

Quels concepts pour le résumé automatique par extraction ?

Clément Tek
Université de Nantes

2 mars 2015

SOMMAIRE

INTRODUCTION

Préface

Qu'est-ce que le résumé automatique par extraction ?

Qu'est-ce qu'un concept ?

ÉTAT DE L'ART

ROUGE

Concepts

Extraction par optimisation linéaire

Modèle de régression

PISTES

Regroupement

Pondération

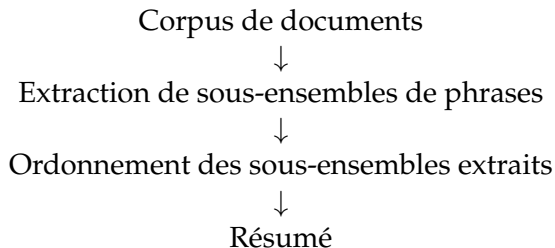
CONCLUSION

PRÉFACE

Travail d'étude et de Recherche par ANTHONY PENA, MARIE LENOUE et CLÉMENT TEK, étudiants à l'Université de Nantes.

Encadré par FLORIAN BOUDIN, maître de conférence à l'Université de Nantes et membre de l'équipe TALN du LINA et HUGO MOUGARD, doctorant au LINA.

QU'EST-CE QUE LE RÉSUMÉ AUTOMATIQUE PAR EXTRACTION ?



QU'EST-CE QU'UN CONCEPT ?

Entité représentant une notion, en général pondérée.

Peut être :

- ▶ Un bigramme
- ▶ Un groupe nominal
- ▶ Un arbre de dépendances
- ▶ ...

ROUGE - RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION

- ▶ Système de mesure pour le résumé automatique
- ▶ Comparaison avec un ensemble de références (résumés) produit par des humains
- ▶ Utilisé dans des compétitions telles que le TAC (Text Analysis Conference) et le DUC (Document Analysis Conference)

CONCEPTS

Uniquement des bigrammes de mots dans la littérature :

- ▶ Traitement rapide donnant de bons résultats
- ▶ Pas de travaux sur l'importance du choix des unités textuelles pour les concepts

Comment peut-on évaluer le choix du type de concept ?

EXTRACTION PAR OPTIMISATION LINÉAIRE

D. GILICK et B. FAVRE : *A scalable global model for summarization*, Juin 2009.

Maximiser la fonction objectif :

$$\sum_i w_i c_i$$

c : fréquence d'apparition du concept

w : poids du concept

MODÈLE DE RÉGRESSION - 1

C. LI , X. QIAN et Y. LIU : *Using Supervised Bigram-based ILP for Extractive Summarization*, Août 2013.

Utilise plusieurs features (traits) indicatives au niveau des mots :

- ▶ Fréquence du bigramme dans le sujet donné
- ▶ Fréquence du bigramme dans les phrases choisies
- ▶ Ratio de Stop Words (mots trop communs ou vides) dans le bigramme
- ▶ Similarité avec le titre du sujet (nombres de tokens communs divisé entre les deux chaines par la longueur de la chaine la plus longue)
- ▶ Similarité avec la description du sujet

MODÈLE DE RÉGRESSION - 2

Features au niveau des phrases :

- ▶ Taux de phrase : Nombre de phrases dans le bigramme divisé par le nombre total de phrases sélectionnées
- ▶ Similarité de la phrase avec la concaténation du titre et de la description du sujet
- ▶ Position de la phrase dans le texte
- ▶ Longueur de la phrase en nombre de mots
- ▶ Début de paragraphe ou non (binaire)

Améliorations peu significatives.

- ▶ Features ou non ?

REGROUPEMENT

Locutions nominales :

- ▶ Point de vue
- ▶ A peu près
- ▶ Suprême de volaille
- ▶ ...

Comptées comme trigrammes, mais ne représentent qu'une entité. Que faire ?

- ▶ Regrouper en un seul mot
- ▶ Ajouter un dictionnaire de locutions

PONDÉRATION

Autre piste pour améliorer, changer la pondération des concepts. Actuellement :

- ▶ Fréquence d'apparition du concept
- ▶ Extraction des phrases ayant le poids le plus important

N'y a-t-il pas mieux ?

- ▶ Problème : Dépendant du corpus
- ▶ Dans notre cas : corpus d'articles journalistiques. Ajouter du poids au début et à la fin ?
- ▶ Pondérer plus les citations ?

CONCLUSION

Des idées, mais pas encore d'implémentation. Des questions encore en suspens :

- ▶ Vaut-il mieux améliorer son score ROUGE ?
- ▶ Ou privilégier l'évaluation humaine ?
- ▶ Faire un compromis ? Mais à quel prix ?