Automatic annotation of medical reports using SNOMED-CT: a flexible approach based on medical knowledge databases

**Damien De Meyere*, Thierry Klein†, Thomas François*,
Jean-Claude Debongnie*, Cristina Radulescu†, Nicole Mbengo†
Maliki Ouro Koura†, Yves Coppieters 't Wallant†, Cédrick Fairon***

*CENTAL, IL&C, Université catholique de Louvain
Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgium

†École de Santé Publique, Université libre de Bruxelles
Route de Lennik 808, 1070 Bruxelles, Belgium

Abstract

This paper presents a rule-based method for the detection and normalization of medical entities using SNOMED-CT which, although based on knowledge stored in terminological resources, allows some flexibility in order to account for the language variation typical of medical texts. Our system is based on the software Unitex and is one of the few to code French medical texts with SNOMED-CT concept identifiers. Our evaluation quantifies the benefits of such a flexible approach, but also emphasizes terminological resource shortcomings for the processing of medical reports written in French. Finally, our methodology is an interesting alternative to supervised training, as the extraction rules require limited development.

1. Introduction

One of the great challenges faced by the healthcare sector is the deployment of electronic health record systems, which aim to gather all the documents produced throughout the clinical pathway of a patient. Even though such aggregation of clinical data can be used for multiple purposes (funding of hospitals, information exchange among health actors, design of clinical trials, epidemiology, etc.), almost 80% of the information consists in free text, which can not be easily processed automatically. While it is already common to structure medical information using reference classifications such as the International Classification of Diseases (ICD) or the International Classification of Primary Care (ICPC) for the creation of the Minimum Basic Dataset (MBDS), this approach has several limitations: first of all, each resource is used for different purposes, which complicates the exchange of information, despite the existence of metathesaurus like the *Unified Medical Language System*. Secondly, the encoding process is still carried out manually by specialised coders.

Many efforts have been made to partially automate the coding of medical documents in relation with the production of the MBDS, and several studies have highlighted the contribution of Natural Language Processing (NLP) techniques towards the automatic analysis of clinical texts to assist the coder with the selection of ICD codes (see section 2). However, we claim that classifications like the ICD are not the most relevant resources for semantic indexing, due to (1) their uniaxiality: they contain a finite set of codes restricted to one aspect of medicine – i.e. 14,000 codes for ICD-9-CM and 69,000 for ICD-10-CM – and (2) their lack of compositionality: each code is mutually exclusive and can not be used in combination with any other code to represent a diagnosis.

On the opposite, the Systematized NOmenclature of MEDicine – Clinical Terms (SNOMED-CT), released by the International Health Terminology Standards Development Organisation (IHTSDO), provides a broad set of 310,000 medical concepts organized into 19 axes, and allows the combination of these codes in order to render a more complex concept. Such a terminology is said to be post-coordinated and multi-axial. It is thus more appropriate to represent the content of medical records, as it is more precise and versatile than other coding systems. To the best of our knowledge, not much research has been carried out so far in relation to automatic SNOMED-CT coding: a few studies exists for English, but even fewer coding systems exist for French (see section 2).

In this context, the research project iMediate (Interoperability of Medical Data through Information Extraction and Term Encoding) aims to deal with medical reports written in French to generate a structured representation of the patient's medical pathway. On that basis, several high-valued applications (a semantic search engine, a classifier to assist MBDS coding, etc.) will be developed.

In this paper, we only focus on one objective of the project, namely the automatic annotation of unstructured texts with SNOMED-CT concept identifiers. To this aim, we present a flexible extraction method based on a set of generic finite-state graphs and existing terminological resources for French (see section 3). The performance of this system is evaluated in section 4 and compared with human judgement on the one hand, a simple string matching method on the other hand. We then discuss the results in section 5 and provide some ways for further improvement.

