

Automatic Mapping of Clinical Documentation to SNOMED CT

Holger STENZHORN^{a,1}, Edson José PACHECO^b,
Percy NOHAMA^b, Stefan SCHULZ^a

^a*Institute for Medical Biometry and Medical Informatics,
University Medical Center, Freiburg, Germany*

^b*CPGEI, Electrical Engineering Department,
Federal Technological University of Paraná (UTFPR), Curitiba, Brazil*

Abstract. Clinical documentation needs to be fine-grained to truthfully represent the history, development, and treatment of a patient. But natural language, as the main information carrier, is characterized by many issues, like idiosyncratic terminology, spelling and grammar errors, and a lack of grammatical structure. Therefore coding systems, like ICD-10, have been introduced, but their use varies highly among physicians, and they are often used incompletely or incorrectly. The almost exponential growth of clinical data is yet another problem. We present a new methodology to process this data: Through combining several natural language processing methods we extract morphemes from clinical texts and map them onto concepts from SNOMED CT. We first performed a manual analysis of clinical texts received from a university hospital and evaluated the issues found in them. Based on this we implemented a prototypical system which incorporates both the OpenNLP and the MorphoSaurus natural language processing systems.

Keywords. ontology, SNOMED CT, information retrieval, natural language processing, electronic medical records

1. Introduction

In clinical documentation, such as findings reports or discharge summaries, fine-grained descriptions are necessary to truthfully represent the history, development, and treatment of a patient. But natural language, as the main information carrier for this purpose, is characterized by several issues: It uses idiosyncratic terminology and contains highly context-dependent and ambiguous expressions, such as acronyms and abbreviations. Spelling errors, grammar violations, and a lack of grammatical structure are also highly common. Therefore, standard coding systems, like ICD-10 [1], have been introduced to disambiguate clinical documentation through univocal codes. But in practice the annotation with such codes varies among physicians and often tends to be incomplete, or even incorrect [2, 3]. Furthermore, if the steadily growing amount of clinical data [4] is also taken into account, then it gets obvious that novel methodologies are necessary for properly organizing this data and which can be easily implemented and used in everyday clinical practice.

¹ Corresponding Author: Holger Stenzhorn, Institute for Medical Biometry and Medical Informatics, University Medical Center, Stefan-Meier-Str. 26, 79104 Freiburg, Germany; E-mail: holger.stenzhorn@uniklinik-freiburg.de.

Below we present one such possible approach that automatically maps clinical narratives onto concepts from the clinical terminology SNOMED CT [5]. To achieve this, we combine several natural language processing (NLP) methods, like stemming or morphological analysis, and map the outcome onto appropriate concepts. They can in turn be applied for various tasks, such as concept-driven search in biomedical databases.

2. Goal and Research Question

The development of our approach is driven by the issues and needs described above: We want to tackle these problems that occur with manual annotation by creating an automated annotation process which is able to create consistent annotations for all documentation texts. This can tremendously improve retrieval of texts with similar or same content since they are annotated with the same concepts and can thus be found by the same concepts as well. Since our approach is intended to also map multilingual texts onto (language-independent) concepts, also texts in different languages can be retrieved by the same concepts. The use of SNOMED CT concepts has another advantage in that its concepts are logically defined. This allows filtering or expanding retrieval results based on restricting or expanding the actual concepts found in the texts.

The core research question that results from this is: How can we abstract from the textual surface of clinical narratives and represent its content through SNOMED CT concepts? In the following, we describe our answer proposal to pragmatically combine automatic natural language processing (NLP) and ontology mapping to solve this task.

3. Methods and Material

3.1. Natural Language Processing (NLP)

As both training and evaluation material, we built a corpus of about 160,000 discharge summaries collected over a three year period in the cardiology department of the University Hospital de Clínicas in Porto Alegre, Brazil. For a first survey of the characteristics of these documents, we manually examined about 1,000 of them.

All summaries were written in Portuguese by physicians and sixth year medical students. The writing style and accuracy varied widely. Whereas many writers reasonably followed spelling and grammar standards, others introduced typing errors, ungrammatical sentences, or idiosyncratic case, punctuation, and accentuation choices.

3.1.1. Acronym Processing

One major issue encountered in the sample was the multitude of abbreviations and acronyms: Before any further processing could happen they needed to be expanded. A first attempt to manually create an acronym list had to be abandoned after obtaining over 3,000 candidates after analyzing only 5% of the documents. Instead we created a set of regular expression rules based on manual analysis to extract acronym candidates. A medical expert iteratively checked the results of the rule application. After further refining the rules we ended up with 2,300 good candidates for the whole sample.

