German Medical Data Sciences: A Learning Healthcare System
U. Hübner et al. (Eds.)

doi:10.3233/978-1-61499-896-9-165

De-Identification of German Medical Admission Notes

Phillip RICHTER-PECHANSKI[a,1], Stefan RIEZLER[a] and Christoph DIETERICH[b,1]
[a] *Department of Computational Linguistics, University of Heidelberg, Heidelberg, Germany*
[b] *Section of Bioinformatics and Systems Cardiology, Klaus Tschira Institute for Integrative Computational Cardiology and Department of Internal Medicine III, University Hospital Heidelberg, German Center for Cardiovascular Research (DZHK) - Partner site Heidelberg/Mannheim*

Abstract. Medical texts are a vast resource for medical and computational research. In contrast to newswire or wikipedia texts medical texts need to be de-identified before making them accessible to a wider NLP research community. We created a prototype for German medical text de-identification and named entity recognition using a three-step approach. First, we used well known rule-based models based on regular expressions and gazetteers, second we used a spelling variant detector based on Levenshtein distance, exploiting the fact that the medical texts contain semi-structured headers including sensible personal data, and third we trained a named entity recognition model on out of domain data to add statistical capabilities to our prototype. Using a baseline based on regular expressions and gazetteers we could improve F2-score from 78% to 85% for de-identification. Our prototype is a first step for further research on German medical text de-identification and could show that using spelling variant detection and out of domain trained statistical models can improve de-identification performance significantly.

Keywords. De-identification, anonymization, medical admission notes, named entity recognition, personal health information

1. Introduction

The task of de-identification of German medical texts is a precondition for any further research on these resources [1]. The main obstacle for research on German medical texts is the lack of shared medical corpora [1]. If medical texts are available, privacy issues make serious research a hard task. Medical texts can only be analyzed without legal restrictions if tokens containing personal information, e.g. personal health information (PHI)[2], had been deleted by de-identification [2].

The objective of this project was to build a de-identification tool based on named entity recognition (NER) and spelling variant detection, regular expressions and

[1] Corresponding Authors, Phillip Richter-Pechanski, Im Neuenheimer Feld 325, 69120 Heidelberg, Germany; E-mail: richter@cl.uni-heidelberg.de or Christoph Dieterich, Oudenarder Straße 16, Building D/04 (1st Floor), 13347 Berlin, Germany; E-mail: christoph.dieterich@uni-heidelberg.de.

[2] In this paper we use PHI specified in the Health Insurance Portability and Accountability Act. A set of eighteen named entities need to be de-identified to comply with this regulation. For further details see: https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act.

gazetteers. Our medical data consisted of medical admission notes as binary MS-DOC format. The data included a time period of 2004-2016. The whole un-annotated data set consisted of 180,000 notes including 132 million tokens in total. The notes contained texts from the cardiology domain and have the following basic structure: i) Header containing addressee, sender, patients name and address; ii) Salutation; iii) Diagnosis, anamnesis, etc. (as free texts, tables, images); iv) Summary.

Previous de-identification approaches mostly targeted English medical texts and were based on rules, gazetteers and/or machine learning models [3,5,6,7,8,9,12]. Most recently Yuwono *et al.* used spelling variant detection for de-identification of unstructured medical texts, based on structured PHI entities in a database [4]. Stubbs *et al.* 2015 extensively reviewed existing de-identification systems [10].

Due to the total lack of shared German medical data the use of supervised learning NLP tools is restricted as they need training data to work properly [1]. There are several commercial research activities on German medical texts, but currently no publications are available.[3]

2. Methods

2.1. Definitions

To analyze medical data the consent of the patient is the legal basis generally preferred by law. Due to the vast amount of data, this is typically not possible to fulfill [2]. If there is no individual consent, medical texts can only be analyzed without legal restrictions if the personal data is fully removed [2].

The task of removing personal data is described with several different terminologies.[4] In this task we use the term de-identification, which is the most general term. De-identification describes the process of removing the association between a set of identifying data and the data subject.[5]

2.2. Three-Step Approach

To benefit from well-established de-identification methods and to gain experience on statistical methods we are using a three-step approach for our task.

Spelling Variant Detection: The header mostly contains the name and address of the patient, the name and address of the recipients of the note and the contact information of the clinic. To parse the medical text for spelling variants we are using the approach of Yuwono *et al.*, who used the minimum edit distance measure to identify PHI tokens [4].

Named Entity Recognition: After investigating NER tools for German we identified the German Stanford NER as the best performing NER tool for the recognition of person (PER) and location (LOC) entities on out-of-domain data [11]. We trained the best model on a combination of three popular NER data sets.[6]

[3] Commercial vendors: Averbis and Statice: https://averbis.com, https://www.statice.io/.

