WIA

SYNTHÈSE : RÉSUMÉ AUTOMATIQUE DE TEXTE

January 19, 2018

KARIMOUNE MOSSI Abdoul aziz

Contents

1 INTRODUCTION

Cette synthèse aborde la problématique du résumé automatique de document. Le résumé d'un document consiste à exprimer en peu de mot le contenu essentiel du document, il doit être le plus court possible. Il permet au lecteur d'accerder directement au contenu le plus important du document en peu de temps. Grace à la croissance importante des documents textuels, nous avons besoin de système qui nous assure cette tache de manière automatique et nous fait gagner beaucoup de temps. Cependant résumer des documents necessite des connaissances linguistiques dont ne dispose pas les systèmes, ce qui rend la tache complexe.

Aujourd'hui il existe deux manières de faire du résumé automatique. La première, le résumé par abstraction consite à reformuler le document en peu de mot et produire un résumé. La deuxième manière est le résumé par extraction qui fournit comme résumé les phrases les plus importantes du document. Dans cette synthèse, nous parlerons du résumé par extraction puisque c'est le plus utilisé et aussi parceque le résumé par abstraction n'est pas encore mature.

1.1 Définition des termes

Extraction d'information : consiste à extraire et à structurer automatiquement un ensemble d'informations précises apparaissant dans un ou plusieurs documents textuels écrits en langue naturelle.
Résumé multi-document : résumé automatique de plusieurs documents de source différentes.
Résumé mono-document : résumé automatiuqe d'un document.
Phrase pertinente : une phrase pertinente est définie par sa position, le nombre de mot, sa longueur, ses mots clés.
Vecteur sémantique : représente un vecteur de bigrammes, dont la dimension est le nombre de bigrammes dans le document et les valeurs du vecteur sont pour chaque bigramme de la phrase on calcule le nombre de fois qu'il est présente dans le document. Les valeurs des bigrammes qui sont pas dans la phrase sont nulles.
Bigramme : bigramme est toute séquence de 2 items, qui peuvent être des lettres, des mots, des étiquettes

1.2 Historique

La tâche de résumé automatique a vu le jour en 1958, grace aux travaux de recherche de Luhn sur le résumé de documents scinetifiques. Mais c'est surtout entre 1958 et 1978 qu'a eu lieu les travaux fondateurs sur le résumé automatique de texte. Après il a fallu attendre 1990 et surtout grace aux travaux de K. Spärek-Jones et J. Kupieck pour

contaster un avancé important. Aujourd'hui la croissance de documents numériques et les avancés en apprentissage automatique font du résumé automatique un domaine important de recherche. C'est notamment un des domaines important du traitement automatique de la langue naturelle (TAL).

1.3 Difficulté de la taches

La principale difficulté qaund il s'agit de faire du résumé par extraction est la caractérisation des unités textuelles importantes. En d'autres termes le critère de selection des phrases pertinentes est la difficulté. Nous verrons dans la suite du document comment répondre à cette question selon les différentes apporches de résumé.
Ensuite la deuxième difficulté est celle de l'évaluation des systèmes qui font du résumé automatique. Les résumés des personnes sur lesquels on se base pour évaluer ne sont pas forcement les memes puisqu'il n'existe pas de résumé ideal.

2 MÉTHODES DE RÉSUMÉ PAR EXTRACTION

2.1 Approche par graphe

Dans cette approche, le document est représenté par un graphe d'unités textuelles (pour le résumé automatique par graphe la phrase est utilisée comme unité textuelle) liées entre elles par le calcul de mesure de similarité. Chaque phrase subit un pre-traitement[1]. Les phrases sont representées par des vecteurs de N dimension où N est le nombre de mots differents dans le document. Ensuite on calcule le poids tf x idf pour chaque composant de la phrase. Ce sont ces poids qui sont utilisés par les méthodes de calcul de similarité, comme la méthode cosinus, pour mesurer la similarité entre les phrases. Elles sont inter-connectées entre elle grâce à ces mesures de similarités.

Le processus extractif est alors considéré comme une identifcation des sommets les plus importants du graphe. Les algorithmes de classement basés sur les graphes, telque PageRank ou lexRank permettent ensuite de décider de l'importance de chaque sommet dans le graphe. Les sommets les plus prestigieux sont alors selectionnés pour produire le résumé.

Cette approche a fourni des bons resultats de résumé. Cependant, il est important de noter que les algorithmes de classements utilisés pour selectionner les sommets dependent de la constrcution du graphe. Comme le graphe est construit à partir des mesures de similarité, donc ces mesures ont un impact considerable sur le résumé.

[1] **Le pre-traitement sur les phrases consite à supprimer les stops words et les ponctuations et à transformer les majuscules en minuscles**

Pour améliorer les performances plusieurs propositions ont étés faites pour définir une bonne mesure de similarité. (Boudin et al., 2008) proposent notamment une mesure qui permet de créer des relations entre deux segments qui même s'ils ne partagent aucun mot,en contiennent des morphologiquement proches.

2.2 Approche par optimisation

L'approche consiste à considérer le problème de résumé par extraction comme un problème d'optimisation. Plusiuers méthodes sont proposées pour modèliser le problème de résumé.

