Combining semantic and lexical methods for mapping MedDRA to VCM icons

Jean-Baptiste Lamy, Rosy Tsopra

Combining semantic and lexical methods for mapping MedDRA to VCM icons

Jean-Baptiste LAMY[a,1] and Rosy TSOPRA[a]
[a] *LIMICS (Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé), Université Paris 13, Sorbonne Université, Inserm, 93017 Bobigny, France*

Abstract. VCM (Visualization of Concept in Medicine) is an iconic language that represents medical concepts, such as disorders, by icons. VCM has a formal semantics described by an ontology. The icons can be used in medical software for providing a visual summary or enriching texts. However, the use of VCM icons in user interfaces requires to map standard medical terminologies to VCM. Here, we present a method combining semantic and lexical approaches for mapping MedDRA to VCM. The method takes advantage of the hierarchical relations in MedDRA. It also analyzes the groups of lemmas in the term's labels, and relies on a manual mapping of these groups to the concepts in the VCM ontology. We evaluate the method on 50 terms. Finally, we discuss the method and suggest perspectives.

Keywords. Terminology as Topic, MedDRA, Computer Graphics, Nonverbal Communication, Icons

1. Introduction

Medical terminologies are essential for semantic interoperability between health information systems. In order to facilitate the use of medical terminologies by clinicians, we have developed for more than ten years VCM (Visualization of Concept in Medicine). VCM is a compositional iconic language that proposes icons for representing the main medical concepts, including symptoms, disorders, risks, antecedents, treatments and follow-up procedure [3]. In VCM, icons are created by assembling colors and iconems (*i.e.* small parts of icons) according to a graphical grammar. There are about 150 iconems, which allow the creation of thousands of icons. These icons do not aim at being as precise as medical texts or terminologies, but rather aim at enriching texts and providing a broader overview or a visual summary, *e.g.* the "heart" pictogram can be easily searched for in a patient record. The semantics of VCM has been formalized with an OWL (Web Ontology Language) ontology [4].

However, these icons can only be included in user interface if mappings exist between medical terminologies and VCM. Mappings allow the display of icons instead of (or in addition to) the term labels. In a previous work, we proposed a semantic method for mapping SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) to VCM icons [5]. However, purely semantic methods can only be used when the two resources mapped are highly structured. This is the case for SNOMED CT, which is a light ontology, and VCM, which has a formal semantics. But many medical

[1] Corresponding Author; E-mail: jean-baptiste.lamy@univ-paris13.fr.

terminologies, such as MedDRA (Medical Dictionary for Regulatory Activities) and ICD10 (International Classification of Diseases, release 10), still lack a formal semantics. For these terminologies, semantic methods is not a valid option.

In this paper, we present a lexico-semantic method for mapping MedDRA to VCM. The method takes advantage of the hierarchical relations of MedDRA, but also analyzes the lemmas present in MedDRA terms and maps these lemmas to concepts in the VCM ontology. These concepts correspond to iconems. For example, the "cardiac" lemma can be mapped to the "heart" concept associated with the "heart" pictogram. We applied the proposed method to the cardiac chapter of MedDRA, and we evaluated a subset of 50 terms. Finally, we discuss the interest of the method before presenting some perspectives.

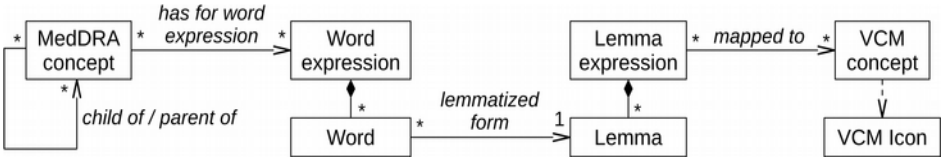

Figure 1. The main classes and relations in the ontology (displayed as a UML class diagram).

