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).