2. Related work

There is considerable research addressing the automatic identification of medical concepts, more specifically diseases and procedures. However, these studies are generally limited to the identification of lexical items and their

membership to general classes of interest such as diagnoses, medical procedures and anatomical parts. They rarely go as far as establishing a link between these arbitrary character strings and a knowledge base used as a conceptual backbone. In this regard, recent developments in the field of information extraction aim to make a reasoned use of ontologies in order to improve the annotation process (Wimalasuriya and Dou, 2010). Systems developers can build upon such knowledge-based annotation to create high-valued applications such as drugs adverse-effects monitoring (LePendu et al., 2012) or decision support tools for clinical trial design (Matskanis et al., 2012).

Several approaches are available to carry out this extraction and normalization task. The first one consists in training a statistical model on an annotated corpus (Kate, 2013). However, the training resource used to this aim must be representative of the (sub)languages, diversity of codes and type of texts the program will process, and collecting such a corpus requires significant efforts. The second approach is based upon the development of extraction rules or lists of terms. From that perspective, a lot of research has been carried out for English, given the rather impressive amount of high-quality linguistic resources available. As already mentioned, these works focus either on the ICD (Friedman et al., 2004; Crammer et al., 2007; Goldstein et al., 2007; Farkas and Szarvas, 2008), the UMLS (Aronson, 2001; Zou et al., 2003; Bashyam et al., 2007; Dai et al., 2008) or SNOMED-CT (Jung et al., 2009; Lee et al., 2010).

As regards the SNOMED-CT encoding, Patrick et al. (2007) used a word-by-word comparison matrix combined with well-known NLP techniques such as stopwords removal, stemming and variants generation on the basis of the UMLS Specialist Lexicon. The challenge of abbreviations and negation detection was also addressed. Although no evaluation has been conducted, the authors noted the strong potential of post-coordination in order to take into account the descriptive nature of medical texts.

More recently, Khare et al. (2012) showed that SNOMED-CT can improve homogeneity and interoperability of medical data encoded through encounter forms. The coding workflow hinges on string-matching techniques and contextual hints typical of data encoded through computer forms. However, this study moves somewhat away from our research, as we focus on authentic texts from medical health records.

Some research has also been carried in the French-speaking community, where there is a growing interest in biomedical NLP to foster interoperability of clinical data (Pereira et al., 2006; Delbecque and Zweigenbaum, 2007; Medori, 2008; Pereira et al., 2009; Kevers and Medori, 2010; Névéol et al., 2012). Again, these works concentrate on ICD, MeSH or UMLS coding, with varying degrees of human-computer interaction.

Pereira et al. (2009) present a methodology for indexing medical texts on the basis of several controlled vocabularies, such as French versions of SNOMED 3.5, ICD-10-CM and MeSH. The established mappings between these resources give the opportunity to select which knowledge resource will actually be used to output the results. This system relies on a dictionary of medical terms enriched with variants and uses a bag-of-words approach combined with several NLP treatments such as stemming. Although the benefits of using multiple terminologies is clearly emphasized, the accuracy of SNOMED-CT coding is not evaluated per se, as the authors compare the ICD-10-CM transcoding with codes produced by an existing commercial software. Such a method is quite problematic, as the ICD is used in financial and epidemiological contexts, and is not as precise as SNOMED-CT. Our own practice showed that some ICD-10-CM codes can group dozens of SNOMED-CT codes.

3. Workflow description

Our annotator aims at coding French medical texts using SNOMED-CT. As outlined above, most of medical terminological components available in French are designed for classification purposes and do not systematically reflect natural language as actually used in medical texts. Consider for example the entry *Diabète sucré, sans précision, avec complications non précisées* "Unspecified diabetes mellitus with unspecified complications", which is unlikely to appear in authentic texts. Other expressions, such as *diabète sucré suivi de complications* or *diabète sucré compliqué* are more likely to occur. In our opinion, this is a rationale to implement a workflow that can deal with such language variation. As we shall show, our rule-based system can be used to recognize medical terms that are not explicitly present in the initial terminological resource.

