Anonymization of German Medical Texts

using Spelling Variant Detection and Named Entity Recognition

Ph. Richter-Pechański

Department of Computational Linguistics at Heidelberg University

December 14, 2017

Outline

1. Introduction
2. Definitions
3. Data
4. Methods
5. Experiments & Evaluation
6. Future Work

Introduction

The task of anonymization of German medical texts is a precondition for any further research on these resources. (Starlinger 2017)

Problem Definition

- Main obstacle for research on German medical texts is the lack of shared medical corpora. (Starlinger 2017)
- If medical texts are available, privacy issues make serious research a hard task.
- Can only be analyzed without legal restrictions if personal details had been deleted by de-identification. (Schlunder 2015)

Anonymization, Pseudonymization, De-Identification

De-Identification

removing association between a set of identifying data and the data subject (ISO/TS 25237:2008)

Anonymization

Direct and indirect personal identifiers have been removed. Data can **never** be re-identified. (EU Directive 95/46/EC)

Pseudonymization

Data can no longer be attributed to a specific data subject without the use of additional information. (GDPR*, Article 4)

*General Data Protection Regulation

Personal Health Information (PHI)

PHI (Protected Health Information) defined in the U.S. Health Insurance Portability and Accountability Act (HIPPA).

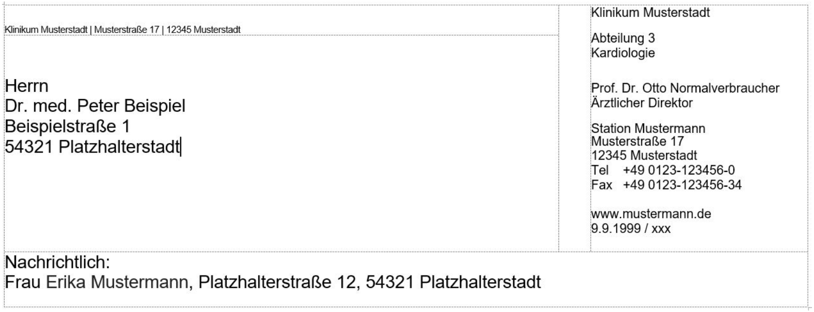
- names
- geographic data
- all elements of dates
- telephone/Fax numbers
- e-mail addresses
- web URIs

Data

- Medical admission notes from cardiological domain.
- Most notes contain a semi-structured header.
- Contain free texts, tables and images.
- No additional annotation.

- MS-DOC 97
- $\approx$ 180,000 notes
- $\approx$ 130 Mio. tokens.

Data (Example Header)



Klinikum Musterstadt | Musterstraße 17 | 12345 Musterstadt

Herrn
Dr. med. Peter Beispiel
Beispielstraße 1
54321 Platzhalterstadt

Klinikum Musterstadt

Abteilung 3
Kardiologie

Prof. Dr. Otto Normalverbraucher
Ärztlicher Direktor

Station Mustermann
Musterstraße 17
12345 Musterstadt
Tel +49 0123-123456-0
Fax +49 0123-123456-34

www.mustermann.de
9.9.1999 / xxx

Nachrichtlich:
Frau Erika Mustermann, Platzhalterstraße 12, 54321 Platzhalterstadt

Figure: Example header in a medical admission note

Data (Example)

Extract of an Admission Note

Sehr geehrte Kollegen,

wir berichten über Ihren Patienten <SALUTE> <PER>, geboren am <DATE>, wh. <LOC>, <PLZ> <LOC> der sich am <DATE> in unserer Ambulanz vorstellte.

Diagnosen:

Aktuell: Operabilität mit leicht erhöhtem Risiko gegeben Z.n. Herzkatheter-Untersuchung, zuletzt $< DATE >$ (anamnestisch)

Z.n. weiterer Eingriff <DATE> (<LOC>), Versuch der ... Anamnestisch OP bei ... <DATE>

Anamnese:

Die Vorstellung des Patienten erfolgte zur präoperativen kardiologischen Abklärung vor ...

