A Named Entity Recognition Shootout for German

Martin Riedl and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{riedlmn,pado}@ims.uni-stuttgart.de,

Abstract

We ask how to practically build a model for German named entity recognition (NER) that performs at the state of the art for both contemporary and historical texts, i.e., a big-data and a small-data scenario. The two best-performing model families are pitted against each other (linear-chain CRFs and BiLSTM) to observe the trade-off between expressiveness and data requirements. BiLSTM outperforms the CRF when large datasets are available and performs inferior for the smallest dataset. BiLSTMs profit substantially from transfer learning, which enables them to be trained on multiple corpora, resulting in a new state-of-the-art model for German NER on two contemporary German corpora (CoNLL 2003 and GermEval 2014) and two historic corpora.

1 Introduction

Named entity recognition and classification (NER) is a central component in many natural language processing pipelines. High-quality NER is crucial for applications like information extraction, question answering, or entity linking.

Since the goal of NER is to recognize instances of named entities in running text, it is established practice to treat NER as a "word-by-word sequence labeling task" (Jurafsky and Martin, 2009). There are two families of sequence models that constitute promising candidates. On the one hand, linear-chain CRFs, which form the basis for many widely used systems (e.g., Finkel et al., 2005; Benikova et al., 2015), profit from hand-crafted features and can easily incorporate language- and domain-specific knowledge from dictionaries or gazetteers. On the other hand, bidirectional LSTMSs (BiLSTMs, e.g., Reimers and Gurevych, 2017) identify informative features directly from the data, presented as word and/or character embeddings (e.g., Mikolov et al., 2013; Bojanowski et al., 2017).

When developing NER tools for new types of text, one requirement is the availability of different resources to inform features and/or embeddings. Another one is the amount of training data: linear-chain CRFs require only moderate amounts of training data compared to BiLSTM. To perform representation learning, BiLSTMs require considerably annotated data to learn proper representations (see, e.g., the impact of training size by Dernoncourt et al., 2016). This consideration becomes particularly pressing when moving to "small-data" settings such as low-resource languages, specific domains, or historical corpora. Thus, it is an open question, whether it is generally a better idea to choose different model families for different settings, or whether one model family can be optimized to perform well across settings.

This paper investigates this question empirically on a set of German corpora including two large, contemporary corpora and two small historical corpora. We pit linear-chain CRF- and BiLSTM-based systems against each other and compare to state-of-the-art models, performing three experiments. Due to these experiments, we get the following results: (a), the BiLSTM system indeed performs best on contemporary corpora, both within and across domains; (b), the BiLSTM system performs worse than the CRF systems for the smallest historical corpus due to lack of data; (c), by applying transfer learning to adduce more training data, the RNN outperform CRFs substantially for all corpora. The final BiLSTM models form a new state of the art for German NER and are freely available.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 120–125
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

2 Model Families for NER

As mentioned above, contemporary research on NER almost exclusively uses sequence classification models. Our study focuses on CRFs and BiLSTMs, the two most widely used choices.

CRF-based Systems. Linear-chain CRFs form a family of models that are well established in sequence classification. They form the basis of two widely used Named Entity recognizers.

The first one is STANFORDNER[1] (Finkel et al., 2005) which provides models for various languages. It uses a set of language-independent features, including word and character n-grams, word shapes, surrounding POS and lemmas. For German, these features are complemented by distributional clusters computed on a large German web corpus (Faruqui and Padó, 2010). The ready-to-run model is pre-trained on the German CoNLL 2003 data (Tjong Kim Sang and De Meulder, 2003).

Benikova et al. (2015) developed GERMANER[2] , another CRF-based NER system. It was optimized for the GermEval 2014 NER challenge and also uses a set of standard features (word and character n-grams, POS) supplemented by a number of specific information sources (unsupervised parts of speech (Biemann, 2009), distributional semantics and topic cluster information, gazetteer lists).

