Evaluation of German NER Tools on Medical Admission Notes

AP: NER tools for German medical text SoSe 2017

Ph. Richter-Pechanski

September 20, 2017

Content

NER in Medical Texts

In comparison to other domains, named entity recognition (NER) on German medical texts is still an almost unexplored research area.
(Sterlinger 2016)

Problem Definition

Reasons for that are:

- Lack of shared medical corpora, as they must be anonymized first
- Lack of annotated corpora
- Still just a few cooperations between medical and computer science institutions

Problem Definition

Additionally there are some major challenges in research on medical texts:

- Medical texts are free texts

Problem Definition

Additionally there are some major challenges in research on medical texts:

- Medical texts are free texts
- Occasionally even hand-written

Problem Definition

Additionally there are some major challenges in research on medical texts:

- Medical texts are free texts
- Occasionally even hand-written
- Sometimes semi-structured

Problem Definition

Additionally there are some major challenges in research on medical texts:

- Medical texts are free texts
- Occasionally even hand-written
- Sometimes semi-structured
- Contain non-standarized and ambiguous abbreviations

Problem Definition

Additionally there are some major challenges in research on medical texts:

- Medical texts are free texts
- Occasionally even hand-written
- Sometimes semi-structured
- Contain non-standarized and ambiguous abbreviations
- Varying and sometimes even locally specific terminology

Problem Definition

This project investigated German NER tools and their performance on medical texts.
Following steps had been accomplished:

- Evaluation of state of the art NER tools for German

Problem Definition

This project investigated German NER tools and their performance on medical texts.
Following steps had been accomplished:

- Evaluation of state of the art NER tools for German
- Comparison of their performance on out-of-domain data sets

Problem Definition

This project investigated German NER tools and their performance on medical texts.
Following steps had been accomplished:

- Evaluation of state of the art NER tools for German
- Comparison of their performance on out-of-domain data sets
- Evaluation of NER tools trained on out-of-domain corpora on available German medical texts

Problem Definition

- Previous work showed that the Stanford NER tool performed best on out-of-domain data sets *(Richter-Pechanski 2017)*
- Today we focus on evaluation of this tool on medical admission notes

Tools

- Python 3 (comfortable handling of UTF-8 text)
- German Stanford NER *(Pado 2010)*
- Stanford PTBTokenizer (Tokenizing the medical texts)
- Bash tools for preprocessing and corpus analysis
- extract_corpus.py (Create new corpus from GermaEval, CoNLL and EP and mixes in medical entities)
- evaluation.py (Evaluation script using Scikit Learn library)
- several converter scripts, for converting GermEval, CoNLL and European Parliament data into Stanford NER compatible data

Non-medical Training Data

- Bypassing lack of annotated medical texts by using out-of-domain data

Non-medical Training Data

- Bypassing lack of annotated medical texts by using out-of-domain data
- Combining existing data sets to get larger training sets.

Non-medical Training Data

- Bypassing lack of annotated medical texts by using out-of-domain data
- Combining existing data sets to get larger training sets.
- Two German training corpora typically used in German NER research plus one manually annotated corpus. (total: around 720.000 tokens)
  - CoNLL 2003
  - GermEval 2014
  - European Parliament

Non-medical Training Data

- Bypassing lack of annotated medical texts by using out-of-domain data
- Combining existing data sets to get larger training sets.
- Two German training corpora typically used in German NER research plus one manually annotated corpus. (total: around 720.000 tokens)
    - CoNLL 2003
    - GermEval 2014
    - European Parliament
- As named entity classes we adapted the classes used in the CoNLL 2003 corpus. (exluding MISC class)

Non-medical Training Data

- Bypassing lack of annotated medical texts by using out-of-domain data
- Combining existing data sets to get larger training sets.
- Two German training corpora typically used in German NER research plus one manually annotated corpus. (total: around 720.000 tokens)
  - CoNLL 2003
  - GermEval 2014
  - European Parliament
- As named entity classes we adapted the classes used in the CoNLL 2003 corpus. (exluding MISC class)

LOC	Locations
ORG	Organizations
PER	Persons

Table: Named entity classes

MEDICATION Class

- As a sub-task we mixed in a list of drugs often used in cardiology
- Randomly inserted fixed amount of drugs **without semantic context** into the data set and annotated these as 'MEDICATION'

Example extract with MEDICATION class

Barack	LOC
Obama	LOC
erhält	O
Aspirin	MEDICATION
Friedensnobelpreis	MISC
.	O

Table: MEDICATION entity token including class label

Medical Test Data

Our medical data in this project consists of medical admission notes.

