Health Informatics Meets eHealth
G. Schreier et al. (Eds.)

doi:10.3233/978-1-61499-645-3-71

Extraction of UMLS® Concepts Using Apache cTAKES™ for German Language

Matthias BECKER[a,1] and Britta BÖCKMANN[a,b]
[a] *Department of Medical Informatics, University of Applied Sciences and Arts, Dortmund, Germany*
[b] *IMIBE, University Hospital Essen, Germany*

Abstract. Automatic information extraction of medical concepts and classification with semantic standards from medical reports is useful for standardization and for clinical research. This paper presents an approach for an UMLS concept extraction with a customized natural language processing pipeline for German clinical notes using Apache cTAKES. The objectives are, to test the natural language processing tool for German language if it is suitable to identify UMLS concepts and map these with SNOMED-CT. The German UMLS database and German OpenNLP models extended the natural language processing pipeline, so the pipeline can normalize to domain ontologies such as SNOMED-CT using the German concepts. For testing, the ShARe/CLEF eHealth 2013 training dataset translated into German was used. The implemented algorithms are tested with a set of 199 German reports, obtaining a result of average 0.36 F1 measure without German stemming, pre- and post-processing of the reports.

Keywords. Natural Language Processing, Information Extraction, Data Mining, Medical records, UMLS.

1. Introduction

Almost all healthcare institutions use Medical Information Systems. These systems tend to replace medical documentation on paper [1]. So often diagnoses, symptoms and medications are recorded and documented in a structured manner. However, a variety of clinical data is still stored in different unstructured formats like text and images. The development of new tools for intelligent IT-based text analysis can make it possible to access this knowledge and use it - for the electronic patient record itself or even for research purposes. Most of the available tools and also terminologies are in English language and cannot be used directly for German text.

There are various approaches internationally. One promising approach is machine learning natural language processing (NLP) [2] with information extracting techniques [3]. Extracting information from clinical notes has been the focus of a growing body of research these past years. There are existing solutions such as an advanced Text Mining pipeline based on Apache UIMA for German language by Averbis [4] and extensive researches on this topic by ID Berlin [5] and Semfinder [6]. This approach is consistent to the mentioned approaches with the difference of the focus on the identification of SNOMED CT codes using open source technology. The Challenges of this approach are to combine and adapt existing English standards to the German language and to evaluate the results without an existing German gold standard corpus.

[1] Corresponding Author: Matthias Becker, University of Applied Sciences and Arts, Dortmund, Emil-Figge-Str. 42 44227 Dortmund, Germany, Matthias.Becker@fh-dortmund.de

One of the fundamental tasks in clinical natural language processing research is to extract clinically relevant entities (e.g. diseases and drugs) using semantic standards such as Concept Unique Identifier (CUI) defined in the Unified Medical Language System (UMLS) [7] and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [8]. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world [9], but it is not available in German language.

Therefore, the idea is to use the English SNOMED-CT by combining it with UMLS, which is available in German language. By extracting the UMLS concepts it is possible, to map this concepts to other non-German speaking classifications and ontologies [10]. To extract the concepts out of German clinical notes, a natural language processing pipeline for the German language is needed.

One of the most proven natural language processing tools is the open source natural language processing system for extraction of information from electronic medical record clinical free-text Apache cTAKES. Apache cTAKES already offers a variety of algorithms for text analysis and information extraction. This system was deployed at the Mayo Clinic and is currently an integral part of their clinical data management infrastructure and has processed over 80 million clinical notes [11]. It can normalize to domain ontologies such as SNOMED-CT using UMLS concepts. The aim of this project is to adjust an open source natural language processing pipeline to the German language to extract the concepts out of German clinical notes to map them with SNOMED-CT.

2. Methods

2.1. Architecture Overview

For the extraction of UMLS concepts from German clinical notes, a natural language processing pipeline with a mapping to the UMLS database is necessary. Once the concepts are identified, it is possible to map to domain ontologies and international terminologies. So Apache cTAKES was adapted to German clinical notes. Figure 1 shows the overview architecture of the system.

To configure the natural language processing pipeline for the objectives, various German OpenNLP models and German UMLS database were integrated into the cTAKES database. The cTAKES standard pipeline *AggregatePlaintextFastUMLSProcessor* was applied to extract the German concepts from the unstructured German clinical notes. For the mapping of the concepts to SNOMED-CT codes, the *ctakessnorx* database of cTAKES was expanded.

