What Matters for Neural Cross-Lingual Named Entity Recognition: An Empirical Analysis

Xiaolei Huang[1*], **Jonathan May**[2], **Nanyun Peng**[2]
[1]Dept. of Information Science, University of Colorado Boulder
[2]University of Southern California Information Sciences Institute
xiaolei.huang@colorado.edu, {jonmay, npeng}@isi.edu *

Abstract

Building named entity recognition (NER) models for languages that do not have much training data is a challenging task. While recent work has shown promising results on cross-lingual transfer from high-resource languages to low-resource languages, it is unclear what knowledge is transferred. In this paper, we first propose a simple and efficient neural architecture for cross-lingual NER. Experiments show that our model achieves competitive performance with the state-of-the-art. We further analyze how transfer learning works for cross-lingual NER on two transferable factors: sequential order and multilingual embeddings, and investigate how model performance varies across entity lengths. Finally, we conduct a case-study on a non-Latin language, Bengali, which suggests that leveraging knowledge from Wikipedia will be a promising direction to further improve the model performances. Our results can shed light on future research for improving cross-lingual NER.

1 Introduction

Named Entity Recognition (NER) is an important NLP task that identifies the boundary and type of named entities (e.g., person, organization, location) in texts. However, for some languages, it is hard to obtain enough labeled data to build a fully supervised learning model. Cross-lingual transfer models, which train on high-resource languages and transfer to target languages are a promising direction for languages with little annotated data (Bharadwaj et al., 2016; Tsai et al., 2016; Pan et al., 2017; Yang et al., 2017; Mayhew et al., 2017; Wang et al., 2017; Cotterell and Duh, 2017; Feng et al., 2018; Xie et al., 2018; Zhou et al., 2019).

Cross-lingual NER models learn transferable knowledge mainly from four sources: translation (Mayhew et al., 2017; Feng et al., 2018), bilingual embeddings (Ni et al., 2017; Xie et al., 2018), phonetic alphabets (Bharadwaj et al., 2016) and multi-source learning (Mayhew et al., 2017; Lin et al., 2018; Zhou et al., 2019; Chen et al., 2019; Rahimi et al., 2019). Translation can be applied at either the sentence level via machine translation or at the word and phrase level via application of bilingual dictionaries. Bilingual embeddings are a form of bilingual dictionaries; they constitute a projection from one pre-trained language representation into the same vector space as the other such that words with the same meaning have similar representations (Conneau et al., 2018). Phonetic alphabets enable different languages to share the same pronunciation characters so that the character-level knowledge is transferable between languages that otherwise have different character sets, such as English and Bengali (Hermjakob et al., 2018). Multi-source learning is effective when multiple or similar language resources are available by learning shareable knowledge. For example, training a model on both English and Hindi can significantly improve the model performance on Bengali than only using English (Mayhew et al., 2017). However, there is little prior work with detailed analysis of how cross-lingual NER models transfer knowledge between languages on different levels.

In this paper, we focus on a single-source zero-shot transfer setting where we transfer from English to target languages that have no annotated data. In our settings, the resources are limited to annotated source language data, bilingual dictionaries, and unlabeled corpora from both source and target languages. We first propose a neural cross-lingual NER model that combines the ideas of translation, bilingual embedding, and phonetic

---

alphabets. Next, we conduct qualitative analyses to answer the following questions on how the model transfers knowledge under the cross-lingual settings: 1) does the source language syntax matter? 2) how do word and character embeddings affect the model transfer? We analyze how F1 scores differ across different entity lengths. Finally, we conduct a case study on Bengali.

2 Model

NER models take a sequence of tokens as input and predict a sequence of labels such as person (PER) or location (LOC). In this paper, we adopt the neural architecture NER model from Lample et al. (2016); Ma and Hovy (2016). The model first combines pre-trained word and character embeddings as token representations, then feeds the representations to one layer of a Bidirectional Long Term Short Memory (Bi-LSTM) (Hochreiter and Schmidhuber, 1997), and finally predicts the outputs via a Conditional Random Field (CRF). We show the model architecture on left of Figure 1. However, languages express the same named entity in different words and characters. To bridge the barriers, we combine three strategies: bilingual embedding, reverse translation, and transliteration.

Bilingual Embedding. Word embeddings are usually trained for each language separately and therefore in different vector spaces (Ruder et al., 2019). To map word embeddings into the same shared space, we use MUSE (Conneau et al., 2018) to build our bilingual embeddings. The bilingual embeddings enable the words with the same meaning to have similar word representations in the shared space.

