Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition

Zenan Zhai, Dat Quoc Nguyen and **Karin Verspoor**
School of Computing and Information Systems
The University of Melbourne, Australia
zenanz@student.unimelb.edu.au, {dqnguyen, karin.verspoor}@unimelb.edu.au

Abstract

We compare the use of LSTM-based and CNN-based character-level word embeddings in BiLSTM-CRF models to approach chemical and disease named entity recognition (NER) tasks. Empirical results over the BioCreative V CDR corpus show that the use of either type of character-level word embeddings in conjunction with the BiLSTM-CRF models leads to comparable state-of-the-art performance. However, the models using CNN-based character-level word embeddings have a computational performance advantage, increasing training time over word-based models by 25% while the LSTM-based character-level word embeddings more than double the required training time.

1 Introduction

Bi-directional Long-Short Term Memory Conditional Random Field models (BiLSTM-CRF), in which a BiLSTM is coupled with a CRF layer to connect output tags, have been shown to achieve state-of-art performance in sequence tagging tasks including part of speech (POS) tagging, chunking, and NER (Huang et al., 2015). The combination of word embeddings and character-level word embeddings has been explored in this context, with Ma and Hovy (2016) using Convolutional Neural Networks (CNNs) to construct character-level word embeddings and Lample et al. (2016) applying LSTM networks. This work showed that the use of character-level word embeddings improves the performance of the models, by contributing the ability to recognize unseen words.

Biomedical Named Entity Recognition (BNER) is a vital initial step for information extraction tasks in the biomedical domain, including the Chemical-Disease Relationship (CDR) extraction task where both chemical and disease entities must be identified (Li et al., 2016). Character-level word embeddings could be particularly significant in this context, given that new entity names are frequently created, and may follow consistent patterns including productive morphology such as common prefixes (e.g., *di-*) or suffixes (e.g., *-ase*). Features that capture word-internal characteristics have been shown to be effective for BNER tasks in CRF models (Klinger et al., 2008).

Lyu et al. (2017) applied a BiLSTM-CRF model with LSTM-based character-level word embeddings to a gene and protein NER task, demonstrating state-of-art performance that outperformed traditional feature-based models. Luo et al. (2018) further improved on this result on a chemical NER task by adding an attention layer between the BiLSTM and CRF layers (Att-BiLSTM-CRF).

In an experiment by Reimers and Gurevych (2017b), optimal hyper-parameters for LSTM networks in sequence tagging tasks were explored, with the finding that incorporation of character-level word embeddings significantly improved performance on NER tasks on general datasets including CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003). However, the choice of CNN-based (Ma and Hovy, 2016) or LSTM-based character-level word embeddings (Lample et al., 2016) did not affect the performance significantly. Since the CNN has fewer parameters to train than BiLSTM network, it is better in terms of training efficiency, and was recommended as the preferred approach.

In this paper, we implement and compare models with each type of word embedding to generate empirical results for the tasks of chemical and disease NER, using the BioCreative V CDR corpus (Li et al., 2016). These BNER categories are the most searched entities in the biomedical literature (Islamaj Dogan et al., 2009), and hence particularly important to study.

The results show that models with CNN-based character-level word embeddings achieve state-of-the-art results comparable to LSTM-based character-level word embeddings, while having

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 38–43
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics

the advantage of reduced training complexity, demonstrating that the prior results also hold for the BNER task.

2 Experimental methodology

This section presents our empirical approach to comparing state-of-the-art neural network models for chemical and disease NER.

2.1 Dataset

In our experiments, we use the BioCreative V CDR corpus (Li et al., 2016). This corpus provides a set of 1000 manually-annotated abstracts (9193 sentences) for training and development, and another set of 500 manually-annotated abstracts (4840 sentences) for test. In particular, we used a pre-processed version of the CDR corpus from Luo et al. (2018),[1] which provides predicted POS-, chunking- and gazetteer-based tags:

- POS and chunking tags are predicted by the GENIA tagger (Tsuruoka et al., 2005).[2]
- Gazetteer tags are encoded in BIO tagging scheme based on matching to the external Jochem chemical dictionary (Hettne et al., 2009).

Following Luo et al. (2018), we randomly sample 10% from the set of 1000 abstracts for development, and use the remaining for training.

