Automatic acquisition of synonyms from French UMLS for enhanced search of EHRs

Natalia GRABAR[a], Paul-Christophe VAROUTAS[b], Philippe RIZAND[b], Alain LIVARTOWSKI[b], Thierry HAMON[c]

[a] *Université Paris Descartes, Faculté de Médecine; Inserm, U872; SPIM, 75006 Paris, France*
[b] *Institut Curie, Département d'Information Médicale, 25 rue d'Ulm, 75005 Paris, France*
[c] *LIPN-UMR CNRS 7030, Université Paris-Nord, Avenue J.B. Clément, 93430 Villetaneuse, France*

Abstract. Currently, the use of Natural Language Processing (NLP) approaches in order to improve search and exploration of electronic health records (EHRs) within healthcare information systems is not a common practice. One reason for this is the lack of suitable lexical resources. Indeed, in order to support such tasks, various types of such resources need to be collected or acquired (*i.e.*, morphological, orthographic, synonymous). In this work, we propose a novel method for the acquisition of synonymous resources, which are mostly missing. This method is language-independent and relies on existence of structured terminologies. It enables to decipher hidden synonymous relations between simple words and terms on the basis of their syntactic analysis and exploitation of their compositionality. Applied to series of synonym terms from the French subset of the UMLS, the method shows 99% precision. The overlap between thus inferred terms and the existing sparse resources of synonyms is very low.
Keywords: Algorithm, Linguistics, Terminology, Data acquisition, Syntax, Lexicons

1. Introduction

The *Prométhée* framework [1], conceived and deployed at the *Institut Curie* [2], enables the aggregation, cross-interrogation, visualization and simple statistical analysis of heterogeneous biomedical data, including the hospital's electronic health records (EHRs). The framework has been launched in 2002 and is used on a daily basis by the institute's healthcare professionals, in various cognitive contexts [3]. Prométhée's current fulltext search engine is limited in that users must, within their queries, employ keywords which, at best, would match words and expressions directly used in health records. Guessing such 'best' keywords is a difficult task: language variation is rich and often unpredictable. Within the field of breast cancer, area of expertise of the *Institut Curie*, the International Classification of Diseases for Oncology (ICD-O) [4] is used to query structured elements of the EHRs. Even so, its terms do not allow to cover all the expressions occurring in EHRs.

In this work, we propose resources to improve the search in French language within *Prométhée*, when based on free-text queries or queries based on ICD-O categories. We shall first describe existing lexical resources suitable for enhancing *Prométhée's* fulltext search capacities. We then propose a novel high-quality method for deciphering elementary synonymous relations between simple terms or words. This method is language-independent and based on the identification of syntactic invariants.

2. Background

Within the NLP field, we distinguish, in respect with term processing levels, among morphological, orthographic and synonymous lexica. We notice that not all of these resources are equally well described for various languages and/or specialized areas. Currently, the morphological description of languages is the most complete and several languages are provided with at least inflectional lexica [5-6]. In the biomedical domain, we can mention the widely used lexicon of the UMLS system [7] for English, and similar resources for other languages [8-9]. However, few lexical resources can be found for the description of synonymous or orthographic relations. WordNet [10] proposes synonym relations for English, but the corresponding resources for other languages are not freely available. Besides, it has been shown that general lexica, for instance WordNet, are insufficient for specialized knowledge extraction [11]. Indeed, additional specialized information is crucial to improve their coverage. In order to provide a solution, we propose to exploit specialized terminologies, as several of them are created and continuously updated in the biomedical area, and decipher the hidden synonyms they contain.

3. Material

In the current work, we use the French subset of the UMLS [7] as the original resource from which elementary synonym relations are inferred. The goal of the UMLS is to merge various biomedical terminologies (currently over 100). It is a multi-purpose resource, which includes concepts and terms from many different source vocabularies developed for very different purposes. The whole UMLS contains currently over 1,3 million concepts, 71,883 of them have labels in French. Concepts can group several labels, with same or very close meaning. In addition to the UMLS, two available sets of French synonyms were used for comparison purpose: (1) set of general synonyms comes from the general-language dictionary *Le petit Robert* [12], which lists 140,141 pairs of single-word synonyms; (2) specialized set of synonyms extracted from the Masson medical dictionary, formerly online [13]; it contains 831 pairs of single-word synonyms.

