Using a cross-language approach to acquire new mappings between two biomedical terminologies

Fleur Mougin[a], Natalia Grabar[b]

[a] LESIM, INSERM U897, ISPED, University Bordeaux Segalen, France
fleur.mougin@isped.u-bordeaux2.fr
[b] STL, UMR 8163, CNRS, University Lille 3, France
natalia.grabar@univ-lille3.fr

Abstract. The exploitation of clinical reports for generating alerts especially relies on the alignment of the dedicated terminologies, i.e., MedDRA (exploited in the pharmacovigilance area) and SNOMED International (exploited recently in France for encoding clinical documents). In this frame, we propose a cross-language approach for acquiring automatically alignments between terms from MedDRA and SNOMED International. We had the hypothesis that using additional languages could be helpful to complement the mappings obtained between French terms. Our approach is based on a lexical method for aligning MedDRA terms to those from SNOMED International. The concomitant use of multiple languages resulted in several hundreds of new alignments and successfully validated or disambiguated some of these alignments.

Keywords: biomedical terminologies, mapping, cross-language methods

1 Introduction

The semantic interoperability among the communicating systems involves the exploitation of terminological resources. However, the alignment[1] between some terminologies is not always available, despite the intensive research studies already performed. Indeed, the pairs of terminologies relevant to a given medical field may not be treated yet. For instance, when we look for the alignment between MedDRA (exploited in the pharmacovigilance area) and SNOMED International (exploited recently in France for the encoding of clinical documents), we can find nearly nothing. It is noteworthy that through the UMLS®, the current mapping between MedDRA and SNOMED International is only 31%, which seriously impedes the situation.

This study falls within the French project RAVEL (Retrieval And Visualization in ELectronic health records) in which it is necessary to link pharmacovigilance databases to information present in clinical patient documents. In this frame, we propose a lexical approach, which exploits cross-language knowledge, for acquiring automatically alignments between terms from MedDRA and SNOMED International.

[1] In this paper, we use equally the terms “alignment” and “mapping”

adfa, p. 1, 2011.

2 Background

MedDRA. The Medical Dictionary for Regulatory Activities[2] (MedDRA) has been designed for the encoding of adverse drug reactions chemically induced by drugs. It contains a large set of terms which are hierarchically structured. The 15.1 version of MedDRA used in this study is available in French, English and Spanish.

SNMI. The Systematized NOmenclature of MEDicine International[3] (SNMI) is a multi-axial terminology providing a very large coverage of the biomedical domain. This terminology is composed of concepts organized hierarchically. The English version of SNMI is included in the UMLS and the Spanish one can be created from the Spanish version of SNOMED CT (also in the UMLS). The French version of SNMI is made available by the national Agency of Shared Health Information Systems[4].

UMLS. The Unified Medical Language System® (UMLS) [1] includes two sources of semantic information: the Metathesaurus® and the Semantic Network. The former integrates over 150 terminologies, including MedDRA and SNMI. The version used in this study (2012AA) contains more than two million concepts which correspond to clusters of terms (and codes) coming from the different terminologies. The Semantic Network is a much smaller network of 133 semantic types organized in a tree structure. These semantic types have been aggregated into fifteen coarser semantic groups [2], which represent subdomains of biomedicine (e.g., **Anatomy**). Each Metathesaurus concept has a unique identifier (CUI) and is assigned at least one semantic type.

Related works. The mapping between terminologies and ontologies is an active research area independently of the application domain. The ontology alignment evaluation initiative[5] gathers a great number of researchers around this topic. In the biomedical area, researchers work also on the alignment of several terminologies. First of all, the existence of the UMLS and its intensive international exploitation testify about it [3]. However, few works have addressed the mapping between MedDRA and other resources, such as SNOMED CT. Four experiences in English aimed at improving the current alignment of these two terminologies by exploiting hierarchical relations [4,5] or simple synonyms and a decomposition of MedDRA terms [6,7]. We are not aware about existing works on the alignment between MedDRA and SNMI.
A few works have studied the alignment of terminologies in a cross-language context. For instance, multilingual resources such as WordNet or UMLS may be exploited in such a way [8,9]. Thus, the existing alignment in one language, which can be more complete than in other languages, may be exploited to sort out the alignment between terms from other languages. With this approach, the implicit information becomes explicit for other languages. Another example of the cross-language alignment exploits parallel corpora [10] in order to build bilingual dictionaries. In this work, the assumption is that if two words are mutual translations, then their more frequent collocates are likely to be mutual translations as well.