[4] De-Identification, anonymization and pseudonymization are often used in this research area. All these terms are not consistently defined in different privacy laws. In summary anonymized data can never be re-identified, while pseudonymized data can only be re-identified with the use of external information.

[5] https://en.wikipedia.org/wiki/De-identification (ISO/TS 25237:2008).

[6] For further info see: https://github.com/MaviccPRP/ger_ner_evals.

Regular Expressions and Gazetteers: For the rule-based approach we combined plain gazetteers for first names, surnames and German cities and towns with gazetteers containing German street names with regular expressions.

In addition to the NE classes used by the NER model with RegexNER we are trying to recognize the following named entity classes: SALUTE (Herr; Frau); EMAIL (ex@xyz.com); PHONE (0123/3456); URI (www.xyz.de); DATE (01.02.1999); PLZ (12345); TITLE (Dr., Prof.).

2.3. Implementation

We implemented the methods using the Stanford Core NLP pipeline in a Java program including the preprocessing and de-identification steps. After preprocessing, parsing the header and annotating the medical text with a rule-based NER and a statistical NER model the algorithm starts the de-identification step. It replaces all spelling variants of tokens found in the header with its NE (e.g. <PER>, <LOC, etc…) label or <O> if it is not recognized as a NE by the NER models. If the token is not a spelling variant but annotated as a NE it will be replaced by its NE label, too. If the token is neither a spelling variant nor labeled with a NE class, the algorithm checks if the token contains location suffixes, pre-defined in a list. If the token contains a suffix it will eventually be de-identified and replaced by <LOC>. If the token is neither a NE, a spelling variant nor contains a suffix it will be directly written to the output.[7]

3. Results

Our baseline is the purely rule-based approach with Stanford RegexNER. We had an annotated test set of 15 medical admission notes with in total 14,134 tokens and 680 annotated named entities. They have been annotated with 10 pre-defined named entity tags. Table 1 shows the results for the binary de-identification task (PHI recognized vs. PHI not recognized). Table 2 shows the results for the NER task, the baseline, and for the full featured model.

Table 1. Evaluation of the binary PHI recognition task. Comparing all three models.

Model	Precision in %	Recall in %	F2-Score in %
Baseline	80	78	78
Baseline + Spelling Variant	79	85	83
Full Featured	71	89	85

Table 2. Evaluation of the NER task (B: Baseline, F: Full-Featured).

NE	Precision B/F	Recall B/F	F2-Score B/F
DATE	98/98	93/93	94/94
LOC	**93/66**	**51/56**	**56/57**
PER	**26/33**	**48/87**	**41/66**
PHONE	100/100	58/58	73/73
PLZ	88/88	100/100	97/97
SALUTE	100/100	98/98	98/98
TITLE	100/100	95/95	97/97

[7] To keep semantic value, we replace the PHI token using their NE class. The replacement mechanism can be adapted to specific needs, E.g. dates can be replaced my its year part, to track courses of time.

4. Discussion

Evaluating our de-identification task, ignoring the different named entity classes, table 1 shows that the spelling variant detector improved our baseline F2-score by 5%, while almost keeping precision. Our NER model decreases precision considerably, while improving recall to a lesser extent. There were 605 correct classified PHI tokens but still we got 251 non-PHI tokens de-identified and 75 PHI-tokens not de-identified.

Evaluation of the NER task showed a high F2-score for purely rule-based classes like PLZ, SALUTE, TITLE and DATE. The PER class improved F2-score by 25% whereas the LOC class made almost no improvement. A closer look at precision and recall in table 2 showed a huge improvement of recall for PER class by almost 40%, while precision could be improved by 6%. The LOC class lost 27% in precision, gaining 5% in recall. Due to ambiguities of entries in LOC and PER gazetteers (e.g. Schöneberg, Rostock, etc.) precision of LOC class was suffering.

To get a better understanding of the false negatives, we analyzed them more precisely. From 165 PER tokens there were ten false negatives, of which eight were one letter abbreviations of a first name and two not de-identificated surnames. From 104 LOC tokens, there were 35 false negatives, majority of which (29) belonged to the term ’Chest Pain Unit’.

The F2-score of the PHONE class was the lowest from the rule-based recognized entities. While precision was 100% recall was 58%. From 26 phone number tokens nine tokens were not recognized. All of them represented the last token of a phone number, not matched by the regular expression.

The precision of the PER class is comparably low, as we got several false positives. We got 165 PER entities in the gold standard but the full featured model classified 294 PER tokens. 68 false positives were ’Die’, because this token was listed in the gazetteer ‘German and international first names’. Further 32 false positives were ’Sehr’ (classified as PER by the statistical model). The same goes for the tokens ’Pain’ and ’Kollegin’.

