Medical Informatics Europe '99
P. Kokol et al. (Eds.)
IOS Press, 1999

Automatic enrichment of the Unified Medical Language System starting from the ADM knowledge base

Franck LE DUFF, Anita BURGUN, Bruno POULIQUEN, Denis DELAMARRE, Pierre LE BEUX
LABORATOIRE D'INFORMATIQUE MÉDICALE, FACULTE DE MEDECINE, RENNES

ABSTRACT : The Unified Medical Language System (UMLS) project aims to provide a repository of terms, concepts and relationships from several medical classifications. This work describes the possibility to enrich automatically with meaningful links the UMLS database by using description of diseases from another knowledge base, in our case ADM (Aide au Diagnostic Medical). In spite of the constraints and the difficulties to qualify the interconcept links, the results show that it is possible to find and create new links from a french knowledge database to the UMLS one. One of the interests of this work is that the automated learning of the connections could be used with others knowledge databases like expert system databases.

1 - Introduction

Representing knowledge remains an essential problem for medical information systems and networks on account of the plethora of the documents produced. It is necessary to obtain reliable and powerful browsers to reach the most relevant information quickly. The work that we approach here is of interest in the construction of these tools for nomenclatures and in the exhaustivity of such tools. There exists many possibilities to carry out a medical nomenclature. According to the method, the result will be connected more with one catalogue or classification [1] i.e. with a tool more or less fixed, or it will give a 'compositional' system largely dependent on the data-processing tool [2] . Many authors [3, 4, 5, 6] show that these classifications instruments can be complementary from each other and that it is possible to enrich a system from another. This idea of re-usable lexicons has been described by [7]. Our choice was made on the analysis and the transposition of the conceptual links between two systems to try to answer the following question: Is it possible to enrich the inter-conceptual links of the UMLS language unified starting from another knowledge base? The ADM Base is a French base developed at the beginning of the Seventies by a team of intermist physicians [8] which currently contains the description of several thousands of diseases . The UMLS is a project of the American National Library of Medicine which has the goal to integrate information coming from different backgrounds. Although the Metathesaurus includes a lot of concepts, some authors remind the necessity to improve the quality and the structure of the database [9]. In our case, we wanted to see if it was possible to extract knowledge from the French ADM data base and then complete the UMLS data base by integrating this information as new relations between concepts rather than adding new concepts. The first goal and interest of this work lies in the creation of links between concepts which belong to different databases, indeed, each base included in the UMLS project, keeps its meaningful links between concepts but there is few relations between data bases. The second goal was to create and display new relations between signs or symptoms and diseases and then improve the requests to the UMLS database.

2 The interchange space

Acquisition of knowledge by accumulation of concepts in the base by automatic methods has been already described in the literature [3,5,6] More interesting for the subject that we treat in this paper, are the articles describing knowledge acquisition by accumulation of links or creations of links between concepts. [10] evokes on this subject the difficulty of updating the successive versions of the Metathesaurus and the possible

conflicts between the original thesaurus and the thesauri locally improved by the teams. Acquisition of knowledge has been also discussed for the description of surgical procedures, [11] recalls the difficulty of exploiting information from the meaningful links on account of the richness of the labels (broader, narrower, other, " is_a ", " part_of " ...). [12] and his team proposes to create relevant associations between concepts by inheritance in the UMLS Semantic Network. The first step consists in defining 'views' that could be classes of network and the second step consists in defining rules of associations. The links between concepts are then represented with conceptual graphs and are functions of defined views and constraints between concepts. The ARIANE project is an example of this work. The purpose of this project consists in modelling and implementing an intelligent and ergonomic interfaces according to the kind of the biomedical information database [13].

2.1 - The UMLS Database

The Unified Medical Language System is a collection of concepts resulting from several data bases elaborated in various places with different languages and different logics. The UMLS knowledge base includes three parts : The Metathesaurus is a collection of terms and concepts from many different biomedical vocabularies and classifications. The Semantic Network is a structure for categorizing objects in a medical domain. The Information Source Map is intended to describe the computerized biomedical sources. Whereas the Metathesaurus contains many concepts, the number of relations between two concepts should be increased.

