Evaluation of German Named Entity Recognition Tools on Medical Admission Notes

AP: NER tools for German medical text SoSe 2017

Phillip Richter-Pechański

September 17, 2017

1 Introduction and Previous Work

After investigating named entity recognition (NER) tools for German in our former task "Evaluation of German Named Entity Recognition Tools"[1] we now examine the performance of these tools on German medical texts.

There are some major challenges in research on NER in medical texts. Most of the time medical texts are free texts, sometimes semi-structured. The texts contain non standardized and ambiguous abbreviations and a varying and sometimes even locally specific terminology.

Medical entity recognition for English and non-German texts have been a research topic since Chang tried to extract medical acronyms from such texts.[2] Since then several approaches targeted on extracting and classifying medical entities from clinical and biomedical texts.[1][7][8][10][15][17]

Due to specific syntactic issues in German (e.g., morphology, capitalization), NER for German is much more challenging than NER for English.[3] The task of medical entity recognition and classification for German medical texts is still an almost unexplored research area.[11]. In Germany there are currently two commercial NER tools like RadMiner and ProMiner available[2]. In addition there are NER researches on local hospital level.[5][9][13]

The main obstacle for research on German medical texts is the lack of shared medical corpora.[11] If medical texts are available, privacy issues and the lack of annoted texts make serious research a hard task. The only major effort on creating a medical corpus for non-commercial NLP tasks is the FRAMED corpus, which contains around 100.000 annotated tokens from various medical genres. However this corpus is not publicly available, too.[14]

[1] *https://github.com/MaviccPRP/ger_ner_evals/*
[2] ProMiner[6], RadMiner[4]

2 Background

This study had the opportunity to use a corpus of around 180.000 medical admission notes with a total of around 132 million tokens due to the cooperation of the Section of Bioinformatics and Systems Cardiology at the Klaus Tschira Institute for Integrative Computational Cardiology under the supervision of Christoph Dieterich and the Database Systems Research Group at the Heidelberg University supervised by Michael Gertz. As these notes are not annotated and just on a very low level structured and standardized this study is designed as a proof-of-concept trying to run state of the art NER tools for German medical texts.
As a preliminary work we first investigated and compared all popular and publicly available NER tools for German. On GitHub[3] we collected all data and scores from the investigation. We finally identified the German Stanford NER[4] as the best performing tool on out-of-domain data and selected this tool for our task for NER on German medical texts.

3 Tools

This project is based on Java and Python libraries. In addition we developed some scripts and tools for processing and converting structured and unstructured corpora data.

3.1 Tools

- Python 3 (comfortable handling of UTF-8 text)
- Stanford NER[3][5]
- Stanford PTBTokenizer (Tokenizing the medical texts)[6]
- Bash tools for preprocessing and corpus analysis
- extract_corpus.py (Create new corpus from GermEval, CoNLL and EP and mixes in medical entities)
- evaluation.py (Evaluation script using Scikit Learn library)
- several converter script, for converting GermEval[7], CoNLL[8] and EP[9] data into Stanford compatible data

[3] *https://github.com/MaviccPRP/ger_ner_evals/*
[4] for details see footnote 6l footnote
[5] *https://nlp.stanford.edu/software/CRF-NER.shtml*
[6] *https://nlp.stanford.edu/software/tokenizer.shtml*
[7] *https://sites.google.com/site/germeval2014ner/data*
[8] *http://cogcomp.org/page/resource_view/81*
[9] *https://nlpado.de/ sebastian/software/ner_german.shtml*

4 Data

4.1 Non-Medical Data

We used two German training corpora typically used in German NER research. As named entity classes we adapted the classes used in the CoNLL 2003 corpus, excluding the MISC class, explained in table 1.

LOC	Locations
ORG	Organizations
PER	Persons

Table 1: Named Entity classes

German CoNLL 2003: Selected texts from German newspaper Frankfurter Rundschau. The training data consists of 206.931 tokens in 12.705 sentences. The named entity (NE) classes are distributed as described in table 2.

