INSTITUT FRANCOPHONIE INTERNATIONAL (IFI)

¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤¤

TRAVAIL PERSONNEL ENCADRE (TPE)

Resume automatique de texte

RAPPORT DE RECHERCHE BIBLIOGRAPHIQUE

Rédigé par :

Hugues Kanda

Encadrant

Ho Tuong Vinh

PROMOTION 22

Année Académique 2018– 2019

I. INTRODUCTION

Mon travail est basé sur le problème pratique de résumer automatique de texte oriente vers la prévention, l'éradication et détection des maladies ravageuses déclenchées sur les réseaux sociaux. De jour en jour, la masse d'informations textuelles sous forme électronique ne cesse d'augmenter, que ce soit sous forme de documents accessibles sur internet, dans les bases de données des entreprises et des gouvernements. Il devient de plus en pins difficile d'accéder aux informations intéressantes sans l'aide d'outils spécifiques.

Dans ce contexte, il est nécessaire de pouvoir accéder au contenu des textes par des moyens rapides et efficaces. C'est la fonction d'un résumé qui constitue un moyen efficace et éprouvé pour représenter le contenu des textes, et ainsi permettre un accès rapide à leur contenu.

La technique du résumé automatique qui permet de choisir un ou plusieurs documents utilement est une réponse à ce besoin.

II. PROBLEMATIQUE

Notre travail relève les problèmes suivants : actuellement les réseaux sociaux (Facebook, Twitter, etc.) sont devenus les sources génératrices des données.

Dans notre travail, nous comptons exploiter ces sources afin de détecter et résumer de manière automatique tous les cas liés aux maladies qui font rage en Afrique à savoir : Ebola, méningite, malaria, et d'autres pestes et fléaux.

Les gens deviennent de plus en plus actifs à rapporter et à poster sur leur situation sanitaire et sur ce qu'ils leur semblent bizarres.

III. TECHNIQUES DE RESUME AUTOMATIQUE DE TEXTE

Depuis lors, de nombreux travaux ont été publiés pour traiter le problème de la résume automatique de texte.

En général, il existe deux méthodes différentes pour le résumé automatique: **l'extraction et l'abstraction**.

La méthode de résumé de texte extractif fonctionne en sélectionnant un sous-ensemble de mots, phrases ou phrases existantes du texte original pour former un résumée, elle utilise une approche statistique pour sélectionner des phrases ou des mots-clés importants à partir du document. Les phrases extraites tendent à être plus longues que la moyenne. Les informations conflictuelles peuvent ne pas être présentées avec précision.

La méthode de résumé de texte abstrait génère une phrase à partir d'une représentation sémantique, puis utilise des techniques de génération de langages naturels pour créer un résumé plus proche de ce qu'un humain pourrait générer. Un tel résumé pourrait contenir des mots qui ne sont pas explicitement présents dans l'original. Il consiste à comprendre le texte original et à le répéter en moins de mots. Il utilise une approche linguistique pour comprendre le texte original, puis génère un résumé.

Les résumés abstractifs sont plus précis que le résumé extractif, mais sont difficiles à générer à cause d'ils besoin d'une compréhension approfondie des tâches du PNL.

La méthode de résumé de texte abstractive et extractive utilise des approches statistiques ou linguistiques ou une combinaison des deux pour générer un résumé.

Différentes approches statistiques sont discutées dans la section ci-dessous :

1. Approches statistiques

Les approches statistiques [1] peuvent résumer un document en utilisant les caractéristiques statistiques de la phrase comme le titre, l'emplacement, la fréquence, l'attribution de poids aux mots-clés, puis le calcul de la phrase et la sélection de la phrase la plus haute. L'importance d'une phrase peut être décidée par plusieurs méthodes telles que:

1.1. Méthode du titre [9]

Cette méthode [9] [2] indique que les phrases qui apparaissent dans le titre sont considérées comme plus importantes et sont plus susceptibles d'être incluses dans le résumé. Le score des phrases est calculé comme le nombre de mots couramment utilisés entre une phrase et un titre. La méthode de titre ne peut pas être efficace si le document n'inclut aucune information de titre.

1.2. Méthode de localisation [9]

Les poids sont affectés au texte en fonction de l'emplacement, que ce soit apparu dans la position principale, médiane ou finale dans un paragraphe ou dans apparaît dans la section proéminente du document, par exemple conclusion ou introduction. Conduire plusieurs phrases d'un document ou dernières phrases ou conclusion est considérés être plus important et inclus dans le résumé. Edmundson [9] ont utilisé cette méthode. L'emplacement méthode repose sur les en-têtes d'intuition suivants: au début et à la fin du texte, le texte est en gras, contenir des informations importantes pour le résumé.

