MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190212

Annotating German Clinical Documents for De-Identification

Tobias Kolditz,[a] Christina Lohr,[a] Johannes Hellrich,[a] Luise Modersohn,[a]
Boris Betz,[b] Michael Kiehntopf,[b] Udo Hahn[a]

[a] Jena University Language & Information Engineering (JULIE) Lab, Friedrich Schiller University Jena, Jena, Germany
[b] Institute of Clinical Chemistry and Laboratory Diagnostics, Jena University Hospital, Jena, Germany

Abstract

We devised annotation guidelines for the de-identification of German clinical documents and assembled a corpus of 1,106 discharge summaries and transfer letters with 44K annotated protected health information (PHI) items. After three iteration rounds, our annotation team finally reached an inter-annotator agreement of 0.96 on the instance level and 0.97 on the token level of annotation (averaged pair-wise F1 score). To establish a baseline for automatic de-identification on our corpus, we trained a recurrent neural network (RNN) and achieved F1 scores greater than 0.9 on most major PHI categories.

Keywords:

Data Anonymization, Confidentiality, Natural Language Processing

Introduction

Natural Language Processing (NLP) research, in general, has greatly benefited from sharing software resources (tools) and datasets (corpora, lexical repositories, etc.), as well as organizing challenge competitions to evaluate this research infrastructure [1]. This role model has also been adopted by biomedical NLP [2], including clinical NLP [3]. The i2b2 series of competitions[1] mark the cornerstone for the development of numerous clinical datasets[2] and the offspring of specialized software solutions[3] that shape the current state-of-the-art in clinical NLP–yet, for the English clinical language only. Still, almost all of these (de-identified) datasets are accessible to the international research community under the legal conditions of Data Use Agreements (DUA) building on less restrictive data privacy law in the US than in many other parts of the world.

National privacy law in Europe and the General Data Protection Regulation (EU) 2016/679 pose a number of obstacles on the way to extramural sharing clinical corpora. Yet, the minimal prerequisite for all shared clinical corpora is de-identification, i.e., the removal of any patient-identifying information from clinical data. In Germany, one of the top representatives of a particularly restrictive data privacy legislation, the heavily funded Medical Informatics research Initative (MII)[4] was launched recently whose goal is, among others, to empower interoperability of and transparent access to data (including clinical document corpora) and software resources within and across different clinical sites at the national level, in conformance with valid data privacy regulations. Our team heads the NLP acitivities within one of the four funded consortia, *Smart Medical Information Technology for Healthcare* (SMITH)[5] [4]. One of the early outcomes of these efforts is 3000PA [5], a German-language clinical corpus made of 3,000 discharge summaries and transfer notes from three national university hospitals located in Jena, Leipzig and Aachen. Moreover, we are also involved in developing automated workflows for *Healthcare Integrated Biobanking* (STAKI^{2}B^2) where de-identification of clinical corpora is a prerequisite when it comes to workflow validation across different biobanks. In this paper, we report on the de-identification of a list of protected health information (PHI) categories reflecting US law criteria (Health Insurance Portability and Accountability Act (HIPAA)),[6] and their adaptation to national requirements and clinical particularities in German hospitals.

Related Work

Many systems for automatic de-identification have been developed, especially for the English language, but only a few large de-identification gold-standard corpora are reported in the literature–even fewer are publicly available, typically protected by DUAs. In this paper, we report on our efforts in creating a de-identification gold-standard corpus for the German language. To put this work in perspective, we focus on previous work on the creation of corpora annotated with PHI categories.

