Automatic Conceptual Indexing of French Pharmaceutical theses

Vincent MARY †, Bruno POULIQUEN ‡, Franck LE DUFF †
Stefan J. DARMONI ≀, Alain SEGUI ∝, Pierre LE BEUX †
† *Laboratoire d'informatique médicale, Faculté de Médecine - Rennes, France*
‡ *European Commission, Joint Research Centre, IPSC Ispra, Italie*
≀ *Direction de l'Informatique et des Réseaux, CHU - Rouen, France*
∝ *Laboratoire de mathématiques et de physique pharmaceutique*
Faculté de Pharmacie - Rennes, France

Abstract. French pharmaceutical theses are rarely quoted. If the main obstacles originate from language or access barriers, proper indexation could also be blamed. Manually extracted keywords don't necessary come from a structured thesaurus. In the following work, this manual indexing method is compared to an automated one, "Nomindex", based on UMLS. The automated method is improved by the addition of a relevance scoring system. The first indexing step consists of downloading, adapting and indexing theses in electronic format. Results will then be analyzed and sorted by relevance, by comparing classic statistical indices (noise/silence/relevance). It was assumed that the manually obtained keywords were always relevant. The silence of manual indexing is nevertheless high : seven new keywords are proposed by Nomindex, whose results are mixed (10 % of silence, but 50 % of noise). Theses results are promising for the first experiment : pharmaceutical document without lexicon improvement. The indexing, if it is currently insufficient for a real life use, could easily be improved by specific updates of the lexicon.

1 Introduction

Over a thousand pharmaceutical theses are published every year in France, which represents a very large amount of academic work. These publications can be based on bibliography, laboratory or research work, and are validated by approved professors. It could therefore be referred to on the same ground as a publication in the classic sense. Pharmacy is involved in many aspects of modern medicine. However, pharmaceutical theses are rarely referred to in French or international publications. The theses catalogue is updated every 6 month, and theses are only exceptionally available online. This adds to the problem of the language.
A new automated indexing method, "Nomindex" [1], is proposed here, which combines a conceptual automatic indexing method, via the UMLS metathesaurus [2] [3], and a simplified search through the metathesaurus semantic network and the computation of the relevance score TFIDF [4]. The indexing is also made available online with a web browser.

2 Objective

The main objective of this work is to compare the manual indexing system to its automated counterpart, on a sample of theses downloaded from the internet. The first step consists in

downloading, adapting and indexing theses in electronic format. Results will then be analyzed and sorted by relevance, by using classic statistical measures (noise/silence/relevance).

3 State of the Art

3.1 Docthèse / Sudoc

Docthese was, up to recently, the system holding the catalogue of French theses. It was updated by CDROM every 6 months and was available at university libraries. Sudoc [1] (managed by ABES [2]) was developed in parallel and aimed to replace Docthese. Available online, SUDOC offers theses details such as author's name, title, director's name and, usually, four or five relevant keywords. These keywords are not necessarily consistent with the MeSH [5], or with an other thesaurus. This indexing is performed by the author and the director, and is validated by a librarian. SUDOC is updated according to the publication of new theses.

3.2 Indexing Engine

The purpose of the indexing engine is to recognize concepts from a text and to use them to create a database allowing to search documents.
The engine use a local lexicon extracted from the ADM [6] knowledge base. Keywords are grouped by concepts considering associated words, coumpound words, prefixes and suffixes. Each concept corresponds to a notion from the French MeSH. We identify each concept using the UMLS CUI (Concept Unique Identifier).
A relevance score, TFIDF [7], is attached to each concept ofound into a document. This score weights the concepts according to their frequency in the document, relatively to their frequency in the whole corpus. After a first search, it is possible to look for the broader and narrow concepts, or to look for the similar documents (i.e. the documents sharing the same concepts).

4 Material and Method

4.1 Downloading the theses, converting to HTML, indexing.

Pharmaceutical theses are available only in their hardcopy forms, making it difficult to convert back to an electronic format. Our catalogue of theses is therefore composed of theses downloaded from personal sites or with the help of professors of pharmaceutical universities. They are usually in two different formats : MSWord document or PDF.
The indexing engine only works with HTML formated documents. Word 97 allows a direct conversion between word documents and HTML. Two modules are offered by Adobe to deal with the PDF documents: make accessible plugin and save as XML plugin. The theses were saved in HTML 3.2 format, without CSS (Cascading Style Sheets). This format retains the H1 and H2 tags, used by the engine to weight the concepts. The theses are fed to the engine, which extracts the concepts. The relevance score is computed once all the theses have been

[1] Système Universitaire de Documentation
[2] Agence Bibliographique de l'Enseignement Supérieur : `http://corail.abes.sudoc.fr`

Table 1: Silence and noise.

Average number of keyword	
Manual (a)	4.95
Nomindex silence (b)	2.05
Manual silence (c)	6.95
Noise (d)	10.1 / 20

indexed. Keywords obtained by the manual indexing method (docthese, suddoc, or within the theses themselves) are then compiled to form the MeSH.

4.2 Comparison

We compared the list of MeSH keywords obtained from the manual indexing performed by the librarian to the list of keywords proposed by Nomindex. The first 20 keywords were arbitrarily retained. It was assumed that the manually obtained keywords (a) were relevant, since they had been defined by competent people and validated by indexing professionals. Thus, by substracting the keywords found by the two indexing methods, the silence of the indexing engine is defined (b). It is equal to the number of manual keywords which have not been found by the engine.
We then select the most relevant terms among those suggested by the engine by reading the thesis content. Relevant keywords are defined as being neither too specific nor generic.
Two measures are therefore defined : in one hand, the silence (c), coresponding to the number of relevant terms found by the engine but not present in the manual indexing. On the other hand, the indexing noise (d), defined as the number of irrelevant terms.

