METHODOLOGY FOR BUILDING THEMATIC INDEXES IN MEDICINE FOR FRENCH

Yalina Alphonse, Pierrette Bouillon
{pierrette.bouillon, yalina.alphonse}@issco.unige.ch
ISSCO / TIM, Ecole de traduction et d'interprétation, Université de Genève, Boulevard du Pont-d'Arve, CH-1211
Genève 4, http://www.issco.unige.ch

Abstract

The aim of this project is to propose a methodology in automatically building thematic index from French medical texts in order to improve the IR process. In this article, we focus on the selection process of relevant terms. Contrary to Bourigault and Charlet (1999) who defend a statistical method followed by human intervention, we propose an automatic method that takes advantage of available a priori medical resources such as the MeSH thesaurus (Lindberg et al., 1993) and GALEN (Rector et al., 1996).

1. INTRODUCTION

The aim of this project is to propose a methodology for automatically building a thematic index from French medical texts in order to improve the IR process. Following Bourigault and Charlet (1999), we take as input the set of terms resulting from the terminological extractor SYNTEX (Bourigault and Fabre, 2000), then move on to two other steps: (1) selection of relevant terms in the set of candidate terms produced by SYNTEX and (2) structuring of the relevant terms in order to link them together (see also Aït El Mekki and Nazarenko, 2003). In this article, we focus on step (1), namely the process of selecting relevant terms. Contrary to Bourigault and Charlet (1999) who defend a statistical method (followed by human intervention), we propose an automatic method that takes advantage of available a priori medical resources such as the MeSH thesaurus (Lindberg et al., 1993) and GALEN (Rector et al., 1996). In the following we summarize our results in this project: we first describe the medical corpus, on which we tested our methodology and the SYNTEX terminological database. In order to evaluate the methodology, a set of candidate terms has been evaluated by doctors. These evaluation data will be described in section 4. Finally we move on to the selection methodology and its evaluation.

2. DESCRIPTION OF THE CORPUS

The corpus is a set of patient reports from Geneva hospital in the field of digestive surgery. It contains about 76,000 words. After anonymization and lemmatisation, we obtain 4,068 distinct words. All the reports are from 1997.

Two characteristics are important to mention here and help to clarify the aim of this research: first, the corpus is very short and second, it is very focussed, since all the reports come from the same division. As a consequence, what we want to do is not to cover by an index all the notions of the digestive system disease field. We want instead to find a way to help doctors and nurses retrieve information in a small specific corpus.

3. DESCRIPTION OF THE SYNTEX TERMINOLOGICAL DATABASE

The result of the step of terminology extraction by SYNTEX is a set of 32,918 terms of different grammatical categories: noun phrases (20,949), verbal phrase (9,571), adjectival and participial phrases (2,178), and adverbial phrases (220). These phrases constitute what we will call a *terminological database* and will be the input for our selection method.

The range of the length of the terms varies from 1 to 5 words. Among them, we have 5,727 simple terms and 27,191 complex terms. All complex terms are already decomposed by SYNTEX in a recursive way into two units: head and expansion, for example, *ablation de l'appendice* is decomposed into *ablation* and *appendice.* This is a very interesting feature from our point of view, but it also means that we will need heuristics to determine which terms should be decomposed and which ones should be kept as they are.

4. DESCRIPTION OF THE SET OF EVALUATION TERMS

In order to develop and evaluate our selection method, we built evaluation data with doctors from Geneva hospital. These data reveal some important points about the SYNTEX database:

1. More than 50% of the candidate terms are relevant and should be kept for the index, but only 10% have an exact correspondent in the Mesh. That means the MeSH is far too general and that a simple matching between the candidate terms and the MeSH is not enough. We need a more complex selection methodology;
2. The term frequency is not a good criterion, since it does not indicate reliably the most important terms. For example, the good term *iléite* has a frequency of 2 whereas the term *doigt* has a frequency of 25. Similarly, even among acceptable terms, the frequency is very variable,

for example, *fistule digestive* has a frequency of 22 while *calcul urétéral* only 2. This criterion alone cannot be used to validate the good terms.

