Université de Montréal
Département de génie informatique et génie logiciel

IFT6010
Introduction au traitement automatique des langues naturelles

Rapport

Soumis par
Ludovic Font,
Lara Haidar-Ahmad

20 Février 2015

Introduction :

Comme pour de nombreux problèmes, il existe deux façons de résumer automatiquement un texte : la première, par “extraction”, consiste à travailler sur les phrases afin de trouver les plus significatives pour les extraire du texte, et former ainsi un résumé. La seconde, par “abstraction”, consiste à générer de nouvelles phrases à partir d’une compréhension du fond du texte. La première est plus facile à implémenter, mais produit des résultats parfois biaisés ou verbeux. La seconde produit des résultats plus intéressants, mais est plus compliquée à mettre en place.
Dans ce projet, nous tenterons de combiner les deux méthodes afin de tirer parti des avantages de chacune. La première, fonctionnant par extraction (LexRank, par Gunes Erkan et Dragomir R. Radev), aura pour but de récupérer un maximum de phrases-clés. La seconde, par abstraction (Opinosis, par Kavita Ganesan, ChengXiang Zhai et Jiawei Han) servira à compresser ces phrases afin d’obtenir un résumé plus fluide et moins redondant. Nos critères d’évaluation utiliseront les métriques ROUGE essentiellement.

Description de l’algorithme par extraction (LexRank) :

Plusieurs approches sont considérées pour construire le résumé d’un texte à partir des phrases qui le constituent. Afin de choisir les phrases les plus significatives, différents algorithmes sont conçus pour attribuer des scores aux phrases du texte. Les phrases ayant les meilleurs scores seront par la suite sélectionnées. Le calcul de score se base sur deux méthodes principales; soit le calcul de centroïdes et le calcul de centralités.

Calcul de centroïdes:

Cette méthode se base sur les mots qui constituent les phrases. On considère que les phrases contenant le plus de mots clefs ayant rapport avec le sujet du texte, sont des phrases centrales. Les mots les plus utilisés dans la langue se verront alors attribuer de faibles scores, et les mots plus rares auront de plus hauts scores. Ainsi, plus le mot est rare, plus il contribue à augmenter le score de la phrase à laquelle il appartient. Cette méthode a alors pour avantage de détecter des mots spécifiques, et les mots clefs qui sont directement reliés au sujet du texte. Les approches se basant sur les centroïdes ont donné des résultats satisfaisants dans les premiers systèmes de création automatique de résumés, ce qui témoigne de l’efficacité de cette méthode.

Calcul de centralités:

Cette méthode se concentre sur les relations entre les phrases plutôt que sur le contenu des phrases. Ainsi, plus une phrase a de relations avec d’autres phrases du texte, plus elle a de chance de contenir des informations importantes ayant un lien important avec le sujet.

Il faut alors pouvoir définir le taux de similarité entre deux phrases, mais encore, il faut aussi pouvoir définir la centralité d'une phrase à partir de ces similarités.
Afin d'extraire les similarités entre les phrases, on a recours à des formules mathématiques se basant sur le taux de mots que les deux phrases ont en commun, ainsi que sur la fréquence d'apparition de ces mots dans le texte. On peut représenter le problème par un graphe pondéré; les noeuds du graphe seraient les phrases du texte, les liens entre les noeuds seraient les similarités entre les phrases, et les coûts des liens représenteraient le taux de similitude entre les phrases. Ainsi, les phrases ayant le plus de liens avec d'autres phrases, serait celle qui contiendrait le plus d'information importante puisqu'elles seraient centralisées et auraient donc plus de chances d'avoir un lien direct avec le sujet du document.
D'autre part, pour mesurer la centralité d'une phrase, on se base sur la centralité des mots de cette phrase. Afin de calculer la centralité des mots dans un texte, on se réfère à la liste des mots les plus fréquents du texte. Plus les mots d'une phrase sont fréquents dans le texte, plus ils contribuent à augmenter le score de la phrase.