- Contain texts from cardiology domain

Medical Test Data

Our medical data in this project consists of medical admission notes.

- Contain texts from cardiology domain
- Let us have a look at an example

How does a Medical admission note look like?

Sehr geehrte Kollegen,

wir berichten über Ihre Patientin <PER>, geboren am <DATE>, wh. <LOC>, die sich am <DATE> in unserer Ambulanz vorstellte.

Diagnosen:
Aktuell: Operabilität mit leicht erhöhtem Risiko gegeben

Z.n. 2x Herzkatheter-Untersuchung, zuletzt <DATE> (anamnestisch
Z.n. HWS-Eingriff <DATE> (<LOC>), Versuch der Stabilisierung
Anamnestisch OP bei intracranieller Blutung <DATE>

Anamnese:
Die Vorstellung der Patientin erfolgte zur präoperativen kardiologischen Abklärung vor geplanter HWS-Operation. Die Patientin berichtet von reduzierter ...

Kardiovaskuläre Risikofaktoren: Arterielle Hypertonie, Nikotin (sistiert vor 20 Jahren, 2PY), ...

Medical Data

Following basic structure in majority of notes

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

Medical Data

Additionally, not always in same order, following subsections appear in a majority of notes:

Medical Data

Additionally, not always in same order, 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

Medical Data

Charecteristics of medical admission notes

- Amount of text varying a lot

Medical Data

Characteristics of medical admission notes

- Amount of text varying a lot
- Sometimes free unstructured text, sometimes tables

Medical Data

Charecteristics of medical admission notes

- Amount of text varying a lot
- Sometimes free unstructured text, sometimes tables
- Often subsections are titled differently, but contain similar informations

Medical Data

Charecteristics of medical admission notes

- Amount of text varying a lot
- Sometimes free unstructured text, sometimes tables
- Often subsections are titled differently, but contain similar informations
- Concluded by salutation and names of involved physicians.

Medical Data

- Time period: 2004-2016
- Total amount of notes: 180 000
- Total amount of tokens: 132 *Million*

Training and Experiments

- Trained three NER models

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

Table: Amount of entity tokens in training set incl. MEDICATION entities

Evaluation

- No automatic evaluation because no annotated data available

Evaluation

- No automatic evaluation because no annotated data available
- Manually annotated one admission note with our four NE classes

Evaluation

- No automatic evaluation because no annotated data available
- Manually annotated one admission note with our four NE classes
- Test data contains 1.049 tokens

Evaluation

- No automatic evaluation because no annotated data available
- Manually annotated one admission note with our four NE classes
- Test data contains 1.049 tokens

