Quality of Life through Quality of Information
J. Mantas et al. (Eds.)
IOS Press, 2012

doi:10.3233/978-1-61499-101-4-245

Inventory of Tools for Dutch Clinical Language Processing

Ronald CORNET[a,1], Armand VAN ELDIK[a] and Nicolette DE KEIZER[a]
[a] *dept of Medical Informatics, Academic Medical Center, Amsterdam, The Netherlands*

Abstract. Automated encoding of free-text clinical narratives using concepts from terminological systems is widely performed. However, the majority of natural language processing (NLP) tools and terminological systems involve the English language. As parts of the NLP process are language independent, and tools for various languages are available, an overview is needed to determine the applicability to performing NLP of Dutch medical texts. To this end an inventory of tools is created. A literature study and internet search were performed to describe available components for a Dutch NLP system, enabling to encode Dutch text as structured SNOMED CT output without the need to translate SNOMED CT in Dutch. We have found 31 papers, describing a variety of NLP frameworks and tools for the various NLP components for processing English and Dutch free text. Most of them are suitable for English free text, some of them are (also) usable for Dutch. To enable automated encoding of Dutch free text narratives, further research is needed to create a spelling checker, a negation detector, a domain-specific abbreviation/acronym list, and a concept mapper (to map Dutch terms to concepts in a terminological system). Furthermore evaluation of performance for the Dutch 'medical' language is needed.

Keywords. Natural Language Processing, Terminological Systems, SNOMED CT, Dutch

Introduction

The clinical information reported in EHRs is mostly in textual form, which hampers the use of these data for purposes such as automated clinical decision support, reimbursement systems or research [1]. These would be feasible when information is stored in a standardized and structured way. Most EHR systems are built with a combination of structured and unstructured data. For example, date of birth, sex or body length can easily be registered in a structured numerical form, unlike entries such as diagnoses, procedures or reasons for admission. These entries are usually recorded using free text, which allows healthcare personnel to express their thoughts in natural language. Alternatively, information can be recorded in encoded form by using clinical classifications, which cannot include all information healthcare providers would like to express, hence limiting expressiveness and introducing the need to use residual categories such as "Other transient cerebral ischemic attacks and related syndromes". Alternatively, information can be encoded with use of terminological systems such as SNOMED CT, which aim at capturing information with maximum detail, and without resorting to residual classes as in the example above.

[1] Corresponding Author: Ronald Cornet, E-mail: r.cornet@amc.uva.nl.

Various studies [2-4] investigated the use of Natural Language Processing (NLP) (often called Medical Language Processing in literature for the clinical domain) to extract clinically relevant data from EHRs, and a number of applications have been developed to achieve this, such as MedLEE [1], HITEx [5], and MetaMap [6]. These applications enable creation of automated mapping from English clinical narratives to concepts from a terminological system such as SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms). The needs for Dutch healthcare organizations to implement SNOMED CT are growing, but unfortunately no Dutch translation of SNOMED CT exists yet. The costs of translating SNOMED CT are high; a manual approach is time consuming and requires skilled specialized translators [7] , as was demonstrated by the Swedish and Danish translations, which took 40-70 man-years. Therefore there are currently no ambitions to fully translate SNOMED CT into Dutch. The NLP applications mentioned earlier are developed for the English domain and are not suitable for processing Dutch text. The aim of this study is to investigate by a literature inventory the possibility to transform Dutch free-text clinical data to a structured form with the use of SNOMED CT concepts. In our scope 'free-text' will consist of short pieces of text that are found in semi-structured sections from an EHR system (e.g., to express 'diagnoses' or 'reason for admission'). The purpose of this inventory is to describe the availability of the components needed for a Dutch NLP system, ultimately leading to the possibility of converting Dutch text to structured output without the need to translate SNOMED CT in full. This inventory distinguishes language-independent and language-dependent components, according to the pipeline depicted in Figure 1.