[2] https://meddramsso.com/

[3] http://www.ihtsdo.org

[4] http://esante.gouv.fr/snomed/snomed/

[5] Ontology alignment evaluation initiative, from: http://oaei.ontologymatching.org

In our work, we propose to exploit the cross-language context differently. We aim at generating novel alignments independently in three languages (French, English and Spanish). We then study the complementarity of the resulting alignments.

3 Methods

Step 1: generating mappings. We designed a lexical approach, which aligns MedDRA to SNMI terms. First, all these English, French and Spanish terms were segmented into words and then normalized according to: punctuation {*Atrioventricular block, complete*; *Atrioventricular block complete*}, variation of word order {*Edema Quincke's*; *Quincke's edema*}, stopwords {*Mycoplasma hominis pelvic inflammatory disease*; *Pelvic inflammatory disease due to Mycoplasma hominis*}, inflectional {*Cough decreased*; *Decreased coughing*} and derived {*Colon perforation*; *Perforation of colon*} forms, but also synonyms {*Angioleiomyoma*; *Angiomyoma*}. With this approach, we exploited several resources in each language (Table 1), in addition to the terms to be aligned: stopword lists, morphological and synonymy resources.

Table 1. Number of terms in MedDRA and SNMI and then in the lexical resources

	English	French	Spanish
MedDRA	72,867	66,092	65,435
SNMI	164,069	150,689	162,699
Stopwords	183	70	209
Morphological resources	90,583	155,468	17,520
Synonyms	101,805	14,914	35,214

Step 2: filtering mappings. The UMLS semantic groups (SGs) propose a partition of the UMLS concepts. We exploited this information for filtering out wrong mappings. We thus compared the SGs to which belong the UMLS concepts of MedDRA and SNMI terms. If they were not the same, we considered the proposed mapping as wrong and eliminated it. For example, a mapping was found between *Body mass index* (MedDRA) appearing in the UMLS concept *Body mass index procedure* (C0005893) and *Body mass index* (SNMI) part of the UMLS concept *Body mass index* (C1305855). This mapping was automatically removed because these two concepts belong to distinct SGs: **Procedures** and **Physiology**, respectively.

Step 3: comparing mappings between languages. We computed the number of alignments which are common between the different languages. We had the hypotheses that cross-language mappings could be helpful for multiple aspects: (1) enrichment: the alignments generated in other languages are exploited to complete the alignments acquired in French; (2) validation: an exact mapping (i.e., a mapping 1-1) found in multiple languages is more likely to be correct; (3) disambiguation: if a mapping 1-N is obtained in a given language while only one of these pairs is encountered in another language, this allows to eliminate the pair(s) which are found in only one language. We calculated the number of mappings, which satisfied our hypotheses.

4 Results

4.1 Mapping results

We distinguished three situations among the resulting alignments (Table 2):

- The aligned terms are part of distinct UMLS concepts, themselves belonging to distinct SGs. An example is the pair *Uroporphyrin* / *Uroporphyrins*. These terms are respectively part of the UMLS concepts C0202193 and C0042093, which belong to the SG **Procedures** and **Chemicals & Drugs**, respectively. Such alignments are automatically removed from the newly generated mappings;
- The aligned terms are clustered in a unique UMLS concept. For example, *Rash acneiform* and *Acneform eruptions* are part of the UMLS concept C0175167. In this situation, the generated alignments can be automatically considered as correct;
- The aligned terms are included in distinct UMLS concepts belonging to a unique SG. One such pair is *May-Hegglin anomaly* / *May Hegglin syndrome*. These terms are respectively clustered in the UMLS concepts C0340978 and C0272184, both belonging to the SG **Disorders**. Such alignments need to be evaluated manually.