2.2 Models

We use the following BiLSTM-CRF-based sequence labeling models:

- Baseline BiLSTM model (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997) which uses a softmax layer to predict NER labels of input words.
- BiLSTM-CRF (Huang et al., 2015) extends the BiLSTM model with a CRF layer which allows the model to use sentence-level tag information for sequence prediction.
- BiLSTM-CRF + CNN-char (Ma and Hovy, 2016) extends the BiLSTM-CRF model with character-level word embeddings. For each word, its character-level word embedding is derived by applying a CNN to the character sequence in the word.

Hyper-para.	**Value**
Optimizer	Nadam
Mini-batch size	32
Clipping	$\tau = 1$
Dropout	[0.25, 0.25]

Table 1: Fixed hyper-parameter configurations.

CNN-based

Hyper-para.	**Value**
charEmbedSize	30
Window size	3
# of filters	30
# of Params.	2,730

LSTM-based

Hyper-para.	**Value**
charEmbedSize	30
BiLSTM layer	1
LSTM size	25
# of Params.	11,200

Table 2: Hyper-parameters for learning character-level word embedding. "charEmbedSize" and "# of Params." denote the vector size of character embeddings and the total number of parameters, respectively.

- BiLSTM-CRF + LSTM-char also extends the BiLSTM-CRF model with character-level word embeddings which are derived by applying a BiLSTM to the character sequence in each word (Lample et al., 2016).

Following Luo et al. (2018), we also consider the impact of extra features including syntactic features such as POS and chunking tags, and a chemical term feature based on matching to an external gazetteer. Figure 1 illustrates the general BiLSTM-CRF model architecture with character-level word embeddings and additional features, while Figure 2 illustrates CNN-based and LSTM-based architectures for learning the character-level word embeddings.

2.3 Implementation details

We used a well-known implementation of BiLSTM-CRF-based models from Reimers and Gurevych (2017b).[3] We used the training set to learn model parameters, the development set to select optimal hyper-parameters, and the test set to report final results. Here, we tune the model hyper-parameters using the performance across both NER categories ("Overall") on the development set.

We employed pre-trained 50-dimensional word vectors from Luo et al. (2018). These pre-trained vectors were derived by training the Word2Vec skip-gram model (Mikolov et al., 2013) on a large text collection of 2 million MEDLINE abstracts.

