

Résumé automatique dans un contexte bilingue ou trilingue

Directeur : Juan Manuel Torres Moreno(MCF HDR, 50%); **Co-encadrant** : Eric SanJuan (MCF, 50%) ;

Mots clefs: Résumé automatique, Extraction d'Information textuelle, Système de Question Réponse, Traitement automatique de la langue naturelle écrite.

Étant donné l'énorme quantité d'information disponible, notamment en ligne sur le Web, lire et comprendre les informations pertinentes sont des tâches très coûteuses. Dans ce scénario, les applications de Traitement Automatique du Langage Naturel (TAL) se présentent comme des solutions très importantes, par exemple, en résumé automatique de textes, la récupération et l'extraction d'information et les systèmes de Questions-Réponses.

Le résumé est la tâche de produire un condensé à partir d'un ou plusieurs textes source. C'est une tâche naturelle pour l'être humain, mais elle représente un grand défi pour une machine. Pour cette raison, elle est un des domaines les plus étudiés actuellement au TAL. En plus de l'application décrite précédemment, les résumés sont utilisés au jour le jour dans de nombreuses tâches, par exemple, les sommaires (ou avis) de livres, les synopsis de films et de romans, le résumé des prévisions météorologiques, les *abstracts* des articles scientifiques, etc.

Le LIA possède une grande trajectoire dans les domaines du résumé automatique, compression de phrases et de l'analyse discursif automatique. En compression, on a travaillé sur la tâche de résumé automatique depuis différents points de vue : a) statistiques, b) symboliques ou c) hybrides. Le LIA a aussi fait des recherches récemment sur l'évaluation automatique de résumés.

Nous proposons la recherche et l'exploration conjointe du résumé automatique dans un contexte bilingue ou trilingue. En effet, sous l'effet de la mondialisation, un nombre croissant d'utilisateurs des usagers du Web est au moins bilingue. Un résumé bilingue ou trilingue consiste à grouper dans une langue par défaut l'information que l'on retrouve en commun dans les langues choisies, et à enrichir ce résumé en allant chercher l'information exprimée exclusivement dans une seule des langues, sans pour autant chercher à la traduire. Il s'agira de constituer un ensemble d'entités que l'on sache identifier sur plusieurs langues et qui suffise à établir si une information est redondante ou pas.

Au départ, nous avons l'intention de nous concentrer sur les méthodes et techniques de résumé mono et multi-document éprouvées, afin de choisir les meilleures pour différentes sélections de paires ou de triplet de langues. En principe, l'approche sera de type extractive, dans le but de produire des résumés composés de segments entiers des textes originaux. Puis, afin de compléter le processus de résumé, les méthodes de compression de phrases seront testées, c'est à dire, les méthodes pour compresser une seule phrase, sans qu'elle perde sa signification et sa grammaticalité. Idéalement, le résumé par extraction suivi par une compression de phrases peut produire de meilleurs résultats que chacune de ces approches séparément. Enfin, nous avons l'intention d'investir dans des outils d'analyse profonde des textes, pour mettre à disposition de s ressources plus sophistiquées pour le résumé.

Il est prévu l'utilisation des textes généraux, ainsi que des textes d'un domaine spécialisé. Dans les textes généraux, les textes journalistiques sont généralement la meilleure référence. Dans le cas de textes d'un domaine spécifiques, nous allons utiliser des textes médicaux.

Boudin, F. (2008). Exploring statistical approaches for automatic text summarization. PhD thesis. Université d'Avignon.

Fernández, S.; SanJuan, E.; Torres-Moreno, J. M. (2007). “Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in text summarization and topic segmentation”. Lecture Notes in Computer Science 4827. Berlín: Springer. 861-871.

Florian B. and Torres-Moreno, J.M. *A Maximization-Minimization Approach for Update Summarization.*. Book chapter in [Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing V](#), Nicolov, Nicolas, Galia Angelova and Ruslan Mitkov (eds.), 143–154, 2009.

SanJuan E., Ibekwe-Sanjuan F. Multi Word Term Queries for Focused Information Retrieval. CICLing 2010: 590-601.