2. Methods

MedDRA is a medical classification often used for coding drug adverse effects. We worked on the French translation of MedDRA 18. MedDRA has 5 levels: system class organs (SOC), high-level group terms (HLGT), high-level terms (HLT), preferred terms (PT) and low-level terms (LLT). It is a multiaxial classification*:* a term can have more than one parent. Since LLT are mostly synonyms of PT, we worked only on the first 4 levels. We used PyMedTermino [6] for accessing MedDRA and Owlready [7] for managing ontologies.

First, we designed an OWL ontology including MedDRA terms with codes, labels and parent-child relations. Figure 1 shows the general structure of the ontology. We also added to the ontology all subsets of consecutive words present in each label (we call them *word expressions*) and we lemmatized them using the French Snowball lemmatizer (leading to *lemma expressions*). For example, the term "Coronary artery disorders" produces 6 lemma expressions: {"coronar", "arter", "disord", "coronar arter", "arter disord", "coronar arter disord"}. Stop words were also removed. Finally, the ontology allows mapping lemma expressions to VCM concepts, from the VCM ontology.

Then, we designed a lexico-semantic method for mapping a MedDRA term to VCM icons in 4 steps (see example for the MedDRA term "Cardiac perforation" in Figure 2). First, VCM concepts are gathered in two sets: containing VCM concepts associated with the parent terms (here, "Cardiac disorders NEC" and "Cardiac and vascular procedural complications") and containing VCM concepts associated with the lemma expressions in the term (here, "cardiac" and "perfor"; no mapping exists for "cardiac perfor"). Second, if some specific anatomo-functional location concepts *L* are present in (*e.g.* Heart), we remove from all location concepts that are not in *L* or one of their children (*e.g.* Vascular). Third, the union is computed. Fourth, one or more VCM icons are produced from *C*, using the VCM ontology and the previously

published method [4]. The resulting icon combines both information found lexically (*e.g.* the "Lesion and perforation" iconem) and through inheritance relations (*e.g.* the "Procedural complication" iconem).

The method is recursive: before treating a given term, it needs to treat all its parents, in order to gather the VCM concepts associated with the parent terms. In addition, when searching for lemma expressions in the label of a MedDRA term, longer lemma expressions are prioritized over shorter ones. For example, "auricular" is mapped to the Ear VCM concept and "auricular fibril(ation)" to the Heart rhythm and Pathology VCM concepts. When "auricular fibril" is matched, it consumes the "auricular" lemma, and thus the Ear VCM concept is not found.

We developed a user interface for manually mapping lemma expressions to VCM concepts. This interface allows (1) browsing MedDRA terms, with their associated VCM icons according to the previously described method, (2) listing the word and lemma expressions found in a given MedDRA term, (3) listing all lemma expressions present at a given depth in MedDRA (*e.g.* SOC or HLT), and (4) mapping a given lemma expression to one or several VCM concepts.

We applied this method to the cardiac SOC of MedDRA. One of the author (JBL) implemented the system and mapped the lemma expressions. To evaluate the system, we randomly chose 50 terms and a medical informatics expert (RT) manually mapped them to VCM. We gave to the expert the list of the 50 terms, with their ancestors, and a lexicon of the VCM language. Then, we compared the expert-produced gold standard with the icons generated by the lexico-semantic method.