For clinical English, several corpora have been compiled and annotated for the purpose of de-identification. Building on manual de-identification work by Douglass et al. [6; 7], Neamatullah et al. [8] assembled an automatically de-identified gold-standard corpus of 2,434 nursing progress notes of patients from intensive care units collected in the MIMIC II project [9]. The final corpus consists of almost 340K tokens, contains 1,779 instances of PHI and is available on PhysioNet.[7]

Two different gold-standard corpora for de-identification were assembled for the i2b2 de-identification challenges in 2006 and 2014. For the first challenge [10], 889 medical discharge summaries consisting of almost 550K tokens were annotated semi-automatically, with almost 20K instances of PHI. For the second challenge [11], a corpus of 1,304 longitudinal medical records consisting of more than 800K tokens was annotated, with more than 28K PHI items. The authors report an averaged token-based F1 score of 0.927 between the annotators and the gold standard. For the CEGS N-GRID shared task on de-identification [12], a corpus of 1,000 psychiatric intake records

[1] https://www.i2b2.org/
[2] Accessible via http://dbmi.hms.harvard.edu/programs/health-care-data-science-program/clinical-nlp-research-data-sets
[3] Most notably, the cTAKES system: http://ctakes.apache.org/
[4] http://www.medizininformatik-initiative.de/
[5] http://www.smith.care/?lang=en
[6] https://www.hhs.gov/hipaa/index.html
[7] https://www.physionet.org/

consisting of more than 1,8M tokens was annotated, with more than 34K PHI instances. Token-level agreement of 0.91 F1 score between the annotators and the gold standard was slightly worse than for the i2b2 annotation campaign in 2014. After surrogate replacement, all three challenge corpora were made publicly available via DUAs. The 2006 i2b2 corpus (training set) and a carefully genre-balanced corpus from the Veterans Health Administration (800 documents, with 5.5K PHI items; test set) formed the basis for an "out-of-the-box" evaluation of five operational text de-identification systems [13].

Deléger et al. [14; 15] annotated 3,503 clinical notes of more than 22 different types (more than 1M tokens) with about 30K PHI. They report an inter-annotator agreement of 0.92 F1 score. In a later publication [16] they modify the corpus such that it can be shared for research purposes. Similar annotated corpora (albeit not always publicly available) exist for a few other languages, such as French [17] and Swedish [18; 19]; synthetic (fictional) data were created for Japanese [20].

Unfortunately, no shareable de-identified corpus exists for the German language–in general, there has been only little work on the de-identification of German clinical data. Seuss et al. evaluate a commercial de-identification tool on German clinical reports [21] and to this end assemble a large semi-automatically annotated corpus. They manually correct the output of the tool on 1,400 documents consisting of 5M tokens, identifying more than 23K PHI items (in nine categories pre-defined by the tool), and report high recall values for the system. However, the manual annotation process is tightly coupled with the tool and there is no evaluation of inter-annotator agreement between the three human annotators. Recently, Richter-Pechanski et al. [22] presented a de-identification tool that combines an off-the-shelf named entity recognizer (German Stanford NER tool, retrained on non-medical data), regular expressions and gazetteers with spelling variation detection. For evaluation purposes, they annotated a set of 15 notes (~ 14K tokens) with ten different categories, identifying 680 PHI entities overall.

Manual Annotation

In this section, we describe in detail the manual annotation campaign we conducted: our data, annotation guidelines, evaluation measures, the rationale for several annotation iterations, and the final annotated corpus. All annotations were created using the *Brat Rapid Annotation Tool* (BRAT) [23].

Data

Annotations were based on the Jena slice of the 3000PA corpus [5], a collection of 1,106 discharge summaries, short summaries and transfer letters. All documents were extracted from EPRs of deceased patients (due to privacy concerns) who were treated in either internistic or ICU units for at least five days between 2010 and 2015. The extraction included a conversion from a proprietary data format to plain text (for details see [24]). Our work on this data was approved by the local ethics committee (4639-12/15) and the Data Protection Officer of the Jena University Hospital.

Annotation Guidelines

Our annotation guidelines are based on the 18 PHI categories defined in HIPAA and subsequent manual and automatic annotations of these PHI types, most notably as part of the i2b2 de-identification challenges [10-12; 25]. We iteratively developed and updated our guidelines in three preliminary annotation rounds, adjusting them to our data, common clinical requirements and the particularities of the German language. The final guidelines define eight broad categories, three of which are further divided into more specific annotation types:

- **Age:** age of patient or relative (any age)
- **Contact**: URL, IP address, email, phone or fax number
- **Date:** any date (excluding single days of the week and times of the day)
  - **Birthdate:** date of birth
- **ID:** any ID or code (patient id, medical record number, codes with unknown semantics that might be PHI)
  - **Typist:** shorthand symbols for medical typists
- **Location:** place names (mainly addresses, including street, house number, zip code, city, district, state)
- **MedicalUnit:** medical units (hospital names, names of hospital departments and ambulant medical units)
- **Person:** names of persons
  - **Patient:** patient names
  - **Relative:** names of relatives
  - **Staff:** names of medical staff
- **Other:** any remaining PHI

Each category is also an annotation type by default; thus, we come up with 13 annotation types. Annotators were required to select only the most specific applicable type for each entity. For instance, if they found the name of a physician, they were instructed to assign it the type *Staff*, not *Person*. Only if the context was unclear about whether a name belongs to a patient, a relative or a member of staff, or if the person denoted by a name belongs to neither of these three groups, the annotators were instructed to assign it the generic *Person* type.

Our scheme collapses several of the HIPAA-defined types of PHI because they are infrequent or indistinguishable in our data: *ID* comprises social security, medical record, account, certificate/license, patient-related serial numbers, and any numbers or codes that may directly or indirectly identify a patient or member of staff. Since we work with plain texts, we dropped the category for photographic images. On the other hand, we introduced some more fine-grained distinctions, e.g., between names of patients, relatives and staff. The *Typist* category was introduced after the second pre-iteration (see below) to resolve systematic disagreement among annotators.

Evaluation Measures

In the following, we report inter-annotator agreement (IAA) as pair-wise averaged F1 score [26], both on instance level and on token level. An instance is a single annotation, possibly spanning multiple tokens, e.g., "*Jane Smith"* denotes one instance of a patient's name comprised of two tokens. Tokens are based on the tokenization script of the TreeTagger[8] with minor post-processing. To obtain token-based agreement, we annotate each token overlapping with an annotation with the annotation's type, sub-token annotations are expanded to full tokens. Due to missing spaces, we got two annotation types per token in 13 cases–we kept both annotation types for each of these tokens when calculating agreement, but used only one as input to the classifier described in the second part of this paper.

Annotation Iterations

We ran three preliminary annotation iterations (summarized in Table 1) before we started the final annotation project. Iterated annotation rounds for building de-identification gold standard corpora have also been suggested by Browne et al. [27] as a strategy to cope with the complexities of this task that are

[8] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

meticulously described in their paper. The number of annotations in this table includes redundant annotations from different annotators on the same documents. In the first iteration, eight medical students and two physicians worked on the full set of 1,106 documents, covering all protected health information with generic annotations according to a first set of guidelines based on the list of PHI items defined in HIPAA and the i2b2 de-identification challenges [11; 25]. We automatically pre-annotated dates with a regular expression covering standard numerical date representations. Annotators at this stage could focus on finding PHI items in the text and were not supposed to categorize them. These first annotations were intended to serve as pre-annotations for later iterations (ensuring that each document was checked for false negatives at least twice) and formed the empirical basis for a refined set of annotation guidelines. After a qualitative analysis of the results, we further specified the extent of each annotation type, defined a list of 12 types of PHI (the types listed in the previous section, excluding *Typist*) and added illustrative examples for each category.

Table 1 – Annotation iteration setups and results

Annotation	Pre-1	Pre-2	Pre-3	Final
# documents	1,106	25	12	1,106
# agr. docs	0	25	12	50
# annotators	10	4	4	5
# ann. types	1	12	13	13
# annotations	40,664	4,347	2,556	51,814
Avg. F1 (token)	-	0.91	0.96	0.97
Avg. F1 (inst.)	-	0.87	0.92	0.96

Using simple patterns and keywords, the generic annotations from the first iteration were automatically categorized according to the new guidelines. These data served as pre-annotations for the second iteration in which four medical students each worked on the same set of 25 pre-annotated documents. Their task was to identify missing PHI items and correct both extent and type of the existing annotations. They achieved an inter-annotator agreement of 0.87 ($\sigma = 0.02$) on instance level and 0.91 ($\sigma = 0.02$) on token level.