While it is often argued that rule-based systems require a lot of development work and maintenance, we avoid this pitfall by designing four generic, but flexible extraction patterns using the corpus-processing toolbox Unitex (Paumier, 2015). These rules are applied in parallel with custom-made dictionaries generated on the basis of circa 212,800 preferred terms and synonyms coming from existing terminological databases[1]. Given the fact that no appropriate resources exist to train a specific machine learning algorithm for such a coding task, we consider our program as an interesting and practical alternative to statistical models.

3.1. Generation of the extraction resource

We first have to generate the dictionary into the format required by the Unitex software (see figure 1). This step relies on processes commonly used in NLP workflows, which will also be performed on the input text: tokenization, POS-tagging (Schmid, 1994), stopwords removal and stemming (Porter, 1980). However, other specific treatments are necessary: removal of generic phrasings used in classification systems (*non classifié ailleurs* "not elsewhere classified", *SAI* "NOS", etc.), conversion of French diacritics into unaccented letters, lower case conversion

[1] As there are only partial French translations of SNOMED-CT, an automatic mapping of existing terminologies in French using the UMLS and other resources disseminated in the Belgian medical community – i.e. ICD-9-CM, ICD-10-CM and ICPC-2 – was carried out. Another part of the project is dedicated to semi-automatic terminological enrichment, which is the subject of a separate communication (Lemaire et al., 2015).

(with the exception of some kind of acronyms[2] and specific tokens that must not be considered as stopwords[3]). Beyond that, some terms may benefit from automatic variants generation: consider the entries containing Roman numerals, which can be converted into Arabic numerals and vice versa. Finally, we make sure that some terms can be written using a space, a dash or being agglutinated (e.g.: *cardiomyopathie*, *cardio-myopathie*, etc.).

As shown in figure 1, each stem in the dictionary is associated with two semantic tags (separated by a +): (1) a key referring to one or more SNOMED-CT concept identifier(s) and (2) the minimum number of tokens needed to complete the term. In other words, it means that a stem bearing the tag "2" must co-occur with another stem sharing the same concept key and also being part of a term composed of 2 lexical stems (e.g.: *choc/réaction anaphylactique/allergique* "anaphylactic/allergic choc/reaction")

```
reaction,.IM19VRIcCKPQ+2
anaphylact,.IM19VRIcCKPQ+2
choc,.IM19VRIcCKPQ+2
allerg,.IM19VRIcCKPQ+2
```

Figure 1: Unitex-compliant dictionary entries for the French terms *anaphylactic reaction* and *allergic choc*

3.2. Rule-based fuzzy string matching

The next step is to develop extraction and normalisation rules. To this end, we harnessed the opportunities offered by the so-called "morphological mode" implemented in Unitex to (1) perform queries using dictionaries and (2) set additional conditions that must be fulfilled to extract the sequence.

Figure 2: Unitex graph for the extraction and coding of sequences of two stems

In Figure 2, for example, the lexical mask <2> involves a dictionary lookup for stems having the semantic tag "2", i.e. stems that can compose a term of two tokens. The dictionary-entry variable `a` can be used to access information from the dictionary, and the pattern `$a.CODE.GRAM$` returns the first semantic code in the dictionary related to the current stem (i.e. the concept key), which is then stored in the `t1` variable. The box `insert` calls a subgraph allowing the insertion of one or more function words, including punctuation and blank spaces. When the second stem is recognized following the same principle, we check whether or not the two stems are associated to the same concept code through the comparison function `$t1.EQUAL=t2$`, which blocks the grammar exploration if the value stored in variables `t1` and `t2` is different, but allows the transducer to output the concept ID otherwise.

There are three other similar graphs, respectively recognizing sequences of 1, 3 or 4+ tokens, but the latter differs in that it includes a recursive step to detect an unlimited number of acceptable tokens until the exploration of the grammar is blocked by an `EQUAL` expression, assuming that the alignment of 4 or more stems linked to the same concept key is not a matter of chance.

Finally, we designed a graph for the detection of negated diseases or diagnoses. In that regard, we found the work of Chapman et al. (2013) very useful, as they provide a multilingual negation lexicon. Nevertheless, the latter has been adapted to fit with some peculiarities of our data.