BiLSTM-based Systems. Among the various deep learning architectures applied for NER, the best results have been achieved with bidirectional LSTM methods combined with a top-level CRF model (Ma and Hovy, 2016; Lample et al., 2016; Reimers and Gurevych, 2017). In this work, we use an implementation that solely uses word and character embeddings.

We train the character embeddings while training the model but use pre-trained word embeddings. To alleviate issues with out-of-vocabulary (OOV) words, we use both character- and subword-based word embeddings computed with fastText (Bojanowski et al., 2017). This method is able to retrieve embeddings for unknown words by incorporating subword information.[3]

3 Datasets

For the evaluation, we use two established datasets for NER on contemporary German and two datasets for historical German.

Contemporary German. The first large-scale German NER dataset was published as part of the CoNLL 2003 shared task (CoNLL, Tjong Kim Sang and De Meulder, 2003). It consists of about 220k tokens (for training) of annotated newspaper documents. The tagset handles locations (LOC), organizations (ORG), persons (PER) and the remaining entities as miscellaneous (MISC). The second dataset is the GermEval 2014 shared task dataset (GermEval, Benikova et al. (2014)), consisting of some 450k tokens (for training) of Wikipedia articles.[4] This dataset has two levels of annotations: outer and inner span named entities. For example, the term *Chicago Bulls* is tagged as organization in the outer span annotation. The nested term *Chicago* is annotated as location in the inner span annotation. However, there are only few inner span annotations. In addition to the standard tagsets also used in the CoNLL dataset, fine grained versions of these entities are marked with suffixes: *-deriv* marks derivations of the named entities (e.g. *German actor* – German is a derived location) and *-part* marks compounds including a named entity (e.g. in the word *Rhineshore* the compound Rhine is location). To compare to previous state-of-the-art methods, we show results on the official metric (a combination of the outer and inner spans) in Section 4. As there are only few inner span annotations, we additionally report results based on the outer spans. To be more conform with the tagsets of the CoNLL task, we focus on outer spans and remove the fine-grained tags in the follow-up experiments (see Section 5 and 6).

Historical German. We further consider two datasets based on historical texts (Neudecker, 2016)[5], extracted from the Europeana collection of historical newspapers[6], a standard resource for historical digital humanities. More specifically, our first corpus is the collection of Tyrolean periodicals and newspapers from the Dr Friedrich Temann Library (LFT), covering around 87k tokens from

[1] http://stanford.io/2ohopn3

[2] http://github.com/tudarmstadt-lt/GermaNER

[3] The source code and the best performing models are available online: http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/german_ner.html

[4] https://sites.google.com/site/germeval2014ner/

[5] https://github.com/KBNLresearch/europeananp-ner/

[6] www.europeana.eu/portal/de

Type	Model	Pr	R	F1
CRF	StanfordNER	80.02	62.29	70.05
CRF	GermaNER	81.31	68.00	74.06
RNN	UKP	79.54	71.10	75.09
–	ExB	78.07	74.75	76.38
RNN	BiLSTM-WikiEmb	**81.95**	**78.13**	**79.99***
RNN	BiLSTM-EuroEmb	75.50	70.72	73.03

Table 1: Evaluation on GermEval data, using the official metric (metric 1) of the GermEval 2014 task that combines inner and outer chunks.

Type	Model	Pr	R	F1
CRF	StanfordNER	80.13	65.43	72.04
CRF	GermaNER	82.72	71.19	76.52
RNN	UKP	79.90	74.13	76.91
–	ExB	80.67	77.55	79.08
RNN	BiLSTM-WikiEmb	**83.07**	**80.62**	**81.83***
RNN	BiLSTM-EuroEmb	76.48	73.54	74.98

Table 2: Evaluation on the test set of GermEval 2014 using the Outer Chunks evaluation schema.

Type	Model	Pr	R	F1
CRF	StanfordNER	74.18	72.50	73.33
RNN	Lample et al. (2016)	-	-	78.76
CRF	GermaNER	85.88	73.78	79.37
RNN	BiLSTM-WikiEmb	**87.67**	**78.79**	**82.99***
RNN	BiLSTM-EuroEmb	79.92	72.14	75.83

Table 3: Evaluation on the test set of the German CoNLL 2003 dataset.

