mEx - An Information Extraction Platform for German Medical Text

Roland Roller, Christoph Alt, Laura Seiffe, and He Wang

German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany
firstname.lastname@dfki.de

Abstract. In recent years, clinical text processing gained a lot of attention. Easing the access to information for medical personnel, combined with the ability to track and forecast a patients development, makes structured information extraction from medical text sources a crucial component. Due to a specialized domain language, existing tools trained on different domains might not yield the desired performance. Clinical data is highly sensitive and therefore a scarce resource. When focusing on non-English languages, the situation is even worse. Besides the limited language resources, hardly any tools are freely available to process clinical text. To address this limitation we present mEx, an Information Extraction system for German medical text. While specialized on the nephrology domain, mEx is powered by generic components that can be adopted to any medical domain. We provide mEx as an online tool that demonstrates various functionalities to process medical text. It can be tested via web front-end or accessed via REST API.

Keywords: Clinical NLP · Information Extraction · Machine Learning

1 Introduction

Progress in Information Extraction from non-English clinical text is hampered by limited access to relevant data sets, mostly due to legal reasons, and the unavailability of tools specialized on this domain. As a result, research groups interested in non-English text have to constantly reinvent the wheel, in many cases starting with the creation of a corpus. This is an extremely time consuming process, yet necessary to be able to train, evaluate, and compare the performance of own models. Starting from scratch also means focusing on the elementary aspects of Information Extraction first. Due to already existing tools (e.g. cTAKES [14], NegEx [1]) and pre-processed data sets (e.g. the corpora provided by the 'Integrating Biology and the Bedside' (i2b2) initiative[1]), researchers focusing on English text are able to address complex problems more quickly.

Conversely for other languages, such as German, this situation looks much worse. Even though various medical datasets exist, none of them is freely accessible up to this point [9]. Considering existing tools the situation is slightly better.

[1] https://www.i2b2.org/NLP/DataSets/

Overall a Part-of-Speech (POS) tagger with integrated tokenizer and sentence splitter (JPOS) [5], some negation detection implementations ([3], [4]), an abbreviation detection system [7], as well as a dependency tree parser [6] exist to process German clinical text.

The state of the development of tools and corpora addressing other European languages is comparable to the situation for German. Sandoval et al. [13] develop a freely accessible online tool for enhancing biomedical related NLP which is based on Spanish, Japanese and Arabic text. The Conference and Labs of the Evaluation Forum (CLEF) eHealth challenge provides datasets also in languages other than English, for example the 2017 challenge for French medical data [10].

In this work we present mEx, an Information Extraction platform for German medical text. mEx unifies multiple NLP components and visualizes processed input text. Beside the testing character of the demo, functionalities can be also requested via REST API. In the following the different NLP components will be briefly introduced, followed by an overview of our demo. The mEx system can be accessed via *http://biomedical.dfki.de/mEx*.

2 Information Extraction Components

mEx combines various NLP (natural language processing) components into a pipeline to extract information from German medical text, particularly of the nephrology domain. We adopt Spacy (http://spacy.io) and its concept of language processing pipelines to implement the components, simplifying adoption and reuse of mEx in different medical domains. This section briefly introduces the components.

Pre-Processing The pre-processing step combines text normalization (e.g. unicode normalization), tokenization, and sentence splitting. Currently, the component uses Spacy's German tokenizer and sentence splitter. Both parts will be replaced by improved versions to address the specific structure of medical text, such as shortened sentences and frequent abbreviations.

Part-of-Speech Tagger The POS tagger assigns each token a part-of-speech (e.g. verb, adjective, or noun). The current demo system integrates the Jena Part-of-Speech Tagger (jPOS) [5] which uses a slightly modified version of the STTS tagset.

Dependency Tree Parser The dependency tree parser infers the syntactic structure of a sentence. In our demo we integrate a re-trained dependency parser optimized for German clinical text [6] which bases on the implementation of Chen and Manning [2]. Our dependency parser takes data from jPOS as input and produces parse trees using the UD tagset.