1.3. Méthode tf-idf [5]

Le terme fréquence fréquence-document inverse est un statistique numérique qui reflète l'importance d'un mot à un document. Il est souvent utilisé comme un facteur de pondération dans l'information récupération et extraction de texte. Tf-idf est utilisé majorly pour les mots d'arrêt filtrage dans une application de résumé et de catégorisation de texte.La valeur de tf-idf augmente proportionnellement au nombre de fois qu'un mot apparaît dans le document. Pondération tf-idf ie schéma est souvent utilisé par les moteurs de recherche notation et classement de la pertinence d'un document donné à un utilisateur question. Le terme fréquence f (t, d) signifie la fréquence brute d'un terme dans un document, que je le nombre de fois que le terme t se produisait dans document d. La fréquence de document inverse est une mesure de si le terme est commun ou rare dans tous les documents. C'est obtenu en divisant le nombre total de documents par le nombre de documents contenant le terme.

1.4. Méthode de mot de repère [9]

Le poids est attribué au texte en fonction de sa signification, comme les poids positifs «vérifiés, significatifs, meilleurs, cet article» et les poids négatifs comme «difficilement, impossible». Les phrases de repère dépendent généralement du genre. La phrase constituée de telles phrases de repère peut être incluse dans le résumé. La méthode des phrases de repère est basée sur l'hypothèse que de telles phrases fournissent un contexte «rhétorique» pour identifier des phrases importantes. L'abstraction de la source dans ce cas est un ensemble de phrases de cue et les phrases qui les contiennent. Au-dessus de toutes les caractéristiques statistiques sont utilisées par la synthèse de textes extractive.

2. Approches linguistiques

Linguistique est une étude scientifique du langage qui comprend l'étude de la sémantique et de la pragmatique. L'étude de la sémantique signifie comment le sens est déduit des mots et des concepts et l'étude de la pragmatique incluse comment le sens est déduit du contexte. Les approches linguistiques reposent sur l'examen de connexion entre les mots et en essayant de trouver le concept principal en analysant les mots.

La synthèse de textes abstrait est basée sur une méthode linguistique qui implique le traitement sémantique pour la synthèse.Les approches linguistiques ont quelques difficultés à utiliser des outils d'analyse linguistique de haute qualité (un analyseur de discours, etc.) et des ressources linguistiques (Word Net, chaîne lexicale, espace vectoriel de contexte, etc.).

Barzilay et Elhadad [4], Miller et al a proposé et développé des concepts forts avec l'aide de caractéristiques linguistiques mais ils nécessitent beaucoup de mémoire pour sauvegarder les informations linguistiques comme Word Net et la capacité du processeur en raison de connaissances linguistiques supplémentaires et de traitements linguistiques complexes.

2.1. chaîne lexicale [4] [7]

Le concept de chaînes lexicales a été introduit par Morris et Hirst [6]. Fondamentalement, les chaînes lexicales exploitent la cohésion parmi un nombre arbitraire de mots apparentés. Chaînes lexicales peut être calculé dans un document source par regroupement (chaînage) ensembles de mots liés sémantiquement. Identités, synonymes, et hypernymes / hyponymes sont les relations entre mots qui pourraient les amener à être regroupés dans le même chaîne lexicale. Les chaînes lexicales sont utilisées pour l'IR et la grammaire corrections d'erreur [4] [7].

En calculant les chaînes lexicales, le les instances nominales doivent être regroupées selon les relations, mais chaque instance de nom doit appartenir à exactement un chaîne lexicale. Il y a plusieurs difficultés à déterminer quelle chaîne lexicale une instance de mot particulière devrait rejoindre.

Les mots doivent être groupés de façon à créer le plus fort chaîne lexicale la plus longue.

2.2. Réseau de mots [10]

Word Net est une base de données lexicale en ligne disponible pour l'anglais la langue. Il regroupe les mots anglais en ensemble de synonyme appelé sy nets. Word Net fournit également une signification courte de chaque relation sy net et sémantique entre chaque sy net.Word-net sert également de thésaurus et de dictionnaire en ligne qui est utilisé par de nombreux systèmes pour déterminer la relation entre les mots. Thesaurus est un ouvrage de référence qui contient une liste de mots regroupés en fonction de la similitude de sens.

Les relations sémantiques entre les mots sont représenté par des ensembles de synonymes, des arbres hyponymies. Word-net sont utilisé pour construire des chaînes lexicales en fonction de ces relations.Word Net contient plus de 118 000 mots différents formes [10]. LexSum est un système de synthèse qui utilise Word Net pour générer la chaîne lexicale.