Output evaluation	# of occurrences
Full detection	***179 (69.38%)***
Exact match	103 (39.92%)
Post-coordination	61 (23.64%)
Less specific	1 (0.39%)
More specific	2 (0.78%)
Incorrect	12 (4.65%)
Partial detection	***79 (30.62%)***
Less specific	30 (11.63%)
Incomplete or wrong	49 (18.99%)
Total	***258 (100%)***

Table 1: Performance analysis for extraction and coding

4. Evaluation

4.1. Terms extraction and accuracy of coding

As defining a reference annotation was necessary, two pairs of medical experts had to highlight all the relevant terms in a set of anonymized discharge letters issued by the gastroenterology service at the University Hospital Saint-Luc (Belgium). We then applied our method and asked the experts to submit a common terms set, possibly including the few automatically extracted terms missing from their initial list. Due to the considerable efforts required by the whole evaluation procedure, we were only able to annotate 5 discharge summaries, providing us with a reference set of 258 clinical terms (including 244 hapaxes). As we will show, evaluating the performance of our annotator using binary metrics of precision and recall may be too restrictive, as they do not take into account the hierarchical nature of SNOMED-CT, or the opportunity to decompose a medical entity in terms of atomic concepts. Therefore, table 1 presents the results in terms of finer-grained categories.

Our system was able to extract and code accurately 39.92% of the sequences. A number of reasons may explain this: above all, existing French terminological components are only partially representative of the language used in medical reports, confirming our intuition about the high degree of linguistic variation in medical reports. We noticed that most fully recognised terms consist in one or

[2]"LES", standing for *lupus érythémateux systémique* "systemic lupus erythematosus", would be confused with the French article *les* "the".

[3]The letter *a* in "vitamine A" might be considered as the 3rd-person singular present indicative of the French verb avoir "to have", which is commonly considered as a stopword. But its systematic removal would result in a loss of information.

two tokens, and are thus not conducive to syntagmatic and paradigmatic variation as we intended to address in this paper. This is why we decided, to better gauge the contribution of our approximate recognition method, to conduct a second evaluation focusing on long terms only (see section 4.2).

The category "Post-coordination" is table 1 shows that our program failed to extract the whole sequence in nearly 25% of the cases, but was able to identify all the constituents. One example is the term *[douleur] au niveau de la [partie supérieure de l'abdomen] en [ceinture] [sous-costale] avec [irradiation] dans le [flanc droit]*. This is quite an unexpected outcome, which shows that an important amount of medical terms tends to use a compositional structure, made of a core concept accepting one or more modifier(s). At the moment, our approach only deals with pre-coordinated expressions, while our error analysis clearly puts forward the issue of automatic post-coordination of SNOMED-CT concept identifiers. In our opinion, post-coordination might alleviate the limited availability of terminological components for French.

In few cases, the given codes were too specific. These errors originate from an incorrect mapping between the initial terminological resources and SNOMED-CT. Finally, incorrect cases can be explained by the multiaxiality of SNOMED-CT: the term *alcool* "alcohol", for example, can refer to the alcohol dependency (*disorder*), a drinking behavior (*finding*), the beverage (*substance*) or even the antiseptic solution (*substance*). At this time, our program does not operate disambiguation, which leads to an increase in the number of incorrect SNOMED-CT codes.

As regards the partial detections, we made the distinction between two cases when accessing a SNOMED-CT code. Either the experts established an "IS-A" relationship (meaning that the sequence to be detected is a subclass of the sequence detected in terms of the SNOMED-CT identifier assigned by the system), or the code was clearly rejected. The first case appears when the main constituent is recognised and correctly coded, as with the sequence *adénocarcinome transitionnel de la vessie* ("transitional adenocarcinoma of bladder"), which is actually coded on the basis on the longest substring encoded in the terminological resource, i.e. *adénocarcinome* ("adenocarcinoma"). In fact, the given code is not totally wrong, as a transitional adenocarcinoma of the bladder is a subtype of adenocarcinoma, but is not sufficient to capture the whole sequence's semantic.

