Decision Support Systems and Education
J. Mantas et al. (Eds.)

doi:10.3233/978-1-61499-921-8-20

Terminology Coverage from Semantic Annotated Health Documents

Marie NDANGANG[a], Julien GROSJEAN[a,b], Romain LELONG[a,b], Badisse DAHAMNA[a,b], Ivan KERGOURLAY [a,b], Nicolas GRIFFON[a,b] and Stéfan J. DARMONI [a,b,1]
[a] *Department of Biomedical Informatics, Rouen University Hospital, Normandy, France*
[b] *INSERM, U1142, LIMICS, Paris, France*

Abstract. Background: Unstructured health documents (e.g. discharge summaries) represent an important and unavoidable source of information. Methods: A semantic annotator identified all the concepts present in the health documents from the clinical data warehouse of the Rouen University Hospital. Results: 2,087,784,055 annotations were generated from a corpus of about 11.9 million documents with an average of 175 annotations per document. SNOMED CT, NCIt and MeSH were the top 3 terminologies that reported the most annotation. Discussion: As expected, the most general terminologies with the most translated concepts were those with the most concepts identified.

Keywords. Big data analytics; Semantic annotator; Semantic Data Warehouse; Terminology server

1. Introduction

Indexing medical documents such as clinical reports is a key to various information retrieval tasks in medical information management. Indexing unstructured health documents (e.g. discharge summaries) is an important task because these information are very often not present in structured data (e.g. Diagnosis Related Group data or lab tests). Automatic indexing may be useful with the increasing amount of new material being produced in biomedical fields that has made manual indexing time consuming and expensive. Various annotating tools are available for English text, mostly based on the Unified Medical Language System (UMLS). Aronson et al. use MetaMap and the tri-gram method to extract UMLS terms, and then refine them to MeSH Descriptors [1]. Natural Language Processing (NLP) techniques can be also applied to annotate documents with UMLS [2]. Gurulingappa et al. use the JSRE system combining Support Vector Machines with different kernels specially designed for the NLP and relation extraction [3]. Vector space model is also a common approach that can be mixed with NLP techniques. Jonnalagadda et al. adopt this approach to identify UMLS concepts in the i2b2/VA concept extraction corpus [4].

[1] Corresponding Author, Stéfan J. DARMONI, Department of Biomedical Informatics, Rouen University Hospital, Normandy, France; E-mail: darmst@chu-rouen.fr.

French-speaking texts do not benefit from such various tools and resources. French is lowly represented in the UMLS [5]. As provided in the 2017AA release, the French UMLS thesaurus manages 11 resources providing a French concept for 143,762 concept unique identifiers (CUI). Since 2007, our team develops the Health Terminology/Ontology Portal (HeTOP; URL: www.hetop.eu) [6], a crosslingual terminology server providing an access to 75 terminologies and ontologies (Knowledge Organisation Systems; KOS) in 32 languages (mainly in English and in French). Some of these terminologies have been partially or totally translated in French. In May 2018, in HeTOP, there is a French concept for 427,912 CUI; therefore, the number of CUI in HeTOP is multiplied by 2.98, when compared to UMLS, although only 17 out of 75 KOS are included in UMLS, which underestimates the number of French concepts.

The aim of this study is to analyze the coverage of 55 terminologies available in French in the HeTOP terminology server on the Rouen University Hospital (RUH) Semantic Health Data WareHouse (SHDW). To do so, a semantic annotator (French acronym: ECMT) has been used on the 11.9 million of health documents present in the RUH SHDW.

2. Methods

The RUH SHDW is composed of four independent layers.

- Layer 0: a NoSQL architecture with a powerful server
- Layer 1: the HeTOP terminology server
- Layer 2: the ECMT semantic annotator [7]
- Layer 3: a semantic multiterminology multilingual search engine, that will not be described in this paper [8].

Only the layer 1 (ECMT) is language dependent: the French, in this use case. The three others are language independent.

2.1. ECMT Semantic Annotator

The ECMT semantic annotator [7] identifies clinical concepts in biomedical documents using terminologies included in HeTOP (URL: http://ecmt.chu-rouen.fr/). ECMT relies on the ”bag-of-words” algorithm and also on pattern-matching designed for discharge summaries, procedure reports or laboratory results which contain symbolic data (presence or absence), numerical data.