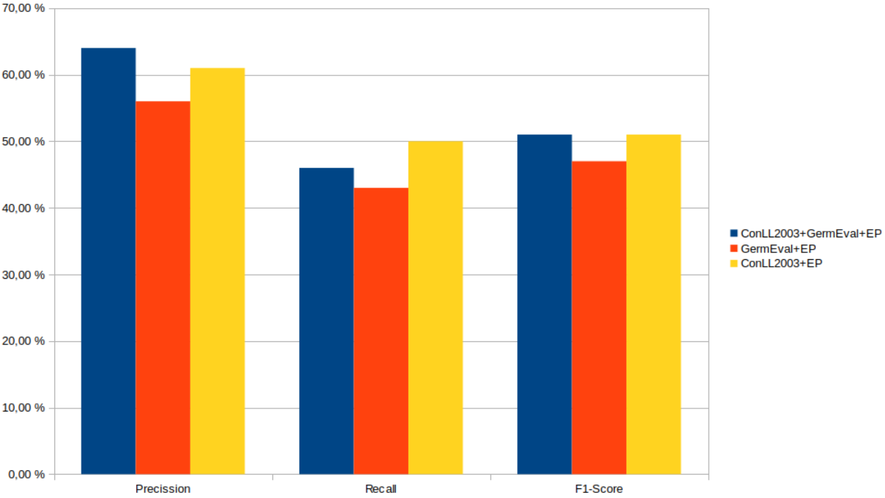
LOC	ORG	PER	MEDICATION
18	25	22	9

Table: Absolute number of tokens per class in test set

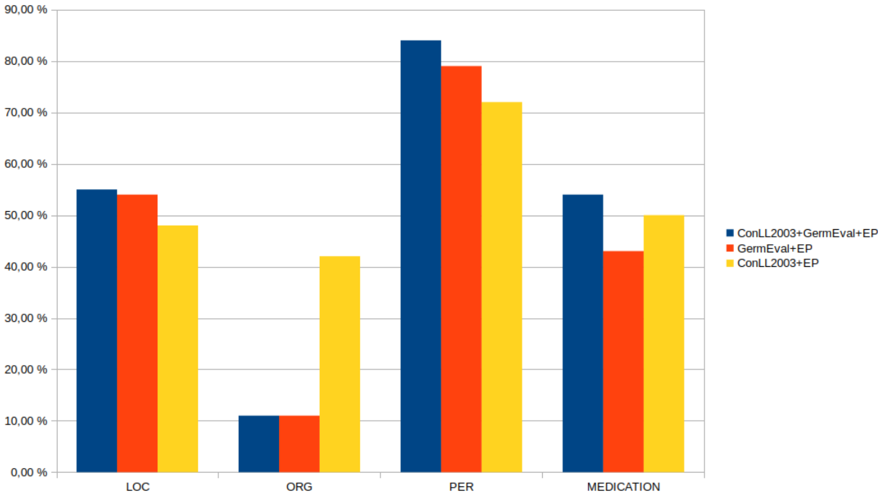
Evaluation

- Used scores are precision and recall with macro average
- Evaluation is token-wise not entity-wise, due to limited data

Evaluation Macro Average



Evaluation per NE



Evaluation

- PER class recognition outperforms all other NEs

Evaluation

- PER class recognition outperforms all other NEs
- LOC and MEDICATION recognition as well reasonable performance

Evaluation

- PER class recognition outperforms all other NEs
- LOC and MEDICATION recognition as well reasonable performance
- ORG recognition score in two data sets very low

Evaluation

Possible explanations for scores:

- PER entities look similar in test and trainings set

Evaluation

Possible explanations for scores:

- PER entities look similar in test and trainings set
- LOC entities are similar to a lesser extend

Evaluation

Possible explanations for scores:

- PER entities look similar in test and trainings set
- LOC entities are similar to a lesser extend
- ORG entities have different shape in newspaper texts and medical texts

Summary

- Out-of-domain trained NER models have reasonable performance on medical texts

Summary

- Out-of-domain trained NER models have reasonable performance on medical texts
- Further research and annotation work on medical texts is needed

Summary

- Out-of-domain trained NER models have reasonable performance on medical texts
- Further research and annotation work on medical texts is needed
- To do meaningful information extraction data need to be anonymized

Summary

- Out-of-domain trained NER models have reasonable performance on medical texts
- Further research and annotation work on medical texts is needed
- To do meaningful information extraction data need to be anonymized
- Goal is to keep as much as possible semantic and syntactic structure

Summary

- Out-of-domain trained NER models have reasonable performance on medical texts
- Further research and annotation work on medical texts is needed
- To do meaningful information extraction data need to be anonymized
- Goal is to keep as much as possible semantic and syntactic structure
- My BA thesis will implement an Anonymizer based on *Yuwono/TouNg* and the NER model trained in this project

References

- Yuwono, Steven Kester and Ng, Hwee Tou (2016):
  Automated Anonymization as Spelling Variant Detection.
- Starlinger, Johannes and Kittner, Madeleine and Blankenstein, Oliver (2016):
  How to improve information extraction from German medical records.
- Richter-P., Phillip (2017):
  Evaluation of German Named Entity Recognition Tools,
  https://github.com/MaviccPRP/ger_ner_evals/.

Acknowledgments

- Section of Bioinformatics and Systems Cardiology at the Klaus Tschira Institute for Integrative Computational Cardiology under the supervision of Christoph Dieterich
- Database Systems Research Group at the Heidelberg University supervised by Michael Gertz

Thank you!