To further train our annotators and achieve even higher agreement values, a third pre-iteration was conducted with four annotators (one of the previous annotators had to leave the team temporarily and was replaced by a new annotator). We defined an additional category, *Typist*, for shorthand symbols identifying medical typists to resolve the ambiguity between *ID* and *Person*[9] and added more examples to the annotation guidelines to resolve cases of systematic disagreement among annotators. Each annotator was provided with six documents with annotations he or she had created in the previous iteration[10] and six documents with semi-automatically generated pre-annotations. Inter-annotator agreement improved on both the old and the new documents. For the full set of documents, the averaged pair-wise F1 score reached 0.92 ($\sigma = 0.04$) on instance level and 0.96 ($\sigma = 0.02$) on token level.

For the main annotation, five annotators worked on the 1,106 Jena documents of 3000PA. A subset of 50 documents (cf. Table 1, '# agr. docs') was annotated by all five annotators and served as the basis for the computation of IAA which improved further to 0.97 ($\sigma = 0.01$) on token level and 0.96 ($\sigma = 0.01$) on instance level. We also noticed an increase in IAA within the final annotation hinting towards a training effect over time.

Final Corpus

The final corpus consists of roughly 1.4M tokens for which our annotators created more than 44K annotations, excluding redundant agreement annotations.[11] Table 2 gives an overview of instance and token frequencies by annotation type for the entire corpus, as well as instance- and token-level agreement by annotation type on 50 multiply annotated agreement documents. Looking at the instance frequencies, we notice that almost half of the PHI items in our corpus are dates, followed by medical units, locations and staff names, each accounting for more than ten percent of all annotations. The large share of dates is partially a consequence of our broad definition of *Date* which only excludes days of the week without month or year and times of the day. At the other end of the frequency spectrum, we find the generic types *Person* and *Relative*. The fact that only few instances fall under the *Person* category indicates that most names could be assigned unambiguously to either *Patient, Relative* or *Staff*. The low frequency of *Relative* is probably an artifact of our data sampling criteria–most of the deceased patients were elderly people for which family history is less important, hence names of relatives occur rarely. The frequency distribution of tokens mirrors that of instances and gives information about the average length of each annotation type. The share of *Date*s is lower as they usually consist of only one token, whereas instances of the types *Location*, *MedicalUnit* and *Staff* often span multiple tokens, raising their relative frequency at the token level.

Table 2 – Number of instances and tokens per annotation type in the final corpus. Instance-level and token-level agreement (Avg. F1) on 50 multiply annotated agreement documents.

Category	Type	Instances		Tokens	
		Frequency	Avg. F1	Frequency	Avg. F1
Age	Age	498	1.00	500	1.00
Contact	Contact	613	0.97	2,009	0.98
Date	Date	20,603	0.98	24,277	0.99
	Birthdate	1,103	1.00	1,103	1.00
ID	ID	398	0.81	424	0.82
	Typist	655	0.86	1,418	0.93
Location	Location	5,429	0.98	11,286	0.99
MedicalUnit	MedicalUnit	6,189	0.90	12,499	0.95
Person	Person	14	-	23	-
	Patient	3,180	0.99	5,167	1.00
	Relative	36	0.80	62	0.88
	Staff	5,231	0.95	10,003	0.97
Other	Other	218	0.28	271	0.26
Total	*	44,167	0.96	69,042	0.97

[9] Shorthand symbols or initials denote persons but are different from typical person names.

[10] The fourth annotator received 12 new documents.

[11] Of all annotations created for the 50 multiply annotated agreement documents, the final corpus only contains those created by the annotator who achieved the highest IAA.

Instance and token-level agreement lie between 0.8 and 1 for most annotation types. However, for *Relative, Other*, and *ID* each annotator found 11 or less instances on average. These values thus cannot be considered reliable. *Person* occurred 14 times in the whole corpus, none of these occurrences are in the 50 agreement documents. Low agreement for *Other* does not come as a surprise (independent of IAA reliability problems), as annotators were supposed to assign this category whenever they found information that they considered confidential, but could not match with any other annotation type. This vague definition led to a number of varying subjective judgements. For many annotation types (especially *MedicalUnit, Typist,* and *Staff*), token-level agreement is higher than instance-level agreement, indicating that there is disagreement about the exact extent of some annotations: partially overlapping annotations raise agreement on token level, but not on instance level. We also observe cases where one annotator decided to create one long annotation while another one created two short ones covering the same tokens. For the task of de-identification, the latter kind of disagreement is less severe, as we are mainly interested in covering all sensitive information.

