Mapping of Clinical Documentation to Ontology

Edson José Pacheco[1], Percy Nohama[1] ,
Stefan Schulz[2]
[1]CPGEI – Electrical Engineering Department –
Federal Technological University of Paraná (UTFPR) – Curitiba– Brazil
[2]Institute for Medical Biometry and Medical Informatics–
University Medical Center– Freiburg– Germany
edson.pacheco@pucpr.br, percy@cpgei.cefetpr.br,
stschulz@uni-freiburg.de

Abstract. *Clinical documentation requires the representation of fine-grained descriptions of patients' history, evolution, and treatment. These descriptions are materialized in findings reports, medical orders, as well as in evolution and discharge summaries. In most clinical environments natural language is the main carrier of documentation. Written clinical jargon is commonly characterized by idiosyncratic terminology, a high frequency of highly context-dependent ambiguous expressions (especially acronyms and abbreviations). Violations of spelling and grammar rules are common. The purpose of this work is to map free text from clinical narratives to a domain ontology. To this end, natural language processing (NLP) tools will be combined with a heuristic of semantic mapping.*

1. Introduction

In clinical documentation, such as findings reports or discharge summaries, fine-grained descriptions are necessary to truthfully represent the patients' history, evolution, and treatment. 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, like acronyms and abbreviations. Spelling errors, grammar violations and a lack of grammatical structure are common.

It is increasingly recognized that the complexity of the health care and life sciences domain demands a consensus on the terms and language used in documentation and communication. This need is driven by the exponential growth of data generated in the contexts of both patient care and life science research. At the moment, this data cannot be fully exploited for integration, retrieval, or interoperability, because the underlying terminology and classification systems (often subsumed under the heading "biomedical vocabularies", see Table 1) are inadequate in various ways. Their heterogeneity reflects the different backgrounds, tasks and needs of different communities – including communities on the side of information technology – and

[1] Corresponding author: Edson José Pacheco, UTFPR/CPGEI, Av. Sete de Setembro 3165, Rebouças – Curtiba - PR, 80230-901, Brazil; edson.pacheco@pucpr.br.

creates a serious obstacle to consistent data aggregation and interoperability of the sort demanded by biomedical research, health care, and translational medicine.

Table 1. Examples of biomedical vocabularies. (NLMb, 2008; MCCRAY et. al, 1995)

Vocabulary	Purpose
ICD-9-CM/ICD-10 [WHO 2008]	disease classification, in health statistics, hospital billing
WHO Drug Dictionary [UMC 2008],	drug classification
ATC [WHOCC 2008],	
RxNorm [NLMa 2008]	
DM+D [NHS 2008]	
NCI Thesaurus and Metathesaurus [NCI 2008]	cancer research
LOINC [REGENSTRIEF INSTITUTE 2008]	inter-laboratory communication
DICOM [MITA 2008]	medical image and imaging process descriptions
MeSH [NLM 2008]	medical literature indexing
SNOMED CT [IHTSDO 2008]	clinical documentation

When combining all those issues, it becomes apparent that novel, intelligent methodologies need to be conceived to properly organize this data. They need to function in a way that can be easily implemented and used in daily clinical practice.

In the following we present one possible methodology: In a nutshell, we are mapping free text from clinical narratives to the clinical terminology SNOMED CT [IHTSDO 2008]. For this, we combine several different natural language processing (NLP) tools, like stemming or morphological analysis, and automatically map the outcome of this to SNOMED CT concepts. The extracted concepts can then be used within various tasks, such as the concept-driven search within biomedical databases. An additional challenge is to exploit SNOMED CT's internal structure, based on Description Logics, in order to match semantically identical expressions.

2. Goal

The development of our approach is strongly driven by real-world applications. We envisage concept-driven search to be one main application to employ the result or our work, having following benefits:

- Improving “searching throughput”. Speed is essential in real-time clinical systems. But storing large amounts of data as simple texts in databases can make retrieval time suboptimal. Thus texts pre-annotated with concepts, can be indexed and then retrieved more efficiently based on those concepts.