Reverse Translation. We use the bilingual dictionaries provided by Rolston and Kirchhoff (2016). Language variations exist among source and target languages and training on monolingual corpus might limit learning the variations. Therefore, we translate the source language, English, to the target language for the training corpus to reconstruct source language sentences and learn source and target languages jointly. However, one English word might have multiple corresponding translations. To select the best translation, we define an empirical score function $F(w, w_{t,i})$ that calculates the cosine similarity between the translated target word ($w_{t,i}$) and the English source

word (w) based on its contextual words ($w_{c,j}$).

$$
\begin{aligned}
F(w, w_{t,i}) = {} & \alpha \cdot \cos(E(w), E(w_{t,i})) \\
& + (1 - \alpha) \cdot \sum_{j=1}^{m} \frac{\cos(E(w_{t,i}), E(w_{c,j}))}{(d_j + 1)^2} \quad (1)
\end{aligned}
$$

where $E(w)$ is the bilingual embedding vector of the word, d_j is the sequential distance between the word and its contextual word j, and α is the trade-off factor that balances impacts of translation pair and contextual words. In this study, we set α to 0.5. We choose the translation pair with the highest similarity score. Note that the reverse only applies during the training step that translates English into the target language.

Target word	Source word	Uroman	Type
grünen	green	gruenen	ORG
ইউরোপের	Europe	iuropera	LOC

Table 1: Examples of Uroman that maps different languages into the same character space. We show two target languages, German (top) and Bengali (bottom). The second column is a translation of the target language. The transliterations show the phonetic similarity between the source and target languages.

Character - Uromanization. Different languages may not share the same characters, but some named entities in the multilingual corpora share phonetic similarities (Huang et al., 2004). To map multilingual corpora into the same character space and connect the phonetic similarity between entities, we employ Uroman[1] (Hermjakob et al., 2018), which transliterates any language into something roughly pronounceable English spellings, while keeps the English words unchanged. We show some examples of Uroman in Table 1.

3 Experiments

In this study, we first evaluate our proposed cross-lingual method on the CoNLL 2002 and 2003 NER datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). We then conduct ablation studies to examine how the model learns transferable knowledge under the cross-lingual settings. Finally, we conduct a case study on Bengali, a low resource language.

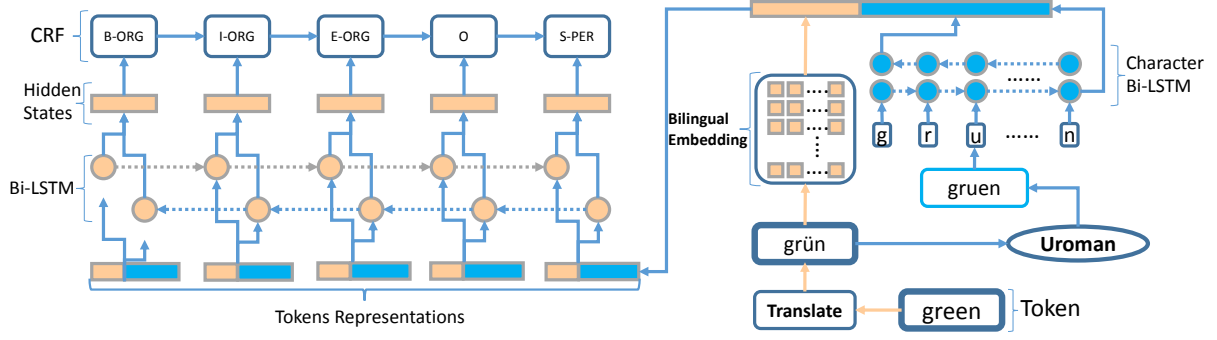[1]https://www.isi.edu/~ulf/uroman.html

Figure 1: Architecture of our proposed model. Each word is translated from the source. Each token representation contains two concatenated parts: a bilingual word embedding (light orange) and character representations (blue). The character representation for each word is generated via the Bi-LSTM. Words are first transliterated and split into individual characters before being fed into the character Bi-LSTM.

3.1 Data

The CoNLL datasets contain four different European languages, Spanish, Dutch, English, and German. The data contains four types of named entities: person (PER), organization (ORG), location (LOC) and MISC. We use Bengali language data from LDC2015E13 (V2.1). To be consistent, we only keep the PER, ORG and LOC tags and ignore MISC tags from the Bengali corpus.

In this study, we choose the BIOSE tag schema instead of standard BIO, where the B, I, O, S, E refer to the beginning, inside, outside, single and end of an entity, respectively. Previous work shows that BIOSE can learn a more expressive model than the BIO schema does (Ratinov and Roth, 2009). We lowercase the data and replace numbers and URLs as "num" and "url" respectively. We use the training data to train our model, the development sets to select the best well-trained model, and the test sets to evaluate performance, where the training data is English and the development and test sets are from the target language.