For further disambiguating the acronyms we automatically created document clusters around each acronym, i.e., all documents containing the acronym. Then those

documents were ranked and divided based on the similarity of the (non-disrupted) tokens neighboring the acronym (+/- 3 tokens) using distance-based weights: Documents were compiled if the tokens before or after an acronym were the same in the documents. The documents were ranked higher the closer a co-occurring token was to the acronym. Through this method we could discover different meanings for one given acronym based on its neighbor words.

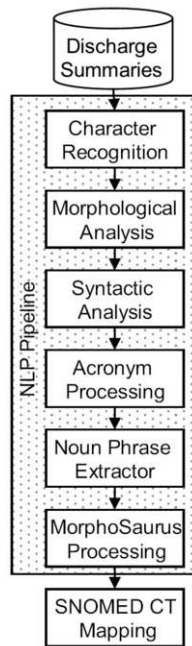


Figure 1. Processing pipeline

We then constructed an unambiguous semantic representation of each acronym meaning. To this end, we did not use the original words as tokens but semantic identifiers provided of the MorphoSaurus system [6] (MIDs). They represent unambiguous, language-independent atomic meanings, of word stems and meaning-bearing word stems, so-called subwords. MorphoSaurus extracts MIDs from different input languages (at the moment English, German, French, Spanish, Portuguese, and Swedish). For reviewing the results of the MorphoSaurus acronym processing, we randomly selected 20% of the documents and let two medical experts manually validate the correctness of the assigned MIDs (and thus meanings). The accuracy was calculated to be at 87.6%.

3.1.2. Noun Phrase (NP) Extraction

As no medicine-specific Portuguese language resource exists (apart from about 140,000 entries in the UMLS Metathesaurus), we decided to semi-automatically extract the NPs since they are supposed to represent a similar granularity level as the descriptions of the SNOMED CT concepts (80% of the texts were used for training, 20% manually annotated with grammatical units to create a gold standard). The NP extraction is based on a combination of the statistics-based OpenNLP toolkit [7] with hand-coded language-dependent NP-building rules to improve precision.

The following task is to extract NPs from clinical documents, like “myocardial infarction”, since the next step consists of mapping them to SNOMED CT. For the initial processing, i.e., sentence detection, tokenization, part-of-speech-tagging, chunking and parsing, named-entity detection, we use the OpenNLP toolkit.

Annotated data is the major prerequisite for any statistical algorithm in natural language processing. But to obtain the necessary amount of human annotations for linguistic data constitutes a labor-intensive knowledge acquisition process. To reduce time – but not at the expense of quality – we adopted a semi-supervised technique, namely active learning (AL) [8], for the part-of-speech tagging and chunking. The AL paradigm is a learning algorithm to control the selection of those examples for which the human annotation is supposed to yield a maximum of information so that the annotation effort can be significantly reduced [9].

3.2. SNOMED CT

We chose the clinical terminology SNOMED CT because it represents an international standard and supports the analysis, encoding and retrieval of data in various medical sub-domains. It constitutes the most extensive terminological resource available (now 311,000 active concepts). In contrast to other, more focused classificatory systems, like

ICD-10, SNOMED CT has a much broader spectrum, including medical procedures, findings or drugs, etc. SNOMED CT consists of hierarchically ordered concepts, identified by a numerical code, name and a set of synonyms to express.

3.3. Combining NLP and SNOMED CT

In the process of mapping of found NPs to appropriate SNOMED CT concepts, the multilingual design of MorphoSaurus is important, as SNOMED CT is only available in English and Spanish but not in Portuguese. To support the mapping, we used the system to map each SNOMED CT concept to a set of MIDs (cf., Table 1). In case more than one MID set was obtained for some SNOMED CT term then the ambiguity was resolved by creating another MID set from the corresponding Spanish term. If this MID set is non-ambiguous then this set is chosen, if not we used all the ambiguous words in a Vector Space Model [10]. Each occurrence of an ambiguous MID is represented as a binary vector in which each position indicates the occurrence or absence of some feature. A single centroid vector is generated for each sense in the training step. The centroids are then compared with the vectors representing the synonyms of the same concept (in English and Spanish) using the cosine metric to compute similarity.

