Development of two linguistic resources to improve the extraction of causal relations in medical language

The objective of this research is to analyse and to automatically extract causal expressions from a medical corpus, built from patient records written by doctors.

This research is based on the idea that causal relations can be expressed either directly (causal connectives, causal verbs, causal expressions [*rendre x* + adj.]) or indirectly by words that potentially have a causal meaning, even if this causal interpretation is not the first and intuitive meaning of these words (connectives [*en cas de*, *et*], syntactic constructions [relative clauses, participles and gerundives]). Based on this premise, we simultaneously form two linguistic resources by means of an exploratory study of the corpus. This anonymised corpus is made of 226 000 medical texts (around 93 million words) that were collected in six services of Brussels hospitals.

First, we design a causal lexicon which is constituted by lexicon and causal adverbials both from standard language and from medical language (such as *post* in 1).

(1) a *amaigrissement de 40 kg **post** pose d'un anneau gastrique en [DATE]*

b weight loss of 40 kg ***post***-putting on a gastric band in [DATE]

Secondly, we develop a program of disambiguation (coded in Perl) which rests upon the method of the automated text categorisation. We propose here to focus on six French connectives that correspond with this causal ambiguity (*et*, *si*, *après*, *quand*, *lorsque*, *dès que*). We adhere to a 4-steps methodology:

(i) Automatic extraction of these connectives (Unitex)
(ii) Manual annotation of causal and non causal meanings by a linguist and by a specialist physician
(iii) Identification of causal contexts
(iv) Lexical formalisation of these contexts

We finally implement these two resources in a program of extraction of causal relations in order to appreciate the value added by these linguistic resources compared to the existing programs in the medical field.

Keywords : Automated text categorisation, medical language, causality, corpus linguistics, information extraction

Bibliography

[1] Gross G. (2009). *Sémantique de la cause*. Peeters, Leuven (Collection linguistique, 96).

[2] Nazarenko A. (2000). *La cause et son expression en français*. Ophrys, Gap (L'essentiel français).

[3] Wall, L., Christiansen, T. & Orwant, J. (2001). *Programmation en Perl*, 3e edition. O'Reilly, Paris.

[4] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Survey*, 34 (1), 1-47.

[5] http://unitexgramlab.org/fr.

[6] Garcia, D. (1998). *Analyse automatique des textes pour l'organisation causale des actions, Realisation du systeme Coatis*. Thèse de doctorat, sous la direction de Prof. Desclés, J.-P., Université Paris IV.

[7] Khoo, C., Chan S. & Niu, Y. (2000). Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. *Proceedings of ACL*, Hong Kong, 2000.