4. Methods: Acquisition of a synonymous lexicon

Within UMLS, terms can show the compositionality (substitution of their components):

C0016627: *{grippe, influenza, peste}* in *Grippe aviaire*, *Influenza aviaire, Peste aviaire*

C0016627*: {...}* in *Avian influenza*, *Fowl plague*

C0038352: *{stomacal*, *gastrique}* in *Contenu gastrique*, *Contenu stomacal*

C0038352*: {gastric, stomach} in Gastric contents, Stomach contents*

We propose a method for the generalization of this observation which allows to acquire a specialized lexicon of elementary synonym relations. We shall refer to the series of synonym terms as *original synonym relations,* and to the series of their substituted components as *induced* or *elementary synonym relations.* As in the given examples, this method exploits the compositional structure of terms and relies on existence of structured

terminologies. The notion of compositionality, central within this method, assumes that the meaning of a complex expression is fully determined by (1) its syntactic structure, (2) the meaning of its parts and (3) the composition function [14]. In order to exploit this principle, terms are first analyzed syntactically into head and expansion components, then specific inference rules are applied. Finally, the obtained results are evaluated.

Preprocessing of terminology: the Ogmios platform. The aim of the terminology preprocessing step is to provide syntactic analysis of terms. Such analysis is crucial for our work: the proposed method exploits syntactic dependency relations and is based on syntactic invariants. We use the Ogmios platform [15,16], which is suitable for the processing of large amounts of data and, moreover, can be customized for a specialized linguistic domain. First, the TagEN [17] tool is used for the recognition of named entities, in order to help the forthcoming segmentation into words and sentences. Then, POS-tagging (assignment of syntactical categories like Noun, Verb, Adjective) and lemmatization (definition of the normalized form of words: *cancers* => *cancer*) was realized with TreeTagger [18]. The step of syntactic parsing of terms is carried with the rule-based term extractor YATEA [19]. The syntactic dependency relations between term components (head and expansion) are computed according to assigned POS tags and parsing rules defined within YATEA.

Acquisition of a synonymous lexicon. The present method is inspired by [14,11]. In [11], authors proposed to apply the semantic compositionality principle for inferring synonymy relations (*rel*) between complex terms. They then postulated that the composition process preserves synonymy and that the compositionality principle holds for complex terms. Roughly, this means that if the meanings *M* of two complex terms *A rel B* and *A′ rel B* are given by the following formulas : $M(A\ rel\ B) = f(M(A), M(B), M(rel))$ and $M(A'\ rel\ B) = f(M(A'), M(B), M(rel))$ for a given composition function *f*, and if *A* and *A′* are synonymous $(M(A) = M(A'))$, then the synonymy of the complex terms can be inferred: $M(A'\ rel\ B) = f(M(A'), M(B), M(rel)) = f(M(A), M(B), M(rel)) = M(A\ rel\ B)$. In the current work, we assume that the inverse function f^{-1} exists and, given synonymous complex terms, can be applied for deducing elementary synonym relations. Our approach takes into account the internal structure of the complex terms. We assume that the syntactic dependency relation between components is preserved through the compositionality principle. Thus, we can infer elementary synonym relations between components of two terms if: (1) analyzed terms are synonymous; (2) these components are located at the same syntactic position (head or expansion) and have the same POS tag; (3) the other components within terms are either synonymous or identical. Parsed terms are represented as a terminological network, within which deduction of the elementary synonym relations is based on the three rules:

Rule 1: If both terms are synonymous and their expansion components are identical, then an elementary synonym relation is inferred between head components. For instance, we can infer the synonym relation {*grippe*, *influenza*} (*influenza*) from the original synonym relation between terms *Grippe aviaire* (*Avian influenza*) and *Influenza aviaire* (*Avian influenza*) where the expansion component *aviaire* (*avian*) is identical.

Rule 2: If both terms are synonymous and their head components are identical, then an elementary synonym relation is inferred between expansion components. For instance, we infer the synonym relation {*gastrique*, *stomacal*} (*gastric, stomach*) from the synonym relation between terms *Contenu gastrique (Gastric contents)* and *Contenu stomacal (Stomach*

contents) where the head component *contenu (contents)* is identical.

Rule 3: If both terms are synonymous and either their head or expansion components are synonymous, then an elementary synonym relation is inferred. For instance, we infer the synonym relation {*pustuleux, vésiculeux*} (*aphthous*, v*esicular*) from the synonym relation between terms *Angine pustuleuse* (*Aphthous pharyngitis*) and *Pharyngite vésiculeuse* (*Vesicular pharyngitis*) where the head components {*angine*, *pharyngite*} (*pharyngitis*) are already known synonyms.

Evaluation. We perform manual validation of the inferred relations between words and simple terms. Each pair is examined, as well as its source series of synonyms. The accuracy of the inferred pairs is thus computed. Moreover, we perform comparison of the acquired synonyms with the existing similar resources: general synonyms from *Le petit Robert* and medical synonyms from *AtMedica* and UMLS and compute the overlap between them.

