Extraction multilingue de termes à partir de leur structure morphologique

Delphine Bernhard*

*TIMC-IMAG
Institut de l'Ingénierie et de l'Information de Santé
Faculté de Médecine
F-38706 LA TRONCHE cedex
Delphine.Bernhard@imag.fr
http://www-timc.imag.fr/Delphine.Bernhard

Les méthodes d'extraction automatique de termes utilisent couramment des patrons décrivant la structure des termes (Ibekwe-Sanjuan et Sanjuan, 2004; Enguehard, 1992; Vergne, 2005). Dans les domaines scientifiques ou techniques comme la médecine (Namer, 2005), de nombreux termes appartiennent au vocabulaire savant et sont construits à partir de formants classiques grecs ou latins situés en début (extra-, anti-) ou en fin de mot (-graphe, -logie). La méthode que nous proposons utilise la structure morphologique des termes en vue de leur extraction et de leur regroupement[1].

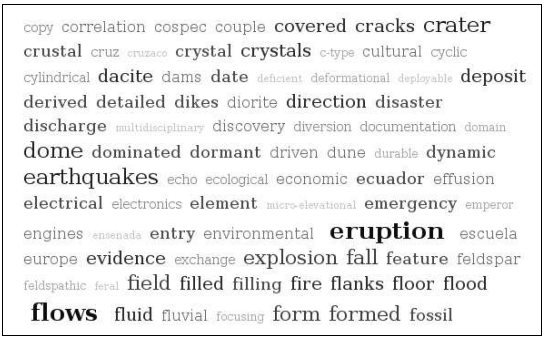
Le système extrait tout d'abord les mots du corpus puis identifie les formants à l'aide de l'expression régulière suivante : **([aio]-) ?(\w{3,}[aio])-**. Même si cette expression régulière est limitée aux formants se terminant par *a*, *i* ou *o*, elle n'est pas uniquement valable pour le français. On trouvera, par exemple, "**chimio-**hormonothérapie" en français, "**chemo-**radiotherapy" en anglais ou "**Chemo-**radiotherapie" en allemand.

Une fois les formants identifiés, les termes sont repérés à l'aide d'un patron qui décrit leur structure morphologique : **F+M** où **F** est un formant et **M** un mot du corpus de longueur supérieure à 3. Le caractère + indique la succession possible de plusieurs formants en début de terme. Lorsque ce patron s'applique à un des mots du corpus, deux termes sont reconnus : le terme de structure **F+M** et le terme de structure **M**. Ainsi, à partir du mot "**radio**thérapie" qui contient le formant "radio", on extrait les termes "radiothérapie" et "thérapie".

Afin de faciliter l'analyse des termes extraits, des familles de termes sont formées en regroupant les termes contenant le même mot M. Le mot M est appelé représentant de la famille. De plus, deux familles sont réunies si leurs représentants ont une chaîne initiale commune de longueur supérieure ou égale à 4 et si l'on retrouve le même formant dans un terme de chaque famille. Le représentant final de chaque famille est le terme le plus fréquent.

Les résultats de l'extraction terminologique sont présentés sous forme de liste pondérée au format HTML (voir figure 1). Ce type de liste se caractérise par l'utilisation d'un code de couleur et d'une taille de police dépendant de la fréquence d'occurrence d'un terme (Véronis, 2005). Seuls les termes représentants de chaque famille sont affichés et le poids d'une famille dans la représentation finale est déterminé par la fréquence cumulée de tous les termes de la famille.

[1]Ce travail a été soutenu en partie par la Commission européenne (projet NOESIS, IST-2002-507960)



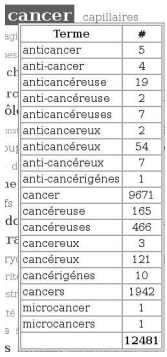


FIG. 1 *Visualisation des termes sous forme de liste pondérée (à gauche) et détail d'une famille de termes (à droite)*

Le systŁme a ØtØ expØrimentØ sur 4 corpus de textes couvrant deux domaines scienti ques distincts, celui de la volcanologie et du cancer du sein, dans deux langues diffØrentes, le français et l'anglais. Les premiers rØsultats obtenus montrent que l'utilisation de la structure morphologique permet de mettre à jour des termes peu frØquents qu'une approche purement frØquentielle ne pourrait identi er. Ces deux approches sont donc complØmentaires. L'algorithme de regroupement permet quant à lui de rassembler les variantes orthographiques, exionnelles et dØrivationnelles des termes dans une mŒme famille.

Références

Enguehard, C. (1992). *ANA Apprentissage Naturel Automatique d'un Réseau Sémantique*. Ph. D. thesis, UniversitØ de Technologie de CompiŁgne.

Ibekwe-Sanjuan, F. et E. Sanjuan (2004). Mining Textual Data through Term Variant Clustering: the TermWatch System. In *Actes de Recherche d'Information Assistée par Ordinateur (RIAO 2004)*, pp. 487 503.

Namer, F. (2005). MorphosØmantique pour l'appariement de termes dans le vocabulaire mØdical : approche multilingue. In *Actes de TALN 2005*, pp. 63 72.

Vergne, J. (2005). Une mØthode indØpendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlØe. In *Actes de CIDE 8*.

VØronis, J. (2005). Nuage de mots d'aujourd'hui. http://aixtal.blogspot.com/2005/07/lexique-nuage-de-mots-daujourdhui.html.

Summary

This articles describes a method for the automatic extraction of terms from corpora of specialised texts. It makes use of morphological elements located at the beginning of words to discover terms and group them in families. Results are displayed as a weighted list.