# Utilisation de graphes sémantiques pour l'extraction et la traduction des idées essentielles d'un texte

Romain André-Lovichi*, Kamel Smaïli*, David Langlois**

*Loria BP 239 54506 Vandoeuvre Lès-Nancy France
**Loria/IUFM de Lorraine
andrelov@loria.fr, smaili@loria.fr, langlois@loria.fr

## 1 Contexte et motivations

Les évolutions technologiques mettent à notre disposition toujours plus d'informations. Dans un tel contexte, pouvoir accéder rapidement aux idées essentielles contenues dans un ou plusieurs documents, éventuellement écrits dans des langues différentes, apparaît comme une perspective de plus en plus séduisante.

Différentes études ont ainsi été menées dans les domaines de la traduction et du résumé automatiques dans le but de transposer de manière automatique un texte d'une langue vers une autre (Hutchins (2001)), ou d'en obtenir une version abrégée (Luhn (1958), Edmundson (1969) ou encore Kupiec et al. (1995)).

La production d'un résumé sous forme textuelle pose cependant différents problèmes (cohérence syntaxique, liens entre les phrases, etc.), c'est pourquoi nous avons essayé dans cette étude de construire une nouvelle représentation des informations contenues dans un texte.

## 2 Graphes sémantiques

### 2.1 Définition

Notre idée est la suivante : représenter les mots les plus importants d'un texte ainsi que leurs contextes. Nous tentons de suivre ainsi ce que nous pensons être la démarche d'un téléspectateur face à un reportage dans une langue qu'il maîtrise mal : identifier les idées essentielles (à partir de leur fréquence d'apparition), et utiliser leurs contextes pour déduire leurs traductions.

Nous appelons donc *graphe sémantique* un graphe constitué de deux types de sommets : les *sommets principaux* qui correspondent aux mots les plus fréquents dans le texte, et les *sommets satellites* qui sont associés aux sommets principaux dont ils constituent le contexte.

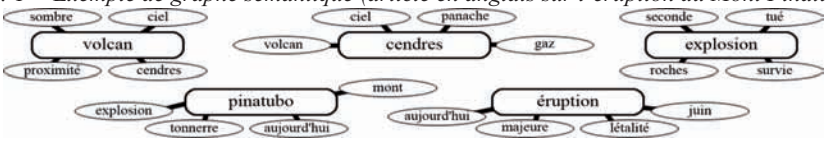
### 2.2 Construction

Pour déterminer ces contextes, nous utilisons la notion de *déclencheurs* (Tillmann et Ney (1996)) : un mot fréquent $x$ est vu comme un *déclencheur* auquel on associe la liste de ses *déclenchés*, c'est-à-dire les mots les plus fortement liés à $x$ au sens de l'information mutuelle.

Pour chaque déclencheur, on traduit alors les paires *déclencheur déclenché*. On en déduit par un vote à la majorité la traduction du déclencheur étant donné son contexte (on obtient un sommet principal), puis les traductions des déclenchés (soit donc des sommets satellites).

### 2.3 Résultats

FIG. 1 – *Exemple de graphe sémantique (article en anglais sur l'éruption du Mont Pinatubo)*




Nous avons pu constater que cette méthode utilise effectivement le contexte d'un mot pour lever une éventuelle ambiguïté sur sa traduction : *échecs* est traduit par *chess* et non par *failures* si le mot apparaît en même temps que *Kasparov* et non *Navratilova*.

En outre, on a aussi pu observer à partir de textes alignés que les résumés-graphes générés dans une langue puis traduits recouvrent la même sémantique que les graphes générés à partir des documents correspondant dans la langue cible.

Notre approche est quasi-indépendante des langues considérées : seule la liste de mots-outils utilisée dans le calcul des déclencheurs dépend de la langue, et même cette dernière peut en fait être générée de façon automatique.

## 3 Perspectives et conclusion

Une amélioration possible de cette méthode est l'utilisation de $n$-grammes : plutôt que de raisonner à l'échelle du mot, on pourrait travailler sur des groupes de $n$ mots consécutifs et favoriser les séries les plus longues. Une autre étape très importante dans le développement de cette méthode sera l'évaluation. L'élaboration d'une mesure sur ces résumés-graphes permettra en effet une évaluation chiffrée des performances de cette approche. L'objectif de notre étude était de présenter l'idée de cette représentation sous forme de graphe. Cette dernière semble effectivement permettre une visualisation schématique du contenu d'un document écrit dans une langue étrangère, répondant ainsi au problème posé.

## Références

Edmundson, H. (1969). New methods in automatic extracting. *Journal of the Association for Computing : Machinery 16*(2), 264–285.

Hutchins, J. (2001). Machine translation over fifty years. *Histoire Epistémologie Langage 23*(1), 7–31.

Kupiec, J., J. Pedersen, et F. Chen (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 68–73. ACM New York, NY, USA.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

Tillmann, C. et H. Ney (1996). Selection criteria for word trigger pairs in language modelling. *Lecture Notes in Computer Science 1147*, 95–106.