To keep the focus of this paper on the technical aspects of de-identification, we used the PHI definition from the Health Insurance Portability and Accountability Act (HIPAA) which specifies more precisely which tokens need to be removed. While comparing “data concerning health” in European General Data Protection Regulation (GDPR) and PHI in HIPAA both terms are very similar.[8] However, PHI classes chosen in this paper do not necessarily comply with the GDPR.[9] Any de-identification task needs to be closely related to specific legal aspects and the intended use of the de-identified data.

5. Conclusion

Though there is a lack in German training data we could achieve satisfactory results for our de-identification and our NER task. While traditional rule-based approaches worked well for numerical and well-structured named entities like DATE, PLZ, SALUTE and TITLE, they achieve worse results for irregular entities like PHONE, LOC and PER. As

[8] For a more detailed comparison between GDPR and HIPAA see: https://iapp.org/news/a/gdpr-match-up-the-health-insurance-portability-and-accountability-act/.

[9] Our approach can be adapted to different legal conditions, as it is based on generic methods. In addition there are situations where full de-identification is not necessary [13].

there was no German corpus available for training a model on the PHONE class, we could use out-of-domain corpora to improve recall for the PER and LOC class. Still there are problems recognizing the LOC and ORG classes. The ORG class had to be excluded from our task, as medical texts contain several abbreviations and acronyms in capital letters, which confuse the out-of-domain trained model and led to very low F2-scores.

Besides the improvement of PHI and multiclass NER recognition scores the NER step added semantic information to our de-identified data, as additional PHI tokens could be replaced by semantically valuable NE labels.

Because of the lack of published research work on de-identification of German medical texts this project understands itself as a proof of concept. Further research needs to be done to improve recall while keeping precision high. As Uzuner *et al.* integrated a precision improvement step using support vector machines this could be done for our medical text set, too [7]. In addition Wellner *et al.* showed that already small annotated training sets can increase recall significantly [5]. Therefore we established an annotation project based on WebAnno[10] to encourage authorized medical researchers to annotate training and evaluation data. An encouraging research for the use of smaller training sets was done by Scheuerwegs who used 100 Dutch medical records for training and Liu *et al.* who trained a bidirectional LSTM model on a training set of 600 annotated mental health records. [8, 12]

6. Conflict of Interest

None.

References

[1] J. Starlinger, M. Kittner, O. Blankenstein, et al., How to improve information extraction from German medical records, *Information Technology* **59** (2017), 171-175.

[2] I. Schlunder, Datenschutzkonforme Lösungen fur die Versorgungsforschung, In *14. DKVF (2015).*

[3] L. Sweeney, Replacing personally-identifying information in medical records, the scrub system,. *Journal of the American Medical Informatics Association, (1996), 333-337.*

[4] S. Kester Yuwono, H.T. Ng, K.Y. Ngiam, Automated Anonymization as Spelling Variant Detection, *Proc. of the Clinical Natural Language Processing Workshop, ClinicalNLP* (2016), 99-103.

[5] B. Wellner, M. Huyck, S. Mardis, et al., Rapidly retargetable approaches to de-identification in medical records, *J Am Med Inform Assoc* **14** (2007), 564–573.

[6] J. Gardner, L. Xiong, HIDE: An integrated system for health info. de-identification, *CBMS* (2008), 254-259.

[7] O. Uzuner, T.C. Sibanda, Y. Luo, et al, A de-identifier for medical discharge summaries, *AI in Medicine* **42** (2008), 13-35.

[8] E. Scheurwegs, De-identification of clinical free text in Dutch with limited training data: A case study, *Proc. of the Workshop on NLP for Medicine and Biology associated with RANLP* (1993), 18-23.

[9] O. Ferrandez, B.R. South, S. Sheng, et al., BoB, a best-of-breed automated text de-identification system for VHA clinical document, *J Am Med Inform Assoc* **20** (2013), 77-83.

[10] A. Stubbs, C. Kotfila, Ö. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives, *J Biomed Inform* **58** (2015), 11-19.

[11] M. Faruqui, S. Padó, Training and Evaluating a German Named Entity Recognizer with Semantic Generalization, *Proc. of KONVENS* (2010), 129-133.

[12] Z. Liu, B. Tang, X. Wang, et al., De-identification of clinical notes via recurrent neural network and conditional random field, *J Biomed* Inform **75** (2017), 34-42.

[13] Schneider, I, U.K., Sekundärnutzung klinischer Daten - Rechtliche Rahmenbedingungen, Medizinisch Wissenschaftliche Verlagsgesellschaft, Berlin, 2015.

[10] https://webanno.github.io/webanno/.