جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Monsieur Lahbib AJALLOUDA

Soutiendra publiquement sa thèse de Doctorat en Informatique

Le Mercredi 26 Juillet 2023 à 11h00 au Grand Amphi à l'ENSIAS

Intitulé de la thèse

KEYPHRASES PREDICTION: A HYBRID MODEL TO UNIFY THE EXTRACTION AND GENERATION MECHANISMS

Président :

Pr. Khalid NAFIL, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Ahmed ZELLOU, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. Hassan SILKAN, PES, Faculté des sciences, Université Chouaib Doukkali, El Jadida

Pr. Abdelmajid HAJAMI, PES, Faculté des sciences et techniques, Université Hassan I, Settat

Pr. Youssouf EL ALLIOUI, PH, Faculté Polydisciplinaire Khouribga, Université Sultan Moulay Slimane, Béni Mellal

Examinateur :

Pr. Taoufiq RACHAD, PH, ENSIAS, Université Mohammed V de Rabat

UNIVERSITE MOHAMMED V - RABAT
ENSIAS

Résumé: La production croissante de données textuelles a créé de nombreux défis pour les tâches de traitement du langage naturel (TAL). La prédiction des phrases clés est la tâche la plus importante en TAL. Une phrase clé est un mot ou une séquence de mots qui définit le contenu et les sujets d'un document sans avoir à l'analyser. Les phrases clés apparaissant dans un document sont appelées présent keyphrases. En revanche, les phrases clés qui ne sont pas mentionnées dans le document sont appelées absent keyphrases. De nombreuses tâches TAL exploitent des phrases clés pour améliorer leurs performances, telles que la synthèse de texte, l'analyse des sentiments, la recherche d'informations et la classification de texte. Pour prédire les phrases clés, les mécanismes d'extraction ou de génération sont utilisés. Dans cette thèse, nous allons passer en revue les méthodes d'extraction et de génération des phrases clés, afin d'identifier les techniques exploitées pour prédire les phrases clés. Aussi, les défis auxquels sont confrontés les deux mécanismes, afin de proposer des solutions pour les surmonter.

Le mécanisme d'extraction repose sur l'utilisation de techniques telles que des calculs statistiques, des représentations graphiques, l'intégration des phrases, les algorithmes d'apprentissage automatique et profond pour prédire les phrases clés. Ce mécanisme nécessite un prétraitement du texte et de la sélection des phrases clés candidates. Le mécanisme d'extraction se caractérise par sa couverture de toutes les parties du document et la diversité des phrases clés extraites. En revanche, il ne peut pas générer de phrases clés non mentionnées dans le document. De plus, ces méthodes ne peuvent pas résoudre le problème de chevauchement. Le mécanisme de génération est basé sur des modèles de sequence to sequence (Seq2Seq). Ce sont des modèles d'apprentissage en profondeur qui ont été utilisés dans de nombreuses tâches telles que la traduction automatique, les réponses automatiques aux questions, l'annotation d'images et la synthèse de texte. Le mécanisme de génération ne nécessite pas un pré-traitement approfondi ni de sélection de phrases clés candidates. Ce mécanisme se caractérise par sa capacité à générer des phrases clés présentes et absentes. D'autre part, la plupart des méthodes de génération de phrases clés ne peuvent pas couvrir toutes les parties du document. De plus, les méthodes de génération de phrases clés ne peuvent pas surmonter le problème de chevauchement. Aussi, la plupart des méthodes qui adoptent ce mécanisme sont supervisées et nécessitent une grande quantité de données d'apprentissage.

La combinaison des mécanismes d'extraction et de génération dans un modèle contribuera à surmonter ces contraintes. Dans cette thèse, nous proposons un modèle hybride qui combine ces mécanismes. Le modèle permet d'exploiter les avantages de l'extraction et de la génération dans un seul cadre. Notre modèle adopte l'hypothèse que les phrases clés doivent être proches des paragraphes principaux d'un document. Pour cela, le modèle proposé se compose de quatre étapes principales. Dans la première étape, nous sélectionnons les phrases clés candidates, sur la base d'une approche que nous avons proposée afin d'améliorer les performances de sélection des phrases clés candidates. Le poids des paragraphes du document est calculé à la deuxième étape à l'aide de la similarité cosinus. Où chaque paragraphe est représenté par la technique d'embedding Universal Sentence Encoder (USE). Nous

