

Nom et Prénom : LAMSIYAH SALIMA

Date de soutenance : 04/12/2021

Directeur de Thèse : LAMSIYAH SALIMA

Sujet de Thèse :

**UNSUPERVISED EXTRACTIVE MULTI-DOCUMENT SUMMARIZATION BASED ON
TRANSFER LEARNING**

Résumé :

Le Résumé Automatique de Textes (RAT) est l'une des tâches les plus importantes du Traitement Automatique du Langage Naturel (TALN) qui a pour objectif de produire une représentation condensée et pertinente à partir d'un ou de plusieurs documents textuels tout en minimisant la redondance. Il existe deux principales approches pour aborder la tâche du RAT : extractive et abstractive. L'approche extractive consiste à extraire des phrases complètes censées être les plus pertinentes du document et à les concaténer de façon à produire un extrait. Alors que l'approche abstractive vise à fournir un résumé en générant des phrases qui ne sont pas nécessairement contenues dans le document original, ceci au travers d'une analyse du contenu sémantique du document.

Ce travail de thèse se penche sur l'approche extractive pour le résumé automatique de textes, et plus précisément sur le résumé générique multi-document (RG-MD) et le résumé orienté-requête multi-document (ROR-MD). Le résumé générique synthétise les documents originaux sans prendre en compte les besoins en information des utilisateurs, par contre, le résumé orienté-requête a pour objectif de ne résumer que les informations répondant à une requête de l'utilisateur. Notre objectif principal est de développer des systèmes robustes et efficaces pour le RG-MD et le ROR-MD qui ne nécessitent aucune connaissance du domaine.

La représentation de textes est une pierre angulaire qui affecte la performance des systèmes du résumé automatique de textes. La représentation en sac-de-mots et les représentations distribuées des mots ont prouvé leur efficacité dans plusieurs tâches du résumé automatique de textes. Cependant, la plupart de ces représentations ne prennent pas en considération l'ordre et les relations sémantiques entre les mots d'une phrase, et par conséquent elles ne reflètent pas le sens complet d'une phrase. Pour pallier ce problème et améliorer la performance des systèmes du résumé extractif multi-document, nous proposons quatre contributions adoptant l'approche non supervisée.

Dans la première contribution, nous proposons une méthode extractive non supervisée pour le RG-MD basée sur l'approche centroïde et les représentations distribuées des phrases (Sentence Embedding). Nous améliorons également la pondération des phrases en combinant trois métriques, à savoir la pertinence de la phrase, la nouveauté de la phrase, et la position de la phrase dans le document. De plus, nous présentons une analyse comparative de plusieurs modèles de représentation distribuée des phrases pour évaluer leur impact sur la performance du résumé extractif multi-document. Pour améliorer davantage la méthode RG-MD proposée, nous exploitons, dans la deuxième contribution, l'apprentissage par transfert à partir du modèle BERT. Dans la troisième contribution, nous suggérons une méthode extractive non supervisée pour le ROR-MD en s'appuyant sur l'apprentissage par transfert à partir des représentations distribuées des phrases, du modèle BM25 et du critère de pertinence marginale maximale. L'usage des représentations distribuées des phrases lors du processus d'extraction et de reclassement des phrases nous permet de produire des résumés adaptés à la requête utilisateur et de minimiser la redondance. En fin, la dernière contribution consiste à améliorer la performance du ROR-MD en combinant l'apprentissage par transfert et le modèle Sentence-BERT qui se base sur l'architecture siamoise et le mécanisme du fine-tuning.

La validation expérimentale de l'ensemble des solutions apportées dans le cadre de cette thèse est effectuée en utilisant la collection standard DUC'2002-2007 et le corpus Multi-News.

Mots clés : Traitement Automatique du Langage Naturel; Résumé Automatique de Textes; Résumés Générique/Orienté-Requête Multi-Documents; Approche Extractive; Apprentissage par Transfert; Apprentissage Multi-Tâche; Représentation Distribuées des Phrases.

Abstract:

Automatic Text Summarization (ATS) is a research area in the context of Natural Language Processing (NLP) whose goal is to automatically process and synthesize texts while preserving their important aspects. ATS approaches are classified into two categories: extractive and abstractive. In the former, summaries are produced by identifying and extracting the most relevant sentences from the source documents, while in the latter, summaries are generated by reformulating and fusing ideas and often by using a new lexicon. This thesis focuses on Extractive ATS systems, and more specifically on Generic Multi-Document Summarization (GMDS) and Query-Focused Multi-Document Summarization (QF-MDS) tasks. G-MDS systems generate summaries that represent all relevant facts of the source documents without considering the users' information needs. Besides, QF-MDS systems produce summaries where the content of the summary is derived from the user's information need or simply the user's query. Our main objective is to develop robust and effective systems for both G-MDS and QF-MDS tasks that require no domain knowledge.

Text representation is a fundamental cornerstone that affects the effectiveness of text summarization methods. Bag-of-words and word embedding have shown to be effective for several text summarization tasks. However, most of these representations do not consider the order and the semantic relationships between words in a sentence, and hence they do not carry the meaning of a full sentence. Therefore, we propose four contributions to deal with this issue and to improve the performance of unsupervised extractive multi-document summarization.

In the first contribution, we propose an unsupervised extractive method for G-MDS based on the centroid approach and the sentence embedding representations. We improve sentence scoring by combining three metrics, including sentence content relevance, sentence novelty, and sentence position. Moreover, we provide a comparative analysis of nine sentence embedding models used to represent sentences as dense vectors in a low dimensional vector space in the context of extractive multi-document summarization. In the second contribution, we improve the aforementioned G-MDS method by leveraging transfer learning from BERT fine-tuning on Natural Language Understanding tasks for sentence representation learning. Specifically, we fine-tune BERT on supervised intermediate tasks from GLUE benchmark using single-task and multi-task fine-tuning. Then, we transfer the learned knowledge to our summarization task. In the third contribution, we propose an unsupervised extractive QF-MDS method based on transfer learning from pre-trained sentence embedding models, BM25 model, and maximal marginal relevance criterion. We combine the BM25 model with the semantic similarity to select a subset of sentences based on their relevance to the query. Moreover, we incorporate sentence embedding representation in the maximal marginal relevance method to re-rank the candidate sentences by maintaining query relevance and minimizing redundancy. In the last contribution, we explore the potential of the recent pre-trained Sentence-BERT (SBERT), based on the Siamese network structure and fine-tuning mechanism, to boost the performance of extractive query-focused multidocument summarization task.

Keywords: Natural Language Processing; Automatic Text Summarization; Generic and Query-Focused Multi-Document Summarization; Unsupervised Extractive Approach; Text Representation; Transfer Learning; Multi-Task Learning; Sentence Embedding.