Aktuelle Medikation:

Marcumar nach INR 1,0 – 2,0

EnaHexal X mg 0-0-1

Simvastatin X 0-0-

Laboratory:

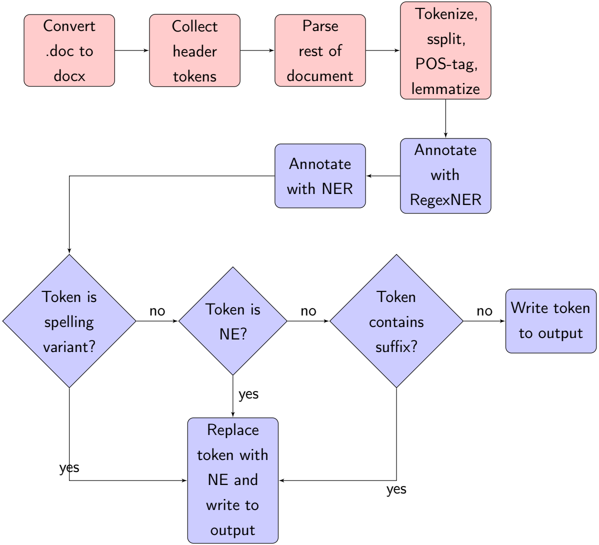
Glucose nue. 111 45-122 mg/dl HbA1c

Haemoglobin 13.9 12-15 g/dl MCV 2.0 2-2,2 /pl

Tools

- Java 8 (for de-identification task)
- Python 3 (for evaluation)
- German Stanford NER (Pado 2010)
- Stanford RegexNER (Manning 2014)
- Stanford Core NLP Pipeline (Tokenizing, sentence splitting, pos tagging and lemmatization)
- LibreOffice convert-to (converting MS-DOC to MS-DOCX)
- Apache POI XWPF (Handle tables/lists in MS-DOCX files)
- scikit-learn (for evaluation)
- mlxtend (for statistical tests)

Our Algorithm



Convert .doc to docx

Collect header tokens

Parse rest of document

Tokenize, ssplit, POS-tag, lemmatize

Annotate with RegexNER

Annotate with NER

Token is spelling variant?

no

Token is NE?

no

Token contains suffix?

no

Write token to output

yes

Replace token with NE and write to output

yes

yes

Regular Expressions and Gazetteers

Stanford RegexNER combines plain gazetteers for first names, surnames and German cities and towns with gazetteers containing regular expressions for street names, phone numbers, dates ....

Gazetteer Regex

```
Haup[sS]tr[\.]?(a(ß|ss)e)? [0-9]{0,3}[a-zA-Z]?
```

Regular Expressions and Gazetteers

In addition to NE classes used by the NER model (PER, LOC) with RegexNER we are trying to recognize the following named entity classes:

Named Entity Classes RegexNER

Class	Example
SALUTE	(Frau, Herr, Herrn)
EMAIL	(max@mustermann.de)
PHONE	(012 - 34 5677)
URI	(www.example.xyz)
DATE	(01.12.2015)
PLZ	14123
TITLE	(Dr., Dr. med.)

Named Entity Recognition

After investigating named entity recognition tools for German in our former task "Evaluation of German Named Entity Recognition Tools" we identified the German Stanford NER as the best performing tool on out of domain data. (Faruqui/Padó 2010, Richter-Pechanski 2017)

Named Entity Recognition

- The best model had been trained on a concatenation of ConLL 2003, GermEval 2014 and an Europarl data set. (Richter-Pechanski 2017)
- We evaluated the model recognizing PER, LOC and ORG tokens on a single test medical admission note containing 1049 tokens and 65 named entities.
- It achieved the following $F_1 - scores$:

LOC	ORG	PER
54%	11%	84%

Table: F_1-scores on medical text

⇒ Exluding ORG from our task.