2.3. Théorie des graphes [8]

La théorie des graphes [8] peut être appliquée pour représenter la structure du texte ainsi que la relation entre les phrases du document. Les phrases dans le document sont représentées sous la forme de nœuds. Les arêtes entre les nœuds sont considérées comme des connexions entre les phrases. Ces connexions sont liées par une relation de similarité. En développant différents critères de similarité, la similarité entre deux phrases est calculée et chaque phrase est notée. Chaque fois qu'un résumé doit être traité, toutes les phrases ayant le score le plus élevé sont choisies pour le résumé.

Dans les algorithmes de classement de graphe, l'importance d'un sommet dans le graphe est calculée itérativement à partir du graphe entier.L'algorithme TextRank est un algorithme basé sur un graphique qui s'applique à la synthèse. Un graphique est construit en ajoutant un sommet pour chaque phrase du texte. Les arêtes entre sommets sont établies en utilisant des interconnexions de phrases.Ces connexions sont définies en utilisant une relation de similarité, où la similarité est mesurée en fonction du chevauchement du contenu.

Le chevauchement de deux phrases peut être déterminé comme le nombre de jetons communs entre les représentations lexicales de deux phrases. La partie itérative de l'algorithme est par conséquent appliquée sur le graphe des phrases. Lorsque son traitement est terminé, les vertices (phrases) sont triés par leurs scores. Les phrases les mieux classées sont incluses dans le résultat.L'extraction de résumé par génération de graphes sémantiques [8] est une méthode qui utilise des triplets SOP (sujet-objet-prédicat) à partir de phrases individuelles pour créer un graphe sémantique du document original.

À l'aide de l'algorithme d'apprentissage support Vecteur Machines, il forme un classificateur pour identifier les triplets SOP à partir du graphe sémantique du document qui appartient au résumé. Généralement, les éléments fonctionnels principaux des phrases et des clauses sont les sujets, les objets et les prédicats. Ainsi, l'identification et l'exploitation des liens entre eux pourraient faciliter l'extraction du texte pertinent. Une méthode qui crée un graphe sémantique d'un document, basée sur la forme logique triplante sujet-prédicat-objet (SPO), et apprend un sous-graphe pertinent qui pourrait être utilisé pour créer des résumés.

2.4. Clustering [3]

Le regroupement est utilisé pour résumer un document en regroupant les données ou les phrases similaires. La méthode indique que le résultat du résumé dépend non seulement des caractéristiques de la phrase, mais aussi de la mesure de similarité de la phrase.MultiGen est un système multidocument dans le domaine des nouvelles. L'une des méthodes de classification des phrases développées par ZHANG Pei-Ying et LI Cun-he [3] est discutée dans l'article [2].

L'algorithme utilisé pour déterminer le nombre de groupes est la méthode K-means. Il aide à regrouper les phrases du document et extrayait les phrases de sujet pour générer le résumé extractif pour le document. De cette manière, les phrases sont regroupées et sélectionnées pour la synthèse. Les approches linguistiques sont plus difficiles à mettre en œuvre alors que les approches statistiques ont plus de succès mais ont peu de limites.

Par conséquent, dans ce travail, nous nous concentrons sur les méthodes de résumé extractif et fournissons un aperçu de certaines des approches les plus dominantes dans cette catégorie. Il y a un certain nombre d'articles qui fournissent des aperçus complets des techniques et des systèmes de synthèse de textes.

Dans ce travail, l'étude se concentrera sur la synthèse de textes extractive et examinera principalement les approches concernant l'extraction de phrases, la synthèse de domaines spécifiques et les méthodes de synthèse multidocuments. La section suivante présente les détails sur les approches d'extraction de phrases. Ensuite, la discussion sur le résumé spécifique au domaine est donnée. Suite à cela, les discussions sur les approches de synthèse multidocuments sont présentées et finalement le document se termine par une conclusion.