[1] https://github.com/lingluodlut/Att-ChemdNER

[2] http://www.nactem.ac.uk/GENIA/tagger

[3] https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

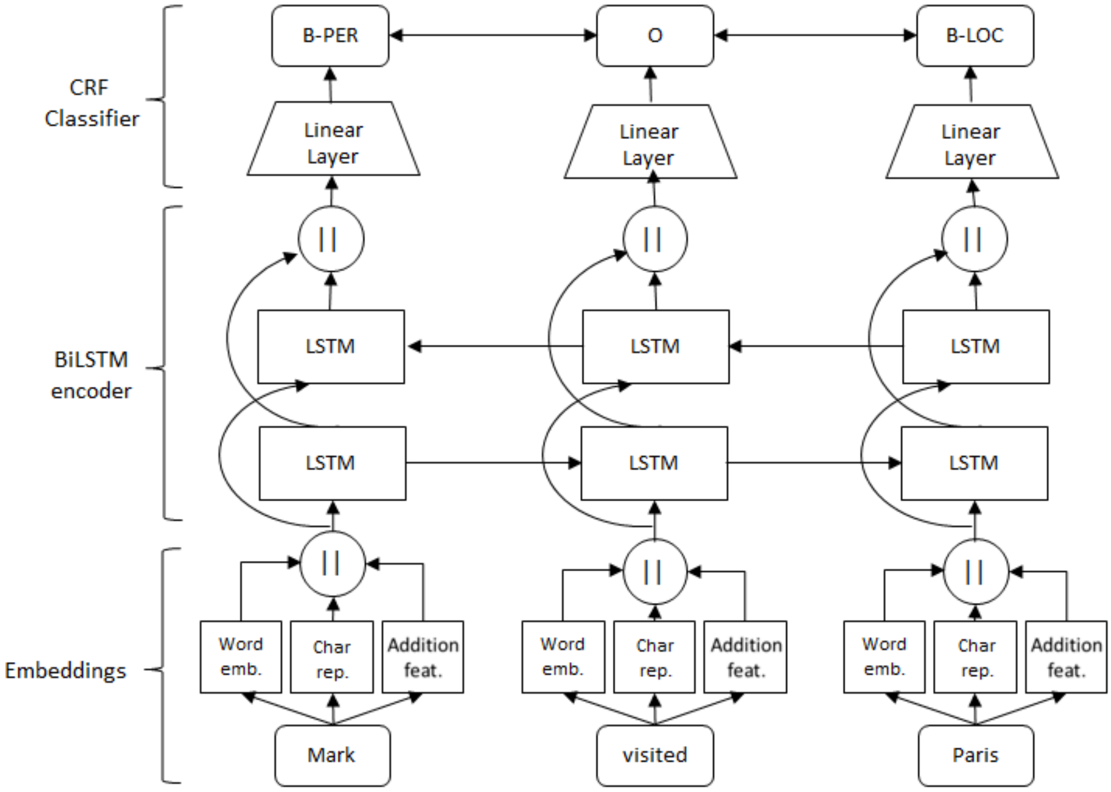
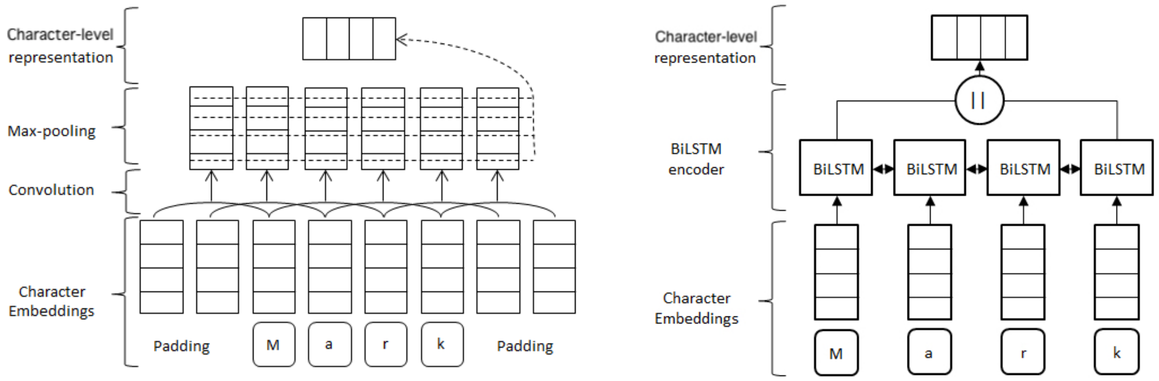


Figure 1: Architecture of BiLSTM-CRF models with character-level word representations and additional features. This figure is adapted from Reimers and Gurevych (2017a).



(CNN-based character-level word representation) (LSTM-based character-level word representation)

Figure 2: Character-level word representations. This figure is also adapted from Reimers and Gurevych (2017a).

Reimers and Gurevych (2017b) showed that the BiLSTM-CRF model achieved best performance with 2 BiLSTM layers. Therefore, in our experiment, we only evaluated models up to 2 stacked BiLSTM layers. The size of LSTM hidden states in each layer was selected from [100, 150, 200, 250]. We achieved the highest F1 score on the development set when using 250-dimensional LSTM hidden states for all models.

By default, each of the additional features (POS, chunking tags, gazetteer match tag) was incorporated into the model via a 10-dimensional embedding. Other hyper parameters were also fixed as in Reimers and Gurevych (2017b) during initialization. See tables 1 and 2 for more details.

In the training process, we used the score on development set to assess model improvement. Early stopping was applied if there was no improvement after 10 epochs. The threshold for a word that was not in the word embedding vocabulary to be added into the embedding was set to 5. The average training time for each epoch was also recorded.

