The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

Method for Automatic Management of the Semantic Network Ambiguity in the UMLS : Possible Application for Information Retrieval on the Web

Vincent Mary, Franck Le Duff, Fleur Mougin and Pierre Le Beux
Medical Informatics Laboratory, 2, rue du Pr. Léon Bernard 35000 RENNES, FRANCE

Abstract

The Unified Medical Language System (UMLS) is an extensive collection of terms and concepts. The UMLS includes biomedical terms from standard classifications. The semantic network (SN) links the concepts, sometimes ambiguously. In this paper we try, on one hand to describe the relationship between concepts more efficiently and on the other hand to find new relationships. Assuming that re-usability and automatic extraction of knowledge from existing thesaurus enables an improvement of the metathesaurus, we cross the SN with linked concepts from the ADM (Assisted Medical Diagnosis). Results are presented and our discussion concerns firstly the use of the SN only; secondly the improvement that allows pre-selection of linked concepts, and thirdly the possibility to coincide with other developments that improve the metathesaurus.

Keywords:

Semantic Network, UMLS, ADM, Metathesaurus Improvement.

1. Introduction

The semantic network of the Unified Medical Language System (UMLS)[1] reduces complexity by grouping concepts according to the semantic types (ST) that have been assigned to them. More than reducing complexity, the SN can also be useful to manipulate the knowledge included in this knowledge base. Because reusing an existing knowledge base can save considerable time and effort we chose to exploit the semantic network to find and extract knowledge to improve the Metathesaurus and particularly the relationship between concepts included in the Metathesaurus. Concerning this kind of relationship, the statement made by Bodenreider [2] is always true and the relationship between concepts is mainly qualified with ambiguous relationships.

The most important obstacle was the difficulty for automatically qualifying these new relationships. Previous research has introduced approaches to facilitate knowledge extraction. Most of them have the same starting point, using the UMLS MRCOC table (who defines the CoOccurring relations between concepts). The authors suppose that co-occurring concepts belong to the same semantic space, defined by the relationship in the metathesaurus : Burgun [3], Mendonca [4] and Zeng [5].

Burgun [3] explored the SN and showed the possibility of finding and representing the semantics of the relationships between two co-occurring concepts. Although her analysis presents some results, she suggests carrying out further research in this domain.

Because we believe that transferring automatically qualified relationships is now more relevant than adding ambiguous relationships between concepts, we wanted to assess the contribution of the SN in the creation of new relevant links in UMLS. Our previous work [6] consisted in looking for new relationships between concepts starting from other

knowledge bases and in particularly from French medical databases. This paper presents an attempt to use the Semantic Network to qualify meaningful links between concepts.

2. Goals

The main aim of this work was to find a method to qualify more efficiently the relationships between concepts. To achieve this aim we specify two particulars goals. The first one was to use the capability of the SN to produce new meaningful links and qualify them at the same time. The second goal was to evaluate the knowledge of the SN to remove the ambiguity of a new relationship between two concepts of the Metathesaurus.

3.Materials and Methods

Selecting concepts

We used two samples of concepts to find new relations.

The first group consisted of randomized concepts from the Metathesaurus. We used this sample to find relationships between concepts and to qualify these links. The 2001 release of the UMLS semantic network represents 134 ST. Concepts are linked by one or more relationships. The Metathesaurus which contains 796,656 concepts proposes also 9,524,132 relationships between those concepts. The most representative links are the SIB (Sibling) relationships (4.459.562) and the RB (Broader) (838.234) relationship.

The second group consist of concepts which have been marked as linked concepts in a knowledge database, the ADM (ADM : Assisted Medical Diagnosis)[7,8]. To select these concepts we searched for some hierarchy or relationships such as descriptions of diseases. We mapped terms from this French data base and terms from the UMLS [6,9]. The work consisted in selecting first, all the terms attached to one disease entity in the French database and all the signs and symptoms included in the same database. Next, we searched for all the UMLS concepts corresponding to the terms found (an English-French translation of the found signs could allow an extension of the results). Finally, for two terms linked in the French database, if we found them in the UMLS, we checked the relationship between the concepts. If there was no link, we created it and we tried to qualify it. This work allowed us to select a set of concepts able to be linked. The same set of concepts was used in the present work to evaluate the contribution of the SN for qualifying the relationships.

Search for links

We also used a research method to find relationships for each group of concepts.

For the first group, we followed the ST to find the related concept (Diagram 1). Starting from CUI1 (Concept unique identifier), a random concept in the Metathesaurus, we collected all the ST_{CUI1} and listed all the ST, relationships and concepts connected with ST_{CUI1}. Either the relationships between semantic types were in the same proportion and then it was not possible to remove ambiguity between CUI1 and CUI2 or one kind of relationship was more important than the others (or there was only one kind of relationship) for the pair of concepts and then it became possible to inherit a relationship type to link these concepts.