Différents types de résumé pourraient être utiles dans diverses les applications et les systèmes de synthèse peuvent être catégorisés basé sur ces types. En plus de l'abstrait et de l'extrait, ils sont différents types de résumés. Une compréhension complète des principales dimensions de la variation, et les types de raisonnement nécessaire pour produire chacun d'eux, est encore une question d'enquête. Cela rend l'étude du texte automatisé résumer un domaine passionnant dans lequel travailler. Diverses méthodes de synthèse peuvent être comparées en fonction du type de résumé et application. Le système de résumé peut être classés dans les catégories suivantes, ils sont:

- Basé sur des approches

Il y a deux stratégies pour résumer ceux qui sont résumés par extraction, qui consiste à extraire phrases source telle qu'il est et en ajoutant dans un résumé, et résumé par abstraction, ce qui implique de générer de nouvelles phrases pour le résumé [1]. Le besoin d'abstraction est particulièrement élevé lorsque les opinions sont diverses. Le système par extractive extrait simplement les phrases du document original et les ajoute au résumé. La méthode extractive est généralement facile à mettre en œuvre et repose sur des caractéristiques statistiques

et non sur une relation sémantique avec des phrases. Par conséquent, le résumé généré par cette méthode a tendance à être incohérente. La synthèse par abstraction nécessite une compréhension du texte original et génère ensuite le résumé qui est sémantiquement lié. Il fournit un résumé plus généralisé mais il est difficile à calculer.

- Basé sur le type de détails

Basé sur le type de détail, le résumé peut être informatif ou indicatif [1]. Un résumé indicatif est utilisé pour une vision rapide d'un long document et il ne fournit que l'idée principale du texte original. Ceux-ci sont généralement petits et encouragent l'utilisateur à lire le document original. Par exemple pendant que l'achat de tout roman un acheteur lit le résumé fournit à l'arrière du roman. Résumé informatif sert de substitution au document original. Il fournit des informations concises sur le document original à l'utilisateur.

- Basé sur le type de contenu

Cette classification est basée sur le type de contenu dans le document original [1]. Le résumé générique est un système qui peut être utilisé par n'importe quel type d'utilisateur et le résumé ne dépend pas de l'objet du document. Toutes les informations sont au même niveau d'importance et ne sont pas spécifiques à l'utilisateur.

La synthèse de requêtes [1] est un type de réponse à une question où le résumé est le résultat d'une requête. Il fournit la vue des utilisateurs et ne peut être utilisé par aucun type d'utilisateur.

- Basé sur la limitation

Le résumé peut être classé en fonction de la limitation de l'entrée texte [1]. Les systèmes spécifiques au genre n'acceptent que des types particuliers de contributions, comme des articles de journaux, des histoires, des manuels, etc. Limités au type de contribution qu'ils peuvent accepter.

Le système indépendant du domaine peut accepter différents types de texte. Ils ne dépendent pas du domaine et peuvent être utilisés par n'importe quel type d'utilisateur. Il y a peu de systèmes qui dépendent du domaine.

- Basé sur le nombre de documents d'entrée

La synthèse peut être classée selon qu'un système accepte un ou plusieurs documents en entrée [1]. La synthèse de document unique ne peut accepter qu'un seul document en entrée. Ils sont généralement plus faciles à produire car ils impliquent la synthèse d'un document unique. La synthèse multidocument accepte plusieurs documents du même sujet en tant qu'entrée. Il est plus difficile à mettre en œuvre car il y a plusieurs documents à résumer.

- Basé sur la langue

Le système monolingue n'accepte que les documents avec un langage spécifique et la sortie est basée uniquement sur cette langue. Les systèmes multilingues peuvent accepter des documents dans différentes langues et produire un résumé de différentes langues. Les tableaux suivants présentent une comparaison de toutes les méthodes de résumé en fonction du type de résumé.

IV. SOLUTIONS POSSIBLES

Nous aurons à construire un prototype pour le résumé de texte et nous allons faire une synthèse des avantages et inconvénients des méthodes les plus pertinentes.

Par ce fait, nous aimerions performer l'analyse textuelle basée sur l'usage d'apprentissage automatique afin d'alerter les gouvernements, institutions ou organisations qui s'en occupent dans le souci de prendre des promptes décisions qui vont impacter positivement les mesures préventives, mobilisation de personnel, de corps médical, de ressource humains, et le déploiement le personnel dans le pays affecte .Résumé automatique de texte oriente vers la prévention, l'éradication et détection des maladies ravageuses déclenchées sur les réseaux sociaux.