Spelling Variant Detection

Header mostly contains:

- name and address of the patient
- name and address of the recipients
- contact information of the clinic

Spelling Variant Detection

- Approach of Yuwono, using minimum edit distance to identify spelling variants in the medical texts. (Yuwono 2016, Wagner/Fischer 1974)
- Edit distance ratio R is defined as: $R = \frac{d}{min(|n|,|w|)}$.
- d is minimum edit distance of n and w.
- Longer names have higher probability of being misspelled than short names.
- R takes into account the length of a string.
- If R $<$ 0.333 w will be de-identified. (Yuwono 2016)

Baseline

- Rules and gazetteers. (Uzuner 2008)
- Stanford RegexNER. (Manning 2014)
- Used for multiclass NE recognition and binary PHI recognition.
    - Effect of spelling variant detection only evaluated in binary PHI recognition, as detector is not labeling additional named entities, just decides weather to de-identify a PHI tag or to keep it.

Test Set

16 medical admission notes with 14.150 tokens.

Named Entity	#Tokens
DATE	241
PER	165
LOC	104
TITLE	75
SALUTE	52
PHONE	26
PLZ	15
ORG	2
URI	-
EMAIL	-
Total	680

Table: Named entity tokens in test set

Scores and Metrics

- Accuracy measure treats all classes the same.
- Accuracy is not feasible for highly imbalanced data. (Sokolova 2006)
- Not recognizing a PHI has a higher cost, than de-identifying a non-PHI.
- We focus on recall, to find out, how many relevant items had been recognized.

Scores and Metrics

- Using precision we control how many selected items are relevant.
- To take into account the higher relevance of recall, we are using $F_2 - score$ for calculating a single quality score for our classifiers (Ferrandez 2013)

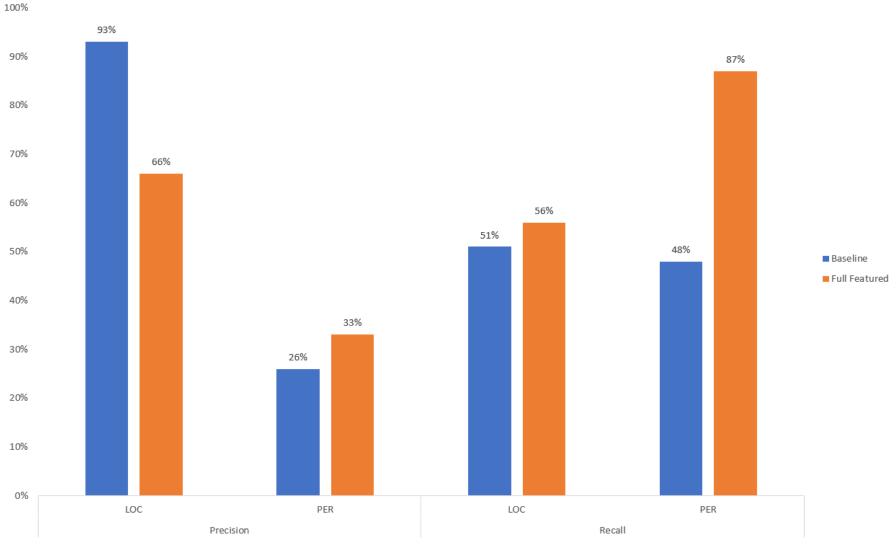
$$F_2 = \frac{5 \cdot (precision \cdot recall)}{(4 \cdot precision + recall)} \tag{1}$$

Multiclass NER Evaluation

Model	DATE	LOC	PER	PHONE	PLZ	SALUTE	TITLE
Baseline	0.94	0.56	0.41	0.73	0.97	0.98	0.97
Full Featured	0.94	0.57	0.66	0.73	0.97	0.98	0.97

Table: F_2-score on medical test set

Multiclass NER Evaluation



Final Thoughts Multiclass NER

- High recall and precision for purely rule-based recognized classes:
    - PLZ
    - SALUTE
    - TITLE
    - DATE

Final Thoughts Multiclass NER

- High recall and precision for purely rule-based recognized classes:
  - PLZ
  - SALUTE
  - TITLE
  - DATE
- Low recall of PHONE class (58%) => Shape of phone numbers are highly variable, thus hard to specify by a regular expression.