Description de l'algorithme par abstraction (Opinosis) :

L'idée générale de cette méthode est de repérer des phrases ou éléments de phrases similaires et de les fusionner. Par exemple, les trois phrases :

- "The iPhone's battery lasts long, only had to charge it once every few days."
- "iPhone's battery is bulky but it is cheap."
- "iPhone's battery is bulky but it lasts long!"

seraient fusionnées en :"'iPhone's battery is cheap, lasts long but is bulky"

Ce procédé s'effectue en plusieurs étapes : sélection de phrases, génération du graphe, exploration, compression et construction du résumé.

- **Sélection de phrases**

Cette étape se fait de manière *ad hoc*, selon le sujet particulier que l'on veut résumer. Par exemple, on pourra choisir, pour poursuivre l'exemple précédent, toutes les phrases en rapport avec la batterie de l'iPhone (toutes les phrases contenant le mot "battery"). Ces phrases sont ensuite annotées par étiquetage morpho-syntaxique (part-of-speech, POS).

- **Génération du graphe**

L'idée est de représenter chaque mot étiqueté par un noeud, et deux mots se suivant dans une phrase sont reliés par une arête orientée.

- **Exploration**

Ici, on parcourt le graphe à la recherche de chemins valides, reliant un début de phrase à une fin de phrase (une ponctuation ou une conjonction de coordination, comme "but" ou "yet").

Les chemins valides doivent également respecter certaines exigences structurelles. Ils reçoivent enfin un score basé sur le nombre de phrases contenant ce chemin.

- **Compression**

Cette étape consiste simplement à réunir certains chemins dont les prédécesseurs dans le texte sont identiques. Par exemple, les deux phrases “the sound quality is really good” et “the sound quality is clear”, peuvent être compressées en “the sound quality is really good and clear”.

- **Construction du résumé**

Il s’agit ici de classer les chemins selon leur score, puis d’éliminer les redondances à l’aide d’une mesure de similarité. On sélectionne enfin les N meilleurs chemins, selon le nombre de phrases que l’on veut dans le résumé.

Méthodologie:

Dans cette partie du rapport, nous allons présenter les étapes de travail à suivre pour le reste de la session. La première étape consistera à implémenter les 2 algorithmes décrits ci-haut; nous auront alors, un premier algorithme qui générera des résumés par extraction de phrases, et un deuxième algorithme qui générera des résumés par abstraction. Nous serons alors en mesure de tester ces deux algorithmes séparément, afin de juger de leurs efficacités respectives. Notre but étant de combiner les deux approches, nous allons nous concentrer par la suite sur l'intégration des deux algorithmes. Ainsi, lors du traitement d’un texte, nous allons tout d’abord extraire les phrases importantes, puis, par la suite, améliorer le résultat par l’algorithme d’abstraction. Nous serrons alors en mesure de tester les résultats obtenus par la combinaison des deux approches et comparer les résultats à ceux des premiers tests. Finalement, nous pourrons conclure sur l’efficacité de notre algorithme.

Afin de pouvoir effectuer les tests requis, nous allons nous servir des informations de la base de données de DUC (Document Understanding Conferences). Nous allons utiliser ces données afin d’effectuer nos tests que nous allons évaluer par la suite à l’aide des métriques ROUGE (Recall-Oriented Understudy for Gisting Evaluation).

Conclusion

Tout au long de notre étude, nous allons nous concentrer sur la génération automatique de résumé, et en particulier sur les méthodes d’extraction et d’abstraction. Nous allons comparer l’efficacité de ces deux méthodes et tenter de les combiner afin de tirer avantage de chacune d’elles.
Le sujet de notre étude étant défini, nous sommes maintenant en mesure de la mettre en place.

Références:

Erkan, G. E., Radev, D. R. (2004). *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.* Tiré de http://www.jair.org/media/1523/live-1523-2354-jair.pdf
Opinosis : abstractive summarization. Tiré de http://dl.acm.org/citation.cfm?id=1873820