A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-26

3000PA—Towards a National Reference Corpus of German Clinical Language

Udo HAHN[a,1] Franz MATTHIES[a] Christina LOHR[a] Markus LÖFFLER[b]
[a] *Jena University Language & Information Engineering (JULIE) Lab*
Friedrich-Schiller-Universität Jena, Germany
http://www.julielab.de
{forename.surname}@uni-jena.de
[b] *Institute for Medical Informatics, Statistics and Epidemiology (IMISE)*
Universität Leipzig, Germany
{forename.surname}@imise.uni-leipzig.de

Abstract. We introduce 3000PA, a clinical document corpus composed of 3,000 EPRs from three different clinical sites, which will serve as the backbone of a national reference language resource for German clinical NLP. We outline its design principles, results from a medication annotation campaign and the evaluation of a first medication information extraction prototype using a subset of 3000PA.

Keywords. German language, clinical text corpus, medication information extraction

1. Introduction

Clinical reports and free text entries in electronic patient records (EPR) are a rich source of medical information with high added value for clinical decision making [1]. Yet, this knowledge is mostly verbally encoded, thus making it hard to be harvested automatically. Fortunately, biomedical natural language processing (NLP) technology has matured significantly over the years and, hence, unstructured knowledge in verbal form can be extracted from clinical narratives for the benefit of patients [2].

Despite the huge amounts of raw text stored in clinical information systems, these textual resources are by no means easily accessible for another reason. Strict legal rules for the protection of patients' data privacy prohibit the transfer of clinical documents from the hospital to external sites, e.g. NLP labs. For instance, the workflow developed for the i2b2 challenge competitions [3] to share clinical documents in conformance with legal rules—*1)* complete pseudonymization of textual occurrences of 18 well-defined Protected Health Information categories (for a complete list, cf. [3]), *2)* approval of the de-identified clinical documents by institutional review boards, and *3)* release of anonymized data on the basis of Data Use Agreements (DUA)—is effective within the context of the Anglo-American legal culture only. (National) European law is much more restrictive, so that, up until now, only very few non-English corpora have been released. This situation has a massive negative effect on improving clinical NLP tools, since

[1] Corresponding Author: Udo Hahn, Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Fürstengraben 30, 07443 Jena, Germany, E-mail: udo.hahn@uni-jena.de

sharable, open-source language resources—corpora and software—are at the heart of NLP research as they play a pivotal role for performance testing and classifier training.

A recently started large-scale national funding initiative in Germany[2] aims at changing this situation fundamentally by the concerted effort of researchers from clinical medicine, medical informatics, and major IT companies. The SMITH consortium,[3] one of the four major players in this initiative, incorporates lead members from the universities and university hospitals in Leipzig (UKL), Jena (UKJ) and Aachen (UKA). As a first result of the collaboration under the SMITH umbrella, the 3000PA corpus has been set up. It will serve as a backbone for a national reference corpus of German language clinical documents to be made accessible on an on-demand basis via Data Integration Centers that act as trustful information brokers for all kinds of service requests.

2. Related Work

In the US, the past decade has seen a series of clinically oriented NLP shared tasks. Prominent examples are the "TREC Precision Medicine / Clinical Decision Support Track" as part of the "Text Retrieval Conference" (TREC),[4] the NLP branch of the "Integrating Biology and the Bedside" (i2b2) initiative,[5] and, since 2015, "Clinical TempEval",[6] a challenge campaign mainly organized by NLP researchers with a focus on temporal orderings of clinical events. From these activities, de-identified and semantically annotated clinical corpora have emerged which can easily be acquired, in a de-identified form, by signing a DUA. Accordingly, for clinical NLP with focus on the English language, there are plenty of resources available. For the non-English language communities, however, less comfortable conditions prevail. Only very few EU countries follow the DUA policy, such as reported for a clinical adverse drug reaction corpus for Spanish [4] or a comprehensive Dutch clinical corpus [5]. Some labs working on non-English languages have announced plans for releasing their resources, e.g., for French [6], Polish [7] or Swedish [8]. Apparently, these plans have not been fully realized as, to the best of our knowledge, none of these corpora is currently DUA-available for the research community. Another source for medical language resources in Europe derives from the "CLEF eHealth" initiative.[7] Established in 2013, this series of health-related challenges led to the preparation of several corpora—mostly for English, but also for other European languages. However, they are typically very small and only available for usage directly related to the respective task, i.e., neither usable later on nor available for the research community independent of the specific CLEF task.