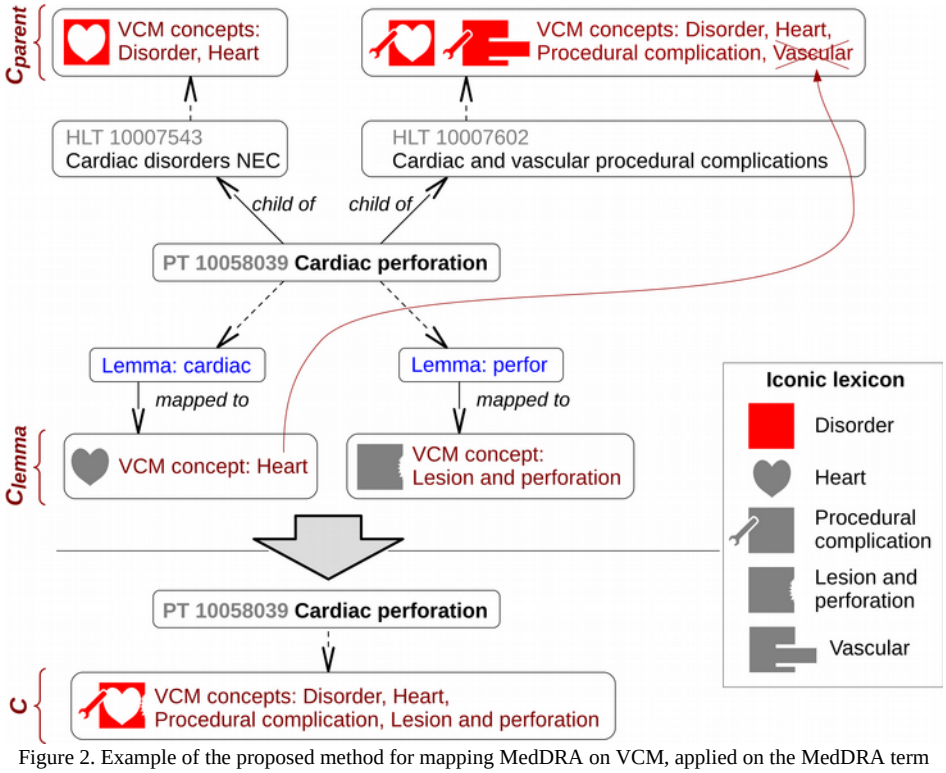

Figure 2. Example of the proposed method for mapping MedDRA on VCM, applied on the MedDRA term "Cardiac perforation".

3. Results

The proposed method was applied to the cardiac SOC of MedDRA, which includes 634 MedDRA terms (excluding LLT). The mapping involved 212 lemma expressions (123 with one lemma, 76 with two, 9 with three and 4 with four). The 634 terms were mapped to 114 different VCM icons. Most terms (541) were mapped to a single icon; 85 were mapped to two icons and 8 to three. For example, "Cardiac hypertensive complications" was mapped to two icons: cardiac disorders and hypertensive disorders.

Out of the 50 terms evaluated, the icons generated by our method was identical to the gold standard for 40 terms. For 6 terms, the generated icons were incomplete, *i.e.* the system generated only some of the icons proposed by the expert, or icons with fewer iconems. In several situations, the expert interpreted the MedDRA term, for example a term classified as procedural complications was interpreted as having potentially other possible causes. In another situation, the expert combined several VCM iconems that cannot be combined due to spatial overlap.

For 4 terms, the results were discordant. One discordance was related to mycoplasma infections, which were wrongly associated with fungal infection in the ontology. Two were related to tumors: the SOC "Neoplasms benign, malignant and unspecified" included the lemma "benign", which leads to classify all tumors as benign. We fixed this problem by mapping the lemma expression "neoplasm benign malign" to the VCM concept tumor without precision on malignancy. Finally, the last discordance was related to Carney complex, which is a rare disease with many clinical manifestations.

4. Discussion

In this paper, we proposed a lexico-semantic method for mapping MedDRA to VCM icons. The method combine the use of hierarchical relations in MedDRA with a mapping between lemma expressions and concepts in the VCM ontology. We worked on the French version of MedDRA; it might be interesting to compare the results when using other languages. The application of our method to the cardiac SOC of MedDRA required to map 212 lemma expressions, instead of 634 terms if a purely manual approach was followed. In addition, lemma expressions are shorter and thus easier to map.