Table 2. Number of generated mappings in each language

	Distinct SGs	Same UMLS concept	New	Total
English	493	3,230	1,135	4,858
French	250	1,506	1,400	3,156
Spanish	148	3,006	351	3,505

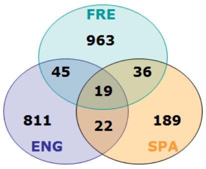

Fig. 1. Comparison of new mappings generated in English, French and Spanish

4.2 Comparing mappings according to the languages

Regardless of the languages, our approach results in 2,085 distinct new mappings between MedDRA and SNMI (Fig. 1). The mappings specific to a unique language complete those obtained in the two other languages. Few mappings overlap between the three languages. Indeed, only 6.2% of mappings found between MedDRA and SNMI terms involve more than one language. An example is the mapping between the MedDRA terms *Infection due to Mycobacterium fortuitum* (FRE: *Infection à Mycobacterium fortuitum*, SPA: *Infección por Mycobacterium fortuitum*) and the SNMI terms *Mycobacterium fortuitum infection* (FRE: *Infection à Mycobacterium fortuitum*, SPA: *Infección por mycobacterium fortuitum*), which are respectively part of the UMLS concepts C0275711 and C0877567. This low overlap is however helpful to validate 77 exact mappings and to disambiguate 42 mappings 1-N, which were found between MedDRA and SNMI terms. The previous example illustrates the "validation aspect". The MedDRA term *Familial tremor* can illustrate the "disambiguation aspect". It was mapped to the following SNMI terms in English: *Essential tremor*, *Persistent tremor* and *Congenital trembles* and in French: *Tremblement grossier* (i.e., *Coarse Tremor*) and *Tremblement essentiel* (i.e., *Essential tremor*). By combining the mappings generated in each language, we can conclude that the mapping between the MedDRA term *Familial tremor* and the SNMI term *Essential tremor* is the best one.

5 Discussion

Overall, the approach presented in this paper provided more than eleven thousands mappings between MedDRA and SNMI terms. 47.7% to 85.8% of these mappings were deemed correct automatically because they belong to a unique UMLS concept. More than two thousands of the remaining mappings are entirely new (because they are part of distinct UMLS concepts). Regarding our hypotheses, the complementarity of the results obtained in each language confirms the interest of using a cross-language approach for mapping purposes. Conversely, the overlap of new mappings according to the languages is very low. We assume this is due in part to the fact that the aligned terms remain specific in each language. We remind that this overlap was however useful to mutually validate or disambiguate some of the generated mappings.

For future works, we would like to exploit the compositional structure of MedDRA terms, as done in previous studies [6,7], for improving the mapping between MedDRA (which has complex and compositional terms) and SNMI (which has syntactically more simple terms). Finally, a manual validation of new mappings should be performed by medical experts.

Acknowledgments. The authors acknowledge the support of the French ANR and the DGA, under grant Tecsan (ANR-11-TECS-012).

References

1. Lindberg, D.A., Humphreys, B.L., McCray, A.T. The Unified Medical Language System. Methods Inf Med. 32(4):281‑291 (1993)
2. Bodenreider, O., McCray, A.T. Exploring semantic groups through visual approaches. J Biomed Inform. 36(6):414‑432 (2003)
3. Fung, K.W., Bodenreider, O. Utilizing the UMLS for semantic mapping between termino-lo-gies. AMIA Annu Symp Proc. 266‑270 (2005)
4. Bodenreider, O. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. AMIA Annu Symp Proc. 45‑49 (2009)
5. Alecu, I., Bousquet, C., Mougin, F., Jaulent, M-C. Mapping of the WHO-ART terminology on SNOMED CT to improve grouping of related adverse drug reactions. Stud Health Technol Inform. 833‑838 (2006)
6. Nadkarni, P.M., Darer, J.D. Determining correspondences between high-frequency MedDRA concepts and SNOMED: a case study. BMC Med Inform Decis Mak. 10:66 (2010)
7. Mougin, F., Dupuch, M., Grabar, N. Improving the mapping between MedDRA and SNOMED CT. In Peleg, M., Lavrač, N., Combi, C. (eds.), AIME 2011, LNCS. Berlin, Heidelberg: Springer-Verlag; pp. 220‑224 (2011)
8. Malaisé, V., Isaac, A., Gazendam, L., Instituut, T., Brugman, H. Anchoring dutch cultural heritage thesauri to WordNet : two case studies. Proc. of the Workshop on Language Technology for Cultural Heritage Data. pp. 57 ‑ 64 (2007)
9. Merabti, T., Soualmia, L.F., Grosjean, J., Palombi, O., Müller, J-M., Darmoni, S.J. Translating the foundational model of anatomy into French using knowledge-based and lexical methods. BMC Med Inform Decis Mak. 11(1):65 (2011)
10. Och, F.J., Ney, H. Improved statistical alignment models. Proc. of the 38th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; pp. 440‑447 (2000)