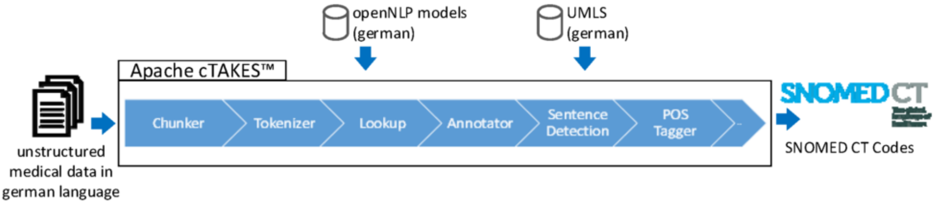

Figure 1. Architecture Overview for German language

2.2. Dataset and Evaluation

In 2013, the ShARe/CLEF eHealth Evaluation Lab (SHEL) organized three shared tasks on natural language processing and information retrieval (IR): 1) clinical disorder extraction and encoding to SNOMED-CT, 2) acronym/abbreviation identification, and 3) retrieval of web pages based on queries generated when reading the clinical reports. To test and evaluate the system, a ShARe/CLEF eHealth 2013 shared task 1 training set of 199 notes was used. Table 1 shows the amount of concepts within the training dataset which is used as gold standard.

Using these notes the pipeline was evaluated with English as well as German clinical notes to compare both results with the gold standard. The 199 English notes have been translated into German using the Google Translator initially [12]. The translated notes were then processed manually. The translated notes were analyzed by the same pipeline, but with an integrated German UMLS database and German OpenNLP models.

Table 1. Statistics of the training dataset:

Type	#Note	#CUI
ALL	199	2798
DISCHARGE	61	1969
ECHO	42	479
RADIOLOGY	42	257
ECG	54	93

The evaluation of the adapted pipeline follows the standard metrics of evaluation for the task using F1 (Equ. 3), i.e. the harmonic mean of recall (Equ. 2) and precision (Equ. 1). This is the same metric used by participants of the ShARE/CLEF 2013 Task 1.

$$P = \frac{true\ positive}{true\ positive + false\ positive} \tag{1}$$

$$R = \frac{true\ positive}{true\ positive + false\ negative} \tag{2}$$

$$F1 = \frac{2PR}{(P+R)} \tag{3}$$

Precision (P) is the number of correct positive results divided by the number of all positive results, and Recall (R) is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.

2.3. Semantic Standards

The Unified Medical Language System (UMLS) integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records. For the natural language processing pipeline, the English and German concepts from the UMLS database were used. The existing mapping between the concepts and SNOMED-CT codes of cTAKES was applied. For the German-language expansion of the system the UMLS version 2015AB was integrated into the cTAKES database.

The primary purpose of SNOMED CT is to encode the meanings that are used in health information and to support the effective clinical recording of data with the aim of improving patient care. SNOMED CT provides the core general terminology for electronic health records and contains more than 311,000 active concepts. It is comprehensive coverage includes: clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimens. In this test series only the diseases where considered.

2.4. Natural Language Processing Pipeline

Natural language processing is an area of research and application that discovers how computers can be used to understand and manipulate natural language text or speech. Most common approaches to natural language processing are based on machine learning, a type of artificial intelligence that examines and uses patterns in data to improve a program's own understanding. For natural language processing and UMLS concept extraction on German clinical notes, different analysis steps in the form of an Analysis Engine (AE) are required. The applied Analysis Engine consists of the steps Chunker, Tokenizer, DictionaryLookupAnnotator, SentenceDetectorAnnotator, SimpleSegmentAnnotator and POSTagger. The Lookup Annotator is linked to the German UMLS database for matching concepts [13].

In this approach, the NLP-Pipeline was implemented in Apache cTAKES. It processes clinical notes and identifies types of clinical named entities – drugs, diseases/disorders, signs/symptoms, anatomical sites and procedures using OpenNLP models [11]. For the preprocessing of German notes, three TIGER data [14] trained OpenNLP models were integrated into the natural language processing pipeline [15]. The TIGER project aims to produce a large syntactically annotated corpus of German newspaper text. This includes a German Maxent Part-of-Speech tagger, Tokenizer and Sentence Detector.

3. Results

The results show the number of UMLS concepts extracted for English and German clinical notes compared to the gold standard.

Table 2 shows the performance of the pipeline with the ShARe/CLEF eHealth 2013 training dataset on English notes. The pipeline achieved an average recall of 0.69 with a moderate average precision of 0.25. The moderate precision has the consequence that the F1 value is attenuated. The best F1 could be achieved with discharge notes, which are the most complex notes with an average CUI of 32.28.

Table 2. Statistics of NLP-Pipeline with English notes:

Type	Recall	Precision	F1
DISCHARGE	0.71	0.27	0.39
ECHO	0.56	0.26	0.35
RADIOLOGY	0.69	0.23	0.35
ECG	0.81	0.25	0.38

Table 3. Statistics of NLP-Pipeline with German notes:

Type	Recall	Precision	F1
DISCHARGE	0.37	0.30	0.33
ECHO	0.47	0.51	0.49
RADIOLOGY	0.37	0.28	0.32
ECG	0.39	0.25	0.30

Table 3 shows the performance of the pipeline with the ShARe/CLEF eHealth 2013 training dataset on translated German notes. The results of the German pipeline have a much lower average recall of 0.4, but a slightly higher average precision of 0.34 compared to the pipeline with English notes. Thereby the average F1 value of 0.36 is almost equivalent to the English average F1 value of 0.37. Especially with echo notes, a much better F1 value could be achieved with German notes.

