Automatic construction of dictionnaries, application to product characteristics indexing

Suzanne PEREIRA[a,c,d], Blandine PLAISANTIN[a], Michelle KORCHIA[a], Nicolas ROZANES[a,b], Elisabeth SERROT[a], Michel JOUBERT[c] and Stefan J DARMONI[d]
[a] *Vidal, Issy les Moulineaux, France* [b] *LINALCO, Paris* [c] *LERTIM, Faculté de Médecine, Université de la Méditerranée, Marseille* [d] *CISMeF & TIBS, LITIS, Rouen*

Abstract. Summary of Product Characteristics (SPC) indexing enables to extract all the information needed to analyze a prescription and find some inappropriate medications. We evaluate a French Multi-Terminology Indexer tool (F-MTI) for SPC automatic indexing. This tool uses a dictionary containing the textual forms that are likely to appear in natural language text for the drug clinical particular terms contained in the Vidal thesaurus (TUV). We developed a method to automatically generate this dictionary. The evaluation showed a precision of 52.9% and a recall of 46.2%. F-MTI will be integrated in a semi-automatic indexing tool.

Keywords. Abstracting and Indexing/methods, Natural Language Processing, Information, Evaluation Study France.

1. Introduction

Computerized drug prescribing alerts can improve patient safety by reducing the use of potentially inappropriate medications [1]. The Summary of Product Characteristics (SPC) is the basis of information for health professionals on how to use the medicinal product safely and effectively. A SPC contains a description of a certain medicinal product's properties and the conditions attached to its use. This document, which is spread by the French health products safety agency (AFSSAPS), provides information on the drug composition, pharmaceutical form and strength, authorized applications (indications), adverse reactions, cautions and safety regulations.

SPC indexing enables to extract all the information needed to analyze a prescription and find some inappropriate medications. The Vidal Company manually indexes SPCs by assigning to them terms from different thesauri. Among them, four thesauri describing indications, contraindications, adverse reactions, precautions for use, have been recently aggregated in the TUV thesaurus (French acronym for Unified Thesaurus of Vidal).

SPC manual indexing is a tedious task provided by pharmacists that need to be helped. Therefore, we developed a French Multi-Terminology Indexer tool (F-MTI). This tool has already been evaluated for Medical record automatic indexing using ICD10 [2] and SNOMED 3.5 [3] and for Web resources automatic indexing using MeSH [4]. In this paper, we show an evaluation of the performances of F-MTI for SPC automatic indexing.

2. Material & Methods

2.1. F-MTI

F-MTI uses several Natural Language Processing approaches to analyze a document to be indexed and translate the emerging concepts into the appropriate controlled vocabulary [2,3,4]. One of these approaches uses a terminology dictionary containing full TUV terms and their variants to extract clinical particular concepts. Dictionary entries contain a specific form of a TUV term that is likely to appear in natural language text (i.e., the actual term, its inflected forms, its synonyms or an inflected form of a synonym) as well as the TUV term, itself. For example, the term "diminution des facteurs de coagulation" («decrease of coagulation factor» in English) is linked to the variant form "diminutions des facteurs de coagulation". This dictionary is applied to a document to be indexed using the NOOJ tool, a linguistic annotation system for corpus processing [5].

2.2. Automatic construction of dictionaries

The manual constitution of a dictionary is a very time consuming task. For the 11,965 terms and synonyms of the TUV terminology, we integrated automatically variant forms coming from several previous projects. We have also elaborated two methods to gather automatically textual forms from corpus and create inflection forms.

First, all the inflection forms and synonyms included in the terminology are integrated automatically in the dictionary and linked to the corresponding preferred term.

Then to complete this list, we analyzed the inflection forms and synonyms from previous projects. We analyzed the UMLF lexicon [6], the MeSH dictionary of MAIF [7] and the lexicons of the VUMeF project [8]. Variant forms linked to equivalent TUV terms have been integrated automatically in the TUV dictionary.

A big part of the inflection forms and synonyms of a term are the result of the combination of the inflected forms or synonyms of the pertinent words combined using link words. Syntactic graphs can extract these textual forms from corpus as they can precise inflection forms and synonyms of a term. For example, the different forms of the term «diminution des facteurs de coagulation » can be represented by the graph presented at figure 1. <diminution>, <facteur> and <coagulation> corresponds to the inflection forms and synonyms contained in the dictionary of words. <MVP> is the lexicon of link words (983 stop words selected for this task including the dash). This graph does not take into account the order of the words.

