Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-710

Automatic Annotation of French Medical Narratives with SNOMED CT Concepts

Christophe GAUDET-BLAVIGNAC [a,1], Vasiliki FOUFI [a], Eric Wehrli [b], and Christian LOVIS [a]
[a] *Division of Medical Information Sciences, Geneva University Hospitals and University of Geneva*
[b] *Laboratoire d'Analyse et de Technologie du Langage, University of Geneva*

Abstract. Medical data is multimodal. In particular, it is composed of both structured data and narrative data (free text). Narrative data is a type of unstructured data that, although containing valuable semantic and conceptual information, is rarely reused. In order to assure interoperability of medical data, automatic annotation of free text with SNOMED CT concepts via Natural Language Processing (NLP) tools is proposed. This task is performed using a hybrid multilingual syntactic parser. A preliminary evaluation of the annotation shows encouraging results and confirms that semantic enrichment of patient-related narratives can be accomplished by hybrid NLP systems, heavily based on syntax and lexicosemantic resources.

Keywords. Interoperability, narrative data, SNOMED CT, NLP

1. Introduction

Medical data is composed of both structured and unstructured data. Narratives are a type of unstructured data that contains crucial semantic information but not readily usable by computers. Moreover, narratives constitute a challenge for interoperability between healthcare systems, hospitals and departments.[1] SNOMED CT (henceforth SCT), created in 2002, constitutes a terminology organized as a directed graph with concepts as nodes and relationships as edges. Currently, SCT contains more than 350,000 concepts. Property rights and developments are held by SNOMED International (London, UK). The goal of SNOMED International is to develop SCT and to ensure that it becomes the "most comprehensive and precise common global language for health terms in the world"[2]. Since SCT supports post-coordination, i.e. a formal grammar that can associate existing concepts, qualifiers, and predicates, it has similar properties to a natural language. In this paper, a method for automatic annotation of French medical narratives with SCT codes is proposed.

[1] Corresponding author. E-mail : christophe.gaudet-blavignac@unige.ch

2. Related work

Processing medical data with various terminologies, and recently SCT, has been a research focus of other studies as well[3], [4]. Those studies have pursued two kinds of goals. The first goal was the classification of documents, such as pathology or radiology reports[5], [6] in categories related to the disease mentioned in the text. The second one was a more general information retrieval task that aimed at extracting codes or annotating free text with concepts[7]. Commonly used terminologies are the ICD10[8], the UMLS[9] or SCT[10]. In some cases, preliminary work involves creating a subset of those terminologies in relation to a specific goal. It is the case with the UMLS because its Metathesaurus contains more than 100 terminologies, classifications and thesauri.

The methods used to annotate or classify free-text documents vary. Rule-based methods need to be manually or semi-manually developed but require no training corpus and can produce very satisfying results when combined in a pipeline[11], [12]. On the other hand, machine learning and statistical methods, such as Naïve Bayes or Support Vector Machine, do not require the manual creation of rules. However, access to large gold standard corpora used as training sets is essential[13]. Hybrid NLP systems integrating both statistical and linguistic approaches have also been proven very efficient at NLP tasks targeting the medical language[13]. The work presented in this article differs in several ways from the studies previously mentioned. First, the language of the free-text documents used in those references is mostly English. Working with another language requires to translate the terms and adapt the rules to the specificities of the target language. Second and most important, the absence of syntactic-semantic parsing of the text to detect terms in different morphological or syntactic structures makes the method presented in this paper innovative. Our system performs analysis of free medical text in French on the morphological, syntactic and semantic level and annotates the recognized terms with SCT concepts simultaneously.

3. Method

In this research, SCT is approached as a natural language. Automatic annotation of narratives with SCT concepts therefore requires the processing of texts using NLP tools.

3.1. Tool

The tool used for this goal is the hybrid multilingual syntactic parser *Fips* [14]. It relies on generative grammar concepts and is made of a generic parsing module which can be refined to suit the specific needs of a particular language or sublanguage. The lexicon is one of the key components of the parser. It contains detailed morphosyntactic and semantic information, selectional properties, valency information, and syntactic-semantic features that influence the syntactic analysis. To achieve automatic annotation of medical narratives, modifications were needed to correctly process the specificities of the French medical language such as abbreviations or technical terms.