For German-language medical corpora the situation is even worse: *all* clinical corpora are *only* available for the research staff within the lifetime of a project and remain inaccessible forever for the outside world. We here briefly mention those which contain at least 300 clinical documents. FRAMED [9], the first published German-language

[2] A brief description of the 120M EUR endowed funding initiative "Medizininformatik" is available at https://www.bmbf.de/de/medizininformatik-3342.html

[3] The goals of the SMITH consortium (Smart Medical Information Technology for Healthcare) are described at http://www.smith.care/

[4] http://www.trec-cds.org/

[5] https://www.i2b2.org/NLP/DataSets/Main.php

[6] http://alt.qcri.org/semeval2017/task12/

[7] https://sites.google.com/site/clefehealth/

medical corpus ever, consists of a mixture of roughly 300 clinical reports, textbook fragments and consumer-related health texts annotated with low-level linguistic metadata (up to the level of parts of speech). In 2012, [10] assembled a corpus of 544 clinical reports from various medical domains (e.g., echocardiography, EEG, lung function, X-ray thorax) for an information extraction (IE) task. From a clinical data warehouse with roughly 70,000 clinical reports, [11] selected 660 de-identified transthoracic echocardiography reports for IE. In 2016, [12] collected 450 surgery reports to build language models adapted to metadata from two German medical thesauri. In the same year, [13] developed an annotation schema for the nephrology domain using 1,725 discharge summaries and clinical notes. A collection of 1,696 de-identified clinical in- and outpatient discharge summaries were assembled from a dermatology department for an unsupervised abbreviation detection procedure [14] and supervised machine learning using an SVM for abbreviation and sentence delineation [15]. Recently, [16] mention a corpus composed of 3,000 chest X-ray reports used for term extraction to support IE.

3. 3000PA Corpus — Corpus Design and Annotation of Medication Information

In a consortium-wide effort within SMITH, we requested from each clinical site involved (UKA, UKJ, UKL) EPRs of deceased patients (for data privacy reasons) for a six-year cohort (2010-2015) treated in either internistic or ICU units for at least 5 days. We then sampled roughly 1,000 clinical documents (mostly, discharge summaries) from these EPRs per site and so created the 3000PA corpus with approximately 3,000 clinical documents. The multi-site composition policy is unique since, up until now, corpora were assembled from single hospital sites only.

After collection, the documents from 3000PA were manually annotated for medication information. This topic had already been investigated in the Third i2b2 Challenge for English clinical documents [17] and we replicated this study for German clinical language using 3000PA. Annotation guidelines were formulated by iteratively adapting the i2b2 instructions for the English language [18], to the German clinical language. Just as for English, our scheme covers *medication* (drugs) experienced by the patient, *dosage* (the amount of a particular drug given to the patient), *mode* (the way the drug was administered), *frequency* (how often each dose of the medication was given), *duration* (over which period of time the medication was given) and the medical *reason* for which the medication was given. At each of the three local SMITH sites, annotation teams were formed (five students of medicine at UKJ, two and one documentation officer/s each at UKL and UKA, respectively) supervised by the Annotation Management Team in Jena. Due to legal restrictions only the staff at a respective site was allowed to engage in the annotation process of their local documents.

Given the time constraints of the pilot, we were able to annotate 960/850/550 discharge summaries from UKJ/UK/UKA, respectively, altogether 2,360 (from a maximum of 3,000) clinical narratives using the BRAT annotation tool.[8] Annotation quality could only be measured for UKJ where we managed to multi-code 52 documents by all five annotators. We used the elastic centroid approach for matching [19], the F1-score metric for assessment (depending on centroid matching criteria) and computed inter-annotator agreement (IAA) values for this document set (cf. Table 1). IAAs ranged in the (higher) nineties for *medication*, *dosage*, and *frequency*, in the lower eighties down

[8] http://brat.nlplab.org

to the sixties for *mode* and (lower) seventies for *duration*, while they plummeted to the forties for *reason*. These tendencies were amplified when the matching criteria were chosen increasingly selectively (see the lower three rows in Table 1).