5. Results and Discussion

Preprocessing of terminology: the Ogmios platform. 156,404 terms, corresponding to 71,883 UMLS concepts have been fully parsed through the Ogmios platform. The original synonym relations were used to infer elementary relations.

Acquisition of a synonym lexicon. The three rules have been applied to the terminological network formed with 76,240 original synonym terms (54,058 original synonym pairs) and generated 1,196 pairs of elementary synonym relations. The general observation is that only three inferred pairs ({*affection*, *maladie*} (*affection*, *disease*), {*maladie*, *syndrome*} (*disease*, *syndrome*) and {*cancer*, *tumeur maligne*} *(cancer, malignant tumor))* are inferred on 10 to 14 series of original synonyms, while the majority of them are supported by singular series of terms. The acquired synonym pairs can be classified according to their linguistic features, for instance:

- Orthographic variants: {*acathisie*, *akathisie*} *(akathisia)*, {*embolie*, *embole*} (*embolus*)

- Abbreviations: {ARN*t, ARN transfert*} (*tRNA, transfer RNA*), {*biop, biopsie*} (*biopsy*), {EEG*, electro-encéphalogramme*} (*EEG, Electroencephalogram*)

- Named entities: {*bartholin*, *duverney*}, {*côlon, valsalva*}, {*saint jean, rhumatismale*}

- Ellipse: {*insuffisance artérielle, insuffisance*} (*artery insufficiency, insufficiency*), {*adrénergiques, récepteurs adrénergiques*} (*adrenergic*, *adrenergic receptor*)

- Scientific *vs* popular words: {*maladie, pathologie*} (*disease, pathology*), {*abcès*, *empyème*} (*abscess*, *empyema*)

- Morphologically related words: {*spasmodique, spastique*} (*spastic*), {*vermiculaire, vermiforme*} (v*ermiform*)

- Most induced synonym pairs link entities for which no common formal features can be observed: {*augmentation volume, hypertrophie*} *(enlargement),* {*grave, sévère*} *(severe),* {*cancer*, *tumeur maligne*} *(cancer, malignant tumor) ...*

Evaluation. Manual evaluation of the totality of generated pairs has shown that 99.3% (n=1188) are correct, 0.08% (n=1) rejected and 0.6% (n=7) not known. The erroneous pair has been generated from the UMLS concept C0038814 and is due to a variation of

preposition semantics. But except this pair, the efficiency of the proposed method is very high. This is certainly due to the use of controlled terminological data. Moreover, the inferring rules strongly exploit the syntactic scheme within the syntactically analyzed terms. These factors contribute to the acquisition of high-quality synonym pairs. These results confirm that French medical terms within UMLS show compositional structure.

As already mentioned, a very large number of elementary synonymous pairs has been deduced on the basis of only one original pair of synonyms. Such small productivity within UMLS can indicate that the inferred pairs are valid in a small number of contexts, but this observation must be verified using other terminologies or corpora. Additionally, the method inferred several pairs composed of named entities (n=26), which use is reduced to the medical area and, possibly, to certain terms only. For instance:

- {*bartholin*, *duverney*} inferred from C0004768 *Glandes de Bartholin* and *Glandes de Duverney (Bartholin Glands, Duverney's gland)*,

- {*saint jean, rhumatismale*} inferred from C0152113 *Chorée rhumatismale* and *Chorée de Saint Jean (Rheumatic chorea)*

Comparison between the induced elementary synonymous pairs and existing synonyms shows that overlap is very low. Thus, we found only 36 common pairs with the directly available synonyms within UMLS, such as {*tumeur maligne*, *cancer*} (*malignant tumor, cancer*) or {*saignement*, *hemorragie*} (*bleeding*, *hemorrhage*). Thus, the proposed method is useful for it deciphers "hidden" synonyms which are otherwise not accessible. Similarly, we found 2 common pairs with the *AtMedica* resource and 105 with the *Le petit Robert* resource. In the first case, results point out the complementarity between different resources of synonyms from the medical area. As for *Le petit Robert* overlap, the overlap is bigger; still it covers only small part of both resources. It proves that general language resources contain specialized medical vocabulary, although it is not very rich. The difference between them is not surprising as their purpose, as well as addressees and aimed applications, are different. For instance, their use for terminology structuring and knowledge extraction has shown that such general lexica are insufficient for specialized domains [11] and should be completed with specialized resources. Indeed, specialized domains make use of concepts too specific to occur within a general language lexicon.