V. COMPARAISON

Nous présentons la comparaison des différentes approches sous forme de tableau ci-dessous :

Type de méthodes de résumé	Sous-type	Concept	Avantages	Désavantages	Application / TravailTerminé
1. Approches Figures	Abstractive	C'est le processus de réduction d'un document texte afin de créer un résumé sémantiquement lié	Bon taux de compression. Texte plus réduit et résumé lié sémantiquement	Difficile à calculer	SUMMRIST [11]
	Extractif	Il consiste à sélectionner des phrases importantes à partir du document original en fonction des caractéristiques statistiques	Facile de calculer parce qu'il ne fait pas traiter avec la sémantique et plus réussi	Souffre de incohérences, manque de équilibre, résultats dans longue résumé	Sommet applet, d esigned par Surrey Université [11]
2. Détails	Indicatif	Seulement présente idée principale de texte à l'utilisateur. Ils peuvent être habitué rapidement décider si un	Encourage-le utilisateurs à lis le principale document en profondeur. Utilisé pour rapide catégorisation et plus facile à produire	Détaillé l'information n'est pas présent	Information présent sur le arrière la film emballer ou des romans Longueur 5 à dix%

		le texte vaut la peine en train de lire			
	Informatif	Donne concis information de la principale texte	Sert comme une substitution pour le document principal	Ne fait pas fournir rapide aperçu	SumUM [11] Longueur 20 à 30%
3. Contenu	Générique	Résumé généralisé quel que soit le type de l'utilisateur. L'information est au même niveau d'importance	Peut être utilisé par n'importe quel type d'utilisateur	Il fournit une vue de l'auteur non spécifique à l'utilisateur	SUMMARIST [11]
	Requête basée	L'utilisateur doit déterminer le sujet du texte original sous forme de requête et le système extrait seulement cette information	Des informations spécifiques peuvent être recherchées. Il reflète l'intérêt de l'utilisateur	Non utilisé par tout type d'utilisateur. Il est basé sur le type d'utilisateur	Mitre est WebSumm [11]
4. Limitation	Dépendant du domaine	Résumer le texte dont le sujet peut être défini dans le domaine fixe	Ils sont conscients du domaine spécial dont ils dépendent	Limité à l'objet du document	TRESTLE [11]
	Genre spécifique	Accepter uniquement le type de texte spécial en entrée	Surmonte le problème de la synthèse hétérogène document	Modèle de limitation du texte	Newsblaster
	Domaine indépendant	Peut accepter n'importe quel type de texte	Tout type de saisie de texte est accepté. Ce n'est pas dépendant du domaine	Difficile à mettre en œuvre	Copier et coller le système [11]
5. Nombre de document d'entrée	Document unique	Peut accepter un seul document d'entrée	Moins de frais généraux	Impossible de résumer plusieurs documents de sujets connexes	Copier et coller le système [11]
	Multi Document	Peut accepter plusieurs documents d'entrée	Plusieurs documents du même sujet peuvent être résumés en document unique	Difficile à mettre en œuvre	SUMMONS Conçu par l'université de Columbia [11]

6. Langue	Monolingue	Peut accepter une entrée uniquement avec une langue et une sortie spécifiques est basé sur cette langue	Besoin de travailler avec une seule langue	Impossible de gérer une langue différente	FarsiSum [11]
	Multilingue	Peut accepter des documents dans une langue différente	Peut gérer plusieurs langues	Difficile à mettre en œuvre.	SUMM ARIST(English, Japanese, Spanish) [11]

Les méthodes de résumé de texte peuvent être classées principalement dans des catégories extractives et abstractives. Texte la récapitulation par extraction extrait simplement quelques phrases du document original en l'ajoutant au résumé.

VI. REFERENCES

LES OUVRAGES ET ARTICLES

[1] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh.A Comprehensive Survey on Text Summarization Systems. 2009 In proceeding of: Computer Science and its Applications, 2nd International Conference

[2] Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J.Res. Develop., 159–165.

[3] ZHANG Pei-ying, LI Cun-he. Automatic text summarization based on sentences clustering and extraction.

[4] Barzilay, R., Elhadad, M.Using Lexical Chains for Text Summarization. In Proc. ACL/EACL'97 Workshop on Intelligent Scalable Text summarization, Madrid, Spain,1997, pp. 10–17.

[5] Youngjoong Koa, Jungyun Seo 2008.An effective sentence-extraction technique using contextual information and statistical approaches for text summarization.

[6] Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics 17(1):21–43.

[7] Silber G.H., Kathleen F. McCoy. Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. Computational Linguistics 28(4): 487-496, 2002.

[8] Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya.Generic Text Summarization Using Word net. Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.

[9] Edmundson, H.P., 1968.New methods in automatic extraction. J. ACM16 (2), 264–285. S.

[10] William P. Doran, Nicola Stokes, John Dunnion, and Joe Carthy, "Comparing lexical chain-based summarization approaches using an extrinsic evaluation," In Proc. Global Word net Conference (GWC 2004), 2004.

[11] Mahak Gambhir1 · Vishal Gupta1 Recent automatic text summarization techniques: a survey Artif Intell Rev (2017) 47:1–66 DOI 10.1007/s10462-016-9475-9, Published online: 29 March 2016