UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTE DES SCIENCES DHAR EL MAHRAZ
FES

AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mr **: ALAMI Nabil**
Soutiendra **: le Samedi 22/12/2018** à **15 H** Lieu **: Centre des conférences**

Une thèse intitulée :

Contributions to the improvement of automatic summarization of Arabic texts

En vue d'obtenir le Doctorat

FD : Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité : Informatique

Devant le jury composé comme suit :

	NOM ET PRENOM	GRADE	ETABLISSEMENT
Président	Pr. SATORI Khalid	PES	Faculté des Sciences Dhar El Mahraz - Fès
Directeur de thèse	Pr. MEKNASSI Mohammed	PES	Faculté des Sciences Dhar El Mahraz - Fès
Rapporteurs	Pr. OUHBI Brahim	PES	Ecole Nationale Supérieure d'Arts et Métiers- Meknès
	Pr. ZARGHILI Arsalane	PES	Faculté des Sciences et Techniques - Fès
	Pr. AOURAGH Silhoussain	PH	Faculté des Sciences Juridiques, Economiques et Sociales - Salé
Membres	Pr. OUATIK EL ALAOUI Said	PES	Faculté des Sciences Dhar El Mahraz - Fès
	Pr. ALAOUI ZIDANI Khalid	PH	Faculté des Sciences Dhar El Mahraz - Fès

Abstract:

This thesis work is part of the considerable effort made in the context of automatic processing of Arabic language. Specifically, it concerns the field of automatic text summarization (ATS) of Arabic texts. We are particularly interested in studying and enhancing approaches that aim at extracting the most important and pertinent information from one or more source documents and producing a shortened version of that source. First, we present a detailed state-of-the-art regarding ATS in general, and then we focus on the existing systems and approaches designed for summarizing Arabic documents. After analyzing these approaches, we provide four contribution in order to deal with the identified limitations and weakness of the existing systems, and therefore, improving the performance of Arabic summarizers. In our first contribution, we propose a new Arabic summarization system based on a two-dimensional graph model and redundancy elimination component. The proposed method ranks sentences using a combination of statistical and semantic analysis, which is not studied enough in Arabic ATS. We also shown that using stemming in the preprocessing phase improve the performance the Arabic summarization system. In the second contribution, we propose a deep learning-based method for Arabic TS. Traditional Arabic text summarization systems are based on bag-of-words representation, which involve a sparse and high-dimensional input data. Thus, dimensionality reduction is greatly needed to increase the power of features discrimination. We adopt a variational auto-encoder as a generative model for unsupervised features learning technique that learns features from the input text in order to generate a distributed latent semantic vector for each sentence. Our proposed system uses the generated new representation in order to rank each sentence and extract the most salient among them. We also investigate the use of the VAE on two summarization techniques, graph-based and query-based approaches. In the third contribution, we first adopt several unsupervised deep neural network models in Arabic ATS. We then propose the use of the distributed representation of Arabic words, which we build from a large Arabic corpus using word2vec technique. After that, we enhance the quality of Arabic ATS by proposing ensemble learning technique-based models that aggregate the information provided from different models. We also evaluate the proposed approaches on English dataset. Finally, in our fourth contribution, we propose to improve the performance of Arabic ATS by clustering the dataset and identifying topics of each cluster. We adopt a document representation model based on the identified topic space. Then we use this representation to learn the abstract features using unsupervised neural networks algorithms and ensemble learning techniques. The performance of all the proposed approaches is evaluated using datasets designed for this purpose. The experimental results illustrate the effectiveness of our approaches.

Keywords: Arabic text summarization, Natural language processing, semantic analysis, deep learning, extreme learning machine, word embedding, ensemble learning.

CONTRIBUTIONS A L'AMELIORATION DU RESUME AUTOMATIQUE DES TEXTES ARABES

Résumé :

Ce travail de thèse s'inscrit dans le cadre du traitement automatique de la langue Arabe. Plus précisément, le domaine du résumé automatique des textes (RAT). Nous sommes particulièrement intéressés par l'amélioration des approches visant à extraire les informations les plus pertinentes d'un ou plusieurs documents pour en produire une version plus réduite. Dans un premier temps, nous présentons l'état de l'art concernant les approches de RAT en général, puis nous nous concentrons sur les systèmes et approches existants conçus pour les textes Arabe. Ensuite, après avoir analysé ces approches, nous présentons quatre contributions afin de traiter les limitations identifiées dans les systèmes existants et, donc, améliorer les performances des systèmes de RAT Arabe. Dans notre première contribution, nous proposons un nouveau système de RAT Arabe en utilisant à la fois un modèle de graphe bidimensionnel et un algorithme d'élimination de la redondance. La méthode proposée, classe les phrases en utilisant une combinaison de l'analyse statistique et sémantique pour identifier les relations sémantique entre les composants textuels. Nous montrons également que l'utilisation de stemming dans la phase de prétraitement améliore les performances du RAT Arabe. Dans la deuxième contribution, nous proposons une méthode basée sur l'apprentissage en profondeur. La représentation actuelle par sac de mots implique des données d'entrée éparses et de grande dimension. Par conséquent, la réduction de la dimensionnalité est un moyen efficace pour accroître la puissance de la discrimination des caractéristiques. Nous adoptons le variational auto-encoder (VAE) en tant que modèle génératif pour la technique d'apprentissage des caractéristiques non supervisées pour générer une représentation abstraite pour chaque phrase. Notre système utilise cette nouvelle représentation pour classer chaque phrase du texte et en extraire les plus saillantes. Nous étudions également l'utilisation du VAE sur une technique basée sur les graphes et une autre basée sur la requête utilisateur. Dans notre troisième contribution, nous adoptons d'abord plusieurs modèles de réseaux de neurones profonds non supervisés pour le RAT Arabe. Nous proposons aussi l'utilisation de la représentation distribuée des mots arabes, que nous construisons à partir d'un grand corpus Arabe en utilisant la technique word2vec. Après cela, nous améliorons la qualité des modèles adoptés en proposant des modèles basés sur des techniques d'apprentissage ensembliste. Nous évaluons également les approches proposées sur un corpus Anglais. Enfin, dans notre quatrième contribution, nous proposons d'améliorer les performances des modèles déjà proposés par l'utilisation des techniques de clustering et d'identification des thèmes. A partir d'un ensemble de clusters construits par l'apprentissage d'une grande collection de documents Arabe, nous identifions l'espace de thèmes pour chaque cluster. Ensuite, pour chaque document à résumer, nous identifions le cluster auquel il appartient. Par la suite, le document est représenté dans l'espace des thèmes associés au cluster du document. Nous utilisons ensuite cette représentation pour apprendre les caractéristiques abstraites à l'aide des algorithmes de réseaux neuronaux non supervisés et de techniques d'apprentissage ensembliste. La performance de toutes les approches proposées est évaluée à l'aide de plusieurs corpus conçus à cet effet. Les résultats expérimentaux illustrent l'efficacité de nos approches.

Mots clés : Résumé automatique des textes Arabe, traitement automatique du langage naturel, analyse sémantique, apprentissage profond, machine d'apprentissage extrême, word embedding, apprentissage ensembliste, clustering, modélisation thématique.