Automatic De-Identification

In order to provide a first baseline for automatic de-identification on our annotated data and to submit the quality of our annotations to an empirical test, we trained and evaluated a neural network for named entity recognition (NER) designed by Sterbak [28] on the final corpus.

Methods

We randomly sampled 80% of the documents for training and used the remaining 20% as test set. Splitting the corpus on documents rather than sentences ensures a representative share of out-of-vocabulary (OOV) tokens. We provide only tokens and characters as input. The network employs a unidirectional LSTM [29] to learn word representations based on character embeddings. These character-based representations are concatenated with word embeddings (all embeddings were initialized randomly). The resulting complex word representations serve as input to a bidirectional LSTM. At each timestep, the output of the bidirectional LSTM is fed into a densely connected layer of units with a softmax activation function. Each unit of the last layer represents one possible classification result (the annotation types and an additional outside tag, *O*, for non-PHI data). The softmax function provides a probability distribution over these possible results. For each token, we predict the type with maximal probability.

To obtain reliable results despite randomly initialized weights, we train five different models on the same train-test split. We compute standard NER evaluation measures for each model and report averages over all five models. Additionally, we determine the percentage of OOV tokens for an annotation type *T*, that is, tokens that occur in the test set labeled as *T,* but do not occur in the training set with the same annotation type *T*.

Results

Table 3 yields the classifier's precision, recall and F1 score on the test set, as well as the support (the number of expected occurrences) and the percentage of OOV tokens for each annotation type and the outside tag (*O*). Totals are averages over all types weighted by support. For most annotation types, we achieve F1 scores between 0.88 and 0.98. 14% of all PHI tokens are OOV, whereas only 4% of all tokens are OOV. Disregarding confusions between different PHI types and only considering the binary distinction between PHI and non-PHI, the classifier achieves a precision of 0.98 and a recall of 0.96.

Table 3 – Automatic de-identification results on the test set: token-level precision, recall, and F1 score, support and percentage of out-of-vocabulary tokens per annotation type.

Type	Prec	Rec	F1	Supp	OOV
Age	0.885	0.910	0.897	89	25%
Birthdate	0.959	0.992	0.975	222	68%
Contact	0.966	0.945	0.956	435	7%
Date	0.959	0.970	0.964	4,969	15%
ID	0.693	0.385	0.480	95	91%
Location	0.977	0.940	0.958	2,188	11%
MedicalUnit	0.967	0.938	0.952	2,424	4%
O	0.998	0.999	0.999	269,337	3%
Other	0.592	0.384	0.463	63	46%
Patient	0.952	0.962	0.957	990	35%
Person	0.000	0.000	0.000	5	80%
Relative	0.000	0.000	0.000	16	100%
Staff	0.975	0.961	0.968	2,015	7%
Typist	0.898	0.862	0.879	245	12%
Total	0.996	0.996	0.996	283,093	4%
Total w/o O	0.959	0.947	0.952	13,756	14%

Discussion

We observe a clear drop in performance for smaller classes, especially *ID*, *Other, Person*, and *Relative*. The latter two together contribute only 85 tokens in the whole corpus, 21 of which are in the test set. So, we cannot expect the network to learn these rather difficult types. The confusion matrix reveals that, on average, around 43 out of 95 *ID* tokens were misclassified as *Date*s which suggests that the internal representations learned by the network are rather shallow–combinations of digits and punctuation marks may be *ID*s or *Date*s depending on the context–and to a large extent explains the low recall for *ID*. Furthermore, we find that low IAA scores are reflected in the classifier's performance for *Other* and *ID.* The overall F1 score (without the outside tag *O*) is slightly below the token-level results reported by Stubbs et al. [25] where the best system achieves an F1 score of 0.961 for the i2b2-PHI categories and 0.976 for the HIPAA-PHI categories. However, this comparison has to be taken with a grain of salt since both our annotation types and the data are different. The percentage of OOV tokens in the last column of Table 3 gives an idea about the recall a lazy learner might achieve by just remembering tokens occurring in the training set. Since our dataset stems from two units of a single hospital, the percentage of OOV tokens for *MedicalUnit* and *Staff* is low which puts the results for these types into perspective. For *Patient*, we observe a high percentage of OOV tokens, but for this type of PHI we sometimes find clear contextual cues (e.g., German equivalents of *Mr.* or *Ms./Mrs.*) which facilitate automatic recognition.