2.2.1 Maximal Marginal Relevance (MMR)

La méthode Maximal Marginal Releance (MMR) de (Carbonell and Goldstein, 1998) définit une fonction de score basée sur la mesure de similarité entre les phrases du document à résumé. L'idée est de choisir les phrases les plus pertinentes tout en minimisant la redondance. Il n'existe pas d'algorithme qui resoud ce problème de manière optimale puisqu'il a été demontré NP-difficile. Ce pendant des algorithmes glouton donnent des solutions approximatives, comme l'algorithme de (Lin and Bilmes, 2010) qui est souvent utilisé.

2.2.2 Méthode basée sur le volume sémantique

La méthode de (Dani Yogatama et Fei Liu, 2015) définit une rélation géomtrique entre le résumé et le document. On calcule pour chaque phrase du document son vecteur sémantique. Le sous espace vectoriel formé par les phrases du résumé définit un volume sémantique. L'idée c'est donc de choisir les phrases qui permettent d'avoir un plus grand volume sémantique. En effet un grand volume permet de couvrir au maximum le document, donc de fournir les phrases pertinentes. La fonction objective est basée sur ce volume sémantique. Puisque le probléme est NP-difficile, l'algorithme glouton itératif défini par les auteurs pour la soultion approchée est : pour la première phrase on prend celle la plus loin du centroide formé par l'ensemble des phrases, ensuite on prend la phrase la plus loin de la première phrase, après on prend la phrase la plus loin du sous-espace formé par les phrases déjà prises. On itère jusqu'à atteindre le nombre de phrase qui correspond au nombre de mots demandés. Pour calculer la distance entre la nouvelle phrase à ajouter et le sous-espace formé par les phrases déjà prises, ils ont utilisé une méthode basée sur l'algorithme de Gram-Schmidt(Laplace, 1812). Cet algorithme permet de construire une base orthonormée à partir d'une base quelconque donnée. Chaque fois qu'on veut ajouter une nouvelle phrase, on l'ajoute de manière à former une base orthonormée avec le sous espace des phrases déjà prises.

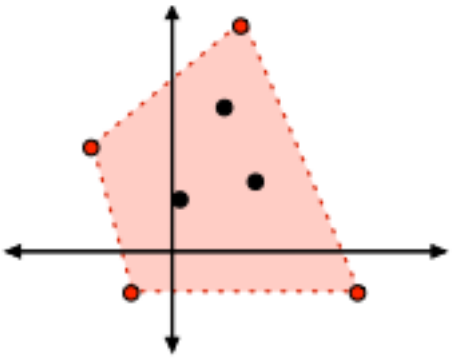

Figure 1

Sur cette figure un document de 7 phrases est representé dans un espace à 2 dimensions. Si on veut par exemple un résumé de 4 phrases. La fonction objective doit fournir les 4 phrases en rouges puisque ce sont ces 4 phrases qui permettent de maximiser le volume sémantique qui correspond dans cet exemple à la surface en deux dimensions.

3 EVALUATION

3.1 Corpus

La campagne d'évaluation Text Analysis Conference (TAC) qui était avant Document Understanding Conference (DUC) est organisée par le National Institute of Standards and Technology (NIST) pour promouvoir les avancés réaliser dans le domaine du résumé automatique de textes et aider les chercheurs à évaluer leurs système.

Pour evaluer leur méthode (Boudin et Torres-Moreno, 2009) ont suivi le même protocole que NIST sur un corpus en francais composé de 20 thématiques différentes. Chaque thématique comprend un ensemble de 10 articles de journaux de source différentes. Pour chaque thématiques 4 résumés de reference ont étés produits par des personnes.

Pour l'anglais, (Dani Yogatama et Fei Liu, 2015) ont évalué leur méthode sur un corpus de résumé non mis à jour de 48 documents TAC-2008 et 44 documents de TAC-2009. Pour chaque thématiques 4 résumés de reference ont étés produits par des personnes.

3.2 Métriques

En résumé automatique de texte la principale métrique utilisé est ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004). Elle compare un résumé candidat à l'ensemble des résumés de référence disponibles.

4 CONCLUSION

L'augmentation des documents numeriques et le manque de temps pour acceder à l'ensemble du contenu des textes montrent de plus en plus l'importance des systèmes de résumé automatique. Actuellement les système les plus perfomants font du résumé par extraction. Les méthodes implementées par les système sont basées sur differents approches.

Grâce aux avancés des recherches dans les reseaux de neuronnes, des méthodes de résumé basées sur l'apprentissage automatique sont apparues. Comme la quantité de données textuelles augmente beaucoup, il serait interresant d'approfondir les recherches sur ces méthodes pour améliorer les performances des système actuels.

REFERENCES

Florian Boudin et Juan-Manuel Torres-Moreno. Résumé automatique multi-document et indépendance de la langue : une première évaluation en français. TALN 2009 – Session posters, Senlis, France, 24–26 juin 2009.

Dani Yogatama, Fei Liu and Noah A. Smith. Extractive Summarization by Maximizing Semantic Volume. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1961–1966, Lisbon, Portugal, 17-21 September 2015.

Güneş Erkan and Dragomir R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22 (2004) 457-479, Ann Arbor, MI 48109 USA, december 2004.