3. The length of the term seems to be an important selection criteria. In particular all the complex terms of more than three words were considered as bad by doctors. For instance, "*adénocarcinome occlusive de colon transverse*" was rejected by the doctors and decomposed into "*adénocarcinome occlusive*" and "*côlon transverse*". However, "*ablation de lame"* was extended to "*ablation de lame ondulée*" by doctors.

All these facts together motivate the selection methodology described in the next section.

5. SELECTION METHODOLOGY

The general idea here is to try to learn automatically what is a good term in three main steps: validation of a set of good terms with the thesaurus MeSH; learning of rules that explain what constitutes a good term and finally application of these rules on the candidate terms in order to validate new candidate terms. This will be done in four stages which are described in the following sections.

5.1. Cleaning the candidate terms

In this first stage and according to our evaluation data, we only select phrases that contain less than 4 lexical words. Also, no composition with numbers (dates, measures such as '250g',...) is allowed. At the end of step 1, the list is composed of 22,726 terms, 4,487 simple and 18,239 complex, as summarized in the table 1:

Categories	Effectives	Simple	Complex
nominal	14.061	2.449	11.612
verbal	6.487	579	5.908
adjectival	2.178	1.459	719
total	22.726	4.487	18.239

Table1: statistics on the grammatical categories in SYNTEX

All the terms were recorded, but only the nominal terms will be the object of our selection methodology since verbal and adjectival terms are not present in MeSH. Nevertheless, these categories will be linked to a nominal entry of MeSH, when it is possible.

5.2. Validation of good terms

In order to validate a set of candidate terms and constitute our learning data, we will compare the set of candidate terms with the MeSH. Since an exact matching is not enough, we will take advantage of morphological and synonymy information, as explained in the following.

Exact matching

If we verify if the candidate terms are contained in the MeSH, we obtain 671 nominal terms exactly recognized and 726 partially recognized. The recognition score for simple terms is better (52%) than for complex terms (only 1% of the list). In order to improve the recall for these nouns, we use morphological relations, as described in the next paragraph.

Adding morphology relation

For this process, we apply Zweigenbaum algorithm (Zweigenbaum et al., 2003) in order to derive from our corpus a list of morphologically related terms (for example, *douleur, douloureux; abdominal, abdomen, abdominalisé*). For each group of related terms, we extract the common stem (here *doul* and *abdom*), and we then verify for each complex term if we can find a MeSH term that contains all the stems of the complex term in any order. In that way, the SYNTEX term *douleur de l'abdomen* is now linked to the MeSH term *abdomen douloureux*. 305 new nominal terms were validated with some very limited noise. The next step is to add synonymy relation in order to recognize SYNTEX candidate terms that are semantically related to a MeSH term, like *ablation de l'appendice* which is a synonym of the MeSH term *appendicectomie*.

Adding synonymy relation

If we want to recognize semantically related terms, we need a tool that is able to split the terms and a synonym dictionary that indicates the link between semantically related morphemes. For this project, we use the Geneva Hospital splitter tool* and synonym dictionary (based on GALLEN). We first split the terms, then replace the morphemes by an identifier that corresponds to a semantic class, for example:

Colostomie→*colo + stomie* → *côlon + abouchement*
iléite →*ilé + ite* → *iléum + inflammation*
vagotomie →*vago + tomie* → *nerf_ pneumogastrique + incision*

We then compare these canonical forms with the MeSH terms handled in the same way. With this method, we validate about 7.000 new terms (partially or exactly recognized).

Representation of the terms

After the validation process, each term is represented in a network, for example, *annexite* will receive the following representation:

annexite_Nom annexite_token annexite_lemma Rel_0 freq=1
Morphology- annex:annexe/NOM
Morphemes- annex|6146 ite|12
Texact- annexite:C13.371.056.114
Classe_Mesh : Female Genital Diseases and Pregnancy Complications
Pattern- annexe inflammation

In this network, Morphology gives the stem of the term, plus all the terms that have a similar stem. Texact indicates that the term was found in MeSH and it gives its reference number. Morphemes contain the result of the

*The splitter is a component of a set of commercial tools developed by D. Baud in 1994.

splitting with the MeSH reference number for each morphemes. Similarly, Pattern contains the semantic identifier for each morpheme that refers to the synonym dictionary.

This database of good terms will be the starting point for the learning process.

