REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي و البحث العلمي
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
جامعة عمّارثليجي بالأغواط
UNIVERSITE AMAR TELIDJI LAGHOUAT
كلية العلوم
FACULTE DES SCIENCES
DEPARTEMENT DE MATHEMATIQUES ET INFORMATIQUE

Mémoire de MASTER

Domain : Mathématiques et Informatique
Filière : Informatiques
Option : Systèmes d'Information et de Décision

Par :
TaibiMessaoudeNadjia

THEME

Implémentation d'une nouvelle approche de résumé de texte biomédicale

Soutenu publiquement le 13-07-2021 devant le jury composé de :

Mr .Younes Guellouma	*Professeur*	*Président*
Mr.Laaradj.Chellama	*M.C. (A)*	*Examinateur*
Mr. Mustapha Bouakkaz	*M.C. (A)*	*Encadreur*

Année Universitaire 2020/2021

Résumé

Le grand volume d'informations textuelles dans le domaine biomédical est toujours un défi qui conduit les chercheurs à développer un nouvel outil de traitement de texte spécifique au domaine. Au cours des dernières décennies, l'extraction de texte biomédicale est devenue en pleine croissance, parmi les taches de l'extraction de textes biomédicaux qui a attiré de nombreux chercheurs. Le résumé automatique biomédical est un domaine de recherche très important car il permet de condenser le texte source tout en préservant les idées importantes ou de répondre aux besoins spécifiques des utilisateurs. Dans notre travail, afin de résoudre le problème de résumé de texte, nous avons proposé une nouvelle contribution basée principalement sur deux étapes. La première étape consiste à nettoyer le texte brute et de calculer le nombre les occurrences de chaque mots clés extrais, puis avec le top mot extrait et notre nouvelle approche nous construisant le résume de texte sous traitement. Afin de mesuré la qualité de notre résumé obtenu, nous avons implémenté des mesures d'évaluation le rappel, précision et F-mesure. Les résultats obtenus ont montré une performance remarquable de notre approche proposée.

Mots clés : Résumé automatique de texte biomédical, Recherche des motifs fréquents, rappel, précision, F-mesure

Abstract

The large volume of textual information in the biomedical field is still a challenge leading researchers to develop a new domain-specific word processing tool . In recent decades, biomedical text extraction has become a growing task, among the biomedical text extraction stains that has attracted many researchers.Biomedical automatic summary is a very important area of research as it allows to condense the source text while preserving important ideas or to meet the specific needs of users

In the part of the work of this theme, in order to solve some problems related to this area, our contributions are mainly based on two axes, The first axis concerns the combination of data-mining techniques (search for frequent reasons) with a conceptual representation of source text By algorithm that calculate the number I enter the article and calculate the occurrence of each word contains in the article, As well as we search for frequent biomedical patterns manually with Excel.

Then with the top word extracted and a proposed algorithm we enter the article with some step we will give the summary of this article.

As for the second axis, we proposed an algorithm that makes the evaluation we entered the automatic summary and summarized it from an expert of the same biomedical article and per some metrics we evaluated the result between the two summaries .

Keywords: Biomedical Text Summary, Frequent itemsets mining,rappel, precision , F-mesure

ملخص

ولا يزال الحجم الكبير للمعلومات النصية في المجال الطبي الأحيائي يشكل تحديا يدفع الباحثين إلى استحداث أداة جديدة لمعالجة الكلمات في مجال محدد، وفي العقود الأخيرة، أصبح استخراج النصوص الطبية الحيوية مهمة متنامية بين مستخلصات النصوص الطبية الحيوية التي اجتذبت العديد من الباحثين.الملخص التلقائي الطبي الحيوي هو مجال مهم جدا من مجالات البحث لأنه يسمح بتكثيف النص المصدر مع الحفاظ على الأفكار الهامة أو لتلبية الاحتياجات المحددة للمستخدمين وفي عملنا، وبغية حل مشكلة ملخص النصوص، اقترحنا إسهاما جديدا يستند أساسا إلى خطوتين، الخطوة الأولى هي تنظيف النص الخام وحساب عدد الحوادث لكل كلمة رئيسية مستخرجة، ثم مع الكلمة العليا المستخرجة ونهجنا الجديد يبني لنا ملخص النص قيد المعالجة.وبغية قياس نوعية الموجز الذي حصلنا عليه، نفذنا تدابير التقييم والتذكير والدقة والتدبير واو.وكانت النتائج التي تم التوصل إليها الأداء المتميز لنهجنا المقترح

الكلمات الرئيسية: ملخص النصوص الطبية الحيوية، البحث عن النمط المتكرر،

Cette thèse est dédiée à

Mes parents,

Ma belle-mère et mon beau-père,

Mon mari,

Mesfrères, et mon beau-frère

Mes sœurs, et ma belle-sœur.

Remerciement :

au nom de Dieu le tout puissant et miséricordieux à qui j'exprime la force morale et physique et nous a permis d'achever ce travail

Je voudrais tout d'abord remercier grandement mon directeur de mon projet Dr. Bouakkaz Mustapha, pour leur confiance, leur encadrement, leurs nombreux conseils, et surtout leur soutien dans les moments difficiles

Mes remerciements s'adressent également à tous les personnes qui ont contribué de près ou de loin avec leurs conseils ou avec leurs encouragements à l'accomplissement de ce travail

Mes remerciements vont à tous les enseignants du département d'informatique que nous respectons beaucoup.