Final Thoughts Multiclass NER

- High recall and precision for purely rule-based recognized classes:
  - PLZ
  - SALUTE
  - TITLE
  - DATE
- Low recall of PHONE class (58%) => Shape of phone numbers are highly variable, thus hard to specify by a regular expression.
- Especially the PER class raises recall from 48 to 87%, but LOC class recall increased by 5%, too.

Final Thoughts Multiclass NER

- High recall and precision for purely rule-based recognized classes:
  - PLZ
  - SALUTE
  - TITLE
  - DATE
- Low recall of PHONE class (58%) => Shape of phone numbers are highly variable, thus hard to specify by a regular expression.
- Especially the PER class raises recall from 48 to 87%, but LOC class recall increased by 5%, too.
- Good results for PER class confirm our initial task „Named Entity Recognition on Medical Admission Notes“. (Richter-Pechanski, 2017)

Final Thoughts Multiclass NER

- High recall and precision for purely rule-based recognized classes:
  - PLZ
  - SALUTE
  - TITLE
  - DATE
- Low recall of PHONE class (58%) => Shape of phone numbers are highly variable, thus hard to specify by a regular expression.
- Especially the PER class raises recall from 48 to 87%, but LOC class recall increased by 5%, too.
- Good results for PER class confirm our initial task „Named Entity Recognition on Medical Admission Notes“. (Richter-Pechanski, 2017)
- Due to ambiguities of entries in LOC and PER gazetteers (Schöneberg, Rostock, u.ä.) precision of LOC class is suffering.

Binary PHI Recognition Evaluation

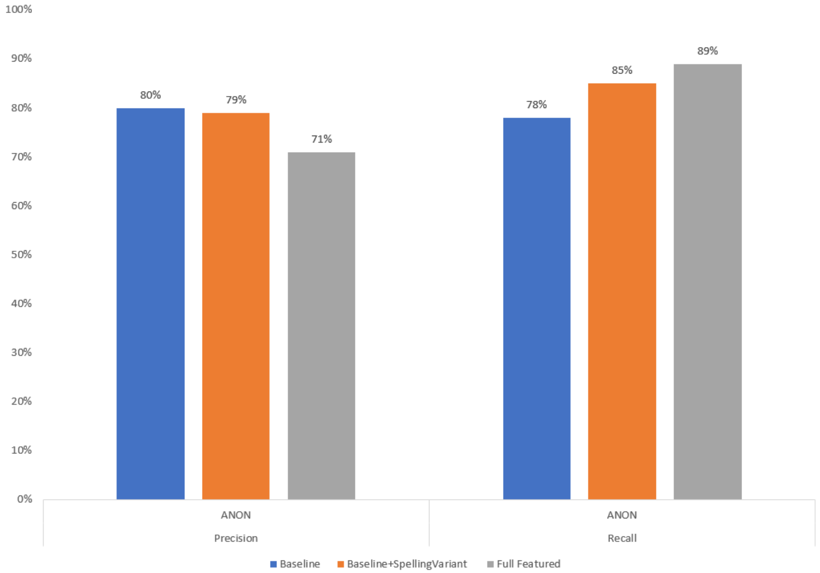
Model	ANON	KEEP
Baseline	0.78	0.98
Baseline + Spelling Variant	0.83	0.98
Full Featured	0.85	0.98

Table: PHI Recognition F_2-score

		Predicted	
		ANON	KEEP
True	ANON	605	75
	KEEP	251	13203

Table: PHI Confusion matrix

PHI Recognition Evaluation



Final Thoughts Binary Evaluation

- Using spelling variant detection our recall could be improved by 7%.