The dictionary of words contains all the inflection forms and synonyms for each pertinent single word of the terminology (38,219 entries) (example: "dimimutions,diminution,X"). This dictionary has been created using medical and general French dictionaries: Morphalou 17[1] (590,020 entries), Lexique 3[2] (55,000 entries), MeSH dictionary [7] (44,856 entries), UNITEX dictionary[3] (683,824 entries), NooJ dictionary [5], UMLF [6] (23,141 entries), VUMeF dictionaries [8] (2,742 entries for drug notions) and the Integral Dictionary by Memodata [9] (540,000 entries).

1 TLFnorme : http://www.cnrtl.fr/lexiques/morphalou
2 http://www.lexique.org
3 http://www-igm.univ-mlv.fr/~unitex/

Inflections for dosage unit and for numbers completed the dictionary of words.

The application of the graph (figure 1) on the corpus using NOOJ tool extracts the inflection forms: <diminution des facteurs de la coagulation> and <diminution du facteur de coagulation>. These inflection forms will complete the TUV dictionary.

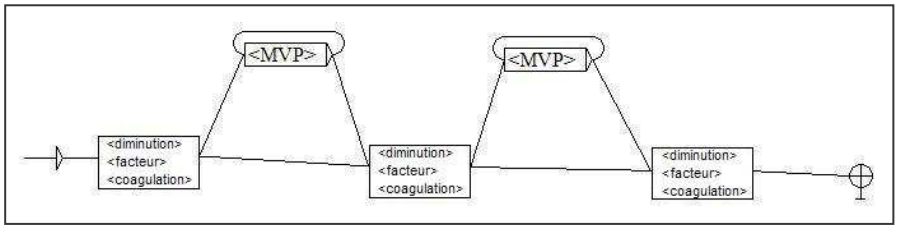

Figure 1. Syntactic graph for the term «diminution des facteurs de coagulation » («decrease of coagulation factor» in English)

More inflection forms (plural and singular) can be obtained by a simple algorithm (add a "s" at the end of the word where there is nothing etc.).
We have produced these inflection forms for terms containing less than three words.

2.3. TUV dictionary

The resulting TUV dictionary contains 40,266 inflection forms and synonyms for 11,965 terms:

- 11,965 terms, added to 23,800 forms found in other lexical resources,
- 11,965 graphs generated automatically using NOOJ interface and applied to a corpus including the integrality of Vidal's SPCs in French (14,104 SPCs), 100 discharge summaries of the Rouen University Hospital, and the integrality of the CISMeF corpus[4] (40,000 Web resources) generated 550 new inflections and synonyms (550 inflection forms and synonyms (55%) have been validated by a TUV expert among 1,007),
- 3,951 inflection forms have been added using our algorithm (3,951 inflection forms (92.7%) have been validated by a TUV expert among 4,279).

2.4. SPC indexing evaluation

SPC indexing consists in applying the TUV dictionary using NOOJ tool. The dictionary is applied once for a corpus of SPCs that need to be indexed. The file obtained includes all the inflected forms and synonyms found in each SPC with their link to the preferred TUV term and their localisation in the SPC. Another graph helps to locate items (*Therapeutic indications, Contraindications*, *Undesirable effects*, *Special warnings and precautions for use*, *Overdose* and *Effects on ability to drive and use machines*) so that we can link each TUV term to an item. For each item, we can deduce the type of each term: the item *Therapeutic indications* is linked to the type "indication". The item *Contraindications* is linked to the type "contraindication". The items *Undesirable effects* and *Overdose* are linked to the type "contraindication". The items *Special warnings and precautions for use* and *Effects on ability to drive and use*

4 Catalogue and Index of Online Health Resources in French : http://www.cismef.org/

machines are linked to the type "caution regulation".

We compared F-MTI's automatic indexing to the manual indexing for a corpus of SPCs. We used a corpus of 5,191 SPCs manually indexed by the Vidal indexers using the previous four thesauri. We converted these RCP to text format. We also converted manually four thesauri indexes to TUV indexes *via* the mapping "four thesauri/TUV" (7,834 entries) to obtain a TUV index for all SPCs. All previous thesaurus terms have a type ("contraindication", "indication", "adverse reaction" or "precaution for use"), which depends on the thesaurus they belong to, these ones have been kept after mapping.

We calculated precision and recall for the comparison between automatic and manual indexing. We considered different categories:

- All types of terms considered separately ("contraindication", "indication", "adverse reaction" and "precaution for use").
- All types of terms are considered (average of the previous category).
- The entire document indexing without taking into account the types of terms.