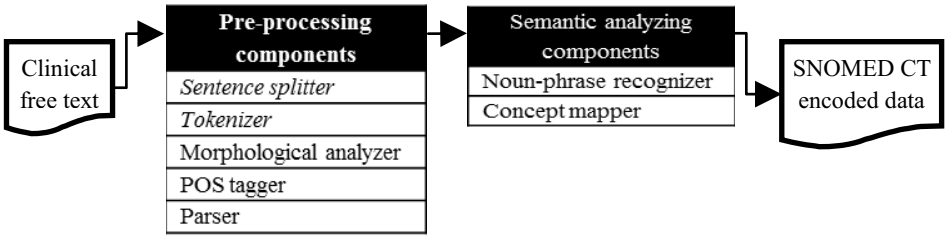

Figure 1. Components of an NLP pipeline. Language-independent components are in italic font.

1. Methods

We have searched in Medline for literature with the following terms: 'natural language processing', 'systemized nomenclature of medicine', 'information retrieval', 'electronic health records', 'automated encoding', 'free text', 'Dutch', 'medical terminology', 'negation detection', 'clinical abbreviations' and 'pipeline'.

We have chosen to include papers which describe complete systems (e.g., MedLEE or HITEx) or methods to automatically encode patient data or the creation of mappings from text (e.g., free, structured, or unstructured) to terminological systems (e.g., SNOMED CT, ICD-9, ICD-10). We excluded papers that are just focusing on non-western languages such as Chinese or Arabic. We also excluded papers that describe or evaluate the techniques from a single component from a processing pipeline such as a 'sentence-splitter' or part-of-speech (POS)-tagger.

From all included papers we extracted information about methodologies on how automated encoding is realized, frameworks and tools. Information about precision and recall rates (quality information) was outside the scope of our study. We did not make distinction between developments in- or outside the medical domain. We did internet search queries (e.g., Google Scholar) to get an overview of the tools discussed in literature.

2. Results

We have found 31 papers about automated encoding of patient data, which are mostly for English purposes, with only 5 addressing Dutch. With the results we were able to make a selection of the tools discussed and used for English and a short overview on initiatives for Dutch language. The results obtained from the internet search queries gave us an overview of the availability of different frameworks mentioned (e.g., GATE[2] and UIMA[3]). With a combination of the different processing parts from these frameworks an inventory of the components usable for Dutch could be made (see Table 1).

Various tools for processing English medical language were found. Other than MedLEE (1), HITEx (5), and MetaMap (6), which were previously mentioned, we retrieved information on MedTAS/MedKAT (8) and cTAKES (9), which are both based on the UIMA framework. These tools can be used for Dutch free text to perform the language-independent tasks of sentence splitting and tokenizing.

Table 1. Overview of tools suitable for processing Dutch natural language.
(int): tool provides the functionality only internally, i.e., the results cannot be retrieved;
+: the tool offers the functionality.

	Language independent		Language dependent				
	Sentence Splitter	Tokenizer	Morphological Analyzer	POS Tagger	Parser	Noun phrase finder	Concept mapper
Apache Lucene[4]		+	Dutch stemmer and analyzer				
TermTreffer[5]		+	Morphological Analyzer; Stopwords; Named entity recognizer; Negation finder	+		Multi-word recognizer	
Alpino[6]	(int)	+	(int)	(int)	+		

[2] http://gate.ac.uk/

[3] http://uima.apache.org/

[4] http://lucene.apache.org/

[5] http://www.inl.nl/tst-centrale/nl/over-de-tst-centrale/projecten/termtreffer

[6] http://www.let.rug.nl/~vannoord/alp/Alpino/

In the past a few projects investigated the use of Dutch natural language processing in the medical domain. The Ménélas project is one of these and is developed for French, German and Dutch (10). The first experimental version of this project was a prototype for French to encode free text to ICD-9-CM-based codes. This system is expanded for Dutch purposes where the language-independent components were reused and the dependent ones were built. These language-dependent components were discussed by Spyns et al. (11), who also evaluated the possibilities for a Dutch medical language processor (DMLP) (12;13). This system is built reusing available components developed by different projects (e.g., Ménélas, LSP-MLP, PUNDIT and PROTON) (11). Beyond the results mentioned above for the Dutch medical domain, very few notable projects did NLP work in this field; these projects can be found in an overview made by Spyns (14).