LOC	ORG	PER
4.363	2.427	2.773

Table 2: Absolute amount of entity tokens per entity class in CoNLL 2003

The **GermEval 2014** data set contains text from German Wikipedia articles and online news texts. The training data consists of 24.000 sentences. The data set contains over 590.000 tokens. The NE classes are distributed as described in table 3.

LOC	ORG	PER
12.791	9.889	12.423

Table 3: Absolute amount of entity tokens per entity class in GermEval

In addition we used texts from the European Parliament annotated by Sebastian Pado following the CoNLL 2003 guidelines.[10]

4.1.1 Preprocessing

Both corpora contain named entities in IOB format[11]. For the sake of simplicity we do not use this extensions in our training set. In addition we converted all training files into a two column simplified CoNLL 2003 style format. The first column contains one token per line, the second column contains the named entity classes. An example is shown in table 4

[10] https://www.nlpado.de/~sebastian/software/ner_german.shtml
[11] *https://en.wikipedia.org/wiki/Inside_Outside_Beginning*

Donald	PER
Trump	PER
works	O
in	O
Washington	LOC

Table 4: Example annotated file

As a sub-task we mixed in a list of drugs often used in cardiology. We simply randomly inserted a fixed amount of drugs without semantic context into the data set and annotated these as 'MEDICATION', so the Stanford NER tool can learn this class using a gazeteer and the tokens in the data set. This resulted in a sentence like in table 5:

Es	O
wird	O
zum	O
Glück	O
evaluiert	O
,	O
Aspirin	MEDICATION
bevor	O
wir	O
uns	O
mit	O
SAVE	ORG
II	ORG
befassen	O
.	O

Table 5: Mixed in MEDICATION entity including its class

4.2 Medical Data

The medical admission notes contain texts from the domain of cardiology. The majority of these notes have the following structure:

- Header
  - Addressee
  - Sender
  - Patients name and address
- Salutation

In the following text, not always in the same order, the following subsections appear in a majority of notes:

- Diagnosis
- Cardiovascular risk factors
- Allergies
- Anamnese
- Physical examination (Körperlicher Untersuchungsbefund)
- Laboratory data (some in tabular structure)
- ECG
- Recommended therapy
- Summary

The amount of text in each subsection is varying a lot. Additionally the subsections contain sometimes free unstructured text, sometimes tables. Occasionally subsections are titled differently, but contain similar informations, e.g., therapy/medication. Often terms are abbreviated, e.g., CRF (Cardiovascular Risk Factors).

The notes are concluded by a salutation and the names of the participating physicians. Listed in table 6 are the quantity of notes, lines and tokens per year detected via bash tools[12].

Year	Notes	Non-blank lines	Tokens
2004	2.272	302.997	1.823.801
2005	2.432	338.047	2.020.605
2006	2.185	300.386	1.872.093
2007	2.163	371.703	2.381.309
2008	1.989	364.386	2.427.136
2009	11.029	1.368.488	8.462.008
2010	44.099	5.095.453	30.330.830
2011	44.969	5.327.624	31.887.364
2013	22.426	2.515.247	14.414.940
2014	45.087	5.845.171	33.803.759
2016	2.249	516.175	3.335.795
Total	180.900	22.345.677	132.759.640

Table 6: Medical admission notes by year

[12]amount of notes: ls -l | wc -l; amount of lines: cat * | sed '/^\s*$/d' | wc -l

4.2.1 Preprocessing

The medical texts are delivered in a binary Microsoft Word doc format. Our predecessor research colleague already converted these documents into a standard text format. As this conversion is not problematic for free text, some issues appeared with MS Word tables. Most tables do not keep its semantic context as rows and columns are not kept accordingly. In this task we did not specifically address this issue, keeping this as a future task. We finally tokenized the notes using the Stanford PTBTokenizer. This resulted in plain text files, containing one token per line.