3.2 Experimental Settings

Bilingual Embedding. We use the 300-dimensional bilingual word embeddings pre-trained by MUSE (Conneau et al., 2018) and 300-dimensional randomly initialized character embeddings. Specifically, we first collect monolingual pre-trained word embeddings from fast-Text (Bojanowski et al., 2017). We then align the embeddings using the MUSE supervised model. We normalize the inputs to MUSE and keep the other hyperparameters as defaults. We merge the aligned word embeddings and remove duplicates. Updating the embedding during training time will change the vector space of partial bilingual embeddings and therefore break the vector alignment. Thus, we freeze the embedding weights during the training step. For OOV (out-of-vocabulary) words, we use a randomly initialized 300 dimension vector within $[-0.1, 0.1]$.

Reverse Translation. We follow the general translation step in Section 2. Specifically, we replace English words with target language words if the translation pairs exist in the bilingual dictionaries (Rolston and Kirchhoff, 2016). If a translation pair does not exist for a word, we keep the word unchanged in the training data. We use the pre-trained bilingual embedding to help select the best choice of polysemy.

Model parameters. We use 300-dimension hidden states for both character and token level Bi-LSTMs. To prevent overfitting, we apply dropout (Srivastava et al., 2014) with a rate of 0.5 on outputs of the two Bi-LSTMs. We follow the Conditional Random Field (CRF) setup of Peng and Dredze (2015). We then randomly initialize the weights of the layers within $[-0.1, 0.1]$. We train the model for 200 epochs and optimize the parameters by Stochastic Gradient Descent (SGD) with momentum, gradient clipping, and learning rate decay. We set the learning rate (lr) and the decay rate (dr) as 0.01 and 0.05 respectively. We update the learning rate by $\frac{lr}{(1+(n+1)*dr)}$ after epoch n. We clip the gradients to the range of [-5.0, 5.0]. We measure performance by F1 score.

3.3 Baselines

In this study, we compare our proposed method with three close works under the cross-lingual settings. We compare our method with the best-reported performance from their works. We briefly

summarize the the three baselines in this section.

- WikiNER (Tsai et al., 2016) first links named entities in the multilingual Wikipedia corpora and extracts page categories as "wikifier" features, and use these features to achieve cross-lingual transfer.

- CTNER (Mayhew et al., 2017) first translates the corpora into English via bilingual dictionaries and multilingual Wikipedia entries on both word and phrase levels, and directly perform NER on the translated target language.

- NCNER (Xie et al., 2018) proposes a neural NER model with attention mechanism. The work is closest to ours. However, our model takes different approaches in obtaining multi-lingual embedding, translation, and transliteration as Xie et al. (2018), while the neural models share the similar architecture of Bi-LSTMs-CRF, and use multi-lingual embeddings to achieve transfer.

3.4 Results

Models	Spanish	Dutch	German
WikiNER*	60.55	61.60	48.10
CTNER*	51.82	53.94	50.96
CTNER*+	65.95	66.50	59.11
NCNER	**72.37**	71.25	57.76
Our Model	64.48	**73.44**	**62.26**

Table 2: F1 score comparisons of cross-lingual models on Spanish, Dutch and German. The "*" indicates the model uses the Wikipedia resources. The "+" means training model by multiple language resources.

As seen in Table 2, our proposed method outperforms the previous works in nearly all cross-lingual tasks. Moreover, the two neural model methods exceed the performance of the other models. These results suggest the effectiveness of neural cross-lingual transfer through multi-lingual embedding space, and show competitive results of our proposed methods.

3.5 Ablation Studies

We conduct ablation analyses to understand how the model transfers learned knowledge from one language to the other. We focus on two aspects: syntax and embeddings.

Syntax analysis. Different languages might not share the same syntax structures. The neural model learns sequential information from the

Models	Spanish	Dutch	German
Full Model	64.48	73.44	62.26
Shuffle	49.44	40.61	27.25
Word-only	53.00	59.60	52.58
Char-only	18.88	12.15	16.46

Table 3: F1 scores of different ablation analyses, compared to our full model.. "Shuffle" means the training data is shuffled. "Word-only" means our proposed model is only fed bilingual word embeddings, and "Char-only" means the model only receives character embeddings as input.

source language and applies the learned knowledge on the target language. The importance of this sequential information is unknown. We shuffle the sentences of training data while keeping the internal order of named entities unchanged. We show the results in Table 3. The results show that after shuffling, the performances of our model decrease and the decreases vary among different languages. This suggests the importance of sequential information for those languages under the cross-lingual settings.

Embeddings analysis. We train our proposed models with both character and bilingual word embeddings, however, how the model values different embeddings is unknown. We feed the model with either bilingual word or character embeddings. Table 3 shows that bilingual word embeddings have better performance than character embeddings across the three languages. The results suggest that the aligned bilingual embeddings are more important than the character embeddings for the three languages.