4.2. Benefits of approximate string matching

As we mentioned earlier, our first evaluation could not properly evaluate the performance gain of our annotator compared with a mere dictionary look-up. Assuming that longer terms are more prone to syntagmatic and paradigmatic variation, we decided to collect a corpus of 124 discharge letters similar to those used for the initial evaluation, and implemented a version of the annotator that extracts sequences of at least 3 content words.

Our program extracted 389 sequences that were evaluated for both term completeness and code correctness, using the same evaluation scheme as explained in section 4.1.

Output evaluation	# of occurrences
Complete term	***383 (98,46%)***
Exact match	253 (65.04%)
Less specific	52 (13.37%)
More specific	37 (9.51%)
Incorrect	41 (10.54%)
Incomplete term	***6 (1.54%)***
Total	***389 (100%)***

Table 2: Performance analysis for sequences of 3 lexical words or more

The results reported in table 2 show that 65% of the codes associated with these long terms are accurately coded, and that in 13% of the cases the code is still correct, though not sufficient to cover all the details entailed in the French term.

Unsurprisingly, the proportion of sequences susceptible to be conceptualised through post-coordinated expressions is much lower, as terms of 3 content words or more are already more complex. The few occurrences we found concern the notion of degree (*déviation* importante *de la cloison nasale* "significant deviation of nasal septum").

A significant proportion of codes are either too specific or more general than the concept actually represented in the text. We found out that some errors were due to an excess of flexibility regarding very frequent terms (cancer, adenocarcinoma, etc.) when generating the dictionary, but, once again, it is also mainly due to problematic mappings to SNOMED-CT.

As a final step, we quantified the gain brought by our flexible recognition method by comparing the sequences extracted by our annotator with those obtained using a baseline approach that only relies on stemming and does not allow any other changes. We found out that the latter is only able to extract 180 occurrences (46%) out of the 389 detected by our method. The sequences only detected by the flexible approach include (1) *cancer bronchique épidermoïde du lobe droit*, recognized on the basis of the entry *carcinome épidermoïde bronchique du lobe moyen du poumon droit* "squamous cell carcinoma of bronchus in right middle lobe", (2) *oedème aigu pulmonaire* on the basis of the entry *oedème pulmonaire aigu* "acute pulmonary oedema", (3) *varices oesophagiennes* on the basis of the entry *varice de l'oesophage* "varice of esophagus" or (4) *les fonctions rénales sont normales* on the basis of the entry *fonction rénale : normale* "renal function: normal".

5. Conclusion

This paper presents a rule-based SNOMED-CT coding workflow for French medical texts. It relies on existing terminologies and on a limited set of generic extraction rules that allow some flexibility during the recognition process. The current system was able to extract and code 40% of terms extracted by medical experts, and opportunities for improvement were also brought forward. Our manual error analysis shows that French terminological components distributed among the medical community are not completely sufficient to support NLP tasks aimed at deal-

ing with all the peculiarities of medical texts, as far as SNOMED-CT coding is concerned. However, our flexible approach turned out to be extremely useful to deal with a part of this variation, as it is able to recognize twice as much long sequences in comparison with a more basic pattern-matching algorithm that only uses stemming. Further perspectives would be to enrich the initial terminological database with variants collected in a corpus of authentic medical texts. Our results also reveal interesting perspectives from the point of view of automatic post-coordination of SNOMED-CT codes on the basis of medical texts.

Acknowledgements

The iMediate project is funded by the Brussels Institute for Research and Innovation (Innoviris) within the framework of the "Bridge – Strategic Platforms 2013 : e-health" program.

References

Aronson, A. R., 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. of the AMIA Symposium*:17–21.

Bashyam, V., G. Divita, D. B. Bennett, A. C. Browne, and R. K. Taira, 2007. A normalized lexical lookup approach to identifying UMLS concepts in free text. *Studies in Health Technology and Informatics*, 129(Pt 1):545–549.

