Automatic methods for mapping Biomedical terminologies in a Health Multi-Terminology Portal

Tayeb Merabti*, Julien Grosjean *
Hocine Abdoune** Michel Joubert** Stefan Darmoni*

*CISMeF, Rouen University Hospital, Normandy &
TIBS, LITIS EA 4108, Rouen University Hospital, Rouen, France
nom.prenom@chu-rouen.fr,
http://www.cismef.org
**LERTIM EA 3283, Marseilles, Faculty of Medicine, Marseilles, France
nom.prenom@ap-hm.fr
http://cybertim.timone.univ-mrs.fr

Abstract. Terminology mapping is an important and crucial task to improve semantic interoperability between health care applications and resources. In 2009, CISMeF created a Health Multi-Terminological Portal (HMTP) to search concepts among all the health terminologies available in French (or in English and translated in French) included in this portal and to browse it dynamically. To map terminologies in the HMTP, two methods are used: (1) conceptual method which exploits various features of the UMLS, (2) lexical method based on natural language processing in French and English. A total of 199,786 mappings were performed between at least two French terms using conceptual method, whereas 266,139 mappings were performed using lexical methods. These mappings were all integrated in the HMTP developed by CISMeF. Conceptual and lexical methods were used to translate some English terminologies into French such as MEDLINEPlus, FMA and SNOMED CT.

1 Introduction

Biomedical terminologies and ontologies have proliferated during the past decade. Due to this proliferation, different health care systems use different biomedical terminologies. In this context, tools and methods to map biomedical terminologies are needed to solve data interoperability problems. The process of terminology mapping consists of identifying relationships or identical (or approximately identical) concepts between terminologies Wang et al. (2008). Various research teams have investigated automatic methods to produce high-quality mappings between terminologies Rocha et al. (1994); Fung and Bodenreider (2005); Bodenreider et al. (1998); Merabti et al. (2010b). The objective of this paper is to describe two types of approaches to map biomedical terminologies in English and French whether or not included into Unified Medical Language System (UMLS)Lindberg et al. (1993). These two approaches are currently implemented into a Health Multi-Terminological Portal (HMTP)Darmoni et al. (2010) developed by the CISMeF team Darmoni et al. (2000).

2 Material

2.1 Unified Medical Language Systems

UMLS Lindberg et al. (1993) integrates over 2 million concepts (2,200,159 in the 2010AA version) from 148 biomedical vocabularies. The UMLS is made up of three main knowledge components, but, for our purpose, we retained only the Metathesaurus: a very large, multi-purpose, and multilingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships between them. Each concept has a unique identifier in the Metathesaurus (Concept Unique Identifier, CUI).

2.2 CISMeF BackOffice & HMTP

The CISMeF BackOffice Darmoni et al. (2010) is a multi-terminological server developed by CISMeF to integrate and manage multiple terminologies. The HMTP[1] is a "Terminological Portal" connected to the CISMeF BackOffice to search concepts among all the health terminologies available in French (or in English and translated in French)[2] included in this portal and to browse it dynamically. A number of 27 terminologies and classifications were included in the CISMeF BackOffice, and therefore in HMTP. Some terminologies and classifications are included in the UMLS meta-thesaurus (n=9) but the majority are not (n=18). Table 1 displays the number of descriptors and relationships included in the HMTP.

Terminologies	27
Concepts	$\succ$ 867,791
Synonyms	$\succ$ 1,837,761
Definitions	223,654
Relations and hierarchies	2,990,365

TAB. 1 – *Main figures of the Health Multi-Terminology Portal*

3 Methods

Two automatic mapping approaches are implemented in the HMTP: conceptual and lexical approach. The conceptual approach uses the UMLS metathesaurus to map the terminologies included into the UMLS. The lexical approaches use some natural language processing tools to map terminologies whether or not included into the UMLS.

3.1 Conceptual Approach

This approach implies that each term to be mapped must be included into the Metathesaurus Joubert et al. (2009). The principle of the method is based on the conceptual construction of the UMLS metathesaurus. Three types of mapping are provided using this method:

[1] `http://pts.chu-rouen.fr/pts_site`

[2] Bilingual graphical user interface (French and English) was developed.

"Exact Mapping", "Broader Mapping" and/or "Narrow Mapping" and "Close Mapping" (see table 2 for some examples), this mapping method is inspired by SKOS (Simple Knowledge Organization System) definitions of mapping properties[3], SKOS language is also used to represent French health terminologies into the French Health Multi-terminological Server Darmoni et al. (2009). The mapping approach is as follows: suppose two terms t1 and t2 of two terminologies T1 and T2, respectively, suppose CUI1 and CUI2, the respective projections of t1 and t2 in the Metathesaurus, then t1 and t2 are mapped if :

- CUI1=CUI2, this corresponds to the "Exact Mapping",
- There is parent of t1 or t2 which maps t2 or t1 respectively, this corresponds th "Broad Mapping" and/or "Narrow Mapping",
- There is an explicit mapping between CUI1 and CUI2, this corresponds to the non-transitive "Close Mapping".

The algorithm is carried out sequentially and stops if a candidate mapping is found. As an application of this, even if an explicit mapping comes from other terminologies, e.g. ICD-9-CM and SNOMED CT Imel (2002) not part of the terminologies under consideration, it still applies to t1 and t2 since it is established between CUI1, to which t1 is attached, and CUI2, to which t2 is attached. In other words, explicit mappings between two terminologies can be "reused" for other terminologies by means of the UMLS concept structure Fung and Bodenreider (2005).

Type of relation	**Source term** (Terminology)	**Target Term** (Terminology)
Exact Mapping	Congenital bladder anomaly (MedDRA)	Congenital anomaly of the bladder, nos (SNMI)
Close Mapping	Diseases of lips (ICD10)	Ulcer of lip (SNMI)
BT-NT Mapping	Hepatic insufficiency (MeSH)	Liver disease, nos (ICPC2)

TAB. 2 – *Examples for each type of conceptual mapping*

3.2 Lexical Approach

In this approach, some Natural Language Processing (NLP) tools (English and French) are used to link terms from different sources. Lexical approach allows to find a term in the target terminology that is the most lexically similar to a given term in a source terminology. Two lexical algorithms were used in French and English to map all HMTP terminologies.

3.2.1 French based approach

This approach uses a French NLP tools and mapping algorithms developed by the CISMeF team to map French health terminologies Merabti (2010); Merabti et al. (2010a,b). These

[3]World Wide Web Consortium Simple Knowledge Organization System: `www.w3.org/2004/02/skos`

tools were used in a previous work and extended to link terms in multiple French Health terminologies:

- Remove stop words: frequent short words that do not affect the phrases such “a”, “Nos”, “of”, etc are removed from all terms in all terminologies.
- Stemming, a French stemmer provided by the “lucene” software library which proved to be the most effective for the F-MTI automatic indexing tools using several health terminologies Pereira (2007), as compared to the stemming tools developed by the CISMeF team.

Mapping used by this approach may provide three types of correspondences between all terms:

- Exact correspondence: if all words composing the two terms are exactly the same.
- Single to multiple correspondence: when the source term cannot be mapped by one exactly target term, but can be expressed by a combination of two or more terms.
- Partial correspondence: In this type of mapping only a part of the source term will be mapped to one or more target terms.

Examples for each type of mapping are given in Table 3. In this work, only the exact correspondence was described.

Type of correspondance	**Source term** (Terminology)	**Target Term(s)**(Terminology)
Exact	Syndrome de Marfan “Marfan Syndrome”(MeSH)	Syndrome de Marfan “Marfan’s Syndrome” (MedDRA)
Single to Multiple	Albinisme surdité “Albinism- deafness syndrome” (ORPHANET)	Albinisme “Albinism” (MeSH) and (+) Surdité “Deafness” (SNMI)
Partial	Chromosome 14 en anneau “Ring chromosome 14” (ORPHANET)	Chromosome humain 14 “Chromosome 14” (MeSH)

TAB. 3 – *Examples of the three types of mappings using the French lexical approach*

3.2.2 English based approach

In this approach we use some NLP tools in English developed by the NLM Browne et al. (2003). These NLP tools are a series of tools designed to aid users in analyzing and indexing natural language texts in the medical field McCray et al. (1994); Peters et al. (2010). They include essentially some tools like:

- LVG (Lexical Variant Generation): a Multi-function tool for lexical variation processing;
- Norm[4]: a program used to normalize English terminologies (UMLS terminology);
- WordInd: a tool used to tokenized terms into word.

In this work we basically used a normalization program ("Norm"). The Normalization process involves stripping genitive marks, transforming plural forms into singular, replacing punctuation, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words into alphabetic order. In the English base approach, only the exact correspondence was used.

4 Results

4.1 Conceptual Approach

A number of 199,786 mappings exists between at least two French terms from UMLS (25,833 (Exact Mapping), 69,085 (Close Mapping) and 104,868 (Broader and /or Narrower Mapping)). In contrast, from the 25,833 terms mapped "Exactly", 15,831 come from SNOMED International where only 296 come from ICPC2 (Table 4). The three types of mappings ("Exact","Broader" and/or "Narrow" and "Close") are included into the HMTP (see Figure 1).

Terminology	**Number of terms mapped**
ICD10	3,282
ICPC2	296
MedDRA	5,700
MeSH	10,637
SNOMED Int.	15,831
WHO-ART	1,392

TAB. 4 – *Number of terms from each terminology exactly mapped (conceptual approach)*

4.2 Lexical Approach

A number of 266,139 mappings exists between at least two terms from HMTP (English and French). Most of these mappings were evaluated in previous work Merabti et al. (2010a); Merabti (2010); Merabti et al. (2010b). For example, in Merabti et al. (2010a) the "Exact mapping" between ORPHANET and some French terminologies was evaluated and considered "relevant" in 98% of cases. Table 5 displays a fragment of the entire matrix mapping between all terminologies of the HMTP. Terminologies included in the HMTP in English and French were mapped using the two lexical approaches. For example, the terminologies: MeSH, SNOMED International, ORPHANET and ATC were mapped using English and French lexical approches. However, some terminologies were mapped using English (SNOMED CT,

[4]National Library of Medicine: Lexical Tools:
`http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2010/docs/userDoc/index.html`

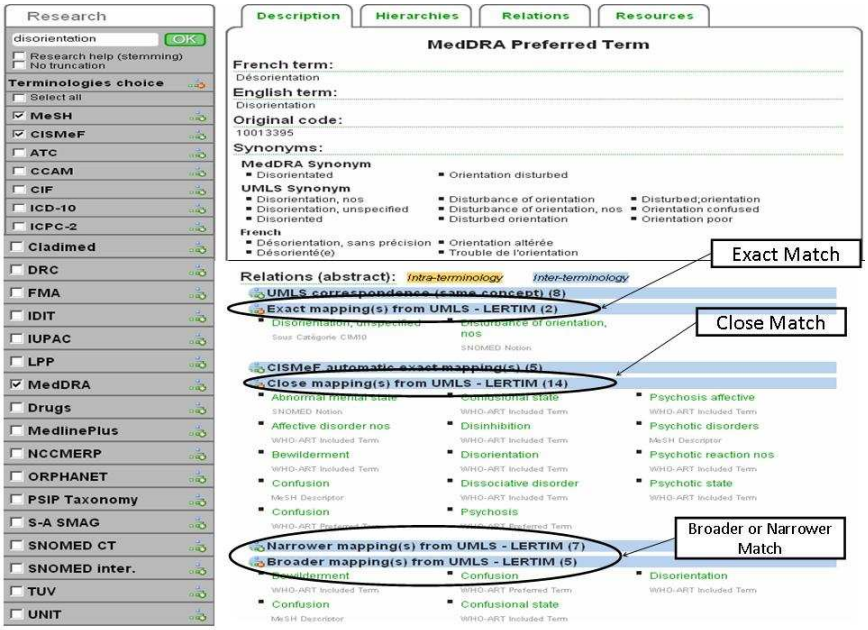

FIG. 1 – *The three types of conceptual approach integrated into the HMTP (Example of the MedDRA term "Disorientation")*

PSIP Taxonomy) or French (CISMeF, DRC) lexical approache alone. All exact mappings are integrated into the HMTP (example of figure 2).

4.3 Comparing the two approaches

As shown in Table 6, the lexical approach was able to find 8,680 more mappings for MeSH and 50,116 for SNOMED International than the conceptual approach. For example, the mapping between the MeSH term "Oral Hygiene" and the SNOMED International "Dental hygienist" was found only by the lexical approach.
The conceptual approach founded 95 more mappings for MeSH and 192 for SNOMED International than lexical approach. For example, the mapping between the MeSH term "Acute-phase proteins" and the SNOMED International term "acute phase reactant" was found only by the conceptual approach.

5 Discussion

The aim of this study was to proposed conceptual and lexical methods to map several biomedical terminologies whatever or not included into UMLS. Methods developed can be applied to map English or French terms. The results obtained through these methods are different according to the type of terminology and the number of target terms used to map the

	FMA	MedDRA	MeSH	ORPHANET	SNOMED International	WHO-ART
CCAM	0	110	305	0	430	5
CISMeF	9	99	517	11	222	17
CISP2	7	138	219	30	254	109
CLADIMED	35	24	258	3	259	4
Codes used for drugs	0	24	1,455	3	302	0
FMA		119	1,745	32	5,777	3
ICD10	10,209	2,380	3,827	947	7,474	1,134
IDIT	0	79	75	0	0	0
IUPAC	58	32	726	8	317	11
LPP	0	0	36	0	22	0
MedDRA	119		3,728	885	5,360	1,278
MEDLINEPlus	34	314	675	138	448	170
MeSH	1,745	3,728		1,805	15,127	1,417
ORPHANET	32	885	1,805		1,635	284
SNOMED International	5,777	5,360	15,127	1,635		1,747
UNIT	0	0	77	0	0	0
WHO-ART	3	1,278	1,417	284	1,747	
WHO-ATC	61	58	3,533	0	1,581	4
WHO-ICF	178	9	294	2	222	7
WHO-ICPS	1	13	159	0	114	6

TAB. 5 – *Fragment of the entire matrix mapping from HMTP*

Terminology	**MeSH**	**SNOMED International**
Number of terms mapped by the two approaches	10,542	15,639
Number of terms mapped only by the conceptual approach	95	192
Number of terms mapped only by the lexical approach	8,680	50,116

TAB. 6 – *The number of MeSH and SNOMED International terms mapped according to each approach*

source terminology. For example, using the conceptual approach, only 10,637 MeSH terms were mapped, whereas 19,222 MeSH terms including the MeSH Supplementary Concepts (n= 186,702) were mapped using a lexical approach. The difference between this two figures can be easily explained by the difference of the target terms used by the two approaches. However, the number of mappings also differs according to the type of terminology. For example, in the

FIG. 2 – *Mapping of the MeSH term "myocardial infarction" according to the lexical approach in HMTP (Exact correspondence)*

lexical approach, there are 1,635 mappings between ORPHANET (terminology for rare disease) and SNOMED International when 15,127 mappings were obtained between MeSH and SNOMED International (see Table 5).

These methods are also used to translate some of English terminologies to French (SNOMED CT Joubert et al. (2009), MEDLINEPlus Deléger et al. (2010)). Lexical approaches are limited in the management of the ambiguous acronyms. For example, the acronym "CMT" corresponds to "Charcot-marie-tooth disease" in MeSH and "Thyroid neoplasms". Another limit of the lexical approach concerns terms lexically close but with a different meaning such as "left" (gauche) and "Gaucher disease" (maladie de Gaucher). Difference in knowledge representation and terminological differences can also cause some problems in the lexical mappings as stressed in Bodenreider and Zhang (2006). For example, there is a mapping between the MeSH term "Marfan syndrome" and the SNOMED International term "Arachnodactyly" because there is a shared synonyms "Dolichostenomelia" between the two terms. However, the same mapping between the ORPHANET term"Marfan syndrom" and the SNOMED international term "Arachnodactyly" was evaluated as false by an ORPHANET expert because "Arachnodactyly" corresponds as a sign of the "Marfan syndrom". In perspective, we a currently working on a third approach based on statistical method (co-occurence).

6 Conclusion

Automatic mapping between biomedical terminologies integrated in the HMTP in English and French was achieved. These mappings were also used to translate English terminologies to French such as FMA, MEDLINEPlus and SNOMED CT.

7 Acknowledgements

Multi-terminology portal was supported in part by the grants InterSTIS project (ANR-07-TECSAN-010); ALADIN project (ANR-08-TECS-001); L3IM project (ANR-08-TECS-00); PSIP project; (Patient Safety through Intelligent Procedures in medication -FP7-ICT-2007-); PlaIR project, funded by FEDER.

References

Bodenreider, O., S. J. Nelson, W. T. Hole, and H. F. Chang (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In *Proc. AMIA Symp. 1998*, pp. 815–819.

Bodenreider, O. and S. Zhang (2006). Comparing the representation of anatomy in the FMA and SNOMED CT. In *AMIA Annu Symp Proc*, pp. 46–50.

Browne, A., D. G, A. Aronson, and M. AT (2003). Umls language and vocabulary tools. In *AMIA Annu Symp Proc*, pp. 798.

Darmoni, S., J. Grosjean, T. Merabti, B. Dahamna, I. Kergouraly, L. Soualmia, and B. Thirion (2010). Health multi-terminology portal: semantics added-value for quality-controlled health gateway. *Journal of Biomedical Semantic*. Submitted.

Darmoni, S., M. Joubert, B. Dahamna, J. Delahousse, and M. Fieschi (2009). Smts: a French Health Multi-Terminology Server. In *Proc. AMIA Symp. 2009*.

Darmoni, S., J.-P. Leroy, F. Baudic, M. Douyère, J. Piot, and B. Thirion (2000). CISMeF : a structured health resource guide. *Methods of Information in Medicine 39*, 30–35.

Deléger, L., T. Merabti, T. Lecroq, M. Joubert, P. Zweigenbaum, and S. Darmoni (2010). A Twofold Strategy for Translationg a Medical Terminology into French. In *Proc. AMIA Symp. 2010*. In press.

Fung, K. and O. Bodenreider (2005). Utilizing UMLS for semantic mapping between terminologies. In *Proc AMIA Symp*, pp. 266–270.
eng

Imel, M. (2002). A closer look: the SNOMED clinical terms to ICD-9-CM mapping. *J AHIMA 73*(6), 66–9; quiz 71–2.

Joubert, M., H. Abdoune, T. Merabti, S. Darmoni, and M. Fieschi (2009). Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. In *Proc. AMIA Symp. 2009*, pp. 291–295.

Lindberg, D., B. Humphreys, and A. McCray (1993). The Unified Medical Language System. *Methods Inf Med 32(4)*, 281–291.

McCray, A., S. Srinivasan, and A. Brown (1994). Lexical methods for managing variation in biomedical terminologies. In *Annual Symposium on Computer Applications in Medical Care*, pp. 235–239.

Merabti, T. (2010). *Methods to map health terminologies: contribution to the semantic interoperability between health terminologies*. Ph. D. thesis, University of Rouen.

Merabti, T., M. Joubert, T. Lecroq, A. Rath, and S. Darmoni (2010a). Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH. *Biomedical Engineering and Research*. In press.

Merabti, T., P. Massari, M. Joubert, E. Sadou, T. Lecroq, H. Abdoune, J. Rodrigues, and S. Darmoni (2010b). Automated approach to map a French terminology to UMLS. In *MedInfo2010*, Cap Town, South Africa. In press.

Pereira, S. (2007). *Muti-Terminology indexing of concepts in health*. Ph. D. thesis, University of Rouen.

Peters, L., J. Kapusnik-Uner, and O. Bodenreider (2010). Methods for managing variation in clinical drug names. In *Proc Annu Symp AMIA 2010*. In press.

Rocha, R., B. Rocha, and S. Huff (1994). Automated translation between medical vocabularies using a frame-based interlingua. In *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, pp. 690–694.

Wang, Y., J. Patrick, G. Miller, and J. O'Hallaran (2008). A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. *BMC Med Inform Decis Mak 8 Suppl 1*, 5.

Résumé

Terminology mapping is an important and crucial task to improve semantic interoperability between health care applications and resources. In 2009, CISMeF created a Health Multi-Terminological Portal (HMTP) to search concepts among all the health terminologies available in French (or in English and translated in French) included in this portal and to browse it dynamically. To map terminologies in the HMTP, two methods are used: (1) conceptual method which exploits various features of the UMLS, (2) lexical method based on natural language processing in French and English. A total of 199,786 mappings were performed between at least two French terms using conceptual method, whereas 266,139 mappings were performed using lexical methods. These mappings were all integrated in the HMTP developed by CISMeF. Conceptual and lexical methods were used to translate some English terminologies into French such as MEDLINEPlus, FMA and SNOMED CT.