Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-734

Sharing Models and Tools for Processing German Clinical Texts

Johannes HELLRICH,[1] Franz MATTHIES, Erik FAESSLER and Udo HAHN
Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena, Germany

Abstract. The automatic processing of non-English clinical documents is massively hampered by the lack of publicly available medical language resources for training, testing and evaluating NLP components. We suggest sharing statistical models derived from access-protected clinical documents as a reasonable substitute and provide solutions for sentence splitting, tokenization and POS tagging of German clinical texts. These three components were trained on the confidential FRAMED corpus, a non-sharable collection of various German-language clinical document types. The models derived therefrom outperform alternative components from OPENNLP and the Stanford POS tagger, also trained on FRAMED.

Keywords. Natural Language Processing, Germany

Introduction

There is a stunning demand for medical natural language processing (NLP) solutions, as expressed in reports from specialized workshops [1,2] or concluded from the outcomes of clinical software challenges like I2B2 [3]. Yet, almost all of these activities and the vast majority of existing medical NLP resources or tools are available for the English language only, as evidenced by the paramount role of the CTAKES [4] initiative and its associated, impressively rich software distribution platform.[2] Clearly, this imbalance does not reflect the language diversity in the pan-European healthcare landscape. Unfortunately, this situation is not going to change rapidly despite some efforts in the Scandinavian countries (see, e.g., [5,6]), in France (see, e.g., [7]), in the Netherlands (see e.g., [8]) or in Germany (see, e.g., [9,10]).

These observations are even more striking, as we witness another marked disproportion between the wide coverage of *general-language* resources and tools for various European languages (mostly based on newspaper documents and other publicly accessible data), on the one hand, and the much more limited, often fragmentary or even entirely lacking infrastructure usable for sublanguage-specific *clinical* NLP in these languages, on the other hand. One main reason for the obvious resource poverty lies in the entirely different "data culture" one encounters in clinical settings. A particular obstacle for continuous progress are legal concerns to fully protect the privacy of the actors involved in clinical activities (patients, clinical staff, hospitals, etc.). These

[1] Corresponding Author.
[2] http://ctakes.apache.org/

extremely restrictive data security regulations imposed on all sorts of clinical data and documents, despite complete anonymization of actor-sensitive data items, create a massive resource bottleneck for clinical NLP. Even if such resources have been created and curated at some local site, transmural sharability is usually blocked by prohibitive legal non-disclosure commitments. Thus, one of the fundamental axioms of modern, empirical NLP—reuse and share existing resources and tools—is broken. Another alternative, namely compromising with solutions originating from other (mostly general-language) domains in the biomedical domain, has already been shown to yield substantial performance drops, at least when applied to the biomedical domain [6,11].

As a way out of this dilemma, we here advocate to *share models* derived from non-disclosable corpora as a *substitute for legally locked raw data*. This approach has recently been suggested as a safe way to collaborate in building complex clinical NLP pipelines [12]. We will illustrate our idea with the JULIE Lab UIMA Component Repository (JCoRe) [13],[3] which we extend here with JPos, a newly developed part-of-speech (POS) tagger. We also augment the existing collection of German models for processing medical text [12], now covering sentence splitting, tokenization and POS tagging. All models were trained on the confidential FraMed mixed-genre medical corpus [14].

1. FraMed—A Confidential German-Language Clinical Corpus

The FraMed corpus consists of a mixture of medical document types such as discharge summaries, pathology reports and medical textbook excerpts, all in German language. It contains 7,000 sentences, with approximately 100,000 tokens, and was manually annotated for sentence boundaries, token segmentation and POS tags. POS annotation in FraMed was carried out with a variant of the Stuttgart-Tübingen-Tagset (STTS) [15] extended by three domain specific tags, namely LATIN (Latin nominatives or genitives in medical terms), ENUM (enumerations) and FDSREF (Reference patterns w.r.t. formal document structure, e.g. *'as described under 2.'*). The LATIN tag is of special importance, as Latin words are part of syntactic constructions uncommon in German (e.g. post-coordinated adjectives) and could thus impede further processing [10].

2. JCoRe and JPos—Core of a Sharable Clinical NLP Pipeline for German

The JCoRe components utilize Maximum Entropy (ME) and Conditional Random Field (CRF) models for machine learning. Accordingly, data are modeled by means of weighted feature functions where the specific feature instances are derived from training data and weights are chosen to fit the model to the data. The feature functions used in these approaches capture information about the actual data to base the classification decision on, including *n*-gram information. Thus, the de-identification of confidential medical data (cf. e.g. [16]) is of vital importance—all person names, dates and addresses in FraMed have been blinded to preclude the identification of individual patients.