3.2. Creation of electronic dictionaries

Specific lexicons have been developed and incorporated in the parser:

a) A French medical language dictionary was created by extracting simple words and collocations from a corpus of discussions of 11,000 discharge summaries from the internal medicine division of the University Hospitals of Geneva during 2012 to 2014. In its current version, the lexicon comprises 4,454 simple words and 5,640 collocations (groups of words) manually processed.
b) A SCT dictionary. To perform automatic annotation of French narratives with SCT codes, the SCT terminology was added as a new language in the parser. 173,067 SCT concepts and their equivalent code were entered in this dictionary.
c) A bilingual French-SCT dictionary. In the aim of automatic annotation, the target language (SCT) must be linked to the source language (French medical language) in a bilingual dictionary. In the current version of the system, 5,842 medical terms have been mapped to SCT concepts.

3.3. Automatic annotation

In this research, the automatic annotation procedure consists of parsing the initial text and recognizing medical terms. Then, the system looks up the dictionaries (both monolingual and bilingual) and proceeds to the SCT code attribution. Terms in medical terminologies can be affected by syntagmatic and paradigmatic variation to different degrees or may be too precise or complex to actually be used in electronic health records[15]. By providing syntactic analysis and a proper recognition of collocations, the parser can detect concepts regardless of the specific morphological or syntactic form under which they appear in the text. Table 1 shows an example of a sentence annotated with SCT concepts:

Table 1. Example of SCT annotation

Initial phrase	SCT Annotation
En raison des douleurs abdominales, un traitement de morphine iv est débuté et les traitements habituels du patient sont poursuivis	{21522001 | douleur abdominale |}, {373529000 | morphine |}, {255560000 | intraveineux |}, {40451002 | habituel |}, {116154003 | patient |}, {266714009 | poursuivre le traitement |}

We can observe that the system is capable of recognizing structures in various forms, i.e. *iv,* the abbreviated form of *intraveineux* ‘intravenous’. It can also identify complex structures even if their constituents do not follow the canonical order and are found in different positions, i.e. the verbal collocation *poursuivre un traitement* ‘continue a treatment’, *les traitements ... sont poursuivis* ‘the treatments … are being continued’.

4. Results

4.1. Automatic annotation

Automatic annotation using the syntactic parser was performed on a corpus of 11,000 discharge summaries. Table 2 below displays the results of the automatic annotation procedure.

Table 2. Automatic annotation of a corpus of 11,000 discharge summaries

Words	4481,191
Annotated terms	892,787
Unique SCT concepts	7,569
Annotated terms per sentence	4,17

4.2. Preliminary evaluation

A preliminary evaluation was completed on a small corpus of five discharge summaries (1,820 words) written by 4 different clinicians, chosen randomly. The corpus was first de-identified (i.e. Protected Health Information (PHI) was removed) and then manually annotated with SCT concepts by one expert. The concepts used for the annotation were selected from the set of codes that are incorporated in the parser's SCT dictionary. The same corpus was processed by the parser and 421 medical terms were automatically annotated. Then, a comparison of the two outputs was performed manually in order to evaluate the system. The performance of the system is very encouraging since precision of 0.7173 and recall of 0.517 were achieved. However, an evaluation on a bigger corpus would allow a more precise measurement of the efficiency of the method.

5. Discussion

5.1. Annotation procedure

The rules used to annotate a narrative with SCT concepts are subject to debate. Since the terminology is structured as a graph with a treelike disposition, there are various levels of granularity for each concept. For instance, *douleur abdominale* 'abdominal pain' could be annotated with a unique SCT code (*21522001*, cf. Table 1) or could be annotated several more specific concepts (*22253000* | *douleur* 'pain' |, *277112006* | *abdominal* 'abdominal' |). At the current stage of the research, the annotation was performed choosing the concept that corresponded to the largest text structure.