1926. Our second corpus is a collection of Austrian newspaper texts from the Austrian National Library (ONB), covering some 35k tokens between 1710 and 1873. These corpora give rise to a number of challenges: they are considerably smaller than the contemporary corpora from above, contain a different language variety (19th century Austrian German), and include a high rate of OCR errors since they were originally printed in Gothic typeface.[7] We use 80% of the data for training and each 10% for development and testing.

4 Experiment 1: Contemporary German

In our first experiment, we compare the NER performances on the two contemporary, large datasets. For BiLSTM, we experiment with two options for word embeddings. First, we use pre-trained embeddings computed on Wikipedia with 300 dimensions and standard parameters (WikiEmb)[8], which are presumably more appropriate for contemporary texts. Second, we compute embeddings with the same parameters from 1.5 billion tokens of historic German texts from Europeana (EuroEmb). These embeddings should be more appropriate for historical texts but may suffer from sparsity.

Table 1 shows results on GermEval using the official metric (metric 1) for the best performing systems. This measure considers both outer and inner span annotations. Within the challenge, the ExB (Hänig et al., 2015) ensemble classifier achieved the best result with an F1 score of 76.38, followed by the RNN-based method from UKP (Reimers et al., 2014) with 75.09. GermaNER achieves high precision, but cannot compete in terms of recall. Our BiLSTM with Wikipedia word embeddings, scores highest (79.99) and outperforms the shared task winner ExB significantly, based on a bootstrap resampling test (Efron and Tibshirani, 1994). Using Europeana embeddings, the performance drops to an F1 score of 73.03 – due to the difference in vocabulary. As the number of inner span annotations is marginal and hard to detect, we additionally present scores considering only outer span annotations in Table 2. Whereas the scores are slightly higher, we observe the same trend as from the previous results shown in Table 1.

On the CoNLL dataset (see Table 3) GermaNER outperforms the currently best-performing RNN-based system (Lample et al., 2016). The BiLSTM again yields the significantly best performance, matching its high precision while substantially improving recall. Again, lower F1 scores are achieved using the Europeana embeddings. In sum, we find that BiLSTM models can outperform CRF models when there is sufficient training data to profit from distributed representations.

5 Experiment 2: Cross-Corpus Performance

A potential downside of BiLSTMs is that learned models may be more text type specific, due to the high capacity of the models. Experiment 2 evaluates how well the models do when trained on one corpus and tested on another one, including historical corpora. To level the playing field, we reduce the detailed annotation of GermEval to the standard five-category set (PER, LOC, ORG, MISC, OTH).

Results for these experiments are presented in

[7] We cleaned the corpora by correcting named entity labels and tokenization. We will make these versions available.

[8] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

		Test data			
Model	Train	CoNLL	GermEval	LFT	ONB
Stanford NER	CoNLL	**72.12**	48.82	39.72	46.36
	GermEval	65.63	**72.09**	45.22	52.21
	LFT	35.25	35.00	**67.26**	52.77
	ONB	34.09	33.96	42.95	**72.42**
Germa NER	CoNLL	**79.37**	60.40	46.53	53.93
	GermEval	71.05	**76.37**	48.05	54.95
	LFT	44.87	45.82	**69.18**	56.38
	ONB	46.56	47.19	48.41	**73.31**
BiLSTM-WikiEmb	CoNLL	**82.99**	66.51	49.28	58.79
	GermEval	78.15	**82.93**	55.99	61.35
	LFT	57.27	53.38	**68.47**	65.53
	ONB	51.42	49.30	49.35	**70.46**
BiLSTM-EuroEmb	CoNLL	**75.83**	55.06	45.30	54.59
	GermEval	70.19	**75.24**	52.15	59.43
	LFT	43.63	43.82	**69.62**	61.10
	ONB	36.33	38.81	46.48	**67.29**

Table 4: Evaluation (F1) for two CRF-based methods and BiLSTM trained and tested on different corpora.