[3] http://www.julielab.de/Resources.html

JCoRe was originally built as a repository of interoperable UIMA[4] components adapted to the specific needs of the analysis of English life sciences literature. It contains self-developed components to account for special phenomena of the biomedical domain, like the notoriously difficult tokenization of protein, gene or chemical names. The three tools described in this paper, a sentence splitter, a tokenizer and a POS tagger, can also be used directly from the command line.

The newly added German JPos tagger utilizes ME models (as implemented in Mallet)[5] for POS tagging. Its features include token suffixes, regular expression-derived token classes (e.g., tokens containing numbers or Greek letters), stemming and *n*-grams of configurable length. Several of its features can be parameterized through a configuration file, e.g., the length of the context window[6] or the n-grams. In various pre-tests a system using a 1-token context window and all other features available in JPos performed best and was thus used for the experiments described in the following section. The resulting models do therefore contain at most bi- and trigram information, which can be assumed to be insufficient to reconstruct original sentences, as they are not specific enough (an important issue for public model sharing).

3. Comparative Evaluation of the NLP Pipeline and the POS Tagger

We here provide a comparison of the sentence splitting, tokenization and POS tagging components of JCoRe with those from the Apache openNLP[7] collection and the Stanford POS Tagger [17].[8] All tools are widely used and written in Java; all POS taggers are based on an ME approach. Both alternative external POS taggers provide special support for German (which we used during our experiments); the Stanford tagger was trained with its 'bidirectional' switch activated. The significance of JPos' relative performance in comparison with the other POS taggers was ensured by performing pairwise *t*-tests on the results for analogous slices during 10-fold cross-validation (p-values < 0.05).

Evaluation results on FraMed are reported in Table 1, sentence splitting and tokenization were already evaluated in more detail in [12]. The specialized JCoRe components and the new JPos tool outperform alternative tools for all three tasks, e.g., JPos decreases the number of wrong POS tags by 23% in comparison to openNLP.

Table 1. Comparison of openNLP and JCoRe components for sentence splitting, tokenization and POS tagging and the Stanford POS tagger by 10-fold cross-validation on FraMed. Sentence splitting and tokenization results are reported by *F*1-Score, POS tagging by accuracy. Cells with a dash (for the Stanford tools) indicate lacking support for this task.

task	openNLP	JCoRe	Stanford
sentence splitting	0.968	**0.994**	—
tokenization	0.995	**0.996**	—
POS tagging	0.968	**0.976**	0.963

[4] https://uima.apache.org/
[5] http://mallet.cs.umass.edu/
[6] We here mean text, tag and other features of adjacent tokens. This causes models to be rather large, at a rate hundreds of MB.
[7] https://opennlp.apache.org/
[8] http://nlp.stanford.edu/software/tagger.shtml

The performance of the three POS taggers was also evaluated on two additional corpora, the German TIGER corpus [18] (newspaper domain) and the English GENIA corpus [19] (biomedical domain), as shown in Table 2. For the TIGER corpus, JPOS performed clearly superior, with 31% fewer wrong POS tags than the Stanford tagger that was ahead in other studies [20]. For the GENIA corpus, all tested tools fall behind the performance figures published for the GENIA tagger [21], which peaks at an accuracy of 0.983. JPOS seems to be under-adapted to English biomedical texts, coming in last, but seems to do well with German texts from both the medical and the newspaper domain.

Table 2. Comparison of OPENNLP, JCORE and Stanford POS taggers by 10-fold cross-validation on the TIGER and GENIA corpora, reported by accuracy.

corpus	OPENNLP	JCORE	Stanford
TIGER	0.969	**0.977**	0.968
GENIA	**0.981**	0.977	0.980

4. Conclusions

Given the lack of publicly available non-English language resources for clinical NLP and its negative effect on the development of specialized tools, we advocate the sharing of tools developed on confidential corpora as a way to sidestep access restrictions for these resources. We demonstrate this idea by providing components from our JCORE repository and accompanying models trained on FRAMED, a confidential German-language clinical corpus. Our solution comprises components for sentence segmentation, tokenization and POS tagging. All components were shown to be far more accurate for annotating German medical texts than general-purpose tools (e.g., OPENNLP) trained on the same corpus.

An obvious extension of our work would include the training (and if necessary adaptation) of more complex NLP components already contained in JCORE, for usage with German-language clinical texts. Yet, currently, FRAMED does not contain any annotation layer above the POS level. There are plans to enrich it with named entity information, which could then be shared as models to help other groups working in this domain. Another important issue is how to combine such external models trained on hidden corpora with existing in-house models or corpora, e.g. by forming ensembles out of isolated components.