5.3 Learning of good terms

In this part, we will try to characterize what is a good term, namely what are the most relevant MeSH classes that should be kept for the index and, inside each class, how can we recognize a good term. In order to do that, we take all the terms from SYNTEX that have an exact correspondent in MeSH (the 671 terms described below). From these terms, we count the frequency of the classes given by MeSH. The result is the following classification:

1- Symptoms and general pathology
2- Surgical procedures
3- Diagnosis
4- Digestive system diseases
5- Body regions
6- Neoplasm
7- Physical sciences
8- Digestive system
9- Therapeutics

This set of classes could be considered as the most important classes to be kept in the index. As predicted, the most frequent classes are related to the digestive system and the surgical procedures. Other important classes concern “Diagnosis” and “Body region”. However, terms from “Human activities” that concern voyages, jobs, etc. are not so frequent in our texts and can be discarded from the index.

One thing is to learn good classes; the other is to recognize the candidate terms that belong to them. In order to do that, we then try to guess, for each of the validated MeSH classes, what are the recurrent semantic combinations, in order to obtain rules like:

a surgical procedure (*abouchement*) followed by an organ (for example, in our candidate terms *iléum, colon, etc.*) is a good term that can be classed into *Surgical procedures, Operative* category.

Up to now, we only retain the most statistically relevant rule for our data, that stipulates:

an organ of the digestive system followed by a pathology is a digestive system disease.

Others will be added in the near future and at that time we will be able to define a way to automatically compare the results. The result of the application of the above rule on the whole corpus gives us a score of 772 new terms, with a precision of about 60%. This methodology for selecting classes and new candidate terms for these classes was applied on the whole corpus of terms. We give a first evaluation in the next section.

6. EVALUATION

In order to build up our methodology, we need to confirm the relevance of each step by an evaluation. The first thing to evaluate is whether the SYNTEX terminological database was pertinent enough. Doctors estimated that from a list of more than 200 terms about 60% of the terms were relevant. We submitted the same list of terms to the MeSH thesaurus and the evaluation revealed that about only 10% of the terms are acceptable. The difference shows that MeSH does not cover all the good terms .

In order to evaluate the methodology, we compare our data with the evaluation data of the doctors. As shown by the table 2, the results are almost concordant .

	Accepted terms	Rejected terms
By doctors	69.5%	30.5%
By our methodology	64.4%	35.6%

Table 2 : statistics on the evaluation of our methodology

Among the accepted terms derived by our rules, the noise is estimated at 21% and the silence at 50%. This latter figure is not always easy to evaluate but the more data we collect from doctors the more our evaluations will be comprehensive. The silence will decrease when we will add more rules; it is also possible to test the method with more MeSH classes. The noise is related to the semantic combinations that we implement into the classes. It can be avoided through the improvement in the quality of these combinations.

7. CONCLUSION

In this paper, we tried to define a methodology to select candidates terms that should be kept for an index on the basis of the MeSH. The methodology contains two steps: (1) selection of MeSH classes that should be kept for the index and (2) detection of candidate terms that belong to these classes. Future work includes the structuring of the terms into each class and the structuring of these classes into the index. The question is whether the user will prefer to access the index via the words or via the MeSH classes.

8. ACKNOWLEDGEMENTS

We would like to thank all those who help us in this work, people from the Geneva Hospital and in particular R. Baud, D. Bourigault, P. Fabry and P. Ruch.

9. REFERENCES

- Aït El Mekki, T. and Nazarenko, A., Le réseau terminologique, un élément central pour les index de « fin de livre », *TIA2003*,Strasbourg, 2003.
- Bourigault, D. and Charlet, J., Construction d’un index thématique de l’ingénierie des connaissances, *IC99*, 1999.
- Bourigault, D. and Fabre, C., Approche linguistique pour l’analyse syntaxique de corpus, *Cahier de grammaire 25*, Université Toulouse le Mirail, 2000.

- Lindberg, D.A.B., Humphreys, B.L., McCray, A.T, The Unified Medical Language System, Meth. Inform. med., 32(4), 1993.
- Rector, A.L., Rogers, J.E, Pole, P., The GALEN High Level Ontology, *MIE'96*, Amsterdam, IOS Press, 1996.
- Zweigenbaum, P., Hadouche, F. and Grabar, N., Apprentissage de relations morphologiques en corpus, *TALN 2003*,