5 Training and Experiments

We trained three models for the Stanford NER. Because the medical admission notes are not annotated, we used popular NER data sets as out-of-domain corpora for training. As our previous evaluations showed, the German Stanford NER performed well on out-of-domain data[13].

To get larger training sets, we trained on combinations of available NER corpora. The first NER had been trained on the GermEval corpus plus the European Parliament corpus (EP) annotated by Sebastian Pado, the second model on the CoNLL 2003 corpus and the EP corpus and the third one we trained on all three corpora combined. Including the mixed-in 'MEDICATION' entities we produced training corpora with the following quantities. The modified ConLL 2003 corpus contains in total 244.543 tokens, the modified GermEval corpus 501.206 and the combined ConLL2003 + GermEval corpus contains 721.394 tokens.

	LOC	ORG	PER	MEDICATION
ConLL2003+GermEval+EP	18.131	14.303	17.036	2.808
GermEval+EP	12.892	10.061	12.541	2.808
ConLL2003+EP	5.340	4.414	4.613	2.808

Table 7: Total amount of NE in Training Corpora incl. MEDICATION entities

6 Evaluation

We could not automatically evaluate the models on a larger medical text test set, because of the lack of annotated data. Due to this limitation and the lack of time, we randomly chose one admission note and manually annotated the file with our four NE classes (PER, LOC, ORG, MEDICATION). The test data contains 1.049 tokens. From these tokens table 8 shows the amount of tokens per entity class.

The results in table 10 are rather hints and require further work, as we only had a small test set. For evaluation we used precision and recall with macro average. Due to

[13]For more detail see: *https* : *//github.com/MaviccPRP/ger_ner_evals/*

LOC	ORG	PER	MEDICATION
18	25	22	9

Table 8: Absolute number of tokens per class

the limited test set, we did not use the more restrictive scores used in the CoNLL 2003 shared task.[14] This means, we evaluate per token, not per entity sequence. To get the harmonic mean of both precision and recall, we evaluate our models using the F1-score:

$$F1 = \frac{2 \cdot precision * recall}{precision + recall}$$

	Precision	Recall	F1-Score
ConLL2003+GermEval+EP	64%	46%	51%
GermEval+EP	56%	43%	47%
ConLL2003+EP	61%	50%	51%

Table 9: Evaluation scores per training set

	LOC	ORG	PER	MEDICATION
ConLL2003+GermEval+EP	54%	11%	84%	54%
GermEval+EP	54%	11%	79%	43%
ConLL2003+EP	48%	42%	72%	50%

Table 10: F1-score per class per data set

With F1-scores between 72 and 84% the recognition of the PER class outperforms all other classes. As well the recognition of MEDICATION and LOC show a reasonable result between 43 and 54%. The annotation of 'MEDICATION' did not achieve better results as the manually annotated admission note contained all drugs as 'MEDICATION' entities, not just the drugs in the gazeteer.

A possible explanation for the good results of the PER class is, that this class contains personal names. Names have quite a similar pattern in newspaper texts, Wikipedia articles or medical admission notes. To a lesser extend but still similar this is true for the LOC class. That is why the class ORG performs worse, as the GermEval and CoNLL corpora contained rather political and cultural organizations than medical ones.

7 Summary and Future Work

After investigating and manually evaluating named and medical entity recognition on the medical admission notes these notes need to be further processed. As Sterlinger pointed out one reason for the lack of research in named and medical entity recognition on German

[14] A named entity is correct only if it is an exact match of the entity in the gold standard.[12].

medical texts is the lack of shared corpora.[11] To further investigate the admission notes and do meaningful information extraction the data need to be anonymized, keeping as much as possible semantic and syntactic structure. In a next step I will cover this topic in my BA thesis. I will adapt the methodology of Yuwono/Tou Ng [16] using header informations to find named entities in the medical texts. In addition I will use the Stanford NER tool, trained on the out-of-domain corpora to double check the anonymization process, as every token to be anonymized is as well a named entity. As we had good reasons to use gazeteers to recognize medical entities, I will further use other types of gazeteers for locations and names, to get a higher precision in NER.

