A Named Entity Recognition Shootout for German

Martin Riedl and **Sebastian Padó**

Institut für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, Germany

{martin.riedl, pado}@ims.uni-stuttgart.de

Universität Stuttgart

1 Introduction

Task: Named Entity Recognition (NER)

Recognition of proper names, e.g. locations, persons, organizations etc.

The	British	Government	sent	Captain	Arthur	Phillip	to	establish	a	colony	in	New	South	Wales
O	B-ORG	I-ORG	O	B-PER	I-PER	I-PER	O	O	O	O	O	B-LOC	I-LOC	I-LOC

Research Questions

- How can we build state-of-the-art performing German NER systems trained on …
  - … big data (contemporary data)?
  - … small data (historic data)?
- What are the performance differences between:
  - „Traditional“ CRF: established, fast, feature engineering, work with few amounts of training data
  - BiLSTM+CRF: representation learning, no feature engineering needed, long distance dependencies, requires large amounts of training data

→ obtain best practice for building NER systems

2 Experimental Settings

Datasets

	Contemporary Texts		Historic Texts	
	CoNLL 2003 [1]	GermEval 2014 [2]	Friedrich Teßmann Library (LFT) [3]	Austrian National Library (ONB) [3]
Text	Newspaper	Wikipedia	Newspaper	Newspaper
Time	~2000	2001-2004	1926	1710-1873
Tokens	220,000	450,000	87,000	35,000

Methods

- CRF-based methods:
  - StanfordNER [4]: CRF + standard features
  - GermaNER [5]: CRF + distributional semantics, gazetteers, …
- Recurrent Neural Network:
  - BiLSTM + CRF [6] using character- and word embeddings using FastText
    - Wikipedia (contemporary encyclopedia)
    - Europeana (historic newspaper from 1703 to 1899)

3 Exp. 1: Which method performs best on the contemporary datasets?

Model	CoNLL			GermEval		
	P	R	F1	P	R	F1
StanfordNER	74.18	72.50	73.33	80.13	65.43	72.04
GermaNER	85.99	73.78	79.37	82.72	71.19	76.52
BiLSTM-WikiEmb	**87.67**	**78.79**	**82.99***	**83.07**	**80.62**	**81.83***
BiLSTM-EuroEmb	79.92	72.14	75.83	76.48	73.54	74.98

BiLSTM outperforms CRFs due to higher recall

4 Exp. 2: How is the performance within and across corpora?

Model	Train	Test data: 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**

- Generalizing to different domains is hard
- NER on small historical domains is even harder
- CRF outperforms BiLSTM on small ONB dataset
- CRF performs similar to BiLSTM on LFT

5 Exp. 3: Can Transfer Learning help?

Same setup as Exp. 2 but with transfer learning and considering only the BiLSTM-based method

Train	Transfer	BiLSTM-WikiEmb: CoNLL	GermEval	LFT	ONB	BiLSTM-EuroEmb: 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**†

Contemporary corpora: minor improvements

GermEval: 82.93 → 82.93

CoNLL: 82.99 → 84.73

Historic corpora: major improvements

LFT: 69.62 → 73.44

ONB: 70.46 → 78.56

6 Conclusion & Future

LSTM + CRF based models outperform traditional CRF if:

- lots of training data is available
- if transfer learning is used

Usage of character- and substring-based embeddings (FastText) solves OOV issues

Future

- Learn multilingual models
- Analyze features learned by LSTM

7 Download

- Source Code
- Models

are freely available

https://github.com/riedlma/sequence_tagging

[1] E. F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proc. of CoNLL-2003. Edmonton, Canada, pages 142-147
[2] D. Benikova, C. Biemann, and M. Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Proceedings of LREC . Reykjavik, Iceland, pages 2524-2531
[3] C. Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In Proc. of LREC. Portoro, Slovenia, pages 4348-4352
[4] J. R.Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proc. of ACL. Ann Arbor, MI, USA, pages 363-370
[5] D. Benikova, S. Muhie Yimam, and C. Biemann. 2015. GermaNER: Free Open German Named Entity Recognition Tool. In Proc. of GSCL. Essen, Germany, pages 31-38
[6] X. Ma and E. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs- CRF. In Proc. of ACL. Berlin, Germany, pages 1064-1074