Université d'Avignon et des Pays de Vaucluse

Sujet de thèse en Informatique — mars 2015

Génération automatique d'*abstracts* par optimisation dans un contexte culturel

Direction : Juan-Manuel Torres (MCF HDR UAPV) - **Co-direction :** Andréa Carneiro Linhares (Prof. Universidade Federal do Ceará, Brésil) et Eric SanJuan (MCF UAPV)

Mots-clefs: Résumé automatique, Traduction automatique, Optimisation, Traitement automatique de langues.

L'interconnexion immédiate ne facilite pas forcément la communication et peut même engendrer l'incompréhension comme l'analyse D. Wolton (1997). A cela s'ajoute la possibilité aujourd'hui d'obtenir des traductions automatiques à partir de tout terminal connecté à internet, et Il est même possible de coupler cette traduction à un résumeur automatique. Ainsi il est techniquement accessible à tous d'obtenir en quelques secondes le résumé dans sa langue de toute œuvre ou document écrit en toute langue et d'avoir l'illusion que le temps de lecture et les barrières linguistiques sont abolis.

En effet, étant donné l'énorme quantité d'information disponible sous forme électronique, 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 (RA), la récupération et l'extraction d'information et les systèmes de traduction. 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 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 de livres, les synopsis de films et de romans, le résumé des prévisions météorologiques, les *abstracts* d'articles scientifiques, l'indexation de documents via leur résumé, etc. Malgré son importance, la génération automatique de résumés présente des problèmes qui restent ouverts : soit les résumés sont informatifs mais leur qualité grammaticale et linguistique sont piètres, soit ils sont lisibles, mais ayant un contenu peu pertinent. La production de résumés personnalisés aux besoins informatifs des utilisateurs et la dimension multilingue des documents sources ajoutent des problématiques supplémentaires au défi du RA.

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

Dans cette thèse nous proposons effectuer la recherche et l'exploration conjointe du résumé automatique dans un contexte *crosslingue* (anglais-français-portugais-espagnol). En effet, sous l'effet de la mondialisation, un nombre croissant d'utilisateurs des usagers du Web est au moins bilingue. Un système de traduction automatique (TA) traduit des phrases dans une langue source vers une langue cible. Un système de résumé extrait l'information importante d'un document. Un résumé *crosslingue* consiste à produire un résumé dans une langue cible, l'information importante que l'on retrouve dans des documents dans l'autre langue. Cependant les deux tâches sont souvent antagonistes. En effet, le système TA traduit plus facilement (avec une meilleure qualité) les phrases courtes qui représente un concept simple. Il ne s'agit pas de développer un traducteur mais d'utiliser les systèmes état de l'art. Un système RA par contre, privilégie les phrases les plus informatives, représentées souvent par les phrases longues. Ainsi, les phrases courtes sont délaissées dans cette démarche. L'enjeu ici es double : d'un coté il s'agit de pondérer les phrases à la fois avec les méthodes RA mais en considérant dans cette optique la qualité liée à la traduction de la phrase. D'un autre coté il s'agit de produire à la fois des résumés lisibles et informatifs. Nous avons l'intention de nous concentrer sur les méthodes et techniques de résumé mono et multi-document qui ont fait leur chemin, afin de choisir les meilleures pour l'ensemble de langues qu'on veut traiter. En principe, l'approche sera de type extractive, dans le but de produire des résumés composés de segments entiers des textes originaux. Mais nous allons pousser la recherche avec l'utilisation de méthodes de paraphrase avancées : la compression, la fusion, l'élimination et la reformulation de phrases par optimisation. Cela dans l'optique de production automatique d'abstraits ayant une qualité comparable à celle produits par les personnes. Ainsi, nous avons prévu de développer de nouveaux algorithmes de fusion multi-phrases (MFS) ; approches qui conjuguent à la fois, des techniques de TAL, d'optimisation (recherche de chemins optimaux dans un graphe) et de méta-heuristiques.

Objectifs

L'objectif de cette thèse est de développer des méthodes de résumés automatiques par abstraction guidés par la langue et le contexte culturel du lecteur. Il s'agit de:

1- Reconnaître les principaux concepts, notions et références des textes sources, principalement à partir du Wikipédia mais aussi de ressources documentaires complémentaires tel que les corpus de presse (New York Times, La Jornada, Le Monde ...). On pourra pour cela s'appuyer sur l'expérience acquise dans le cadre du projet CAAS et de la thèse de R. Deveaud.
2- Evaluer pour chacun d'entre eux la difficulté à leur trouver un correspondant dans la langue et le contexte culturel cible, ce contexte pouvant être déduit des favoris de navigation et de lecture de l'usager. Il s'agira ici d'alignement d'ontologies et de ressources documentaires tel que WordNet et DBpedia.
3- Générer un résumé contextualisé qui articule la traduction de passages extraits des sources avec des extraits de ressources documentaires de la langue cible pour chaque concept et référence qui ne soient pas directement traduisible. Cet aspect se trouve dans le prolongement des tâches de contextualisation de textes courts initiés dans le cadre de CLEF-INEX.
4- Evaluer la fiabilité du résumé obtenu vis-à-vis de la source par des mesures automatiques d'informativité étendues à richesse conceptuelle.

Un résumé culturel multilingue peut ainsi être plus long que le texte d'origine car il tient compte de la connaissance implicite à la langue source. La longueur du résumé obtenu est en soit un indicateur de la différence culturelle. Il ne peut y avoir de réelle communication sans connaissance de ces différences. Des nouvelles mesures d'évaluation peuvent être proposées pour estimer la qualité des résumes.

Candidatures. Un très bon candidat a été identifié pour ce sujet. Il est actuellement inscrit en Master inscrit à l'UFC/Brésil et il effectue son stage de recherche au LIA (févier-juin 2015), sous la thématique de résumé extractif par optimisation.

Références

Boudin, F., Huet, S., and Torres-Moreno, J-M: *A Graph-based Approach to Cross-language Multi-document Summarization*. Polibits 43:113-118 (2011)

Cormen, T.H.; Leiserson, C.E.; Rivest, R.L. & Stein, C., *Introduction to Algorithms*. 3ª Ed. USA: MIT Press, 2009.

Fernández, S.; SanJuan, E.; Torres-Moreno, J.M. *Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in text summarization and topic segmentation*. LNCS 4827. Berlín: Springer. 861-871, 2007

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.

Pontes, E., Linhares, A.C., Torres-Moreno, J.M.: SASI: sumarizador automático de documentos baseado no problema do subconjunto independente de vértice. In: Proc. of the XLVI Simpósio Brasileiro de Pesquisa Operacional (2014).

Talbi, El-G. Metaheuristics: From desgin to implementation. Wiley, First edition, 2009.

Torres-Moreno, J. M. (2014). "Automatic Text Summarization". First edition, Wiley & Sons 2014.

Saggion, H., Torres-Moreno, J.M, da Cunha, I., SanJuan, E., Velázquez-Morales, P.: Multilingual Summarization Evaluation without Human Models. COLING 2010:1059-1067

Dominique Wolton, "Penser la communication", Flamarion 1997.