References

[1] C. Friedman, T.C. Rindflesch, M. Corn. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *J Biomed Inform*, 46(5):765–773, 2013.

[2] V.M. Pai, M. Rodgers, R. Conroy, J. Luo, R. Zhou, B. Seto. Workshop on using natural language processing applications for enhancing clinical decision making: an executive summary. *J Am Med Inform Assoc*, 21(e1):e2–e5, 2014.

[3] W.W. Chapman, P.M. Nadkarni, L. Hirschman, L.W. D'Avolio, G.K. Savova, Ö. Uzuner. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*, 18(5):540–543, 2011.

[4] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–513, 2010.

[5] M. Skeppstedt, M. Kvist, G.H. Nilsson, H. Dalianis. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *J Biomed Inform*, 49:148–158, 2014.

[6] V. Laippala, T. Viljanen, A. Airola, J. Kanerva, S. Salanterä, T. Salakoski, F. Ginter. Statistical parsing of varieties of clinical Finnish. *Artif Intell Med*, 61(3):131–136, 2014.

[7] L. Deléger, C. Grouin, P. Zweigenbaum. Extracting medication information from French clinical texts. In *MedInfo 2010—Proceedings of the 13th World Congress on Medical Informatics*, pages 949–953, Cape Town, South Africa, September 12-15, 2010.

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

[9] U. Hahn, M. Romacker, S. Schulz. medSynDiKATe: a natural language system for the extraction of medical information from findings reports. *Int J Med Inform*, 67(1–3):63–74, 2002.

[10] H.U. Krieger, C. Spurk, H. Uszkoreit, F. Xu, Y. Zhang, F. Müller, T. Tolxdorff. Information extraction from German patient records via hybrid parsing and relation extraction strategies. In *LREC 2014—Proceedings of the 9th Language Resources and Evaluation Conference*, pages 2043–2048, Reykjavik, Iceland, May 26-31, 2014.

[11] K. Tomanek, J. Wermter, U. Hahn. A reappraisal of sentence and token splitting for life sciences documents. In *MedInfo 2007—Proceedings of the 12th World Congress on Health (Medical) Informatics*, pages 524–528, Brisbane, Australia, August 20-24, 2007.

[12] E. Faessler, J. Hellrich, U. Hahn. Disclose models, hide the data: how to make use of confidential corpora without seeing sensitive raw data. In *LREC 2014—Proceedings of the 9th Language Resources and Evaluation Conference*, pages 4230–4237, Reykjavik, Iceland, May 26-31, 2014.

[13] U. Hahn, E. Buyko, R. Landefeld, M. Mühlhausen, M. Poprat, K. Tomanek, J. Wermter. An overview of JCoRe, the JULIE Lab UIMA component repository. In *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP Workshop @ LREC'08*, pages 1–7, Marrakech, Morocco, May 31, 2008.

[14] J. Wermter, U. Hahn. An annotated German-language medical text corpus as language resource. In *LREC 2004—Proceedings of the 4th International Conference on Language Resources and Evaluation*, volume 2, pages 473–476, Lisbon, Portugal, May 24-30, 2004.

[15] A. Schiller, S. Teufel, C. Stöckert, C. Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart und Seminar für Sprachwissenschaft, Universität Tübingen, 1999.

[16] S.M. Meystre, Ó. Ferrández, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore. Text de-identification for privacy protection: a study of its impact on clinical text information content. *J Biomed Inform*, 50:142–150, 2014.

[17] K. Toutanova, D. Klein, C.D. Manning, Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL 2003—Proceedings of the 2003 Human Language Technology Conference and the 3rd Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259, Edmonton, Canada, May 27 - June 1, 2003.

[18] S. Brants, S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, H. Uszkoreit. TIGER: linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620, 2004.

[19] J.D. Kim, T. Ohta, Y. Tateisi, J. Tsujii. GENIA corpus: a semantically annotated corpus for bio-text-mining. *Bioinformatics*, 19(Suppl 1):i180–i182, 2003.

[20] E. Giesbrecht, S. Evert. Part-of-speech tagging: a solved task? An evaluation of POS taggers for the Web as corpus. In *WAC5—Proceedings of the 5th 'Web as Corpus' Workshop*, pages 27–35, Donostia-San Sebastián, Basque Country, Spain, September 7, 2009.

[21] Y. Tsuruoka, J. Tsujii. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *HLT/EMNLP 2005—Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*, pages 467–474, Vancouver, B.C., Canada, October 6-8, 2005.