- Improve search quality. We can use logic-based concept definitions and restrictions in SNOMED CT to filter or expand the search result.

- Multilingual search. As SNOMED CT concepts are language-independent, texts in different languages but with similar or the same content can be annotated automatically with the same concepts (by our proposed methodology) and thus be found via those concepts.

The resulting research question now is: how can we abstract from the textual surface and represent the contents by properly connected standardized identifiers such as given by SNOMED CT concepts? Our proposal is to pragmatically combine natural language processing (NLP) and ontology mapping techniques. In the following we will report on ongoing activities, focusing on selected tasks.

3. Methods and Material

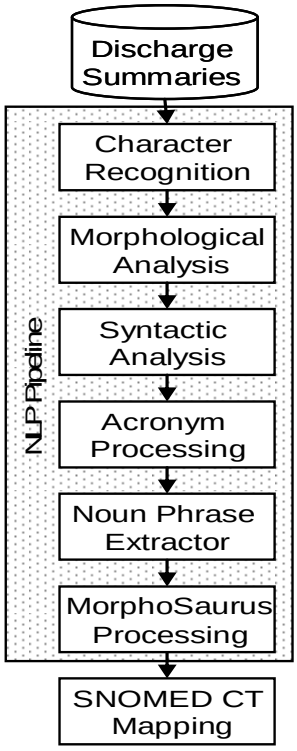

Figure 1. Processing Pipeline.

As both training and evaluation material for our research, we constructed a corpus of about 160,000 discharge summaries that had been 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 1000 of the summaries.

All summaries were written in the Portuguese language and authored by physicians and sixth year medical students. The writing style and the accuracy varied widely. Whereas many writers reasonably followed spelling and grammar standards, other abounded with typing errors, ungrammatical sentences, as well as idiosyncratic case, punctuation, and accentuation choices.

One major issue encountered in the whole sample was the multitude of acronyms and abbreviations. Before any further processing could happen they needed to be expanded. A first attempt to manually prepare a list of acronym had to be abandoned after we obtained over 3000 candidates after analyzing only 5% of the documents. Instead we created a set of regular expression rules based on the manual analysis to extract acronym candidates. A medical expert iteratively checked the results of applying those rules. After a further refinement of the rules we ended up with 2300 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 stop) tokens neighboring the acronym (+/- three tokens) using distance-based weights: Documents were put together 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.

We then had to construct an unambiguous semantic representation of each acronym meaning. To this end we did not use original words as tokens but semantic identifiers provided of the MorphoSaurus system [Markó 2008], so-called MIDs. MIDs represent unambiguous, language-independent atomic meanings, of word stems and meaning-bearing word stems, so-called subwords. MorphoSaurus extracts MIDs from sveral different input languages (at the moment English, German, French, Spanish, Portuguese, and Swedish).

For reviewing the results of the MorphoSaurus processing of the acronyms, 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 87.6%.

3.1. Noun Phrase (NP) Extraction

As there are no medicine-specific Portuguese language resourced (apart from about 140,000 entries in the UMLS Metathesaurus), we decided to semi-automatically extract the noun phrases since they are supposed to represent a similar level of granularity as the descriptions attached to SNOMED CT concepts (we used 80% of the texts for training and 20% as a gold standard. For this standard we manually annotated all the texts with grammatical units). The extraction of the NPs is based on a mixed approach which combines the statistics-based OpenNLP toolkit [OpenNLP 2009] together with hand-coded language-dependent NP building rules to improve precision.

The following task is to extract noun phrases from clinical documents, such as “myocardial infarction” because the subsequent step consists in mapping them to SNOMED CT. For the initial document processing, like sentence detection, tokenization, part-of-speech-tagging, chunking and parsing, named-entity detection, we are employing 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) [Cohn et al. 1996], 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 [Tomanek et al. 2008].

