) we get an F1 score of 0.59 (compared to the 0.16 transfer learning result of Yang et al. (2017)).

Table 2 shows the NER results, where we get competitve results in ConLL02 and improve our baseline in English ConLL03. Table 4 shows how our method generalizes well for low resource transfer learning in POS. Notice the great improvement between our baseline as shown in Fig. 4 and our method shown in Fig. 1. Table 3 demonstrates the performance on POS, where we get the largest improvement on Romanian, which is a low resource language (with fewer labels). Table 5 exhibits the advantage of our regularization for few-shot learning compared to Yang et al. (2017) and Lin et al. (2018). Finally, Table 6 and Table 7 presents the results of our approach for "one-shot" learning compared to Lin et al. (2018) and "zero-shot" learning. A major improvement compared to our baseline is apparent also here. We found for the case of few-shot and one-shot learning that it is better to share the base BiLSTM because it does not see enough examples to train.

Language	Baseline	Method
Spanish	0	**57**
Dutch	0	**55**

Table 7: F1 scores on CoNLL 2002 for zero shot training.

5 Conclusion

This work demonstrates the power of sentence reconstruction for transferring knowledge from a rich dataset to a sparse one. It achieves competitive results with a relatively simple baseline. We also show its strength in few-shot and one-shot learning.

We believe that using the proposed sentence ℓ_2 reconstruction may contribute as an auxiliary loss for other tasks. Also, we have demonstrated our model with MUSE, since it provides word alignment across languages. Yet, our approach can be applied also with other more recent language models that have stronger context-based embeddings.

Acknowledgment. This work was supported by Wipro. We thank Parul Chopra and Amrit Bhaskar for their assitance.

References

Heike Adel, Anton Bryl, David Weiss, and Aliaksei Severyn. 2018. Adversarial neural networks for cross-lingual sequence tagging. *CoRR*, abs/1808.04736.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven pretraining of self-attention networks. *CoRR*, abs/1903.07785.

Miceli Barone and Antonio Valerio. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany. Association for Computational Linguistics.

Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1997. Nymble: a high-performance learning name-finder. In *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for Chinese named entity recognition with self-

attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 182–192, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-supervised sequence modeling with cross-view training. *CoRR*, abs/1809.08370.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2015. Multilingual language processing from bytes. *CoRR*, abs/1512.00103.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. 2018. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 774–784, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2019. Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy. Association for Computational Linguistics.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yufan Jiang, Chi Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2019. Improved differentiable architecture search for language modeling and named entity recognition. 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 3583–3588, Hong Kong, China. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809, Melbourne, Australia. Association for Computational Linguistics.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2019a. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal. Association for Computational Linguistics.

Badri N. Patro, Vinod K. Kurmi, Sandeep Kumar, and Vinay P. Namboodiri. 2018. Learning semantic sentence embeddings using pair-wise discriminator. *CoRR*, abs/1806.00807.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *ACL*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

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 (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. Option comparison network for multiple-choice reading comprehension. *CoRR*, abs/1903.03033.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971.

Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. Deep affix features improve neural named entity recognizers. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 167–172, New Orleans, Louisiana. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *ArXiv*, abs/1906.08237.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *ArXiv*, abs/1703.06345.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 976–986, New Orleans, Louisiana. Association for Computational Linguistics.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics.