MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

From French vocabulary to the Unified Medical Language System: A preliminary study

Olivier Bodenreider[a], Alexa T. McCray[b]

[a]*SPI-EAO Laboratory, Henri Poincaré University, Nancy, France*
[b]*National Library of Medicine, Bethesda, MD, USA*

Abstract

The Unified Medical Language System (UMLS) is an extensive source of biomedical knowledge developed and maintained by the U.S. National Library of Medicine (NLM). The UMLS began to include biomedical terms in other languages a few years ago. However, providing foreign terms for existing concepts is only the first step for the UMLS to become international. The current limits of the use of the UMLS in French are analyzed (partial translation, unique source of the translated concepts, improper character set, and absence of lexical resources for lexical matching tools). Some suggestions are given for French to be better integrated into the UMLS, especially for adapting the lexical resources to French. Once completed, our present work is expected to give the UMLS the capability to be effectively queried in French.

Keywords

Unified Medical Language System; French language; Natural Language Processing.

Introduction

The Unified Medical Language System (UMLS) project was initiated in 1986 by the U.S. National Library of Medicine (NLM). The goal of this long-term project is to help health professionals and researchers use biomedical information from different sources [1]. The UMLS can be described as a source of information for biomedical concepts and a collection of lexical tools allowing users to perform various searches on the terms used to name the concepts. Both data and tools are integrated in the UMLS Knowledge Source Server[1] [2] and used in a large variety of applications (e.g. Internet Grateful Med[2]). The first edition of the UMLS Knowledge Sources was released in 1990. French terms appeared for the first time in the third release of the UMLS.

The use of the UMLS in any particular language starts with the mapping of a query term to one or more UMLS concepts. Then, the semantic information (i.e. the knowledge) can be retrieved and processed. Powerful lexical matching techniques based on the UMLS were developed in the past few years (see, for example, [3, 4]). However, these techniques are currently available only for English.

In this paper, we present an analysis of some of the problems encountered in the UMLS when dealing with languages other than English and propose some specific solutions for better integrating French into the UMLS.

Background

There are four UMLS Knowledge Sources:

1. The UMLS Metathesaurus (MT) provides a common structure for more than 30 biomedical vocabularies. It is organized by concept or meaning. A concept is defined as a cluster of terms representing the same meaning (synonyms, lexical variants, translations). The 1997 version of the MT contains 331,756 concepts named by 739,439 different terms. Interconcept relationships, concept categorization, and information on the co-occurence of concepts in MEDLINE are also included [5]. About 6% of the concepts are translated into French, German, Portuguese and Spanish.
2. The UMLS Semantic Network (SN) defines and organizes the semantic types assigned to each Metathesaurus concept [5]. No translation is provided for the SN.
3. The SPECIALIST lexicon (SL) is an English language lexicon with many biomedical terms. Information for each entry includes base form, spelling variants, syntactic category, inflectional variation of nouns and conjugation of verbs. This information is used by the lexical tools [6].
4. The Information Sources Map (ISM) is a database that describes information sources in terms of content, scope and access conditions [7].

The NLM currently provides two different ways to access the UMLS data:

1. Files from the CD-ROM distribution to be integrated into a local system, and
2. On-line access to the UMLS Knowledge Source Server through a command line interface, through an applica-

1. http://umlsks.nlm.nih.gov
2. http://igm.nlm.nih.gov

tion programming interface (API) and through a Web-based application [2].

The on-line method not only provides access to the data, but also makes it possible to transparently use a large number of tools which are not yet available in the CD-ROM distribution, especially tools allowing users to search terms in the MT.

Given one input term (one or more input words), the MT terms retrieved by one query depend on several criteria summarized in Table 1. The most complex query — called approximate matching — gives a ranked list of MT terms computed by the MetaMap algorithms, and takes into account synonyms, expansion of acronyms and abbreviations, and inflectional and derivational variation [4].

Table 1 - Search criteria for the following queries: Normalized String Index (ns), Normalized Word Index (nw), Word Index (wd), Approximate Matching (am).

Criteria	ns	nw	wd	am
all input words (IW) must be present in the retrieved term (RT)	yes	no	no	no
RT can contain words which do not appear in input term (IT)	no	yes	yes	yes
RT can differ from IT in word order, punctuation, or inflectional variation	yes	yes	no	yes
RT can contain synonyms of IW instead of the original words	no	no	no	yes
RT can differ from IT in derivational variation	no	no	no	yes

The lexical matching techniques not only make heavy use of the SL, but are also based on a syntactic analyzer and on rules for inflectional and derivational variation. Since these resources have only been developed for English, querying the MT in languages other than English is limited to exact matches (RT can not differ from IT in word order, punctuation, or inflectional variation).

Unless otherwise specified, further citations of UMLS will refer to the 8th edition [8].

Analysis of the problems

Since it is composed of concepts and interconcept relationships, the MT is, in principle, a language-independent representation of medical knowledge. However, words are used to name the concepts, and, at the term level, the MT is language-dependent.

The following problems currently limit the use of the UMLS in French:

Quantitative issue

The MT contains 25,932 French terms corresponding to 18,277 concepts: only 5.5% of the MT concepts have one or more name in French.

Qualitative issues related to the translation

While UMLS concepts come from more than 30 vocabularies, French terms in the UMLS come from a unique source: MeSH [9]. Moreover, only main headings from the MeSH were translated. MeSH was translated into French for indexing and retrieval purposes, so that the selection of French terms in the UMLS is not necessarily suitable for other purposes.

Articles and prepositions were often omitted in the French terms translated from MeSH (e.g. "MALADIE HODGKIN" instead of "maladie *de* Hodgkin" for "Hodgkin's Disease"). Furthermore, nouns are often used instead of adjectives (e.g. "TUMEUR CERVELET" instead of "tumeur *cérébelleuse*" for "Cerebellar tumor"). Suppression of stopwords like articles and prepositions, and nominalization (using a noun instead of an adjective) are lexical techniques frequently used for matching purpose but generally not shown to end-users. These transformations only occur in the French translation: the original English terms are syntactically correct, and terms translated into other languages as well (e.g. "DOENCA DE HODGKIN" in Portuguese, "NEOPLASMAS CEREBELOSOS" in Spanish). Although they do not affect the understanding of the meaning by humans, these transformations make it difficult or even impossible for lexical tools like syntactic analyzers to handle the terms correctly.

Qualitative issues related to the representation of the characters

All non-English terms in the UMLS are in upper-case and do not contain any diacritic mark (acute, grave and circumflex accents, tilde, cedilla, dieresis) or ligatures (connected letters).

The character set used for representing characters in UMLS terms (7-bit ASCII) only has an entry for every alphabetic character (from a to z, lower- and upper-case) and for punctuation marks. Diacritic marks are not currently present in the terms that the NLM receives from the French, Spanish, Portuguese and German translators: they are removed from terms in French, Spanish and Portuguese, and are replaced by pairs of characters in German. For example, French "érysipéloïde" (Erysipeloid) becomes "ERYSIPELOIDE", and German "Überdosis" (Overdose) becomes "UEBERDOSIS". In French, the removal of diacritic marks can result in ambiguity. For example, the two words "côte" (rib) and "côté" (side) would be transformed into "cote" (quotation).

Moreover, non-English terms are in all upper-case letters, so that it is difficult for acronyms, abbreviations and symbols to be identified and expanded (e.g. "MG" could stand for both milligram [mg] or Magnesium [Mg]), and for proper nouns to be discovered (e.g. there is no clue for "POMPE" to be interpreted as "pompe" [pump, in "infusion pump"] rather than as "Pompe" [proper noun, in "Pompe's disease"]).

Availability of the lexical resources

Lexical matching techniques require lexical items to be identified and transformed into their base form. Then, derived forms can be computed. Figure 1 shows a simplified implementation of the lexical variant generation (LVG) programs in the UMLS: LVG computes inflectional and derivational variants by applying a set of rules and facts to the base forms [6]. These lexical resources do not currently exist for French in the UMLS.

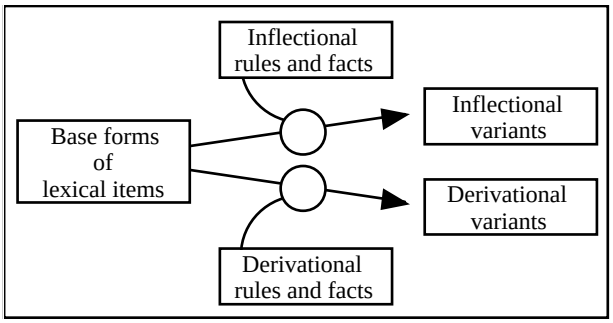

Figure 1 - Lexical variant generation in the UMLS.

Natural language processing also requires other lexical tools like a stemmer and a lexical analyzer, which are language-dependent. Versions of these tools are currently available in several languages.

Possible solutions

Three solutions could be considered in order to better integrate French into the UMLS.

Moving to a more suitable character set

8-bit character sets like ISO 8859-1 are suitable for a large number of Western-European languages, especially for French, German, Spanish and Portuguese, currently present in the UMLS. As an exception, ISO 8859-1 does not provide any representation for the "œ" ligature (lower-cased or upper-cased), which is needed in some French words (e.g. "œil" [eye], "œuf" [egg]). However, this limit is more typographical than linguistical. ISO 8859-1 is the standard currently used by the Hypertext Markup Language (HTML) and a large number of multilingual lexical tools.

But, looking toward the future, as it becomes more international, the UMLS will need a larger character set to overcome this limitation. The Unicode is a 16-bit character set which provides the representation of virtually all existing character sets [10]. However, only a few computer systems and applications currently support Unicode characters, so that Unicode encoded characters would have to be translated into a familiar 7 or 8-bit character set before they could be used by application developers.

Adding French terms in the UMLS Metathesaurus

In a previous study, we have shown that the UMLS does not really require more concepts for the description of medical procedures [11]. We have also found that about 66% of the UMLS concepts, needed for the description of medical procedures, have a name in French within the MT. Nevertheless, terms — and especially French terms — could be added to existing concepts to facilitate information retrieval by providing different expressions for the same concept.

The easiest way to rapidly add French terms to the UMLS is to merge the French translation of vocabulary already existing in the MT in English. Unlike merging a new English vocabulary, the addition of a French translation can be done without any additional work, because it is at the term level and not at the concept level. The French translation of the International Classification of Diseases (ICD-9-CM) has been available for many years. The French translation of SNOMED International is expected to be available soon.

Other valuable sources of French terms are international nomenclatures like ICD-10. ICD-10 is the official nomenclature for diagnosis coding in France and several other European countries. Although an electronic version of ICD-10 is currently available in English and French, ICD-10 is not yet integrated into the UMLS.

Adapting UMLS lexical resources to the French language

We started to adapt UMLS lexical resources and tools to the French language.

A French biomedical lexicon was built from existing electronic lexicons (e.g. French ispell lexicon) and vocabulary extracted from standard French nomenclatures (Catalogue des Actes Médicaux) and French translations of the international nomenclatures (ICD-9-CM, ICD-10). The TreeTagger, a syntactic analyzer, was used to extract the vocabulary and to find out the lexical category and the base form of each lexical item [12]. Unknown items were reviewed in order to filter misspellings and to assign a lexical category to relevant biomedical terms.

Like those used by the lexical variant generation programs, grammatical rules and facts were extracted from French grammar books and converted into appropriate tables. Since the terms to be analyzed are mainly noun phrases, rules and facts are limited to the inflection of nouns (gender, 46 rules), of adjectives (gender and number, 100 rules), and to the past participle of verbs (55 rules).

The lexicon includes 104,000 basic forms (15,000 of which are medical terms). After generating inflectional variants, the lexicon contains 228,000 entries.

The original version of the UMLS lexical tools (lexical variant generation, and normalization programs) were written in C. These tools are currently being rewritten in Java to be used in a distributed and platform-independent environment. The support of multiple dictionaries and sets of rules and facts for variant generation will be part of the next version of these tools as well.