Named Entity Recognition Named entity recognition (NER) detects mentions of pre-defined entities in text, such as drug mentions, body parts, or diseases. However, in our case we consider the recognition of concepts in general, which

also include more abstract ideas, such as detailed medical specifications (e.g. *chronische Niereninsuffizienz*, 'chronic kidney disorder'), diagnostic/laboratory procedures, and treatments. Our NER component utilizes a Bi-LSTM with CRF [8], similarly as described in Roller et al. [12].

Factuality Detection Negations and vague descriptions are a vital part of clinical documentation, as doctors often speculate on the presence of diseases, or discard certain conditions with high probability. In order to process extracted information correctly, its factuality must be taken into account. NegEx [1] is a simple yet popular approach to detect negations and speculations (hedges) in clinical text in context of diseases. mEx integrates a re-implemented python module, based on the work of Cotik et al. [3] and is applied to *medical conditions*.

Relation Extraction A relation describes a particular relationship between concepts or entities, such as a *medical condition* occurs in a particular *body part* or a *dosing*, which connects a *measurement* with its corresponding *drug*. The mEx relation extraction (RE) component bases on a convolutional neural network (CNN), as described in Roller et al. [12]. Both, the NER and RE component were trained on a large data set of nephrological clinical notes and discharge summaries.

Concept Normalization In medical documentation, different entity mentions can refer to the same concept, e.g. 'pain in the head' and 'cephalia' both refer to the concept of 'headache'. mEx employs a two-step concept normalization against the Unified Medical Language System (UMLS) [11]. First, a cross-lingual candidate search is performed for each entity mention found by NER. If the search on the German subset of UMLS yields no result, the English subset is queried, followed by a search on the English subset with a translated entity mention. In the second step, ambiguous terms are scored by the densest sub-graph algorithm computed on UMLS co-occurrences [16].

3 A Medical Information Extraction Platform

Figure 1 presents the main part of the mEx GUI. In the upper part of the center (1) is the text input field located. Here the given input text can by typed in and adjusted. On the left side (2) various synthetic German clinical notes from the nephrology domain are presented. Clicking on one of those documents will show the content in the text input field. On the right side of the window (3) a control panel including the different functionalities are presented and can be modified. The control panel includes *NER*, *RE*, *Negation Detection* and *Normalisation*.

Using the buttons in the center of the GUI (4) the semantic and syntactic text processing can be triggered. While *semantics* includes the different information extraction components on the right side, *syntax* applies POS tagging and dependency tree parsing. The processed results are then presented in the lower part of the center (5) using the brat annotator [15]. Beside some labels and relations, the demo visualises additional information within a popup of each label, such as negations or UMLS candidates.

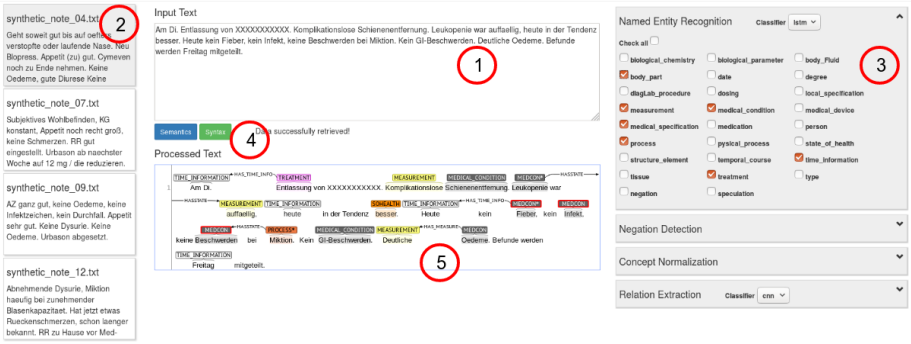

Fig. 1. mEx Graphical User Interface (*http://biomedical.dfki.de/mEx*)

4 Conclusion

We presented mEx, a medical Information Extraction platform for German. The system includes multiple NLP components, such as NER, RE or dependency parsing and is specialized for the nephrology domain. mEx can be accessed via web browser as well as via REST API (details will follow on the webpage). Considering the fact that hardly any tool exist to process German medical text, mEx could be used as a baseline or to compare your own developed models. Moreover, the platform can be easily updated so that additional components and models be integrated in future versions.