After the run of the ECMT semantic annotator on health documents of the RUH SHDW, a manual filtering process was applied based on the top 5,000 most frequent medical concepts automatically extracted: e.g. the concept “university hospital” present more than 27 million times in the 11.9 health documents was filtered as irrelevant, because this information is present elsewhere in the RUH SHDW.

2.2. SHDW Corpus of Health Documents

The RUH SHDW contains all the health documents from the clinical information system of this University Hospital from the beginning of 2000 till July 2017: discharge summaries, clinical notes, clinical reports, procedure reports, drug prescriptions.

3. Material

In the context of big data in health, as SHDW need to be used in the daily practice, there is an urgent need to minimize the response time of all the SHDW layers (HeTOP, ECMT in particular). Therefore, our semantic approach relies on:

- The use of NoSQL database, a new paradigm in data sciences [9]. Based on several internal benchmarks, we choose to implement In Memory Data Grid (IMDG).
- One server with a lots of RAM (about one Terabyte), and 144 cores.

4. Results

From 11,928,168 health documents present in the RUH SHSW, the ECMT semantic annotator has found 5,043,731,628 annotations before the filtering process and 2,087,784 055 after the filtering process. The process time was about 24 hours. This result is very important as it proves the ECMT annotator scalability, which can be reused each week if necessary.

Table 1 displays the terminology coverage (for the Top 20 terminologies) before and after filtering and the filtering factor, which is defined as the ratio for one terminology of the number of annotation before and after the filtering process. The Top 5 KOS are: SNOMED CT, NCIT, MeSH, SNOMED 3.5 (in French) and the TSP, a French Public Health Thesaurus. The average number of annotations per document is 423 before filtering and 175 after filtering. In addition, faced with the redundancy of concepts and the lack of specificity, a filtering step proved to be relevant.

Table 1. Terminology coverage (Top 20) based on the number of annotations.

Terminologies	Number of annotations (after filtering)	Number of annotations (before filtering)	Filtering factor
SNOMED CT	394,133,994	881,884,314	2.2
NCIt	319,853,952	843,195,067	2.6
MeSH	295,537,298	1,024,585,229	3.5
SNOMED 3.5	219,706,745	440,228,408	2.0
TSP*	179,747,539	454,354,922	2.5
MedDRA	137,653,806	225,100,880	1.6
Vidal Thesaurus**	106,616,463	171,231,559	1.6
RADLEX	80,197,479	150,406,338	1.9
FMA	55,350,010	123,777,281	2.2
CISMeF***	51,051,547	138,204,239	2.7
Drug list	33,355,617	37,554,068	1.1
ICNP	30,775,599	50,502,115	1.6
PASCAL	28,520,543	38,917,274	1.4
ICD-10	27,688,468	41,208,901	1.5
HPO+	27,526,442	40,038,338	1.5
MEDLINEplus	18,951,750	24,168,261	1.3
ICD-9	13,445,266	24,023,060	1.8
DRC	13,319,633	17,741,845	1.3
IUPAC	11,942,501	31,473,471	2.6
Cladimed	9,867,474	27,025,227	2.7

* French Public Health Thesaurus; **Drug Thesaurus; *** Medical Specialties Thesaurus; +Human Phenotype Ontology

Table 2 displays the terminology coverage (for the Top 10 terminologies) based on the number of unique identified concepts. It also displays: (a) the number of concepts in the (sub)terminologies (MedDRA and MeSH appears twice as we have separated MeSH descriptors and MeSH Concepts, MedDRA preferred terms and MedDRA Lowest Level Terms (LLT)); (b) the number of translated concepts in French and its percentage; (c) finally, it provides a terminology coverage ratio, which is the ratio between the number of unique identified concepts and the number of translated concepts in French. The Top 5 KOS in terms of number of unique identified concepts are: SNOMED CT, SNOMED 3.5 (in French), MedDRA LLT and MeSH Descriptors. For these Top 10 terminologies, the terminology coverage ratio varies from 12.4% for MeSH Concepts to 85.9% for TSP, a French Public Health Thesaurus.