3. Results

F-MTI performances shows a precision of 57.6% and a recall of 43.4% when compared to manual indexing (see Table 1). When we consider the performances depending on the items, the results are very different. The best performances are obtained for the "Adverse effect" type with a precision of 77% and a recall of 59.4%.

Table 1. Results of the TUV terms indexing for SPC evaluation

Categories	Precision (%)	Recall (%)
Indication	48.1	21.7
Contraindication	46.1	23.5
Adverse effect	77.0	59.4
Precaution for use	28.4	49.3
All types of terms	52.9	46.2
The entire document indexing	57.6	43.4

4. Discussion

F-MTI automatic indexing of SPCs - The manual indexers found the performances obtained satisfying and useful. The results depend on the type of term considered. Indeed, the size and the complexity of the terms depend on the type of the terms. Terms with the type "indication" or "contraindication" are more complex than the others. Then, they are more difficult to identify and the recall is lower.

The recall can also be explained by the fact that some items (like *Qualitative and quantitative composition*) that can contain precaution for use terms have not be taken into account. Moreover, we have considered that each item could contain terms of one type but it is not true for all items. For example, terms from the item *Pregnancy and lactation* can have the type "contraindication" or "precaution for use".

Most of the indexing errors (influencing recall) are due to insufficient coverage of the inflected forms and synonyms for TUV terms in the dictionary. Other methods should be conceived to complete the dictionary. Some other errors can be linked to the conversion of the documents into text that enables to find some TUV terms. The restitution of the titles is sometimes bad that leads to the non recognition of the right type. The tables are not converted but they can contain some terms to be indexed.

Advantages and disadvantages of the dictionary of terms method - The dictionary of terms method is very quick for indexing documents, less than one minute for our entire corpus. Moreover, inflection forms and synonyms are validated before indexing. That enables to generate a minimum of indexing errors. Unfortunately, the quality of the indexing depends on the coverage of the different forms for the terms of a terminology. Our dictionary is actually insufficient.

Perspectives - To enhance the quality of SPC indexing, we planned to integrate TUV indexing rules that exist manually but are not implemented in the tool. Moreover, the items not taken into account will be in the tool. The implementation of new graphs would permit to disambiguate the type of the terms in case of multi-types assignation. Works on XMLization of SPCs will solve the conversion problems.

5. Conclusion

In this study, we developed a method of automatic dictionary construction and evaluated a French Multi-Terminology Indexer to index SPCs with TUV terms. The results seem sufficient to use F-MTI for semi-automatic indexing. Then F-MTI will be integrated in BIBLIS, an helping tool for manual SPC indexing. The indexer expert would be helped by the automatic indexing proposition made by F-MTI.

References

[1] J.F. Bergmann and al. Good medical practice for drugs. Definition, guidelines, references, fields of action and applications, *Therapie* **63** (2008), 4:267-80.

[2] S. Pereira, P. Massari, M. Joubert, E. Serrot, S.J. Darmoni. Exploring Multi-terminology Indexing of Discharge Summaries, *Stud Health Technol Inform* (2008).

[3] S. Pereira, P. Massari, A. Buemi, B. Dahamna, E. Serrot, M. Joubert, S.J. Darmoni. Evaluation of two French SNOMED indexing systems with a parallel corpus. KR-MED 2008 - Representing and sharing knowledge using SNOMED International Conference, Phoenix, AZ, USA, June, 2008.

[4] S. Pereira, A. Névéol, G. Kerdelhué, E. Serrot, M. Joubert, S.J. Darmoni. Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. *AMIA Annu Symp Proc* (2008), 586-90.

[5] M. Silberztein. NooJ: a linguistic annotation system for corpus processing, *Proceedings of HLT/EMNLP* Human Language Technology Conference (2005), 10-11.

[6] P. Zweigenbaum, R. Baud, A. Burgun, F. Namer, E. Jarrousse and al. UMLF: a unified medical lexicon for French. *Int J Med Inform* **74** (2005), 2-4:119-24.

[7] A. Névéol, A. Rogozan, S.J. Darmoni. Automatic indexing of online health resources for a French quality controlled gateway. *Inf Process Manage* **42** (2006), 3: 695-709.

[8] S.J. Darmoni, E. Jarousse, P. Zweigenbaum, P. Le Beux, F. Namer and al. VUMeF: Extending the French part of the UMLS, AMIA *Annu Symp Proc* (2003), 824.

[9] D. Dutoit, P. Nugues, de Torcy T. The Integral Dictionary: a lexical network based on computational semantics. *Springer Ed*, Proceedings of ICCSA International Conference on Computational Science and its Applications (2003).