In the literature, four categories of method have been proposed for mapping medical terminologies [9]: (1) designing the mapping manually, (2) chaining several existent mappings, typically using UMLS (Unified Medical Language System), (3) using lexical methods for matching identical or similar terms, and (4) using semantic ontology alignment. Many works have been published on terminology alignment. Recently, Souvignet *et al.* [11] chained alignments present in UMLS for mapping MedDRA to SNOMED CT. Fung *et al.* [2] compared lexical and semantic approaches on SNOMED CT and ICD10-PCS (Procedure Coding System), and showed that there is an interest in combining both. Mary *et al.* [8] combined lexical and semantic methods for extending an existing alignment between LOINC (Logical Observation Identifiers Names and Codes) and SNOMED CT, using the "anchor flooding" strategy.

Semantic methods usually consist of ontology alignment [10]: concepts are aligned when they have similar relationships. More advanced ontology alignment methods

combine structural and lexical approaches, such as BOAT (Biomedical Ontologies Alignment Technique) [1]. Lexical methods often use the "bag-of-words" model, in which the word order is not taken into account, contrary to the method we proposed. Lexical methods typically consist of aligning terms when they have identical or similar labels. However, this approach is not possible when aligning with icons such as VCM. We proposed a slightly different approach, in which the lemma expressions from a non-formalized terminology are mapped to the concepts of a formalized terminology. This approach could also be used for non-iconic terminologies, *e.g.* an ICD10 to SNOMED CT mapping could be designed by mapping lemma expressions from ICD10 to SNOMED CT organ and morphology concepts.

Perspectives of this work include the application of the method to the entire MedDRA terminology, the use of VCM icons in pharmacovigilance software, and the reuse of the mapping between lemma expressions and VCM concepts for mapping other terminologies to VCM, such as ICD10, or for generating VCM icons from free text.

5. Acknowledgments

This work was supported by the French National Research Agency (ANR) through the Pegase project [grant number ANR-16-CE23-0011].

References

[1] Chua, W.W.K., Kim, J.J.: BOAT: automatic alignment of biomedical ontologies using term informativeness and candidate selection. J Biomed Inform **45**(2), 337–49 (2012)

[2] Fung, K.W., Xu, J., Ameye, F., Gutiérrez, A.R., D'Havé, A.: Leveraging lexical matching and ontological alignment to map SNOMED CT surgical procedures to ICD-10-PCS. AMIA Annual Symposium proceedings **2016**, 570–579 (2016)

[3] Lamy, J.B., Duclos, C., Bar-Hen, A., Ouvrard, P., Venot, A.: An iconic language for the graphical representation of medical concepts. BMC Medical Informatics and Decision Making **8**, 16 (2008)

[4] Lamy, J.B., Soualmia, L.F.: Formalization of the semantics of iconic languages: An ontology-based method and four semantic-powered applications. Knowledge-Based System **135**, 159–179 (2017)

[5] Lamy, J.B., Tsopra, R., Venot, A., Duclos, C.: A Semi-automatic Semantic Method for Mapping SNOMED CT Concepts to VCM Icons. Stud Health Technol Inform **192**, 42–6 (2013)

[6] Lamy, J.B., Venot, A., Duclos, C.: PyMedTermino: an open-source generic API for advanced terminology services. Stud Health Technol Inform **210**, 924–928 (2015)

[7] Lamy JB: Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artif Intell Med **80**, 11–28 (2017)

[8] Mary, M., Soualmia, L.F., Gansel, X.: Usability and improvement of existing alignments: The LOINC-SNOMED CT case study. In: European Knowledge Acquisition Workshop (EKAW), Lecture Notes in Computer Science, vol. 10180, pp. 145–148 (2016)

[9] Saitwal, H., Qing, D., Jones, S., Bernstam, E.V., Chute, C.G., Johnson, T.R.: Cross-terminology mapping challenges: a demonstration using medication terminological systems. J Biomed Inform **45**(4), 613–25 (2012)

[10] Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. In: IEEE Transactions on Knowledge and Data Engineering, vol. 25, pp. 158–176 (2012)

[11] Souvignet, J., Declerck, G., Asfari, H., Jaulent, M.C., Bousquet, C.: OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval. J Biomed Inform **63**, 100–107 (2016).