Finally, in order to minimize word-sense ambiguity, the French MeSH terms have to contain diacritic marks and to be correctly cased. We developed a method to correct the spelling of French MeSH terms semi-automatically by comparing every lexical item of French MeSH to any entry in our lexicon. 3% of the lexicon entries have two or more different diacriticized forms, which can be derivational variants (e.g. "contrôle" [control] and "contrôlé" [controlled]) or not (e.g. "force" [strength] and "forcé" [forced]). 3% have two or more different cased forms (e.g. "carré" [square] and "Carré" [proper noun]). 38% have only one possible diacriticized form. And 56% have no diacritics. French MeSH terms could also be classified. Considering only ambiguity related to diacritics, 82% of the 19,447 terms (excluding chemicals) do not lead to ambiguity and could be properly cased and diacriticized (e.g. "TEST EPICUTANE" to "test épicutané" [Patch tests]). 765 lexical items with two or more diacriticized forms were found in the 18% remaining. These terms need review for resolving word-sense ambiguity. In most cases, the meaning of the lexical item is not affected

(e.g. "CONTROLE" in "essai clinique contrôlé" [Controlled Clinical Trials] and "contrôle de qualité" [Quality Control]). Otherwise, the absence of diacritics can change the meaning (e.g. "FORCE" in "débit expiratoire forcé" [Forced Expiratory Flow Rate] and "force de préhension" [Hand Strength]).

Discussion

Unlike the medical literature, electronic patient records are not primarily written in English. For this reason, multilinguality is a major issue for the international use of medical terminologies. However, the use of medical knowledge does not always require language-dependent support.

Multilinguality

Because it is a way to share and reuse knowledge, multilinguality is a common feature of all major medical terminologies (UMLS, SNOMED International, GALEN, ICD). Although all of these terminologies have a language-independent underlying model, only a few of them offer real multilingual support.

The most translated terminology is certainly the ICD: ICD-10 was currently translated in more than twenty languages. SNOMED International and the UMLS are partially available in an increasing number of languages. GALEN, the youngest of the medical terminologies, is multilingual by design (a name in each language for each concept), and currently provides some 6,000 concepts in five European languages (English, French, German, Italian and Dutch) [13].

No terminology fully offers broad coverage, a strong underlying model and multilingual support [14].

Using medical knowledge

Since we refer to concepts using names, interfaces based on lexical matching techniques are generally used as an entry point to medical knowledge. Although they are not directly related to the knowledge, lexical resources are needed to access it, and should be developed for each language used to name the concepts.

On the other hand, knowledge representation also relies on interconcept relationships and on concept categorization. This information is language-independent, so that no lexical resource is needed to navigate the knowledge (Figure 2). The structure of the knowledge can be made explicit by showing interconcept relationships graphically, as well as the relationships between concepts and the semantic network [15].

The UMLS Knowledge Source Server provides users with powerful tools to access medical knowledge in English. A browser of concepts (graphical or hypertext-based) would be a useful tool for all users, whatever their language is, to navigate the knowledge by following interconcept relationships.

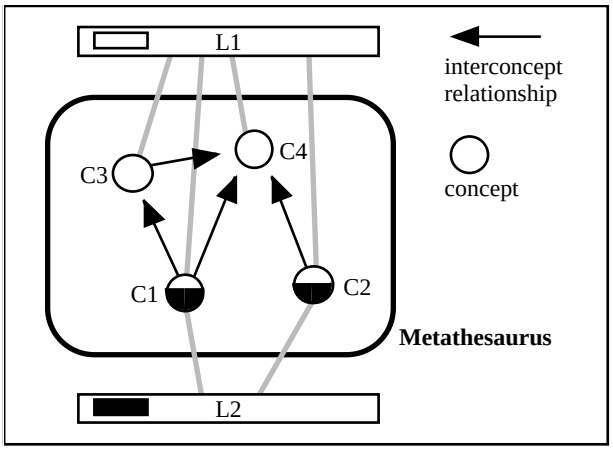

Figure 2 - Using medical knowledge: only two concepts (C1 and C2) have a name in both languages (L1 and L2). C3 and C4 can be reached in the L1 language only, but users can navigate from C1 to C3, whatever their language.

Conclusion

Three major problems have been identified as limiting the use of French in the UMLS. Most of these problems and the solution we propose apply to other non-English languages as well.