Crammer, K., M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll, 2007. Automatic code assignment to medical text. In *Proc. of the Workshop Biological, Translational, and Clinical Language Processing*, BioNLP '07. ACL.

Dai, M., N. H. Shah, W. Xuan, M. A Musen, S. J. Watson, B. D. Athey, F. Meng, et al., 2008. An efficient solution for mapping free text to ontology terms. *AMIA Summit on Translational Bioinformatics*, 21.

Delbecque, T. and P. Zweigenbaum, 2007. MetaCoDe: A Lightweight UMLS Mapping Tool. In R. Bellazzi, A. Abu-Hanna, and J. Hunter (eds.), *Artificial Intelligence in Medicine*, volume 4594 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 242–246.

Farkas, R. and G. Szarvas, 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9(3):S10.

Friedman, C., L. Shagina, Y. Lussier, and G. Hripcsak, 2004. Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of AMIA*, 11(5):392–402.

Goldstein, Ira, Anna Arzumtsyan, and zlem Uzuner, 2007. Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. *AMIA Annual Symposium Proceedings*, 2007:279–283.

Jung, S., S. Kim, S. Yoo, and J. Choi, 2009. Toward the Automatic Generation of the Entry Level CDA Documents. *Journal of Korean Society of Medical Informatics*, 15(1):141–151.

Kate, R. J., 2013. Towards Converting Clinical Phrases into SNOMED CT Expressions. *Biomedical Informatics Insights*, 6(Suppl. 1):29–37.

Kevers, L. and J. Medori, 2010. Symbolic Classification Methods for Patient Discharge Summaries Encoding into ICD. In H. Loftsson, E. Rögnvaldsson, and S. Helgadòttir (eds.), *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 197–208.

Lee, D., F. Lau, and H. Quan, 2010. A method for encoding clinical datasets with SNOMED CT. *BMC Medical Informatics and Decision Making*, 10(1):53–64.

Lemaire, N., T. François, J.-C. Debongnie, D. De Meyere, B. Fauquert, T. Klein, C. Fairon, and M. Van Campenhoudt, 2015. L'enrichissement terminologique d'usage du projet iMediate : une collaboration tripartite terminologie/TAL/sciences de la santé. In *Second Congrès international du Réseau de Lexicographie (RELEX)*. Universidade de Santiago de Compostela.

LePendu, P., S. V. Iyer, C. Fairon, and N. H. Shah, 2012. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. *Journal of Biomedical Semantics*, 3(Suppl 1):S5.

Matskanis, N., V. Andronikou, P. Massonet, K. Mourtzoukos, and J. Roumier, 2012. A Linked Data Approach for Querying Heterogeneous Sources - Assisting Researchers in Finding Answers to Complex Clinical Questions. In *Int. Conf. on Knowledge Engineering and Ontology Development*.

Medori, J., 2008. From Free Text to ICD: Development of a Coding Help. *Proc. of the Louhi Workshop, Turku*, 8.

Névéol, A., J. Li, and Z. Lu, 2012. Linking Multiple Disease-related Resources Through UMLS. IHI '12. New York, NY, USA: ACM.

Paumier, S., 2015. Unitex 3.1.Beta. User manual. http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf.

Pereira, S., P. Massari, A. Buemi, B. Dahamna, E. Serrot, M. Joubert, and S. J. Darmoni, 2009. F-MTI : outil d'indexation multi-terminologique : application l'indexation automatique de la SNOMED Int. In *Risques, Technologies de l'Information pour les Pratiques Médicales*, number 17 in Informatique et Santé. Springer Paris, pages 57–68.

Pereira, S., A. Névéol, P. Massari, M. Joubert, and S. Darmoni, 2006. Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. *Studies in Health Technology and Informatics*, 124:845–850.

Porter, M. F., 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of Int. Conf. on New Methods in Language Processing*, volume 12. Manchester, UK.

Wimalasuriya, D. C. and D. Dou, 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323.

Zou, Q., W. W. Chu, C. Morioka, G. H. Leazer, and H. Kangarloo, 2003. IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. *AMIA Annual Symposium Proc.*, 2003:763–767.