2.2 - Description of the ADM knowledge

The ADM knowledge base is a base covering most of the medical specialities [8]. It contains 15600 diseases or syndroms descriptions and a lexicon of about 110000 terms. The goals were, first to help physicians to find diagnosis and next to provide a quick access to the medical information by using telematics. Recently, ADM has been brought on the Internet network. The knowledge base uses a relational data base system and physicians can query it with a web natural language interface (by the end of 1998 october, morether 2000 users were registered and morether 50000 requests had been done since the beginning of the year). The reason why we choose the ADM database was the structure of the diseases descriptions. Indeed, by using this database, we made the hypothesis that it was possible to complete relations between signs, finding, symptoms and diseases into the UMLS Metathesaurus.

2.3 - Methods

As we have seen ADM is built mainly on a hierarchical semantic model and provides descriptions of diseases whereas UMLS carries out a grid between concepts and categories semantics. The hierarchical structure of the ADM base presents relations between several labels which were validated by experts. Our hypothesis is that it is possible to extract some meaningful links from the first database for integrating it in the UMLS metathesaurus similar meaningful links between concepts.

Stages

We decomposed the work in several process (Fig1). First of all, we sought compatible concepts of UMLS (noted U1n) with the labels of descriptions of the ADM database (noted A1n). This search consisted in finding the correspondences i.e. we searched a link between terms from the ADM labels and terms from the UMLS concepts to match terms from one database to another. The second stage consisted in searching all the labels attached to the first one that had been selected in the first stage and then extract all the symptoms from the ADM description. Next, we searched all the UMLS concepts corresponding to the labels found (an English-French translation of the found signs could allow an extension of the results). One by one, we took the labels of the signs (Sij) and symptoms of the ADM database and searched in the UMLS database if the concepts existed(Cij). The last stage consisted in seeking Lij links between the concept Ui and the signs Cij. If the link didn't exist, we created it (enrichment of UMLS base).

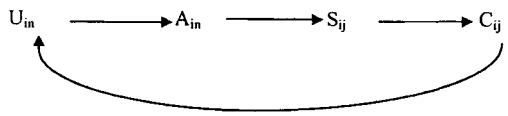

Fig 1 - Ring of search

Common space search

The fields covered by ADM and UMLS are not the same. This constraint forced us to work on the intersection of the fields covered by both databases. To expand this space, we have used an automaton to translate the english concepts terms into french concepts terms. Indeed, ADM strictly French-speaking leads us to work only on the French part of UMLS (this part concerns only the french MeSH, in the 1998 UMLS version). To be relevant and lead to a result, it is necessary and essential that the described labels and the signs attached to these labels, are all in French in the ADM/UMLS intersection. To enlarge the common space between the two bases, we needed to translate automatically all the UMLS preferred concepts and then we used an utility developped to translate english words in french words. The choice to translate UMLS from english to french rather than to translate ADM from french to english was made after analysis of the ADM labels. Whereas the Metathesaurus has a structured wording, the ADM wording is often expressed with many details shortened or not. In addition, there is no accent in the ADM labels and the translation from French to English can be ambiguous [14]. The few following examples illustrate this characteristics, particularly the French word 'FORCE' which can be translated into 'STRENGTH' with an accent or into 'FORCED' without an accent (example : S48271 MOUVEMENTS ARTICULAIRES ANORMAUX GENOU BAILLEMENT INT. EN VALGUS FORCE (FLEXION ROTATION EXT.)

Correspondences search

Once our translation is carried out, we finally built a table of correspondences between ADM codes and UMLS codes starting from the whole translated and not translated French wording (Fig. 2). At this level we had to use a last module for automatic processing of the wording, available in the development tools of ADM. The enrichment can be summed up in four steps. First, we take a concept from the UMLS Metathesaurus. Then we look for an equivalence in the ADM data base. Third, we glance through the ADM description disease and for each entity included in the description we look for an equivalence in the Metathesaurus. Fourth, if we find an equivalence, we take the code of the UMLS concept and we seek if there is a link between this code and the first one which we had started.

3 - Results

Identical french words in the databases

The bases are not similar. French wording are so different that the intersection between ADM and UMLS is a small percentage of the whole bases.