Table 1. Overview of average Inter-annotator Agreement (IAA) of five annotators μ_{IAA} (highlighted in bold) (with standard deviation σ_{IAA}) and the performance of JUMEx in terms of F1 scores (highlighted in bold), Precision (P) and Recall (R) (t describes the threshold and b the boundary of the centroid algorithm [19]).

		Medication	Dosage	Frequency	Mode	Duration	Reason
		F1 (P/R)	F1 (P/R)	F1 (P/R)	F1 (P/R)	F1 (P/R)	F1 (P/R)
t = 2, b = 0	μ_{IAA}	**.93** (.95/.92)	**.95** (.95/.94)	**.95** (.95/.96)	**.80** (.83/.78)	**.70** (.71/.70)	**.52** (.51/.54)
	σ_{IAA}	.02 (.02/.03)	.02 (.01/.02)	.03 (.03/.03)	.07 (.07/.08)	.04 (.05/.06)	.12 (.11/.15)
	JUMEx	**.64** (.76/.55)	**.85** (.83/.86)	**.81** (.87/.76)	**.55** (.58/.51)	**.38** (.36/.41)	–
t = 3, b = 2	μ_{IAA}	**.93** (.92/.94)	**.96** (.96/.97)	**.95** (.93/.96)	**.81** (.77/.85)	**.74** (.68/.82)	**.39** (.33/.53)
	σ_{IAA}	.03 (.02/.03)	.02 (.02/.02)	.03 (.03/.03)	.06 (.07/.07)	.04 (.04/.06)	.07 (.09/.09)
	JUMEx	**.65** (.76/.56)	**.84** (.81/.87)	**.82** (.87/.77)	**.56** (.57/.56)	**.41** (.35/.49)	–
t = 4, b = 3	μ_{IAA}	**.88** (.82/.95)	**.91** (.87/.96)	**.92** (.88/.96)	**.59** (.49/.73)	**.64** (.52/.83)	**.28** (.20/.60)
	σ_{IAA}	.01 (.01/.03)	.01 (.01/.02)	.02 (.02/.02)	.02 (.04/.03)	.03 (.05/.04)	.09 (.09/.08)
	JUMEx	**.65** (.75/.58)	**.83** (.80/.87)	**.82** (.87/.77)	**.58** (.55/.62)	**.40** (.33/.50)	–

4. Medication Information Extraction from the 3000PA Corpus

The semantic metadata from 3000PA were taken to build a pilot system that automatically extracts medication information from this corpus and to evaluate its performance. The first prototype of the medication extractor we developed, JUMEx, is based on the MedXN system [20]. Its rule base was adapted to German, exploiting "Rote Liste"[9] as a task-specific terminological resource for the German language. The F1 scores for JUMEx are also depicted in Table 1. We achieved good coverage for *frequency* and *dosage* (in the eighties), mediocre quality for *medication* (in the mid-sixties) and *mode* (in the upper fifties), and rather low performance for *duration* (in the forties). The attribute *reason* (for administering the medication) was too complex to be adequately covered by the current version of JUMEx. Note that these results, although they comprise the first ones published for German clinical documents, are not really competitive because of the limited time we could spend on prototype development (just one week).

5. Conclusions

We briefly described the activities to set up 3000PA, a preliminary version of the first national reference corpus for German clinical documents composed of roughly 3,000 clinical reports from three different German hospitals. To demonstrate its usability, we annotated around 2,400 of these documents at all sites for medication information (involving six attributes per medication statement). We then set up JUMEx, a preliminary medication extraction prototype for German. This endeavour replicated work for the German language that had originally been conducted for English in the i2b2 Challenge. Future work will not only focus on extending 3000PA, but also enriching it by additional clinically relevant metadata (entities such as diseases, diagnoses, therapies).

[9] https://www.rote-liste.de/

Acknowledgements. This work has been funded by the German *Bundesministerium für Bildung und Forschung* (BMBF) under grant no. 01ZZ1803G.

References

[1] E Ford, JA Carroll, HE Smith, DS Scott, and JA Cassell (2016). Extracting information from the text of electronic medical records to improve case detection: A systematic review. *JAMIA*, **23**(5):1007–1015.

