Contrasting French nominal terms to common language NPs – towards a rule-based term extractor

Ágoston Nagy

According to the traditional approach (e.g. Wüster 1976), terms are lexical units having the following characteristics: they are related to a domain (e.g. informatics, physics), they are connected to one and only one concept that they denote. According to Justeson & Katz (1991), in computational terminology, nominal terms are in the centre of interest since these are the terms having the most complex syntactic structure ranging from simple nouns to extremely long nominal compositions (e.g. in French *donnée* 'data' and *système de gestion de base de données* 'database management system'). Getting to know the internal composition of terms as precisely as possible is of crucial importance when elaborating rule-based automatic term extractors. Term extraction (TE) tools are created to automatically extract technical terms from written technical corpora. TE is usually realised by combining rule-based (linguistic) and statistical methods. In general, statistical methods are used to find all term candidates in a text, and this list is then filtered by rule-based methods. (Cabré *et al.* 2001)

The aim of the presentation is to point out the specificities of nominal terms, especially with respect to prepositional complements (including the preposition+noun sequences of nominal compounds) and adjectival adjuncts. As the main aim of a nominal term extractor is to extract nominal terms, and therefore exclude common language noun phrases (NP), the internal structure of nominal terms will be contrasted to that of ordinary NPs. This contrastive analysis will be based on French grammar books or articles about the structure of NPs, like Riegel *et al.* (2009) and Anscombre (1991), respectively.

The corpora I use for the analysis consist of the description parts of patents written in French. For this presentation I chose ten descriptions of an IT domain (more precisely the G06F patent class) and ten descriptions from a Human necessities domain (namely the A23L class) in which I had manually annotated all terms. The average length of a description is nearly 3500 tokens.

The presentation also reveals how efficient this contrastive analysis is since the patterns gained in this way will be used in a term extractor based on nominal term patterns: my hypothesis is that rule-based approach to term extraction from French corpora can already be efficient without statistical methods since (1) in French, terms tend to have internal structures that are not typical of common language nouns or nominal compositions and (2) patents represent a discourse type that corresponds to nearly all prerequisites of a scientific text (e.g. impersonal style, excessive usage of terms).

On the basis of Cabré *et al.* (2001), I hypothesised that rule-based approaches result in high recall and lower precision. The results confirmed this hypothesis: using only rule-based approach, I achieved a recall of 0,78 and a precision of 0,59. This rule-based method is then complemented with a rule-based filter based on a stopword list containing connectives (e.g. *par exemple* 'for example') and proper names like person names. The filtering resulted in the increase in both the precision (0,66) and the recall (0,83).

The most important message of the results is that term extraction can be efficient not only with the help of statistical methods but also with linguistic methods, especially when recall values are more important. However, the above mentioned term extractor will be later on complemented with statistical filtering methods in order that precision values be improved.

References

Anscombre, M. J-C., 1991, L'article zéro sous préposition, *Langue française* (91), 24-39.

Cabré, M. T., Bagot, R.E., Vivaldi Palatresi, J., 2001, Automatic term detection. A review of current systems. In: Didier Bourrigault, Christian Jacquemin, Marie-Claude L'Homme (eds.) *Recent advantages in Computational Terminology*. Amsterdam/Philadelphia, John Benjamins Publishing Co., 53-87.

Justeson, J. S., Katz, S. M., 1995, Technical terminology : some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1): 9-27.

Riegel, M., Pellat, J-Ch., Rioul, R. (2009) *Grammaire méthodique du français* (4th edition). Paris: PUF.

Wüster, E. 1976. La théorie générale de la terminologie, un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et la science des objets. Actes du colloque international de terminologie, Québec 5-8 octobre 1975, Québec, L'Éditeur officielle du Québec.