• Among the 354 579 concepts available in the UMLS base and the 96991 labels referred in the ADM base, we could only highlight 1988 identical concepts. This result corresponds to a very low rate of recovery (2,01 % on the ADM database ands 0,56 % on the UMLS).

• On the whole set of UMLS synonyms, we located the 785 identical wording between ADM base (96991) and the table of synonyms of the UMLS (168 629). The rate of recovery is lower than 1 % in both cases (0,80 % for ADM and 0,46 % for UMLS).

Translation

The translation let us double the quantity of concepts included into the common space. The totality of the wording automatically translated to extend the working area gave us the following results :

• 52 803 preferred terms (15 % of the total) and 31 334 synonyms or inflections were translated (that is to say 18,5 %). Once the correctly translated wording obtained, an automatic search on the present words was carried out.

• On the whole, 1415 occurrences (including 960 synonyms) among the 84 137 (52803+31334) were translated (1,7 %), i.e. which it was possible to us to bring closer 2830 codes (half ADM, half UMLS).

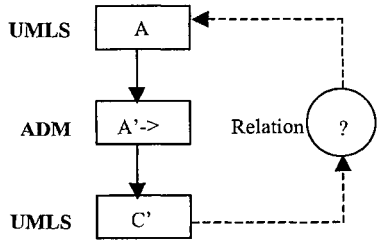

Fig. 2 – The matching Algorithm

New links proposals
Once correspondences between UMLS codes and ADM codes had been made up, we sought if there were missing links. On account of the small common space between the two bases, we did not wish to limit search to a particular nosology and we look up through all the table.

Our table of analysis included 4187 lines and allowed us to locate 14144 links not yet existing in UMLS base. In addition, it should be noted that about 100 % (99,97 %) of the codes processed by this work permits to create at least one additional link in the table of links M_REL of the UMLS table.

CCPT1	LCPT1	CCPT2	LCPT2
C0017658	GLOMERULONEPHRITIS	C0027712	NEPHROLOGY
C0017658	GLOMERULONEPHRITIS	C0018965	HEMATURIA
C0017658	GLOMERULONEPHRITIS	C0235220	ARTERIAL HYPERTENSION
C0017658	GLOMERULONEPHRITIS	C0013604	OEDEMA
C0017658	GLOMERULONEPHRITIS	C0028961	OLIGURIA
C0017658	GLOMERULONEPHRITIS	C0027726	NEPHROTIC SYNDROM

Table 1 : Example of correspondences between the code relating to 'GLOMERULONEPHRITIS' of the ADM description and the codes of the UMLS concepts where there is no relation in the Metathesaurus.

Knowing that UMLS proposes 1 711 054 links, the result of our work allowed an enrichment of the UMLS Metathesaurus from approximately 1 % (0,82 % exactly).

Semantic network
We examined the semantic categorisation of the concepts which help us to create new links.
Three points can be pointed out :

- Concerning the concepts at the origin of the new links, we can note that the semantic type T047 is the most important type compared to the others with 9650 occurences. We find next the semantic type T109 'Organic chemical', T121 'Pharmacologic substance', T033 'Finding', T019 'Congenital abnormality'.
- Concerning the concepts where the new links are pointed we can notice that it is also the semantic type T047 which is the most important type, just before the type T0 to 3 'Finding', T184 'Sign or symptom', T032 'Organic attribute' and T091 'Biomedical occupation or discipline'.
- If we study the T047 type and T033, we can see that the part of the concepts belonging to these semantic types which allowed us to create new links in the whole concept base represents about 30 per cent. 30887 concepts belong to the type T047 and we found 9650 new links with concepts of this type (that is 31,25% and 29,16% with the semantic type T033. These important results can be explained by the structure of the ADM data base).

Nosologies
If we analyse the nosologies related to the UMLS concerned with our work we notice that the links that could be created, are mainly concentrated in four fields : urology, orthopedy, oto-rhino-laringology, dermatology.
If we seek the origin fields of ADM from which these links were found, we find also four fields: Infectious diseases (bacteria, virus, mycoses) (N00006), constitutional diseases and/or congenital (N00007), metabolic diseases and nutrition (N00014), toxic diseases or diseases by chemical agents (N00022). The distribution of the concepts for which one or more links were located starting from ADM base shows that UMLS nosologies concerned with enrichment are not the same as the ADM ones. Cardiology for example, is sparsely concerned with an enrichment in the UMLS but if we look at cardiology (nosology N00003) of ADM database, we can see on the other hand that about 600 concepts contributed to the enrichment of the Metathesaurus. This nosology comes ninth on twenty nine.