[2] K Kreimeyer, M Foster, A Pandey, N Arya, G Halford, SF Jones, R Forshee, M Walderhaug, and T Botsis (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *JBI*, **73**:14-29.

[3] A Stubbs and Ö Uzuner (2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth Corpus. *JBI*, **58**(Suppl):S20–S29.

[4] M Oronoz, K Gojenola, A Pérez, A Diaz de Ilarraza, and A Casillas (2015). On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *JBI*, **56**:318-332.

[5] Z Afzal, E Pons, N Kang, M Sturkenboom, M Schuemie, and J Kors (2014). ContextD: An algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC Bioinformatics*, 15:#373.

[6] L Deléger, AL Ligozat, C Grouin, P Zweigenbaum, and A Névéol (2014). Annotation of specialized corpora using a comprehensive entity and relation scheme. *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 1267–1274.

[7] M Marciniak and A Mykowiecka (2014). Towards morphologically annotated corpus of hospital discharge reports in Polish. *BioNLP 2011 — Proceedings of the Workshop on Biomedical Natural Language Processing @ ACL-HLT 2011*, pp. 92–100.

[8] H Dalianis, M Hassel, and S Velupillai (2009). The Stockholm EPR Corpus: Characteristics and some initial findings. *ISHIMR 2009 – Evaluation and Implementation of e-Health and Health Information Initiatives. Proceedings 14th Intl. Symposium for Health Information Management Research*, pp. 243–249.

[9] J Wermter and U Hahn (2004). An annotated German-language medical text corpus as language resource. *LREC 2004 — Proceedings 4th Intl. Conference on Language Resources and Evaluation*, pp. 473–476.

[10] G Fette, M Ertl, A Wörner, P Klügl, S Störk, and F Puppe (2012). Information extraction from unstructured electronic health records and integration into a data warehouse. *Informatik 2012 — Proceedings 42. Jahrestagung der Gesellschaft für Informatik (GI)*, pp. 1237–1251.

[11] M Töpfer, H Corovic, G Fette, P Klügl, S Störk, and F Puppe (2015). Fine-grained information extraction from German transthoracic echocardiography reports. *BMC Med Inform Decis Mak*, 15:#91.

[12] C Lohr and R Herms (2016). A corpus of German clinical reports for ICD and OPS-based language modeling. *CLAW 2016 — Proceedings of the 6th Workshop on Controlled Language Applications @ LREC 2016*, pp. 20–23.

[13] R Roller, H Uszkoreit, F Xu, L Seiffe, M Mikhailov, O Staeck, K Budde, F Halleck, and D Schmidt (2016). A fine-grained corpus annotation schema of German nephrology records. *ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016*, pp. 69–77.

[14] M Kreuzthaler, M Oleynik, A Avian, and S Schulz (2016). Unsupervised abbreviation detection in clinical narratives. *ClinicalNLP 2016 — Proceedings of the Clinical Natural Language Processing Workshop @ COLING 2016*, pp. 91–98.

[15] M Kreuzthaler and S Schulz (2015). Detection of sentence boundaries and abbreviations in clinical narratives. *BMC Med Inform Decis Mak*, 15(Suppl 2):S4.

[16] J Krebs, H Corovic, G Dietrich, M Ertl, G Fette, M Kaspar, M Krug, S Störk, and F Puppe (2017). Semi-automatic terminology generation for information extraction from German chest X-ray reports. *German Medical Data Sciences. Proceedings of the 62nd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (gmds)*, pp. 80–84 (*Stud Health Technol Inform*, 243).

[17] Ö Uzuner, I Solti, and E Cadag (2010). Extracting medication information from clinical text. *JAMIA*, **17**(5):514–518.

[18] Ö Uzuner, I Solti, F Xia, and E Cadag (2010). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *JAMIA*, **17**(5):519–523.

[19] I Lewin, Ş Kafkas, and D Rebholz-Schuhmann (2012). Centroids: Gold standards with distributional variation. *LREC 2012 — Proceedings 8th Intl. Conf. on Language Resources & Evaluation*, 3894-3900.

[20] S Sohn, C Clark, SR Halgrim, SP Murphy, CG Chute, and H Liu (2014). MEDXN: An open source medication extraction and normalization tool for clinical text. *JAMIA*, **21**(5):858–865.