A truly multilingual UMLS would be possible by moving to a character set allowing diacritics and other special characters to be represented. However, suitable character sets are not currently supported by all systems or applications. On the other hand, there is no reason for non-English terms to be systematically uppercased.

Although the French translation of MeSH terms currently offers a reasonable coverage of the medical domain, more concepts should be translated. Furthermore, the addition of new French terms to already translated concepts would better reflect the diversity of the biomedical language.

Finally, access to UMLS knowledge through names in a given language requires lexical matching techniques which can be used only if the corresponding lexical resources are available. We developed a French lexicon with many biomedical terms from several French vocabularies. The implementation of matching algorithms is being modified to use lexicons and rules for inflectional and derivational morphology in non-English languages.

Once completed, our present work is expected to give to the UMLS the capability to be queried powerfully in French.

Acknowledgments

We gratefully acknowledge Helmut Schmid and Achim Stein (Stuttgart University, Germany) for providing us with a modified version of the TreeTagger.

This work was supported in part by 3M Laboratories and by the following associations: Association des Utilisateurs des Nomenclatures Nationales et Internationales de Santé (AUNIS), Collège des Praticiens Spécialistes en Information et Communication Médicales (COPSICOM), Information Médicale et Gestion des Établissements (Groupe IMAGE) and Contribuer à la Recherche et à l'Innovation au Service du Traitement et de

l'Analyse des Langages utilisés dans les Systèmes de Soins (CRISTAL'S).

References

[1] Lindberg D, Humphreys B, McCray A. The Unified Medical Language System. *Methods Inf Med* 1993;32(4):281-91.

[2] McCray A, Razi A, Bangalore A, Browne A, Stavri P. The UMLS Knowledge Source Server: a versatile Internet-based research tool. *Proc AMIA Annu Fall Symp* 1996:164-8.

[3] Nelson S, Olson N, Fuller L, Tuttle M, Cole W, Sherertz D. Identifying concepts in medical knowledge. *Medinfo* 1995:33-6.

[4] Aronson A. The effect of textual variation on concept based information retrieval. *Proc AMIA Annu Fall Symp* 1996:373-7.

[5] McCray A, Nelson S. The representation of meaning in the UMLS. *Methods Inf Med* 1995;34(1-2):193-201.

[6] McCray A, Srinivasan S, Browne A. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care* 1994:235-9.

[7] Masys D, Humphreys B. Structure and function of the UMLS Information Sources Map. In: Lun K, Degoulet P, Piemme T, Rienhoff O, eds. *MEDINFO 92*: North-Holland Publ Comp, 1992:1518-21.

[8] *UMLS Knowledge Sources*. (8th ed.) Bethesda (MD): National Library of Medicine, 1997.

[9] *Thésaurus biomédical Français/Anglais [French translation of MeSH]*. Paris: Institut pour la Recherche Médicale, 1997.

[10] The Unicode Consortium. *The Unicode Standard, Version 2.0*. Addison-Wesley, 1997.

[11] Bodenreider O, Burgun A, Botti G, Fieschi M, Beux PL, Kohler F. Evaluation of the use of the Unified Medical Language System as a knowledge source for a terminology server of French medical procedures. *J Am Med Inform Assoc* (submitted).

[12] Schmid H. Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*. Manchester, UK, 1994:44-49.

[13] Rector A, Salomon W, Nowlan W, Rush T, Zanstra P, Classen W. A medical terminology server for medical language and medical information systems. *Methods Inf Med* 1995;34(1-2):147-57.

[14] Baud R, Rassinoux A, Lovis C, et al. Knowledge sources for Natural Language Processing. *Proc AMIA Annu Fall Symp* 1996:70-4.

[15] Tuttle M, Cole W, Sheretz D, Nelson S. Navigating to knowledge. *Methods Inf Med* 1995;34(1-2):214-31.

Address for correspondence

Olivier Bodenreider
Laboratoire SPI-EAO,
Faculté de Médecine de Nancy, B.P. 184,
54505 Vandoeuvre-lès-Nancy Cedex
France
(e-mail: boden@spieao.u-nancy.fr)