Conclusions

We annotated 13 PHI categories on the Jena slice of the German-language 3000PA corpus (1,106 documents) based on annotation guidelines that evolved over three annotation pre-iterations. The annotation process ensured that each document was checked twice for PHI and high inter-annotator agreement scores suggest that the resulting annotations are of very good quality. The neural baseline model we trained on the final annotated corpus achieved promising results, indicating that the annotation types defined in our guidelines lend themselves well to a data-driven approach to automatic de-identification. Future work will be directed at incorporating the 1,000 document

slices from the other two clinical sites, Leipzig and Aachen. With these more heterogeneous data, the classifier will be further tuned to maximize recall while ensuring generalizability to provide reasonable input for the planned pseudonymization engine as a resilient basis for the DUA-controlled distribution of a large corpus of German clinical documents.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the SMITH project under grant 01ZZ1803G and the Deutsche Forschungsgemeinschaft (DFG) under grants HA 2079/8-1 and KI 564/2-1 within the STAKI^{2}B^2 project.

References

[1] M. Nissim, L. Abzianidze, K. Evang, R. van der Goot, H. Haagsma, B. Plank, and M. Wieling, Sharing is caring: the future of shared tasks, *Comput Linguistics* **43** (2017), 897-904.

[2] C.-C. Huang and Z. Lu, Community challenges in biomedical text mining over 10 years: success, failure and the future, *Brief Bioinform* **17** (2016), 132-144.

[3] W.W. Chapman, P.M. Nadkarni, L. Hirschman, L.W. D'Avolio, G.K. Savova, and Ö. Uzuner, Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions, *JAMIA* **18** (2011), 540–543.

[4] A. Winter *et al.*, Smart Medical Information Technology for Healthcare (SMITH). Data integration based on interoperability standards, *Methods Inf Med* **57** (2018), e92-e105.

[5] U. Hahn, F. Matthies, C. Lohr, and M. Löffler, 3000PA: towards a national reference corpus of German clinical language, in: *MIE 2018*, Gothenburg, Sweden, 2018, pp. 26-30.

[6] M. Douglass, G. Clifford, A. Reisner, G. Moody, and R. Mark, Computer-assisted de-identification of free text in the MIMIC II database, in: *Comput Cardiol 2004*, Chicago, IL, USA, 2004, pp. 341-344.

[7] M. Douglass, G. Cliffford, A. Reisner, W. Long, G. Moody, and R. Mark, De-identification algorithm for free-text nursing notes, in: *Comput Cardiol 2005*, Lyon, France, 2005, pp. 331-334.

[8] I. Neamatullah, M. Douglass, L.-w. Lehman, A. Reisner, M. Villarroel, W. Long, P. Szolovits, G. Moody, R. Mark, and G. Clifford, Automated de-identification of free-text medical records, *BMC Med Inform Decis Mak* **8** (2008), 32-47.

[9] M. Saeed, M. Villarroel, A. Reisner, G. Clifford, L.-w. Lehman, G. Moody, T. Heldt, T. Kyaw, G. Moody, and R. Mark, Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II): a public-access intensive care unit database, *Crit Care Med* **39** (2011), 952-960.

[10] Ö. Uzuner, Y. Luo, and P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, *JAMIA* **14** (2007), 550-563.

[11] A. Stubbs and Ö. Uzuner, Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UT-Health corpus, *J Biomed Inform* **58 Suppl** (2015), S20-S29.