3. Discussion

The literature review makes clear that NLP is a research area with continuing developments. However, it also shows that the majority of these developments apply to English, and that progress in development of processing Dutch (medical) language is limited. Internet search indicated that this is comparable to other languages, where French and Swedish are among the languages with still ongoing progress in this area.
As shown in Table 1, the components for a Dutch NLP pipeline are covered by some tools suitable for Dutch usage. For the language-independent components there is a widespread variation of tools available which are usable for a Dutch pipeline. Unfortunately, the availability of tools further down the pipeline is limited, as these components are language dependent and hence have to be developed for each individual language. Especially the components for semantic analysis are scarcely implemented. The complete pipeline is covered apart from the concept mapper, which is an essential component for linking free text with a terminological system.

In our search for solutions, we have focused on scientific literature and internet search. In scientific literature, the only research applying to Dutch medical language is the work on Ménélas by Spyns et al. in the late nineteen nineties. Internet search revealed some companies providing tools or services for processing Dutch language, but we have not looked into details of these tools and services beyond the available descriptions. We also have not analyzed the availability of tools for other languages. Consequently, the results of this study are specific for processing Dutch language, but the method can be useful for other languages.

To create a complete NLP pipeline suitable for Dutch, four important issues need to be solved. Firstly, the system must be able to handle idiosyncratic Dutch language used by healthcare practitioners. For example, the grammatically incorrect phrase ‘verdenking pulmonale infectie’ lacks a preposition ‘op’ after ‘verdenking’ (English translation: ‘suspicion pulmonary infection’ instead of ‘suspicion of’ or ‘suspected’). Without correction the sentence will be parsed incorrectly which results in wrongly detected noun phrases, and hence wrong interpretation of the sentence.

Secondly, it is necessary to create an abbreviation and acronym list suitable for the medical specialty the processor is intended to be used for. This list could be used as a lexicon by the morphological analyzer which expands the abbreviations and detects acronyms. Xu (15) describes how to extract an abbreviation list from clinical data.

Thirdly, a negation detector is needed to ensure correct interpretation of a sentence. Creation of negation detectors is widely investigated and described in literature (16).

Finally, a concept mapper should be created which can map Dutch noun phrases to SNOMED CT concepts. As our intent is to map to SNOMED CT, but no Dutch translation thereof exists, ways need to be found to perform such a mapping. Use of the UMLS Metathesaurus could be considered, because it contains Dutch translations e.g., of ICD-10 and ICPC. Likewise, a medical dictionary could be used for obtaining an English translation, which could then be mapped via existing concept mappers for the English domain.

This study provides insight into the availability of tools that implement one or more NLP component, but does not address the performance of the various tools or the possibility of integrating the tools to realize an implementation of an NLP pipeline. Further work in this area includes practical experimentation to integrate the tools, assess their performance, and develop ways to perform concept mapping of Dutch (medical) noun phrases to SNOMED CT.

References

[1] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392-402.
[2] Lussier YA, Shagina L, Friedman C. Automating SNOMED coding using medical language understanding: a feasibility study. Proc AMIA Symp 2001;418-22.
[3] Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. Comput Inform Nurs 2009;27(4):215-23.
[4] Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc 2008;247-51.
[5] Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. AMIA Annu Symp Proc 2009;2009:442-6.
[6] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001;17-21.
[7] Deleger L, Merkel M, Zweigenbaum P. Translating medical terminologies through word alignment in parallel text corpora. J Biomed Inform 2009;42(4):692-701.
[8] Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen P. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009;42(5):937-49.
[9] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute Ch. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-13.
[10] Zweigenbaum P. MENELAS: Coding and Information Retrieval from Natural Language Patient Discharge Summaries. Amsterdam: IOS Press; 1995 p. 82-9.
[11] Spyns P, Willems JL. Dutch medical language processing: discussion of a prototype. Medinfo 1995;8 Pt 1:37-40.
[12] Spyns P, De MG. A Dutch medical language processor. Int J Biomed Comput 1996;41(3):181-205.
[13] Spyns P, De MG. A Dutch medical language processor: part II: evaluation. Int J Med Inform 1998;49(3):273-95.
[14] Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996;35(4-5):285-301.
[15] Xu H, Stetson PD, Friedman C. A study of abbreviations in clinical notes. AMIA Annu Symp Proc 2007;821-5.
[16] Agarwal S, Yu H. Biomedical negation scope detection with conditional random fields. J Am Med Inform Assoc 2010;17(6):696-701.