Acknowledgements

This research was supported by the German Federal Ministry of Economics and Energy (BMWi) through the project MACSS (01MD16011F).

References

1. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics **34**(5), 301–310 (2001)
2. Chen, D., Manning, C.: A Fast and Accurate Dependency Parser using Neural Networks. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 740–750. Association for Computational Linguistics, Doha, Qatar (2014)
3. Cotik, V., Roller, R., Xu, F., Uszkoreit, H., Budde, K., Schmidt, D.: Negation Detection in Clinical Reports Written in German. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016). pp. 115–124. The COLING 2016 Organizing Committee, Osaka, Japan (December 2016)
4. Gros, O., Stede, M.: Determining Negation Scope in German and English Medical Diagnoses. In: Taboada, M., Trnavac, R. (eds.) Nonveridicality and Evaluation: Theoretical, Computational and Corpus Approaches, pp. 113–126. Interactive Factory, Leiden, Netherlands (2014)

5. Hellrich, J., Matthies, F., Faessler, E., Hahn, U.: Sharing Models and Tools for Processing German Clinical Texts. MIE 2015 - Digital Healthcare Empowering Europeans **210**, 734–738 (2015)
6. Kara, E., Zeen, T., Gabryszak, A., Budde, K., Schmidt, D., Roller, R.: A Domain-adapted Dependency Parser for German Clinical Text. In: Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018). Vienna, Austria (September 2018)
7. Kreuzthaler, M., Oleynik, M., Avian, A., Schulz, S.: Unsupervised Abbreviation Detection in Clinical Narratives. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP). pp. 91–98. The COLING 2016 Organizing Committee, Osaka, Japan (December 2016)
8. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. In: Proceedings of NAACL-HLT. pp. 260–270 (June 2016)
9. Lohr, C., Buechel, S., Hahn, U.: Sharing Copies of Synthetic Clinical Corpora without Physical Distribution - A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. pp. 1259–1266. LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan (May 2018)
10. Névéol, A., Robert, A., Anderson, R., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Rondet, C., Zweigenbaum, P.: CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French. In: Working Notes for CLEF 2017 Conference. Dublin, Ireland (September 23-26 2017)
11. Roller, R., Kittner, M., Weissenborn, D., Leser, U.: Cross-lingual Candidate Search for Biomedical Concept Normalization. In: Proceedings of Multilingual BIO. Miyazaki, Japan (May 2018)
12. Roller, R., Rethmeier, N., Thomas, P., Hübner, M., Uszkoreit, H., Staeck, O., Budde, K., Halleck, F., Schmidt, D.: Detecting Named Entities and Relations in German Clinical Reports. In: Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology. German Society for Computational Linguistics and Language Technology, Berlin, Germany (September 2017)
13. Sandoval, A.M., Llanos, L.C., Herrero, C., Zorita, J.M.G.M., Martínez, A.G., Samy, D., Takamori, E.: An Online Tool for Enhancing NLP of a Biomedical Corpus. In: Input a Word, Analyze the World, pp. 25–38. Cambridge Scholars Publishing (2006)
14. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): rchitecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5), 507–513 (2010)
15. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations Session at EACL 2012. Association for Computational Linguistics, Avignon, France (April 2012)
16. Weissenborn, D., Roller, R., Xu, F., Uszkoreit, H., Perez, E.G.: A Light-weight & Robust System for Clinical Concept Disambiguation. In: Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine, SMBM. pp. 85–89. Potsdam, Germany (August 2016)