Model	Chemical			Disease			Overall		
	P	R	F_1	P	R	F_1	P	R	F_1
BiLSTM	87.48	91.61	89.50	78.22	83.54	80.80	83.26	87.97	85.55
BiLSTM + CNN-char	90.65	90.70	90.67	79.34	82.66	80.97	85.44	87.07	86.25
BiLSTM + LSTM-char	90.47	91.64	**91.05**	79.43	83.97	**81.64**	85.37	88.18	**86.76**
BiLSTM-CRF	90.75	90.96	90.86	80.74	83.75	82.21	86.15	87.71	86.92
BiLSTM-CRF + CNN-char	91.64	92.24	**91.94**	81.42	84.67	**83.01**	86.95	88.83	**87.88**
BiLSTM-CRF + LSTM-char	92.08	91.79	**91.94**	81.48	84.22	82.83	87.20	88.38	87.79
BiLSTM-CRF$_{\text{+Gazetteer}}$	92.26	91.01	91.63	81.87	82.19	82.03	87.53	87.03	87.28
BiLSTM-CRF$_{\text{+Gazetteer}}$+ CNN-char	92.62	92.03	**92.32**	80.72	85.28	**82.94**	87.07	88.99	**88.02**
BiLSTM-CRF$_{\text{+Gazetteer}}$ + LSTM-char	92.11	92.33	92.22	82.13	83.66	82.89	87.57	88.42	87.99
Att-BiLSTM-CRF (LSTM-char) (Luo et al., 2018)	92.88	91.07	91.96	-	-	-	-	-	-
Att-BiLSTM-CRF$_{\text{POS+Chunking+Gazetteer}}$ (LSTM-char)	93.49	91.68	**92.57**	-	-	-	-	-	-
TaggerOne (Leaman and Lu, 2016) [♠]	94.2	88.8	**91.4**	85.2	80.2	**82.6**	-	-	-
tmChem (Leaman et al., 2015) [♠]	93.2	84.0	88.4	-	-	-	-	-	-
Dnorm (Leaman et al., 2013) [♠]	-	-	-	82.0	79.5	80.7	-	-	-

Table 3: Results (in %) on the test set. [♠] denotes results reported on a 950/50 training/development split rather than our 900/100 split. As indicated, Att-BiLSTM-CRF used LSTM-char word embeddings.

3 Main results

3.1 Baseline results

Table 3 presents our empirical results. The first three rows show the performance of baseline models without the CRF layer, the next three rows show the performance of BiLSTM-CRF models without additional features, and then the next three rows show the results for BiLSTM-CRF models with additional gazetteer features.

As the empirical results in Table 3 show, the model with CNN character-level embeddings (CNN-char) and the model with LSTM character-level embeddings (LSTM-char) achieved similar overall F1 scores (87.88% and 87.79%, respectively), outperforming BiLSTM-CRF by approximately 1% in absolute terms. In particular, on chemical NER, both BiLSTM-CRF-based models with character-level word embeddings obtained the same F1 score (91.94%), while on disease NER the model with CNN-char obtained slightly higher performance (83.01%) than the model with LSTM-char (82.83%). All models with the CRF layer outperformed their respective baseline BiLSTM models in F1 scores for all entity categories.

3.2 Effect of additional features

When incorporating additional POS and chunking features into three baseline BiLSTM-CRF-based models, we found that no performance improvement based on the baseline models was observed.

On chemical NER, the additional gazetteer feature improved the baseline BiLSTM-CRF by about 0.8% while it only improved the baselines BiLSTM-CRF + CNN-char and BiLSTM-CRF + LSTM-char by about 0.3%, thus clearly indicating that character-level word embeddings can capture unseen word information. Considering both NER categories together ("Overall"), the best performance was also obtained when the gazetteer feature was added, reaching overall F1 scores of 88.02% and 87.99%, respectively, for the two CNN-based and LSTM-based character-level embedding models.

3.3 Comparison with prior work

The performance comparison between our BiLSTM-CRF-based models and other machine learning approaches to the two studied NER tasks is also shown in Table 3. The pattern of chemical NER outperforming disease NER is consistent across all tools.

The Att-BiLSTM-CRF model (Luo et al., 2018) used a BiLSTM-CRF model with LSTM character-level word embedding and an additional attention layer. It achieved an F1 score of 91.96% on chemical NER without additional features. The positive effect of a gazetteer feature was also observed in their results; the model with syntactic and gazetteer features reached an F1 score of 92.57%. Note that the datasets used in this paper might not be exactly the same as ours due to random sampling.

The last three rows of Table 3 show the results presented in Leaman and Lu (2016), where 950 of the abstracts were used for training and 50 for development (cf. our 900/100 split). Dnorm (Leaman et al., 2013) is a model based on pairwise learning to rank on disease name normalization, which achieved F1 score of 80.7% on disease

NER. The tmChem (Leaman et al., 2015) is based on CRF; using numerous hand-crafted features it reached an F1 score of 88.4% on chemical entities. As a semi-Markov model with a richer set of features for NER tasks, TaggerOne (Leaman and Lu, 2016) achieved F1 score of 91.4% and 82.6% on chemical and disease entities, respectively.

Compared to previous non-deep-learning methods using CRFs, the BiLSTM-CRF models have significant advantage on F1 score of both chemical and disease entities, primarily due to improvement on recall.