For the evaluation of the system, only concept codes were used from the dataset. diseases with no CUI Code (CUI-less) were not considered. Consequently, the experimental results are not directly comparable with those achieved by systems participating in the ShARE/CLEF Tasks.

4. Discussion

In this research a clinical disorder recognition and encoding system combining a machine learning based approach for entity recognition with UMLS concept mapping for German language was developed. The dictionary lookup approach on German clinical notes with terminological content from the UMLS for detecting disease was successful with a moderate F1 value.

The first results presented here seem promising even without German stemming, good results could be achieved and English SNOMED-CT codes could be derived. There are several reasons why the recall value of the German pipeline is significantly lower than the English pipeline. The UMLS database is not as extensive for the German Language (196842 entries) like the English (5571374 entries) UMLS database. That is the reason why the English pipeline is much better with complex notes such as discharge and receives more CUIs. Nevertheless, the German pipeline has a better average precision because it also identifies less false positive concepts. This also explains why the F1 value on German echo notes is higher than the value of the English pipeline. The Echo notes with an average CUI of 11.4 are not complex and have less different CUIs compared to the discharge notes. The German pipeline identifies less CUIs but with a higher precision. One more reason why the English pipeline achieved a much higher recall is that the applied cTAKES Analysis Engine is optimized for the English language. The AE contains stemming only for English language and the machine learning OpenNLP models are trained on English expert-annotated gold standard data.

In addition, no German stemming has been integrated into the pipeline at the time of the results publication and the German OpenNLP models have not yet been trained to medical notes. Generally, both results can be improved by pre- and post-processing of the notes. The first Analysis of the result dataset and the translated German notes show that some incorrectly or not identified concepts can be caused by translation errors. The quality of the translation has a significant impact on suitability for evaluation and has to be analyzed and optimized.

The next steps are to implement a Context Analyzer and German stemming to improve the results. Even an extension of the UMLS database and training of the models will be accomplished to see whether and how the results can be improved. Despite the fact that the models were used out-of-the-box without extra training the first results are promising. The pipeline for the German notes also lacks Negation Detection, which is also important to improve the results. Future attempts shall focus on optimizing the natural language processing pipeline for German language.

References

[1] A. Winter, R. Haux, E. Ammenwerth, B. Birgl, N. Hellrung, F. Jahn, Hospital Information Systems, in: Health Information Systems: Architectures and Strategies (Health Informatics). Springer, 2010. pp. 33-36

[2] G. Heyer, U. Quasthoff, T. Wittig, Wissensverarbeitung gestern und heute, in: Text Mining: Wissensrohstoff Text. W3L, Witten, 2012. pp. 12-18

[3] Franke J., Nakhaeizadeh G, Renz I., XML Retrieval and Information Extraction, in Text Mining – Theoretical Aspects and Applications, Physica-Verlag, Germany, 2003. pp. 29-32

[4] Averbis text analytics, https://averbis.com/branchen/pharma/, last access: 18.03.2016.

[5] ID SEMGuide, http://www.id-berlin.de/de/products/1-codierung/2-id-semguide/, last access: 18.03.2016.

[6] Semfinder DE, http://www.semfinder.com/produkte/semfinder-de/eigenschaften.html, last access: 18.03.2016.

[7] U.S. National Library of Medicine, Unified Medical Language System (UMLS), https://www.nlm.nih.gov/research/umls/, last access: 18.01.2016.

[8] Varghese J., Dugas M., Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT. PubMed 54(1) (2015),83-92

[9] T. Benson, Clinical Terminology, Principles of Health Interoperability HL7 and SNOMED (Health Information Technology Standards), Springer, 2012, pp. 201-212

[10] McInnes BT, Pedersen T, Carlis J. Using UMLS Concept Unique Identifiers (CUIs) for word sense disambiguation in the biomedical domain. In: AMIA Annual Symposium Proceedings, Volume 2007. American Medical Informatics Association: (2007), 533

[11] Apache, Apache cTAKES, http://ctakes.apache.org/, last access: 18.01.2016.

[12] Pecina P., Dušek O., Goeuriot L., Hajič J., Hlaváčová J., Jones G.J., Kelly L., Leveling J., Mareček D., Novák M., Popel M., Rosa R., Tamchyna A., Urešová Z., Adaptation of machine translation for multilingual information retrieval in the medical domain, Artif Intell Med 61(3) (2014),165-85

[13] Divita G, T Zeng Q., Gundlapalli A., Duvall S., Nebeker J., Samore M., Sophia: A Expedient UMLS Concept Extraction Annotator, Journal of the American Medical Informatics Association (2014), 467–476 developed to extract UMLS concepts from clinical text. Among them, cTAKES

[14] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association (2010), 17(5):507–13

[15] Institut für Maschinelle Sprachverarbeitung, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html, last access: 18.03.2016.

[16] Apache, Sourceforge OpenNLP Tools Models, http://opennlp.sourceforge.net/models-1.5/, last access: 18.01.2016.