Table 1. MIDs for all descriptions of the SNOMED CT concept “Congestive heart failure (disorder)”

SNOMED CT Concept Description	MIDs
Congestive heart failure	#abund #cardiac #deficien #static
Congestive heart disease	#abund #cardiac #disorder static
Congestive cardiac failure	#abund #cardiac #deficien #static
CCF – Congestive cardiac failure	#abund #cardiac #ccf #deficien #static
CHF – Congestive heart failure	#abund #cardiac #chf #deficien #static

To map the NPs onto SNOMED CT concepts we use a co-occurrence relation, controlled by the distance between occurrences of MIDs: two vertices are connected if their corresponding lexical units co-occur within a window of n MIDs with $2 \leq n \leq 10$. Co-occurrence links [11] express the relations between different syntactic elements. They were very useful in the mapping process because with their help we could extract the most correct mapping between NPs and the concepts. After manual evaluation, five was found to be the best value for n (cf., Table 2). To validate the accuracy of our approach, 20% of the corpus is currently manually annotated with SNOMED CT concepts. For example, only with a window of five MIDs the whole expression “insuficiência cardíaca congestiva” (congestive heart failure) can be covered: “{#deficien, #disorder} #cardiac {#static, #abund}”.

Table 2. Percentage of correctness for different MID window sizes based on 25 manually processed documents

MID Window Size	Correctness
2	66%
3	71.4%
4	80.1%
5	89.3%
6	79%
7	79.5%
8	75.4%
9	45.2%
10	25.4%

4. Conclusion and Outlook

In the above we presented a methodology to map free natural language texts from the clinical domain onto SNOMED CT concepts. We highlighted the various issues that we experienced when we were trying to process the texts, like the frequent use of acronyms. Thus, we presented a proposal to solve those issues within a practically implemented system. So far, our work is ongoing and the implemented components are still on a prototypical level. Therefore, additional work is needed to optimize both the process and the implementation in terms of quality as well as speed. But still, the results we have reached so far are highly encouraging.

Thus, our next steps are to further train the system. For this, we will use a large text corpus that has been automatically processed and mapped and then let the medical experts annotate the texts manually with SNOMED CT concepts as well. We believe that by comparing the manual and automatic annotations, we can detect problematic points in our processing pipeline to deduce solutions for them.

To further validate our approach we plan to create a prototypical search system based on the automatic mapping approach introduced above. This system should make possible to enter clinical texts as input and get back texts annotated with the same or similar concepts (and thus content). We believe that by validating such a system, we can detect weaknesses in our mapping approach and in turn improve it.

Acknowledgements. We would like to thank Mariza Kluck for the corpus provision, and Roosevelt Leite de Andrade, Jeferson Bitencourt and Píndaro Cancian for their collaboration.

Our work is funded by the International Bureau of the German Ministry of Education and Research (BRA 07/022) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

- [1] World Health Organization, Geneva, Switzerland, *International Statistical Classification of Diseases and Related Health Problems (ICD)*, <http://www.who.int/whosis/icd10>.
- [2] Sackett, D. (1980) Clinical disagreement. How often it occurs and why. *Canadian Medical Association Journal* 123:499–536.
- [3] Barnum, J. (1989) The misinformation era: The fall of the medical record. *Annals of Internal Medicine* 110:482–484.
- [4] Pestotnik, S. et al. (2000) Medical informatics: Meeting the information challenges of a changing health care system. *Journal of Informed Pharmacotherapy* 2:1.
- [5] International Health Terminology Standards Development Organisation (IHTSDO), *Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)*, <http://www.ihtsdo.org/snomed-ct>.
- [6] Markó, K. (2008) *Foundation, Implementation and Evaluation of the MorphoSaurus System*. Doctoral Dissertation, Freiburg, Germany.
- [7] OpenNLP, <http://opennlp.sourceforge.net>.
- [8] Cohn, D., Ghahramani, Z., Jordan, M. (1996) Active learning with statistical models. *Journal of Artificial Intelligence Research* 4:129–145.
- [9] Tomanek, K., Hahn, U. (2008) Approximating learning curves for active-learning-driven annotation. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, <http://www.lrec-conf.org/proceedings/lrec2008/>.
- [10] Agirre, E., Edmonds, P. (Eds.) (2006) *Word Sense Disambiguation: Algorithms and Applications*. Series: Text, Speech and Language Technology, Vol. 33, Springer, Heidelberg.
- [11] Mihalcea, R. (2004) Graph-based ranking algorithms for sentence extraction, applied to text summarization. *Proceedings of the Meeting of the Association for Computational Linguistics (ACL 2004)*, Association for Computational Linguistics, Morristown, <http://dx.doi.org/10.3115/1219044.1219064>.