3.2. Combining NLP and 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.

In the process of mapping of found noun phrases to appropriate SNOMED CT concepts, the multilingual design of MorphoSaurus is of outmost importance, as SNOMED CT is only available in English and Spanish but not in Portuguese. To support the mapping, we used MorphoSaurus to map each SNOMED CT concept to a set of MIDs: For instance, “Myocardial infarction” was expressed as “#muscul #cardiac #infarct”, using a Semantic Normalization (cf. Figure 2).

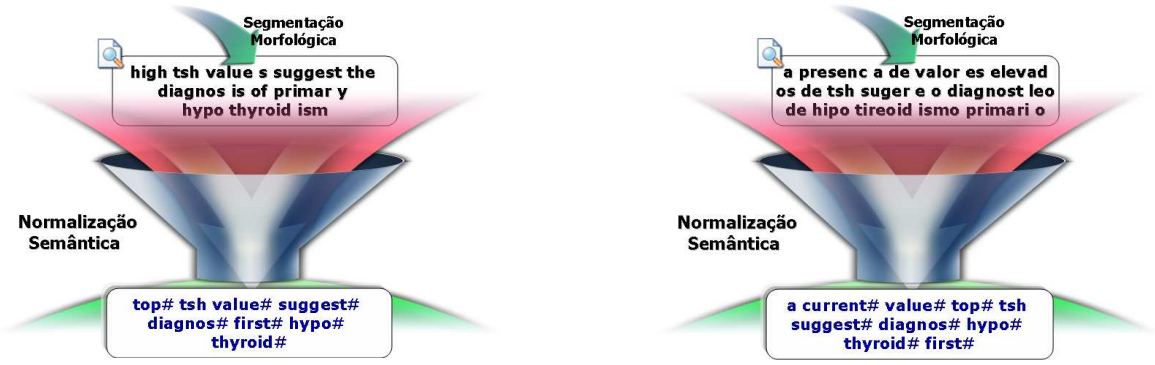

Figure 2 - Semantic Normalization

In case that more than one MID set was obtained for some given 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 [Agirre 2007]. 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. These centroids are then compared with the vectors that represent the synonyms of the same concept (in English and Spanish) using the cosine metric to compute similarity.

To map the noun phrases onto SNOMED CT we are using a co-occurrence relation, controlled by the distance between MID occurrences: two vertices are connected if their corresponding lexical units co-occur within a window of maximum *n* words, where *n* can be set to anywhere from two to ten. Co-occurrence links [Mihalcea 2004] express the relations between syntactic elements. We found them very useful in the mapping task because then we extract the most correct mapping between the noun phrase and the concepts. After manual evaluation, 5 was found to be the best value for *n* (cf. Table 2). To validate the accuracy of our approach, we are currently manually annotating 20% of the corpus with SNOMED CT.

Table 2. Percentage of correctness based on 25 manually processed documents.

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 (cf. Table 2). 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.

Therefore, the next steps for us are to further train the system. For this we will take a large set of texts that have been automatically been processed and mapped and then let the medical experts annotate those texts manually with the concepts from SNOMED CT as well. Like during the initial evaluation, we believe that by comparing the manual with the automatic annotation, we can detect problematic points in our processing pipeline and deduce solutions for them.

Furthermore, we mentioned above that our approach is driven by the use case of concept-based search. To further validate our system we are planning to create a prototypical search system that will be based on the automatic mapping methodology introduced in this article. This system should make possible to enter clinical texts as input and get back texts annotated with the same or similar concepts. We believe that by validating the results of such a system, we can detect weaknesses in our mapping and in turn allow us to improve our mapping approach.