[12] A. Stubbs, M. Filannino, and Ö. Uzuner, De-identifica-tion of psychiatric intake records: overview of 2016 CEGS N-GRID Shared Tasks Track 1, *J Biomed Inform* **75 Suppl** (2017), S4-S18.

[13] Ó. Ferrández, B. South, S. Shen, F. Friedlin, M. Samore, and S. Meystre, Evaluating current automatic de-identification methods with Veteran's Health Administration clinical docu-ments, *BMC Med Res Methodol* **12** (2012), #109.

[14] L. Deléger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser, L. Stoutenborough, and I. Solti, Large-scale evaluation of automated clinical note de-identification and its impact on information extraction, *JAMIA* **20** (2013), 84-94.

[15] L. Deléger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, and I. Solti, Building gold standard corpora for medical natural language processing tasks, in: *AMIA 2012*, Chicago, IL, USA, 2012, pp. 144-153.

[16] L. Deleger, T. Lingren, Y. Ni, M. Kaiser, L. Stoutenborough, K. Marsolo, M. Kouril, K. Molnar, and I. Solti, Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research, *J Biomed Inform* **50** (2014), 173-183.

[17] C. Grouin and A. Névéol, De-identification of clinical notes in French: towards a protocol for reference corpus development, *J Biomed Inform* **50** (2014), 151-161.

[18] S. Velupillai, H. Dalianis, M. Hassel, and G.H. Nilsson, Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial, *Int J Med Inform* **78** (2009), e19-e26.

[19] H. Dalianis and S. Velupillai, De-identifying Swedish clinical text: refinement of a gold standard and experiments with conditional random fields, *J Biomed Semant* **1** (2010), #6.

[20] M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, and E. Aramaki, Overview of the NTCIR-10 MedNLP Task, in: *NTCIR-10*, Tokyo, Japan, 2013, pp. 696-701.

[21] H. Seuss, P. Dankerl, M. Ihle, A. Grandjean, R. Ham-mon, N. Kaestle, P.A. Fasching, C. Maier, J. Christoph, M. Sedlmayr, M. Uder, A. Cavallaro, and M. Hammon, Semi-automated de-identification of German content sensitive reports for big data analytics, *Röfo* **189** (2017), 661-671.

[22] P. Richter-Pechanski, S. Riezler, and C. Dieterich, De-identification of German medical admission notes, in: *GMDS 2018*, Osnabrück, Germany, 2018, pp. 165-169.

[23] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Anania-dou, and J.i. Tsujii, BRAT : a web-based tool for NLP-assisted text annotation, in: *EACL 2012: Demonstrations*, Avignon, France, 2012, pp. 102–107.

[24] C. Lohr, S. Luther, F. Matthies, L. Modersohn, D. Ammon, K. Saleh, A. Henkel, M. Kiehntopf, and U. Hahn, CDA-compliant section annotation of German-language discharge summaries: guideline development, annotation campaign, section classification, in: *AMIA 2018*, San Francisco, CA, USA, 2018, pp. 770-779.

[25] A. Stubbs, C. Kotfila, and Ö. Uzuner, Automated sys-tems for the de-identification of longitudinal clinical narra-tives: overview of 2014 i2b2/UTHealth Shared Task Track 1, *J Biomed Inform* **58 Suppl** (2015), S11-S19.

[26] G. Hripcsak and A.S. Rothschild, Agreement, the F-measure, and reliability in information retrieval, *JAMIA* **12** (2005), 296-298.

[27] A. Browne, M. Kayaalp, Z. Dodd, P. Sagan, and C. McDonald, The challenges of creating a gold standard for de-identification research, in: *AMIA 2014*, Washington, D.C., USA, 2014, pp. 353-358.

[28] T. Sterbak, Enhancing LSTMs with character embed-dings for named entity recognition, in: *Depends on the defini-tion, URL: https://www.depends-on-the-definition.com/lstm-with-char-embeddings-for-ner/*, Last Access: Nov 3, 2018.

[29] S. Hochreiter and H.J. Schmidhuber, Long short-term memory, *Neural computation* **9** (1997), 1735-1780.

Address for correspondence

Tobias Kolditz, Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Fürstengraben 27, 07743 Jena. Email: tobias.kolditz@uni-jena.de