5 Results

The distribution of the noise, relevance and silence is normal (Kolmogoroff - Smirnoff tests).
The engine found a - b = 2.90 keywords (2.9 / 4.95 = 58.6 %) among the first 20 terms and 75,9 % of the terms are detected beyond the 20 first.
It brings 6.95 new keywords (c).
The noise (d) and relevance are respectively of 50,5 % (10,1 / 20) and of 49,5 % (9,9 / 20).
The theses were indexed in few minutes each one, and the final calculation of the relevance in about 20 minutes (at the Sun workstation).
On a total of 21 theses, only 8 could be found in the Sudoc. In fact, an indexing was generally found in the thesis itself (9 theses) or has been provided by the university library (4 theses).

6 Discussion

Nomindex works by using a medical lexicon. Pharmacy is multidisciplinary and includes knowledge beyond the medical domain, often stepping into more scientific subjects, such as physics or mathematics. However the normality of the distribution of the noise and the silence shows that the engine correctly manages this semantic disparity.
The engine found 58 % of the manual keyword among the 20 proposed terms (this seems to

be few), but more than 75 % of all terms. We can explain this relatively low rate (58%) by the too high noise : only half of the 20 concepts are pertinent. Thus, even the notion of limit seems a posteriori not pertinent.
It would be possible to work differently to know if the keywords selected by the human indexer, which are the reference, where found back well. For instance, an evaluation of the engine would allow to confirm that we really found the theses from their own keywords, or it would be possible to seek the acceptable threshold of noise from which the totality of the manual keywords appears ahead of the automatic indexation.
This classification problem can be explained by a performant extraction: on average, the engine brings almost 7 new relevant keywords.
Nevertheless, the ratio noise/silence is correct if we consider that:

- the lexicon covers mainly the domain of medical diagnosis, and not the whole medical domain.
- it is the first experiment on pharmaceutical documents, and the lexicon has never been improved in that domain, as for general medical document. It's enrichment of specific pharmaceutical terms will certainly reduce the silence.
- the noise comes too often from the same concepts which are wrongly recognized. A small specific modification of the lexicon would suppress a lot of these mistakes. In the case of framework on a semi-automatic indexing by the librarian, those ones could improved the lexicon, or use terminological constitution tools.[8].

Silence became weak when compared to Sudoc, which finds only one third of theses . It would seem that the theses aren't put directly on the Sudoc system. Some theses,older than six months did not appear. In this respect, Nomindex could solve these problems. The university recovers the thesis electronic version at the same time as their registration. The user could have the benefit of a powerful search engine, with an search extension making use of a validated metathesaurus.
The computational time of indexing is compatible with an online use (on a central web server for example). Only the computing time of the TFIDF could be problematic when several years of theses will be indexed. Initially, it would be possible to apply this engine to the abstract only.
As previously mentioned, the sample of theses was obtained from the Internet, and might not be a representative group of French theses. These limits had an impact on our work, but we will continue on a more significant number of theses, with the assistance of tha administration of pharmaceutical faculties.

7 Conclusion

It is up to the universities, professors and librarians to amphasize the work realized within their university [9]. During this work, we proved that our semantic indexation would allow a powerful use of the UMLS (metathesaurus and semantic networks) and an effective extraction of the knowledge. The indexing, if it is currently insufficient for a real-life use, could easily be improved by specific updates of the lexicon.
Remain the inconvenient impossible to circumvent of an indexing based on the ADM lexicon, which don't distinguish the word "aine" from "aîné" [10]; may be will it necessary to proceed

a preliminary syntaxic parsing.
Thereafter, an improvement either on the query expansion techniques [11] or improvment on the UMLS itself [12] will allow to put as provision of the research workers and the students an provided medical knowledge, simple of access.

References

[1] B. Pouliquen and D. Delamarre and P. LeBeux, Indexation de textes médicaux par extraction de concepts et ses utilisations. In: 6th International Conference on Statistical Analysis of Textual Data 2002. Proceedings to be printed.

[2] D.A. Lindberg and B.L.Humphreys and A.T. McRay, The Unified Medical Language System. In: Methodes Inf Med 1993 Aug **32** (4):281–91.

[3] P. Nadkarni and R. Chen and C. Brandt, UMLS concept indexing for prodution database: a feasibility study. In: J Am Med Inform Assoc. **8**(1) (2001) 80–91.

[4] G. Salton and C. Buckley, Term weighting approaches in automatic text retrieval. In: Information Processing and Management **24**(5) (1988) 513–523.

[5] National Library Of Medicine, Medical Subject Headings (1986).

[6] P. Lenoir and J.R.Michel and C. Frangeul and Chales G, Réalisation,développement et maintenance de la base de données A.D.M. Médecine informatique. 1981; **6** 51–6.

[7] D.C. Berrios, Automated Indexing for Full Text Information Retrieval. In: Proc AMIA Symp (2000) 71–5

[8] D. Bourigault and C. Jacquemin, Construction de ressources terminologiques, Ingénierie des langues. Hermes (2000).

[9] Ministère de l'éducation, Bulletin Officiel du ministère de l'Education Nationale et du ministère de la Recherche N°34. (2000) 430–4

[10] F. Le Duff and A. Burgun and B. Pouliquen et al., Automatic enrichment of the unified medical language system starting from the ADM knowledge base. In: Stud Health Technol Inform bf 68 (1999) 881–6.

[11] W. Hersh and S. Price and L. Donohoe, Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. In: Proc AMIA Symp (2000) 344–8

[12] W.T. Hole and S. Srinivasan, Discovering Missed Synonymy in a Large Concpt-Oriented Metathesaurus. In: Proc AMIA Symp (2000) 354–8