To calculate the relation per concept pair all over the thesaurus and save process time, we used the MRSTY (Semantic Types table) and SRSTRE2 (Relations between semantic types) tables : T002 regroups 32.832 concepts, which are linked on one hand to 2001 others (T003) with one relationship (T142) and the other hand to 94 (T001) with two relationships

(T142, T186).

For the second group, we listed all the semantic types of the concepts and looked for all the relationships between the semantic types (Diagram 2). If there was a UMLS relationship or not between CUI1 and CUI2, we supposed that the relationship could inherit from the relationship between the semantic types (R1 and R2). Once again, either it was not possible to remove ambiguity between CUI1 and CUI2 because the number of relationships R1 was near / equal to the number of relationships R2, or a semantic type of relationship was important enough to characterize the relationship between the pair of concepts.

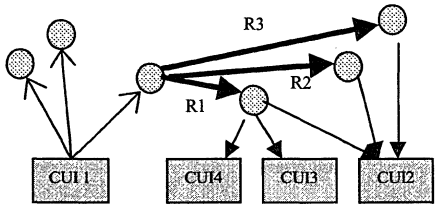

Diagram 1

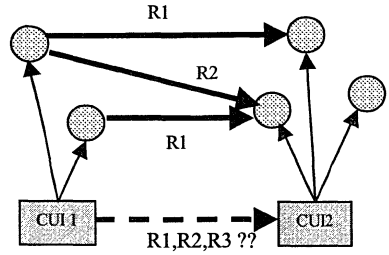

Diagram 2

4.Results

With the first method, the results showed that the number of concepts related to other concepts using the Semantic Network was extremely sizeable : 441 billions couples related by 770 billions relationships (1.63 relationships per couple)(Table 1).

For instance, 'diuretic' (C0012798)(Table 2) is linked with 526,919 other concepts and has one million relationships whereas the only ST of this concept is T121 (pharmacological substance). The ST T121 is linked, itself, to 58 other ST with 11 different relationships. For instance, T142 (interacts with, 390,997 relationships), T149 (complicates, with 110,408 relationships) or T154 (treats, with 110,168 relationships).

In the same way, the concept 2 hydroxy-progesteron is linked through 3 ST (T110,121,125) with 533,018 concepts via 2,816,132 semantic relationships : concepts that belong to the same ST have the same relationship proportion and the same relationships itself.

Table 1

	a	b	c
Starting point	776.940 concepts	3386 couples	57 couples
Founded couples	441 billion	1,546	57
Relationships	716 billion	3,640	143
Link per couple	1.62	2.35	2.5

Table 2

Concept	Nb of linked Concepts	Semantic Type	Nb of Relation ships
Duiretic	526,919	T121 Ph subs.	1,025,787
2 OH Pg	533,018	T110 Steroids	1,025,787
		T121 Ph subs.	876,029
		T125 Hormone	914,358

The second part of this study (Table 1), we selected pair of concepts to find useful relationships in the SN for qualifying the relationships between these concepts.

We analyzed 3,443 couples : 57 (c) of them where already linked in the UMLS (essentially sibling or RB) and 3,386 (b) were not. Throughout the corpus, 45 % (1546) of couples could be linked, with 3,640 semantic relationships (2.35 link for a couple). For the pairs which were already linked in the UMLS, we found all the 57 couples again with 143 semantic relationships.

- For some concepts, it was not possible to reduce ambiguity of relationships between concepts by starting from the SN : for instance, 'duodenitis' has a sibling relationhip with 'nausea', but we found the three semantic relationships 'co-occured with', 'associated with' and 'degree of' with the same occurrence.
- For other concepts, we could reduce ambiguity starting from SN :
    - 'appendicitis' has a sibling relationship with 'nausea', and we found only one relationship, 'associated with'.
    - finally for other concepts, there were several relationship types between the two concepts but one type was more frequent than the others.

5. Discussion

The first experimental results were difficult to exploit. The mass of noise contained a very large number of relevant new links : 'appendicitis' (C0003615), whose ST is 'disease' (T047) has an 'associated with' (T166) relationship with 'inflammation localized' (C0522570), whose ST is 'finding' (T033). Although there is a discriminatory sort criterion (the number of links per relationship is relevant : the nearest are the concepts, the more relationships per couple we found)(Table 2), we were confronted with another problem. Two different concepts which, for instance, have the same single semantic type would have formed the same couples and would have the same relationships : 'migraine' and 'hepatitis' have only one semantic type (T047) and will also inherit the same relationships. It would not be relevant to treat these two different concepts equally. Therefore it appears necessary to start from 'linked' concepts. We preselected concepts from an external thesaurus in the second experiment for three reasons :

- Starting from couples of UMLS, linked concepts did not enable us to find new relationships between concepts.
- Secondly, previous work [3] has already been done on UMLS co-occurring concepts.
- Thirdly, previous work, such as Joubert's [10], proposed conceptual representation of the UMLS based on conceptual graphs. One possible exploitation from our work was to help in the creation of views.