6. Conclusion and Perspectives

Within healthcare information systems, exploration of EHR content is a current and challenging field. Although the NLP approaches could be suitable to address this need, there is a huge need in various types of linguistic resources. For instance, semantic resources such as synonymous lexica are missing especially in specialized domains. In this paper, we propose a novel method for filling in this gap and inferring synonymous relations between words and simple terms. This method exploits the compositionality principle and relies on existence of structured terminologies. It applies a set of rules based on syntactic dependency analysis within terms. The proposed method has been applied to the UMLS subset of French terms. It provides high-quality results: the manual evaluation showed that over 99% of the inferred relations are correct. The comparison with the available resources of synonyms, such as those directly available in UMLS and sparse resources like *AtMedica*

and *Le petit Robert*, is very low. The observed differences seem to indicate that these resources should be combined within NLP tools.

In the near future, we plan to apply the inferred resource to the Institut Curie's EHR corpora (accessible via *Prométhée),* thus further evaluating it during the detection of new synonymous relations between terms used by the institute's healthcare professionals. We shall then predict their possible impact for users, through analysis of the *Prométhée* query logs. Finally, the detected (and validated) synonyms will be implemented within the *Prométhée* fulltext search engine, where they will play a dual role: (1) during the query parsing phase (the question submitted by the user is then compiled into a query, understandable by the search engine) they will permit textual query expansion and/or normalization, (2) during the query results visualization phase, they will support on-demand exploratory analysis and classification of textual results.

As for the method itself, we would like to mention again the fact that it is language-independent. This enables its application to other languages, knowledge domains or terminologies (*ie*, Snomed CT), as long as (1) the required linguistic processing can be realized and (2) synonym relations between complex terms are available.

References

[1] Varoutas PC, Rizand P, Livartowski A.: Using category theory as a basis for a heterogeneous data source search meta-engine: the *Prométhée* framework. In: AMAST 2006

[2] www.curie.net

[3] eHealth impact: study on Economic and Productivity Impact of eHealth, commissioned by the European Commission, Directorate General Society and Media. Case study: Institut Curie, Paris, France: Elios and Prométhée, www.ehealth-impact.org (2006)

[4] Fritz AG, Percy C, Jack A, Sobin LH, Parkin MD.: International Classification of Diseases for Oncology (ICD-O). (2000). OMS, Geneva Switzerland

[5] Burnage G.: CELEX - A Guide for Users. Centre for Lexical Information, University of Nijmegen (1990)

[6] Hathout N, Namer F, Dal G.: An experimental constructional database: the MorTAL project. In Boucher, P., ed.: Morphology book. Cascadilla Press, Cambridge, MA (2001)

[7] NLM: UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland. (2007) www.nlm.nih.gov/research/umls/.

[8] Schulz S, Romacker M, Franz P, Zaiss A, Klar R, Hahn U.: Towards a multilingual morpheme thesaurus for medical free-text retrieval. In: Medical Informatics in Europe (MIE). (1999)

[9] Zweigenbaum P, Baud R, Burgun A, Namer F, Jarrousse E, Grabar N, Ruch P, Duff L, Thirion B, Darmoni S.: Towards a Unified Medical Lexicon for French. In: Medical Informatics in Europe (MIE). (2003)

[10]Fellbaum C.: A semantic network of english: the mother of all WordNets. Computers and Humanities. EuroWordNet: a multilingual database with lexical semantic network 32 (1998) 209-220

[11]Hamon T, Nazarenko A.: Detection of synonymy links between terms: experiment and results. In: Recent Advances in Computational Terminology. John Benjamins (2001) 185-208

[12]Le petit Robert. (1990), Paris

[13] www.AtMedica.com

[14]Partee BH. In: Compositionality. (1984)

[15]Hamon T, Nazarenko A, Poibeau T, Aubin S, Derivière J.: A robust linguistic e platform for efficient and domain specific web content analysis. In: RIAO 2007, Pittsburgh, USA (2007)

[16]/www-lipn.univ-paris13.fr/~hamon/ALVIS/Debian/testing, part of the EU project Alvis www.alvis.info

[17]Berroyer J.F.: TagEN, un analyseur d"entités nommées : conception, développement et évaluation. Mémoire de D.E.A. d'intelligence artificielle, Universit Paris-Nord (2004)

[18]Schmid H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: International Conference on New Methods in Language Processing. (1994) 44-49

[19]Aubin S, Hamon T.: Improving term extraction with terminological resources. In: Advances in Natural Language Processing FinTAL 2006. Number 4139 in LNAI, Springer (2006) 380-387