3.4 Discussion

In our experiment on the effect of additional features, we found that syntactic features such as POS and chunking information did not have clear positive effect on the performance. In contrast, the match/partial match between words and entries in the chemical gazetteer is a good indicator for the presence of chemical entities. Since the Jochem dictionary contains only chemical entities, it is not surprising that the performance on diseases was not substantially impacted by adding the gazetteer feature, although some small variations in performance can be observed, likely due to changed influences from neighboring terms.

The empirical results shown that models using either CNN-char or LSTM-char achieve a similar overall F1 score on chemical and disease NER. The results are further comparable with other state-of-the-art models. This indicates that these character-level models have sufficient complexity to learn the generalizable morphological and lexical patterns in biomedical named entity terms.

On the other hand, as shown by the substantial differences in the number of parameters in Table 2, CNN (LeCun et al., 1989) has the advantage of reduced training complexity as compared to the LSTM models (Hochreiter and Schmidhuber, 1997) under similar experimental settings. In our experimental environment, the execution time of the model with LSTM-char increased 115% relative to the baseline BiLSTM-CRF model, while it only increased by 25% for with CNN-char, as detailed in Table 4. Therefore, consistent with prior results on general NER, we conclude that CNN-based embeddings are preferable to LSTM-based embeddings for BNER.

Model	Avg. Runtime per Epoch (seconds)	Δ
BiLSTM-CRF	106	0
+ CNN-char	**134**	**+26%**
+ LSTM-char	229	+115%

Table 4: Training time of best performing models (2 BiLSTM layers and 250 LSTM units), computed on a Intel Core i5 2.9 GHz PC.

We analyzed the error cases of the CNN-char and LSTM-char models without additional features: 3326 and 3271 words were incorrectly predicted using CNN-char and LSTM-char, respectively, with 2138 mistakes in common. In errors which only was made by one of the two models, we found that CNN-char made more false positive predictions and fewer false negative predictions, while LSTM-char made approximately an even number of the two kinds of false predictions.

The relationship between the length of words and these errors was also explored. For words less than 20 characters in length, the distribution of errors is almost identical for the two models. However, for longer words, the model with LSTM-char tends to make more mistakes. This supports prior observations that LSTM can be difficult to apply to long sequences of input (Bradbury et al., 2017). In approximately 50% of error cases, the word length is short, less than 5 characters. Short biomedical named entities are usually abbreviations and tend to be out-of-vocabulary terms, and are therefore particularly difficult for the character-level word embedding models to capture (Habibi et al., 2017).

4 Conclusion

We compared the performance of BiLSTM-CRF models with CNN-based and LSTM-based character-level word embeddings for biomedical named entity recognition. We confirmed previously published results on chemical and disease NER that demonstrate that character-level embeddings are helpful. We further show empirically, generalizing prior results for general NER to the biomedical context, that there is little difference between the two approaches: both types of character-level word embeddings achieved identical F1 score on the chemical NER task, and similar performance on disease NER (with CNN-char showing a slight performance advantage). However, the CNN embeddings show a substantial advantage in reduced training complexity.

Acknowledgments

This work was supported by the ARC Discovery Project DP150101550 and ARC Linkage Project LP160101469.

References

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2017. Quasi-Recurrent Neural Networks. In *Proceedings of the 2017 International Conference on Learning Representations*.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Kristina M. Hettne, Rob H. Stierum, Martijn J. Schuemie, Peter J. M. Hendriksen, Bob J. A. Schijvenaars, Erik M. van Mulligen, Jos Kleinjans, and Jan A. Kors. 2009. A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22):2983–2991.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint*, arXiv:1508.01991.

Rezarta Islamaj Dogan, G. Craig Murray, Aurélie Névéol, and Zhiyong Lu. 2009. Understanding PubMed user search behavior through log analysis. *Database*, 2009:bap018.

Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. 2008. Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics*, 24(13):i268–i276.

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.

Robert Leaman, Rezarta Islamaj Doan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839–2846.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(1):S3.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.*, 1(4):541–551.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Chen Lyu, Bo Chen, Yafeng Ren, and Donghong Ji. 2017. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinformatics*, 18(1):462.

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)*, pages 1064–1074.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Nils Reimers and Iryna Gurevych. 2017a. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. *arXiv preprint*, arXiv:1707.06799.

Nils Reimers and Iryna Gurevych. 2017b. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348.

M. Schuster and K.K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics*, pages 382–392.