Table 1 shows that our method is interesting for two reasons :

- we retrieve the existing concept relationships in UMLS, but with a new kind of relationship
- we found new relevant medical relationships between concepts : for the 3.386 couples, our work proposes at least one new relationship in 45 % (1546 / 3386) of cases.

For some concepts, it was not possible to reduce ambiguity of relationships starting from the semantic network, but allowed to limit the ambiguity by suggesting other relationship types. However, if the ratio noise / signal was acceptable, the problem of relevance remains. By the number of concepts, we did not find an automatic method to evaluate or select the most relevant relationship types for the pairs of concepts. The end-users have to take into account the medical relevance of the new relationship.

Joubert improved the relevance by choosing his concepts with a view to a specific medical domain. It could be interesting to use the two techniques : a first selection of concepts linked in another thesaurus could be a starting point for an automatic creation of view. Another starting point would be to select a thesaurus (like ADM) where the concepts could be selected from their nosology.

Since we could not find a method for selecting the best relationship between concepts, the

ambiguity remains and then it is not possible to integrate the type in the metathesaurus. However, another possible application is to make searching on the Web more efficient. Indeed, nowadays search engines (like PubMed or Nomindex [11]) consider individually different concepts but do not interpret them. Consequently, the retrieved Web pages consist of appropriate pages, lost in not directly related ones (the noise). Our tool could implicitly add new relevant relationship(s) between concepts. For instance, if a user wants information about 'fever' and 'appendicitis', the fact of associating the two words with 'sign of' will improve the result of this search. So, our work could provide a more efficient system by adding semantics to searches on the Web. Moreover, this function would be transparent to the users. This approach is similar to that of Semantic Web, as the search becomes intelligent, although it is not based on the usual technologies.

6. Conclusion

This method proposes extending the relationship between concepts. In addition to the classical inter concept relationship, it could be useful to use the inter semantic relationship of the semantic network because the relationship already exists in the UMLS. So it is not necessary to import new knowledge and to increase the complexity of the metathesaurus.

In this way, we reuse an existing knowledge base to improve the metathesaurus. This study has enabled us to confirm that it is possible to acquire knowledge from an external thesaurus if by chance we can map the two dictionaries.

7. References

[1] Lindberg C. The Unified Medical Language System (UMLS) of the National Library of Medicine. J Am Med Rec Assoc, 1990. **61**(5): pp. 40-2.

[2] Bodenreider O, et al. Evaluation of the Unified Medical Language System as a medical knowledge source. J Am Med Inform Assoc, 1998. **5**(1): pp. 76-87.

[3] Burgun A and Bodenreider O. Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. Medinfo, 2001. **10**(Pt 1): pp. 171-5.

[4] Mendonca EA, James MD and Cimino JJ. Automated Knowledge Extraction from MEDLINE Citations. Proc AMIA Symp, 2000: pp.575-9.

[5] Zeng Q and Cimino JJ. Automated Knowledge Extraction from the UMLS. Proc AMIA Symp, 1998: pp.568-72.

[6] Le Duff F, et al. Knowledge acquisition to quality Unified Medical Language System interconceptual relationships. Proc AMIA Symp, 2000: pp. 482-6.

[7] Lenoir P, Michel JR, Frangeul C and Charles G. Réalisation, developpement et maintenance de la base de données ADM. Medecine Informatique. 1981. 6: pp. 51-6.

[8] Seka LP, Fresnel A, Delamarre D, Riou C, Burgun A, Pouliquen B and Le Beux P. Computer Assisted Medical Diagnosis using the Web. Int J Med Inf, 1997. **47**(1-2): pp. 51-6.

[9] Le Duff F, et al. Automatic enrichment of the unified medical language system starting from the ADM knowledge base. Stud Health Technol Inform, 1999. **68**: pp. 881-6.

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

[11] Pouliquen B, Delamarre D and Le Beux P. Indexation de textes médicaux par extraction de concepts et ses utilisation. 6th International Conference on statistical Analysis of Textual Data. 2002. Proceeding to be printed

8. Address for correspondence

MARY Vincent : Vincent.mary@univ-rennes1.fr

Laboratoire d'Informatique Médicale, 2 rue du Pr. Léon Bernard, 35000 Rennes, France