Table 2. Terminology coverage (Top 10) based on the number of unique identified concepts

Terminology	Total number of concepts	Number of translated concepts (%)	Number of unique identified concepts	Terminology coverage ratio (%)
SNOMED CT	326,946	194,611 (59.5)	59,330	30.5
SNOMED 3.5	100,908	100,908 (100)	36,229	35.9
NCIt	93,925	68776 (73.2)	25,315	36.8
MedDRA LLT	44,226	44,226 (100)	22,711	51.4
MeSH descriptor	28,329	28,329 (100)	18,288	64.6
MedDRA PT	21,612	21,612 (100)	13,580	62.8
MeSH Concept	365,731	102,116 (27.9)	12,625	12.4
FMA entity	81,041	16,629 (20.5)	8,084	48.6
Radlex	42,313	10,259 (24.2)	6,114	59.6
TSP	7,087	7,087 (100)	6,089	85.9

5. Discussion

As expected, the most general terminologies with the most translated concepts were those with the most concepts identified. However, there was a redundancy of concepts identified by different terminologies but also a lack of relevance for some others. The use of a filtering methodology helped to refine certain results. In addition, certain identified concepts were only related to the type of document, thus providing only a small amount of relevant information.

Table 1 has provided unexpected result: ICD10, which is widely used in the world to index health documents, mainly for DRG purposes is only ranked 14th. Hoping that 11th version of ICD will provide better results in the near future.

Based on 11.9 million health documents, only 59,330 SNOMED CT concepts out of 194,611 were extracted at least once in the RUH SHDW. This figure is coherent with the experience of Prof. Olivier Lemoine in Erasme University Hospital, Brussels, Belgium, who has only translated around 90,000 SNOMED CT concepts in French in the last four years to be used a priori in clinical questionnaires in his Electronic Health Record.

The response time around 24 hours using a NoSQL architecture and a powerful server (1 Terabytes of RAM with 144 cores) to extract medical concepts from 11.9 million health documents proves the scalability of our semantic approach. These figures showed that this project refers to big data analytics. Since September 2017, a PhD student is studying a hybrid approach for the ECMT, associating NLP and deep learning; the first step is to learn an unsupervised word embedding model based on the corpus of 11.9 health documents, comparing Word2Vec and GloVe, since these approaches appeared

to regularly and substantially outperform traditional Distributional Semantic Models. This is a preliminary analysis with an innovative approach based on a multi-terminological semantic annotation of the RUH health documents. However, in front of the identification of poorly informative concepts and the absence of a formal evaluation, it is mandatory to complete this work before a full operational implementation in the RUH SHDW.

References

[1] A. Aronson, J. Mork, C. Gay, S. Humphrey, W. Rogers, The NLM Indexing Initiative's Medical Text Indexer, *Stud Health Technol Inform* **107** (Pt 1) (2004), 268–272.
[2] A. B. Abacha, P. Zweigenbaum, Automatic extraction of semantic relations between medical entities: a rule based approach, *J Biomed Semantics* **2** (5) (2011), S4.
[3] H. Gurulingappa, A. Mateen-Rajput, L. Toldo, Extraction of potential adverse drug events from medical case reports, *J Biomed Semantics* **3** (1) (2012), 15.
[4] S. Jonnalagadda, T. Cohen, S. Wu, G. Gonzalez, Enhancing clinical concept extraction with distributional semantics, *J Biomed Inform* **45**(1) (2012), 129-140.
[5] A. Névéol, J. Grosjean, S.J. Darmoni, P. Zweigenbaum, Language Resources for French in the Biomedical Domain, *LREC* (2014).
[6] J. Grosjean, T. Merabti, N. Griffon, B. Dahamna, S.J. Darmoni, Teaching medicine with a terminology/ontology portal, *Stud Health Technol Inform* **180** (2012), 949-953.
[7] C. Cabot, L.F. Soualmia, J. Grosjean, N. Griffon, S.J. Darmoni, Evaluation of the Terminology Coverage in the French Corpus LiSSa, *Stud Health Technol Inform* **235** (2017), 126-130.
[8] R. Lelong, L. Soualmia, B. Dahamna, N. Griffon, and S.J. Darmoni, Querying EHRs with a Semantic and Entity-Oriented Query Language, *Stud Health Technol Inform* **235** (2017), 121-125.
[9] R. Cattell, Scalable SQL and NoSQL data stores, *Acm Sigmod Record* **39** (4) (2011), 12-27.