Table 4. Unsurprisingly, the best results are gained when testing on the same dataset as the training has been performed. GermaNER consistently outperforms StanfordNER again, highlighting the benefits of knowledge engineering when using CRFs.

Interestingly, these benefits also extend to the historical datasets for which the CRF features were presumably not optimized: overall F1-scores are only a few points lower than for the contemporary corpora, and the CRFs significantly outperform the BiLSTM models on ONB and performs comparable on the larger LFT dataset. The type of embeddings used by BiLSTM plays a minor role for the historical corpora (for contemporary corpora, Wikipedia is clearly better). In sum, we conclude that BiLSTM models run into trouble when faced with very small training datasets, while CRF-based methods are more robust (Cotterell and Duh, 2017).

6 Experiment 3: Transfer Learning

If the problems of BiLSTM from the last section are in fact due to lack of data, we might be able to obtain an improvement by combining them. A simple way of doing this is transfer learning (Lee et al., 2017): we simply start training on one corpus and at some point switch to another corpus. In our scenario, we start by training on large contemporary "source" corpora until convergence and then train additional 15 epochs on the "target" corpus from the domain on which we evaluate. The results in Table 5 show significant improvements for the CoNLL dataset but performance drops for GermEval. Combining contemporary sources with historic target corpora yields to consistent benefits. Performance on LFT increases from 69.62 to 74.33 and on ONB from 73.31 to 78.56. Cross-domain classification scores are also improved consistently. The GermEval corpus is more appropriate as a source corpus, presumably because it is both larger and drawn from encyclopaedic text, more varied than newswire. We conclude that transfer learning is beneficial for BiLSTMs, especially when training data for the target domain is scarce. We applied the same procedure to the CRFs, but did not obtain improvements for the "target" data.

7 Data Analysis

Besides OCR errors, the lower F1 scores for the historic data are largely due to hyphens used to divide words for line breaks. The lowest F1 scores are achieved for the label organization. Evaluating on the ONB dataset, we obtain an F1 score for that label of 50.22 using GermaNER, 48.63 for the BiLSTM using Europeana embeddings and 61.48 using transfer learning. We observe a similar effect for the LFT dataset. Often, the annotations for the organization category are not entirely clear. For example, the typo "sterreichischen Außenministerlum" (should be "Außenministerium", *Austrian foreign ministry*) is manually annotated in the data but not detected by any of the models. However, "tschechoslowakischen Presse" (engl. *Czechoslovakian press*) is detected as organization by all classifiers but is not manually annotated.

8 Related Work

BiLSTMs that combine neural network architectures with CRF-based superstructures yield the highest results on English NER datasets in a number of studies (Ma and Hovy, 2016; Lample et al., 2016; Reimers and Gurevych, 2017; Lin et al., 2017). However, only few systems reported results for German NER, and restrict themselves to the "big-data" scenarios of the CoNLL 2003 (Lample et al., 2016; Reimers and Gurevych, 2017) and GermEval (Reimers et al., 2014; Christian Hnig, 2014) datasets.Sutton and McCallum (2005) showed the capability of CRFs for transfer learning by joint decoding two separately trained sequence models. Lee et al. (2017) apply transfer learning using a BiLSTM for medical NER using two similar tasks

		BiLSTM-WikiEmb				BiLSTM-EuroEmb			
Train	Transfer	CoNLL	GermEval	LFT	ONB	CoNLL	GermEval	LFT	ONB
CoNLL	GermEval	78.55	**82.93**	55.28	64.93	72.23	**75.78**	51.98	61.74
CoNLL	LFT	62.80	58.89	72.90	67.96	56.30	51.25	70.04	65.65
CoNLL	ONB	62.05	57.19	59.43	**76.17**	55.82	49.14	54.19	73.68
GermEval	CoNLL	**84.73**†	72.11	54.21	65.95	**78.41**	63.42	52.02	59.28
GermEval	LFT	67.77	69.09	**74.33**†	70.57	55.83	57.71	**72.03**	70.36
GermEval	ONB	72.15	73.18	62.52	76.06	64.05	64.20	57.12	**78.56**†

Table 5: Results for different test sets when using transfer learning. † marks results statistically significantly better than the ones reported in Table 4.

with different labels and show that only 60% of the data of the target domain is required to achieve good results. Crichton et al. (2017) yield improvements up to 0.8% for NER in the medical domain. Most related to our paper is the work by Ghaddar and Langlais (2017) which demonstrates the impact of transfer learning of the English CoNLL 2003 dataset with Wikipedia annotations.