5.2. Limitations and future work

Medical documents contain sensitive information, as a consequence access to corpora and in particular annotated corpora, is a well-known challenge in this field. This is especially true for languages other than English. The size of the evaluation corpus is one of the major limitations of this paper. In addition, evaluation of medical free-text annotation must be performed in a specific setting to affirm that the results are reliable. The manual annotation task, in particular, should be performed by at least two annotators not directly involved in the development of the automatic annotation tool to avoid bias. Having more than one annotator is important to compute the inter-annotator agreement and set an upper-bound on the annotation task. The annotation of French narratives with SCT concepts is a first step toward the ultimate goal which is the complete representation of patient-related narratives into a formal language. The next step in this research will be the processing of post-coordinated concepts according to the SCT compositional grammar. Post-coordination will enable the storage of the full information contained in the text into SCT post-coordinated sentences.

6. Conclusion

In this paper, a method to annotate French medical free-text with SCT concepts is proposed. This method relies on a syntactic-semantic parser specifically modified to meet the needs of this task. Lexico-semantic resources (monolingual and bilingual dictionaries as well as grammar rules) were constructed taking into consideration the specificities of the French medical language. A preliminary evaluation has shown encouraging results with a precision of 0.7173, a recall of 0.5171 and an F-score of 0.6009. Further research is needed to produce post-coordinated structures and full representation of medical narratives into SCT.

Acknowledgements

This research has been financed by the “Réseau Thématique Langage & Communication” from the University of Geneva.

References

[1] S. M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, and C. U. Lehmann, “Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress.”

[2] “Support : SNOMED International.” [Online]. Available: https://ihtsdo.freshdesk.com/support/home. [Accessed: 20-Mar-2017].

[3] J. Patrick, Y. Wang, and P. Budd, “An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology.”

[4] P. Ruch, J. Gobeill, C. Lovis, and A. Geissbühler, “Automatic medical encoding with SNOMED categories.,” *BMC Med. Inform. Decis. Mak.*, vol. 8 Suppl 1, p. S6, 2008.

[5] G. Zuccon *et al.*, “Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology.,” *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.*, vol. 2013, pp. 300–4, Jan. 2013.

[6] A. Nguyen, J. Moore, G. Zuccon, M. Lawley, and S. Colquist, “Classification of pathology reports for Cancer Registry notifications,” *Stud. Health Technol. Inform.*, vol. 178, no. May 2014, pp. 150–156, 2012.

[7] M. Torii, K. Wagholikar, and H. Liu, “Using machine learning for concept extraction on clinical documents from multiple data sources.,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 18, no. 5, pp. 580–587, 2011.

[8] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson, “Automatic ICD-10 classification of cancers from free-text death certificates,” *Int. J. Med. Inf.*, vol. 84, pp. 956–965, 2015.

[9] B. Riedl, N. Than, and M. Hogarth, “Using the UMLS and Simple Statistical Methods to Semantically Categorize Causes of Death on Death Certificates.,” *AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp.*, vol. 2010, pp. 677–81, 2010.

[10] P. Ruch, J. Gobeill, C. Lovis, and A. Geissbühler, “BMC Medical Informatics and Decision Making Automatic medical encoding with SNOMED categories.”

[11] D. De Meyere *et al.*, “Automatic annotation of medical reports using SNOMED-CT: a flexible approach based on medical knowledge databases,” 2015.

[12] A. R. Aronson, “Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program.”

[13] R. J. Kate, “Towards Converting Clinical Phrases into SNOMED CT Expressions.,” *Biomed. Inform. Insights*, vol. 6, no. Suppl 1, pp. 29–37, 2013.

[14] E. Wehrli and L. Nerima, “The fips multilingual parser,” in *Language Production, Cognition, and the Lexicon*, Springer, 2015, pp. 473–490.

[15] C. Hansart, D. De Meyere, P. Watrin, A. Bittar, and C. Fairon, “CENTAL at SemEval-2016 Task 12: a linguistically fed CRF model for medical and temporal information extraction,” *Proc. 10th Int. Workshop Semantic Eval. SemEval-2016*, pp. 1286–1291, 2016.