3.6 Entity length analysis

While we can observe how models perform overall on the named entities, we can not know how models differ from short to long entities. Particularly, probing if models hold strong biases towards shorter entities can help interpret the process of cross-lingual transfer learning. To make the comparison, we first categorize entities into three levels: single token, two tokens and greater or equal to three tokens. We then calculate the F1-score of the three entity lengths in the test set and among correctly predicted entities. Finally we summarize the F1-score in Table 4.

With comparing to the overall performance across the three languages in Table 2, we can observe that the single token shows relatively closer

Language	1	2	$\geq$3
Spanish	68.93	68.21	47.57
Dutch	69.49	82.46	60.85
German	59.50	70.76	37.22

Table 4: F1 scores of different lengths of entities across the three languages: Spanish, Dutch and German. The number 1 refers to the entities with single token, the 2 means the entities with two tokens and the $\geq$ 3 indicates the entities with not less than three tokens.

to the performance, the entities with two tokens achieve the higher scores, while the entities with more than 2 tokens decrease significantly ranging from 12.59 to 25.04 absolute percentages of F1 scores. The observation indicates that entities longer than two tokens are more difficult to infer. This might encourage us to balance the weight of long entities in our current evaluation method which ignores entity length when datasets have high volumes of long entities.

4 Case Study: Bengali

Models	F1 score
CTNER	30.47
CTNER+	31.70
CTNER*	46.28
CTNER*+	45.70
Our Model	34.29

Table 5: F1 score comparisons of translation-based models on Bengali. The "*" indicates the model uses Wikipedia resources. The "+" means a model is trained with multiple language resources.

The previous cross-lingual settings were only for European languages, which share similar alphabets. However, many languages use non-Latin orthography. In this work, we present a case study on Bengali, which does not use a Latin alphabet. We compare our proposed method with the translation-based method, CTNER (Mayhew et al., 2017). The results in Table 5 show that our model outperforms the previous methods without Wikipedia.

The results suggest multilingual Wikipedia is critical for future performance improvements beyond simple transfer. This is to be expected; a domain discrepancy exists between the source and target language data and therefore many named entities are missing or mismatched: partial data sources of the Bengali come from social media

and online forums (Cieri et al., 2016). By contrast, the only data source of CoNLL data is from news articles (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). While transfer can help provide universal context clue information across languages there is no substitute for a resource of actual names.

Language	Type	Token
Spanish	2.5	0.8
Dutch	2.5	0.9
German	2.3	1.5
Bengali	18.2	12.2

Table 6: OOV rate (percentage) in our bilingual word embeddings across the four languages. Type indicates unique words, and token refers to counting token numbers.

The colloquium words from social media may cause the issue of out of vocabulary (*OOV*) and further impact the transferring process. We summarize the percentage of missing words in our pre-trained bilingual word embeddings in Table 6. The OOV rate of Bengali is significantly higher than that of the other three languages. This suggests that the high OOV rate and the discrepancy of data sources may hurt the effectiveness of transfer.

5 Conclusion

We have presented a simple but efficient method to adapt neural NER for the cross-lingual settings. The proposed method benefits from multiple transferable knowledge and shows competitive performances with the state of the art using limited resources. We examine multiple factors that impact the transfer process and conduct an ablation study to measure their influences. Our experiment on Bengali shows that leveraging knowledge from Wikipedia will be a promising direction for future research.

6 Acknowledgments

References

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 91–96, Taipei, China. Asian Federation of Natural Language Processing.

Xiaocheng Feng, Xiachong Feng, Bing Qin, Zhangyin Feng, and Ting Liu. 2018. Improving low resource named entity recognition using cross-lingual knowledge transfer. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4071–4077, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal romanization tool uroman. *Proceedings of ACL 2018, System Demonstrations*, pages 13–18.

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

Fei Huang, Stephan Vogel, and Alex Waibel. 2004. Improving named entity translation combining phonetic and semantic similarities. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics, Association for Computational Linguistics.

Leanne Rolston and Katrin Kirchhoff. 2016. Collection of bilingual data for lexicon transfer learning. Technical report, Technical Report UW-EE-2016-0001.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, Edmonton, Canada. Association for Computational Linguistics.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228, Berlin, Germany. Association for Computational Linguistics.

Dingquan Wang, Nanyun Peng, and Kevin Duh. 2017. A multi-task learning approach to adapting bilingual word embeddings for cross-lingual named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 383–388, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *International Conference on Learning Representations*, Palais des Congrès Neptune, Toulon, France.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3461–3471, Florence, Italy. Association for Computational Linguistics.