générons des phrases clés absentes via le modèle CopyRNN dans la troisième étape. Dans la quatrième étape, le modèle calculera la similarité cos entre les paragraphes du document et les phrases candidates, ainsi que les phrases générées. Les phrases ayant des meilleurs scores seront considérées comme phrases clés.
Nous avons évalué notre modèle sur trois datasets, chacun représentant un type de texte. Le dataset Inspec contient des résumés d'articles scientifiques, le dataset Semeval2010 contient des articles scientifiques. Alors que le dataset KPTime contient des articles de presse. Les métriques d'évaluation les plus populaires sont utilisées telles que Precision, Recall et F1.Score. Les performances de notre modèle ont été comparées aux performances de trois autres modèles. Key2Vec et EmbedRank, qui sont basés sur des techniques d'embedding de phrases. Le troisième modèle est CopyRNN. Les résultats de l'évaluation ont montré que notre modèle obtenait de meilleures performances, en particulier dans les deux datasets, Semeval et KPTimes, par rapport aux autres modèles.

Mots-clés: Absent keyphrases, Candidate keyphrases, Méthodes de génération de phrases clés, Méthodes d'extraction des phrases clés, Modèle CopyRNN, Poids des paragraphes du document, Present keyphrases, Technique d'embedding de phrases, Traitement du langage naturel.

Abstract: The increasing production of textual data has created many challenges for natural language processing (NLP) tasks. KeyPhrases prediction is the most important task in NLP. A keyphrase is a word or sequence of words, which identify the content and topics of a document without having to analyze it. Keyphrases appear in a document are called present keyphrases. In contrast, keyphrases that are not mentioned in the document are called absent keyphrases. Many NLP tasks exploit keyphrases to improve their performance, such as text summarization, sentiment analysis, information retrieval, and text classification. To predict keyphrases, the extraction or generation mechanisms are used. In this thesis we have reviewed extraction and generation methods, to identify the techniques exploited to predict keyphrases. Also, the challenges faced by both mechanisms. And propose solutions to overcome them.
The extraction mechanism relies on the use of techniques such as statistics, graphs, embedding, machine and deep learning algorithms to predict keyphrases. This mechanism requires pre-processing the text and candidate keyphrases selection. The extraction mechanism is characterized by its coverage of all document parts and the diversity of the extracted keyphrases. On the other hand, it cannot generate absent keyphrases. Also, Keyphrases extraction methods cannot overcome the overlap problem. The generation mechanism is based on sequence-to-sequence (Seq2Seq) models. They are deep learning models that have been used in many tasks such as machine translation, question answering, image annotation, and text summarization. The generation mechanism does not require extensive text processing or candidate keyphrases selection. This mechanism is characterized by its ability

to generate present and absent keyphrases. On the other hand, most keyphrase generation methods cannot cover all document parts. Also, Keyphrase generation methods cannot overcome the overlap problem. Most methods that adopt this mechanism are supervised and require a large amount of training data.

Combination of the extraction and generation mechanisms in one model will contribute to overcome these constraints. In this thesis, we propose a hybrid model that combines extraction and generation mechanisms. The model enables the exploitation of extraction and generation advantages in one framework. Our model adopts the hypothesis that keyphrases should be proximate to main paragraphs in a document. For this, the proposed model consists of four main steps. In the first step, we select candidate keyphrases, based on an approach we proposed to improve the performance of candidate keyphrase selection. Document paragraph weight is calculated in the second step using cosine similarity. Where each paragraph is represented by the universal sentence encoder (USE) sentence embedding technique. We generate absent keyphrases based on the CopyRNN model in the third step. In the fourth step, the similarity cosine of the candidate and generation phrases with the document paragraphs is calculated. Phrases with the highest score are considered as keyphrases.

We evaluated our model on three datasets, each representing a type of text. The Inspec dataset contains scientific papers abstracts, the Semeval2010 dataset contains scientific papers. While the KPTime dataset contains news articles. The most popular evaluation metrics are used such as Precision, Recall and F1.Score. The performance of our model has been compared with the performance of three other models. Key2Vec and EmbedRank, which are based on sentence embedding techniques to extract keyphrases. The third model is CopyRNN. The results of the evaluation showed that our model achieved better performance, especially in the two datasets, Semeval and KPTimes, compared to the other models.

Keywords: Absent keyphrases, Candidate keyphrases selection, CopyRNN model, Document paragraph weight, Keyphrases extraction methods, Keyphrases generation methods, Natural language processing, Present keyphrases, Sentence embedding technique.