4 - Discussion

According to [3], it appears clearly that the UMLS aims at integrating already existing thesauri, than create a standardized vocabulary. *'The UMLS is not intended to serve have as a standard vocabulary, but rather has a means of mapping between existing vocabularies'*. It is not a question to limit the richness of the medical vocabulary but on the contrary to gather in the same base the totality of the expressions and terms indicating the same object. From this point of view, the work what we carried out respects the philosophy that the researchers of the Medical Library of Medicine. Indeed, we sought to transfer and integrate into the UMLS, a

knowledge acquired from the ADM data base. The construction of the UMLS initially aims to reach a high level of integration of most significant medical terms within a Metathesaurus and not to work out an exhaustive grid of links between concepts. This fact allowed other research in particular the inheritance of links by the semantic network as we saw with Joubert and coll. Locating and extracting links from ADM fit in the logical continuation of the construction of a concepts repository like UMLS. The other important point we have to approach relates to the automation of the task. Indeed, the work that we carried out concerned more than 1,7 million of links and several thousands of concepts. This quantity of information justified the development of a data-processing tool to obtain a result. An automatic data-processing is possible when data are homogeneous and standardized. Then we have to look at the quality of UMLS tables, the relevance of their construction and their composition.

Limits related to the structure of base ADM

The results we obtain should not however occult many skews and constraints which have limited our work in a significant way. Considering the quantity of information that we had to process, our work does not concern the majority of the UMLS database (less than 1% of the total) and it is legitimate to see if this work is representative of the database. Four limits must be stated here :

• *The linguistic barrier.*

Work that we carried out was possible after the translation of the UMLS concepts to widen the field of the concepts common to both bases. This constraint shows obvious the limits of search since we were subordinated to the performances of an automat. As we showed in the method, the choice to translate the UMLS from english to french rather than french to english does not seem to have an alternative because of the structure of ADM labels.

• *The lexical barrier.*

This second limit is inherent to the automatic translation and processing of the languages. The labels of the ADM data base do not always correspond to the terms of the diseases used in the Metathesaurus so that common diseases were not all located. The power of the specialized lexicon, completely in English could not help us. The reasons which explain the difficulties of working on the translation of the ADM labels are, on the one hand that ADM database has been existing since more than twenty years and same of terms seems to have aged, and on the other hand because ADM was not conceived so that the labels were reusable. Many links seem unusable or without interest.

• *The fields covered by the ADM labels.*

The descriptions integrated into the data base ADM do not cover the totality of medical knowledge. The maintenance of the base indeed requires time and efforts which are not always easy to mobilize. It results from that an inequality in the nosologies covered by the base and a delay or a lack of updating the data and information included in the descriptions. We can notice in addition that the knowledge of ADM base is the fruit of several participants over several years, which could be at the origin of a certain heterogeneity between descriptions suggested (we can find in the base as an example an entity for 'GLAUCOMA LEFT' and an entity for the disease 'GLAUCOMA RIGHT' and sometimes few descriptions have some English words). This heritage has probably limited the probability of matching between ADM labels and UMLS concepts.

Qualification of the links

Our work was restrained first to seek and second to show if it were possible to automatically create links in the UMLS, starting from another knowledge base. During this work, the problem of the qualification of the links arose quickly. According with the literature, we have been confronted ourselves mainly with two problems:

- the search for an automatic routine of qualification (implied standardized) of the links seems difficult to carry out because of the great quantity of combinations between links [11, 9]. A manual stage for confirmation seems necessary.
- the choice of the method of qualification does not appear obvious. Before adopting the heritage of a qualification via the semantic network like the one shown by Joubert and coll, we tried to see whether it was possible to deduce the new link starting from adjacent relations. It is as an example of what [Lehman 92] describes in a precise way when it presents the Cyc project and the algebra of Huhns and Stephens where the relation of the 'hypotenus' is deduced from the different relations.