Table 2. Sample of mapping. Concept ID is a SNOMED CT code, descriptions are available online in http://snomed.vetmed.vt.edu/sct/menu.cfm

Chunk	Concept ID
Paciente	
com quadro demencial,	52448006
com relato de	
Doença de Alzheimer,	26929004
mas com capacidade de comunicação.	307083000
Cardiopata isquêmica	414545008
em tratamento clínico.	
Interna	
neste hospital	
com quadro de	
febre	386661006
(até 39 graus),	
depressão do sensório e	225454000
sinais radiológicos de consolidação,	95436008
sugestivos de pneumonia.	233604007
Iniciou	
uso de	
levofloxacina	233604007
empírico,	371070000

Chunk	Concept ID
sendo posteiormente trocado	
para cefuroxima,	372833007
tendo a paciente apresentado	
progressiva melhora clínica,	281013003
bem como	
remissão do	
quadro febril.	386661006
Conforme discutido	
com médico assistente,	
paciente	
recebe	
alta	306689006
com sinais vitais estáveis e	72970002
melhora do sensório.	
Orientações aos familares de	
retorno ao hospital	183635001
se febre ou	386661006
piora do estado geral.	285384003

Acknowledgement

We thank Mariza Kluck for the corpus provision, and Roosewelt Leite de Andrade, Jeferson Bitencourt and Píndaro Cancian for their collaboration. Our work is funded by the Internat. Bureau of the German Ministry of Education and Research (BRA 07/022) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

D. Cohn et al., Active learning with statistical models, Journal of Artificial Intelligence Research, 1996:4, 129–145.

E. Agirre and P. Edmonds (eds.) 2007. Word Sense Disambiguation: Algorithms and Applications. Text, Speech and Language Technology, Springer, Heidelberg, Germany, 2007.

IHTSDO (International Health Terminology Standards Development Organisation). Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), 2008. Available from: http://www.ihtsdo.org/snomed-ct. Last accessed: 20 March 2009.

K. Markó, Foundation, Implementation and Evaluation of the MorphoSaurus System, Doctoral Dissertation, Freiburg, Germany, 2008.

K. Tomanek et al., Approximating learning curves for active-learning-driven annotation, In Proceedings of the Language Resources and Evaluation (LREC 2008), Marrakesh, Morocco, 2008.

MITA (Medical Imaging and Technology Alliance). Digital Imaging and Communication in Medicine (DICOM), 2008. Available from: http://medical.nema.org. Last accessed: 20 March 2009.

NCI (National Cancer Institute). NCI Enterprise Vocabulary Services (EVS), 2008. Available from: http://www.cancer.gov/cancertopics/terminologyresources. Last accessed: 20 March 2009.

NHS (World Health Organization). Dictionary of Medicines and Devices (dm+d), 2008. Available from: http://www.dmd.nhs.uk. Last accessed: 20 March 2009.

NLM (United States National Library of Medicine). Medical Subject Headings (MeSH), 2008. Available from: http://www.nlm.nih.gov/mesh. Last accessed: 20 March 2009.

NLMa (United States National Library of Medicine). RxNorm, 2008. Available from: http://www.nlm.nih.gov/research/umls/rxnorm. Last accessed: 20 March 2009.

OpenNLP, http://opennlp.sourceforge.net. Last accessed: 20 March 2009.

R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization, In Proceedings of the 42nd Annual Meeting of the Association for Computational Lingusitics (ACL 2004), Barcelona, Spain, 2004.

REGENSTRIEF INSTITUTE. Logical Observation Identifiers Names and Codes (LOINC), 2008. Available from: http://loinc.org. Last accessed: 20 March 2009.

UMC (Uppsala Centre for International Drug Monitoring). WHO Drug Dictionary Enhanced, 2008. Available from: http://www.umc-products.com. Last accessed: 20 March 2009.

WHO (World Health Organization). International Classification of Diseases (ICD), 2008. Available from: http://www.who.int/classifications/icd. Last accessed: 20 March 2009.

WHOCC (WHO Collaborating Centre for Drug Statistics Methodology). Anatomical Therapeutic Chemical Classification System (ATC), 2008. Available from: http://www.whocc.no/atcddd. Last accessed: 20 March 2009.