9 Conclusion

Our study fills an empirical gap by considering historical datasets and performing careful comparisons of multiple models under exactly the same conditions. We have investigated the relative performance of an BiLSTM method and traditional CRFs on German NER in big- and small-data situations, asking whether it makes sense to consider different model types for different setups. We found that combining BiLSTM with a CRF as top layer, outperform CRFs with hand-coded features consistently when enough data is available. Even though RNNs struggle with small datasets, transfer learning is a simple and effective remedy to achieve state-of-the-art performance even for such datasets. In sum, modern RNNs consistently yield the best performance.In future work, we will extend the BiLSTM to other languages using cross-lingual embeddings (Ruder et al., 2017).

Acknowledgments

This study was supported by the DFG grant Oceanic Exchanges (KO 5362/2-1) and the CRETA center funded by the German Ministry for Education and Research (BMBF).

References

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of LREC*. Reykjavik, Iceland, pages 2524–2531.

Darina Benikova, Seid Muhie Yimam, and Chris Biemann. 2015. GermaNER: Free Open German Named Entity Recognition Tool. In *Proceedings of GSCL*. Essen, Germany, pages 31–38.

Chris Biemann. 2009. Unsupervised part-of-speech tagging in the large. *Research on Language and Computation* 7(2):101–135.

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.

Stefan Bordag Stefan Thomas Christian Hnig. 2014. Modular classifier ensemble architecture for named entity recognition on low resource systems. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*. Hildesheim, Germany, pages 113–116.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of IJCNLP*. Taipei, Taiwan, pages 91–96.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics* 18(1):1–14.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24(3):596–606.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS*. Saarbrücken, Germany, pages 129–133.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*. Ann Arbor, MI, USA, pages 363–370.

Abbas Ghaddar and Phillippe Langlais. 2017. WiNER: A Wikipedia Annotated Corpus for Named Entity Recognition. In *Proceedings of IJCNLP*. Taipei, Taiwan, pages 413–422.

Christian Hänig, Robert Remus, and Xose de la Puente. 2015. ExB Themis: Extensive Feature Extraction from Word Alignments for Semantic Textual Similarity. In *Proceedings of SemEval*. Denver, CO, pages 264–268.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing*. Prentice Hall, 2nd edition.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT*. San Diego, CA, pages 260–270.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Transfer learning for named-entity recognition with neural networks. *CoRR* abs/1705.06273. http://arxiv.org/abs/1705.06273.

Bill Y. Lin, Frank Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Copenhagen, Denmark, pages 160–165. http://aclweb.org/anthology/W17-4421.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL*. Berlin, Germany, pages 1064–1074.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of ICML*. Scottsdale, AZ, pages 1310–1318.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of LREC*. Portoro, Slovenia, pages 4348–4352.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, Jungi Kim, and Iryna Gurevych. 2014. Germeval-2014: Nested named entity recognition with neural networks. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*. Hildesheim, Germany, pages 117–120.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of EMNLP*. Copenhagen, Denmark, pages 338–348.

Sebastian Ruder, Ivan Vuli, and Anders Sgaard. 2017. A survey of cross-lingual embedding models. *CoRR* abs/1706.04902. http://arxiv.org/abs/1706.04902.

Charles Sutton and Andrew McCallum. 2005. Composition of conditional random fields for transfer learning. In *Proceedings of HLT-EMNLP*. Vancouver, BC, pages 748–754.

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 CoNLL-2003*. Edmonton, Canada, pages 142–147.