5 - Conclusion

Our work has shown that it is possible to extract automatically information meaning as conceptual links from one knowledge data base in order to integrate it into another knowledge data base . Although many obstacles are opposed to match the information contained in each of the two bases, this enrichment of the UMLS starting from a knowledge base in French and which concerns a little more than 14000 links, and confirms in addition the possibility of integrating another knowledge and other languages that the English language into the UMLS and then increases the opportunities to use the Unified Medical Language System in other countries.

The difficulties and the constraints which we encountered persuade us on the one hand to pursue a reflexion on the ADM base and its contents and on the other hand to consider this same type of work on other knowledge bases. When [9] evokes the reinforcement of the UMLS structure ('More versus Better'), a similar analysis should be carried out on the ADM. This basic work will consist in taking again the entirety of the contents to model it, to purify it in order to eliminate ambiguities from descriptions. This work could finally extend at other French or English based knowledge bases. An extraction of knowledge of these bases could probably quickly enrich the grid between concepts in the Metathesaurus. Finally further work must be done to determine a powerful and relevant tool for the qualification of the links lately highlighted.

References

[1] Degoulet P., Fieschi M. Informatique médicale in : Masson (éd.) *Abrégés Paris* 1997 : 240 pages.

[2] Rector A., Nowlan W.A. The GALEN project *Comput Methods Progr* Octobre 1994; 45 (1-2) : pp. 75-78. [Salomon 94 - 1] Salomon D. Documentation overview *GALEN documentation* Novembre 1994; A : 5 pages.

[3] Cimino JJ, Octo Barnett G Automated translation between medical terminologies using semantic definitions *M D Computing* 1990; 7 (2) : pp. 104-9.

[4] Joubert M., Fieschi M., Botti G., et al. Représentation de concepts médicaux pour la recherche d'information : réalisation d'une maquette à partir d'UMLS in : Springer Verlag France (éd.) *Informatique et santé* 1992 (5) : pp. 3-17.

[5] Lindberg DAB._, Humphreys BL., Mc Cray AT., et al. The Unified Medical Language System *Meth Inform Med* 1993; 4 (32) : pp. 281-91.

[6] Mc_Cray AT., Srinivasan S., Browne A.C. Lexical methods for managing variation in biomedical terminologies in : Proceedings 18 Anual Symposium on medical applications in medical care (éd.) *J Am Med Informatics Assoc* 1994 : pp. 235-39.

[7] Musen MA Dimensions of knowledge sharing and reuse. Computer and Biomedical Research, 1994, **25**, pp 435-467.

[8] Lenoir P, Michel JR, Frangeul C, et al. Réalisation, développement et maintenance de la base de donées A.D.M. *Med. Inform.* 1981; 6 (1) : pp. 51-56.

[9] Bodenreider O, Burgun A, Botti G et al. Evaluation of the Unified Medical Language System as a Medical Knowledge Source *JAMIA* 1998; I (5) : pp. 76 - 87.

[10] Tuttle MS, Sherertz DD, Erlbaum MS, et al. Adding your terms and relationships to the UMLS Metathesaurus in : Hanley & Belfus, Inc (éd.) *AMIA* 1992 : pp. 219-33.

[11] Burgun A, Bodenreider O, Denier P, et al. Knowledge acquisition from the UMLS Sources : Application to the description of surgical procedures in : Greenes et al. (éd.) *Medinfo Proceedings* 1995 : pp. 75-79.

[12] Joubert M, Miton F, Fieschi M, Robert JJ A conceptual graphs modeling of UMLS components in : Greenes et al. (éd.) *IMIA Proceedings* 1995 : pp. 90-94.

[13] Joubert M., Fieschi M., Robert JJ, et al. UMLS-based Conceptual Queries to Biomedical Information Databases : an overview of the Project ARIANE *JAMIA* Jan/Fév 1998 (1) - 5 : pp. 52 - 61.

[14] Simard M Automatic Restoration of Accents in French Text, Centre d'innovation en technologies de l'information, Laval, Canada 1996 : 9 pages.