References

[1] Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 185–192, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[2] Jeffrey Changm and Hinrich Schütze. Creating an online dictionary of abbreviations from medline. *Operations Research*, pages 612–620, 2002.

[3] Manaal Faruqui and Sebastian Padó. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany, 2010.

[4] Axel Gerstmair, Philipp Daumke, Kai Simon, Mathias Langer, and Elmar Kotter. Intelligent image retrieval based on radiology reports. *European Radiology*, 22(12):2750–2758, Dec 2012.

[5] Udo Hahn. Medsyndikate - a natural language system for the extraction of medical information from findings reports. 2002.

[6] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6(S-1), 2005.

[7] Min Jiang, Yukun Chen, Mei Liu, S Trent Rosenbloom, Subramani Mani, Joshua C Denny, and Hua Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011.

[8] Antonio Jimeno-Yepes, Ernesto Jiménez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga Llavori, and Dietrich Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(S-3), 2008.

[9] Hans-Ulrich Krieger, Christian Spurk, Hans Uszkoreit, Feiyu Xu, Yi Zhang, Frank Müller, and Thomas Tolxdorff. Information extraction from german patient records via hybrid parsing and relation extraction strategies. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[10] Maria Skeppstedt, Maria Kvist, Gunnar H. Nilsson, and Hercules Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49:148–158, 2014.

[11] Johannes Starlinger, Madeleine Kittner, and Oliver Blankenstein. How to improve information extraction from german medical records. pages 171–179, 2016.

[12] Erik F. Tjong Kim Sang and Fien De Meulder. 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 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[13] Martin Toepfer, Hamo Corovic, Georg Fette, Peter Klügl, Stefan Störk, and Frank Puppe. Fine-grained information extraction from german transthoracic echocardiography reports. *BMC Med. Inf. & Decision Making*, 15:91, 2015.

[14] Joachim Wermter and Udo Hahn. An annotated german-language medical text corpus as language resource. 01 2004.

[15] Yan Xu, Junichi Tsujii, and Eric Chang. Named entity recognition of follow-up and time information in 20,000 radiology reports. May 2012.

[16] Steven Kester Yuwono and Hwee Tou Ng. Automated anonymization as spelling variant detection. 2016.

[17] Shaodian Zhang and Noémie Elhadad. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6):1088 – 1098, 2013. Special Section: Social Media Environments.