Final Thoughts Binary Evaluation

- Using spelling variant detection our recall could be improved by 7%.
- Adding our out of domain NER model is raising the recall by another 4%.

Final Thoughts Binary Evaluation

- Using spelling variant detection our recall could be improved by 7%.
- Adding our out of domain NER model is raising the recall by another 4%.
- Precision is suffering by 9%.

Final Thoughts Binary Evaluation

- Using spelling variant detection our recall could be improved by 7%.
- Adding our out of domain NER model is raising the recall by another 4%.
- Precision is suffering by 9%.
- Analyzing false negatives (not de-identified PHI):
  - 29x tokens of '*Chest Pain Unit*'
  - 2x surname (high cost)
  - parts of phone numbers, one letter name abbreviations

Significance Test Binary Model

Testing significance between binary models

		Full featured	
		correct	wrong
Baseline	correct	13726	82
	wrong	125	201

Table: McNemar contingency table binary

chi-squared: 8.52173913043

p-value: 0.00350928970475

McNemar's Test

- p-value (multi-class) $= 0.01676 < 0.05$
- p-value (binary) $= 0.00350 < 0.05$
- The multiclass classifiers and the binary classifiers predict significantly different.
- The null hypothesis $H_0 : p_b = p_c$ can be rejected
- b $<$ c $\Rightarrow$ full featured classifier has lower error rate than baseline classifier.

Feature Work

- As other research studies discovered, already small annotated training sets ($<$ 100 medical texts) can increase recall significantly. (Scheuerwegs 2013, Wellner 2007)
- As lack of training data is a pitfall word embeddings trained on wikipedia articles might help to improve precision. (Dernoncourt/Lee 2016)
- Due to huge gazetteers the speed of RegexNER could be improved by parallelizing our de-identification pipeline.
- NER models trained on wikipedia corpora perform well on corpora with narrower domains.(Nothman 2009, Balasuriya 2009)
- optimize regex based gazeteers, use context information to catch housenumbers, names, etc.

Acknowledgements

- Department of Computational Linguistics at Heidelberg University supervised by Stefan Riezler
- 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

References I

S. K. Yuwono, H. T. Ng, and K. Y. Ngiam.
Automated anonymization as spellingvariant detection.
Clinical Natural Language Processing Workshop (ClinicalNLP 2016), page 99, 2016.

J. Starlinger, M. Kittner, O. Blankenstein, U. Leser.
How to improve information extraction from German medical records.
it - Information Technology, page 171, 2017.

I. Schlunder.
Datenschutzkonforme Lösungen fur die Versorgungsforschung.
14. Deutscher Kongress fur Versorgungsforschung, 2015.

M. Farouqi, S. Padó.
Training and Evaluating a German Named Entity Recognizer with Semantic Generalization.
Proc. of KONVENS 2010.

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, P. Inc, S. J. Bethard, D. Mcclosky.
The Stanford CoreNLP Natural Language Processing Toolkit.
In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, page 55, 2014.

Ph. Richter-Pechanski.
https://github.com/MaviccPRP/ger_ner_evals, 2017.

R. A. Wagner, M. J. Fischer.
The String-to-String Correction Problem.
J. ACM, page 168, 1974.

References II

O. Uzuner, T. C. Sibanda, Y. Luo, P. Szolovits.
A de-identifier for medical discharge summaries.
Artificial Intelligence in Medicine, page 13, 2008.

M. Sokolova, N. Japkowicz.
Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation.
AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, page 1015, 2006.

O. Ferrandez, B. R. South.
BoB, a best-of-breed automated text de-identification system for VHA clinical documents.
Journal of the American Medical Informatics Association, page 77, 2013.

J. Y. Lee, F. Dernoncourt.
Feature-Augmented Neural Networks for Patient Note De-identification.
Computing Research Repository, 2016.

B. Wellner, M. Huyck, S. Mardis.
Rapidly retargetable approaches to de-identification in medical records.
Journal of the American Medical Informatics Association, page 564, 2007.