8 Appendix

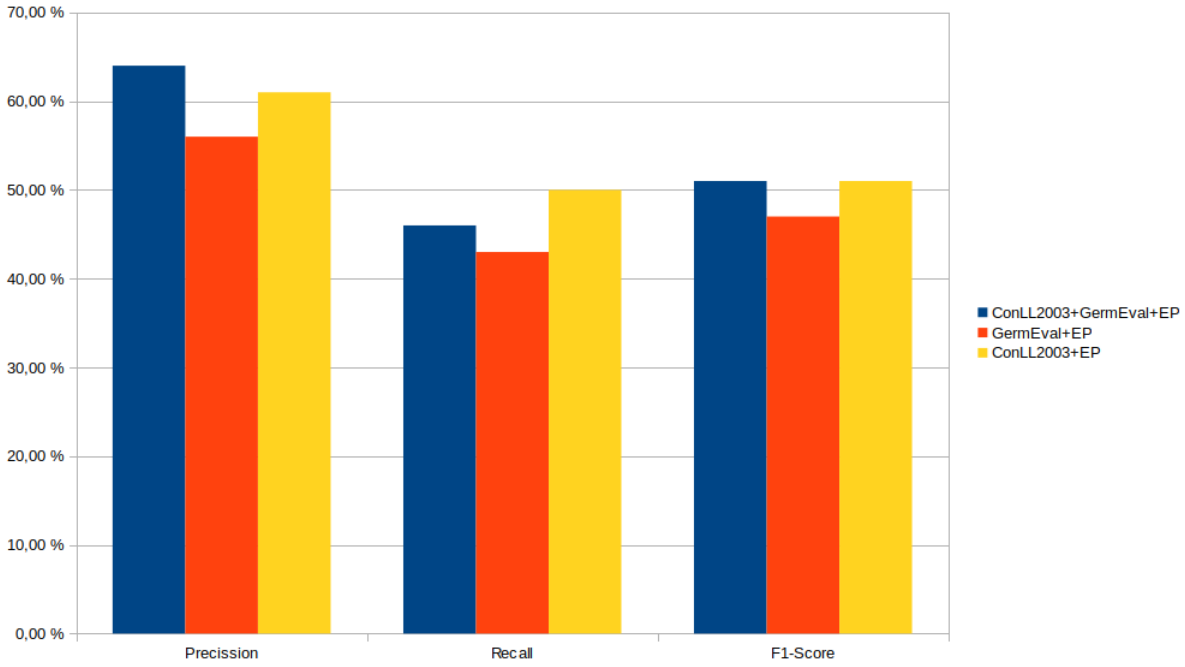

Figure 1: Evaluation scores per training set

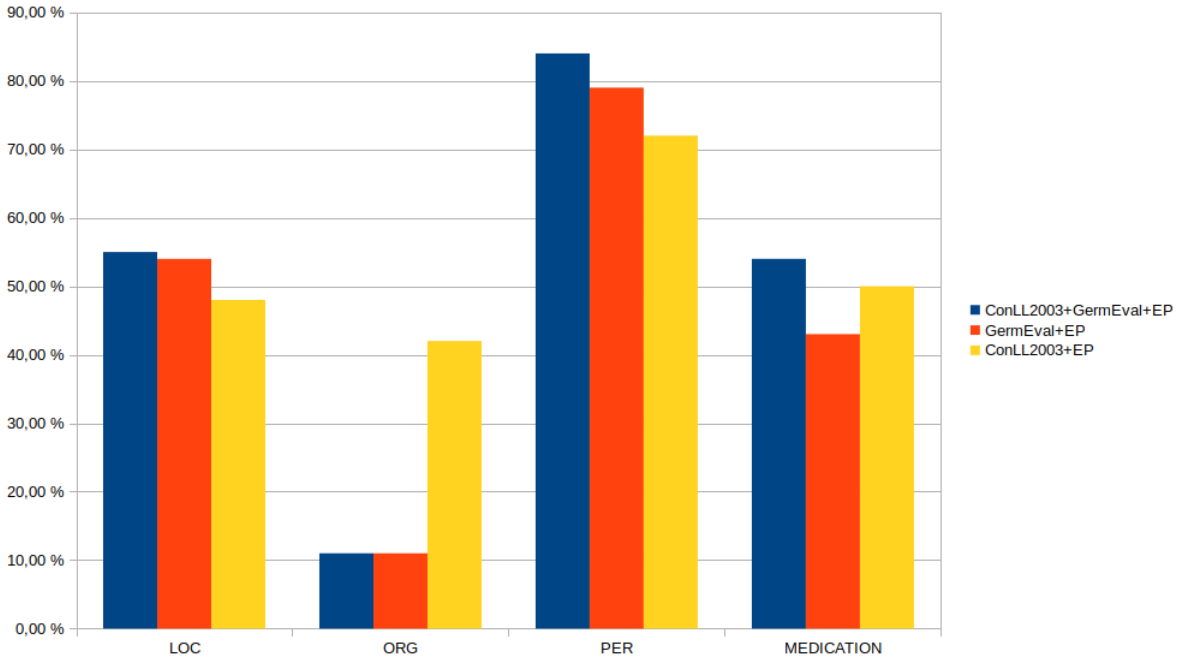

Figure 2: F1-score per class per data set