Enfin, je souhaiterais adresser des remerciements plus particuliers à toute ma famille.

Table des matiéres

Table des figures

Liste des tableaux

Introduction générale

Contexte générale

Actuellement, 80% de l'information dans le monde est stockée sous forme de texte . Les données textuelles sont présentes dans les livres, les nouvelles, les articles, les blogs, les réseaux sociaux, etc. L'information médicale peut provenir de différentes sources, issue des dossiers médicaux informatisés, des entretiens de patients parlant de leur maladie, des symptômes, de la littérature biomédicale et d'Internet. Les progrès technologiques permettent d'informatiser les dossiers médicaux, les entretiens de patients, les articles médicaux, d'adapter les questionnaires de qualité de vie en version électronique, et d'échanger de l'information sur Internet.

La littérature biomédicale fournit une richesse d'informations pour les chercheurs, elle peut servir de point de départ pour évaluer l'état de l'art dans un domaine particulier ou comme source d'informations pouvant être utilisée pour construire des hypo mémoires de recherche qui peuvent ensuite être validée expérimentalement. De plus, cette base de connaissances peut servir de source pour l'interprétation des résultats expérimentaux

Le nombre d'articles qui sont ajoutés à ces bases de données bibliographiques augmente exponentiellement, par exemple la base d données bibliographique MED-LINE, accessible via le moteur de recherche PubMed contient plus de 28 millions de références à des articles de revues, avec presque de 1 million entrées ont été ajoutées en 2019. les résultats d'une recherche PubMed en utilisant des termes qui décrivent des maladies, des médicaments et des modèles d'organismes , Dans tous les cas, le nombre d'articles publiés sur ces sujets a augmenté de façon exponentielle. En plus, le taux de croissance du nombre des données expérimentales produites a également augmenté exponentiellement avec cet énormevolume, le processus de la recherche et de l'extraction d'informations pertinentes dans ces bases de données bibliographiques et la combinaison de ces informations avec des résultats expérimentaux, prend beaucoup du temps.

Au cours des dernières années, il y a eu une accroissement de travaux de recherche sur le domaine de textmining (TM), Le but est de détecter automatiquement, récupérer et extraire des informations dans un corpus de textes, combinant des approches impliquant la linguistique, des

statistiques et de l'informatique, en utilisant les techniques de l'intelligence artificiel, le traitement automatique de la langue et la fouille de données.

Fouille de textes biomédicale est un processus d'extraction de structures (connaissances) inconnues, valides et potentiellement exploitables dans les documents textuels, à travers la mise en œuvre de techniques statistiques ou de machine learning. Mais d'autres applications spécifiques aux textes sont possibles : résumé automatique, extraction d'information, etc…, il y'a plusieurs domaines de recherche : linguistique computationnelle, la bio-informatique, sciences de l'information médicale, Dans cette mémoire nous avons concentré nos travaux sur le domaine de résumé automatique de texte biomédical

Nos contributions

Les premiers travaux sur les résumés automatiques de textes datent des années 50. Pour extraire les phrases pertinentes nécessaires à la construction d'un résumé, considère des caractéristiques comme la fréquence d'occurrence des termes, des mots de titres la longueur et la position de la phrase, Eu égard à ce qui précède, l'objectif général de cette mémoire porte sur l'adaptation des techniques de fouille de texte pour la génération automatique des résumés de texte biomédical

Avec l'avènement de l'Internet et de moteurs de recherche de plus en plus performants, l'importance d'informations condensées du type résumé est devenue nécessaire pour faire ressortir l'information pertinente. De ce fait le résumé automatique a inspiré de nouvelles orientations, plusieurs nouvelles approches ont commencé à être explorées en linguistique (basée sur l'analyse du discours et de sa structure) et en statistique (basée sur la distribution des occurrences des mots)

Il ya plusieurs approches de résumé du texte, Mais il y a peud'approches de résumé de texte adaptées aux textes biomédicaux

Nous avons proposé deuxapproches publiées sous forme des articles de recherche :

Dans notre article intitulé « ***TextSummarization in the Biomédical Domain »*** présenté dans la « 6 aout 2019 » Le but premier de ce article est de passer en revue les efforts de recherche les plus importants accomplis au cours de la décennie actuelle en vue de nouvelles méthodes de synmémoire de textes biomédicaux.À mesure que les principales

parties de ce chapitre sont abordées, les tendances actuelles sont discutées et de nouveaux défis sont introduits.

Dans notre article intitulé "***Combine Clustering and FrequentItemsetsMining to EnhanceBiomedicalTextSummarization***" publié dans la revue "Expert Systemswith Applications (Elsevier)" nous proposons un nouveau système de synmémoire de textes biomédicaux qui combine deux techniques populaires d'exploration de données : le regroupement et l'extraction fréquente d'éléments.Le papier biomédical est exprimé sous la forme d'un ensemble de concepts biomédicaux utilisant le métathésaurus UMLS.L'algorithme K-means est utilisé pour regrouper des phrases similaires.Ensuite, l'algorithme d'Apriori est appliqué pour découvrir les items fréquents parmi les phrases groupées.Enfin, les phrases saillantes de chaque cluster sont sélectionnées pour construire le résumé en utilisant les items fréquemment découverts

Organisation du manuscrit de mémoire

Le travail présenté dans cette mémoire est organisé en quatre chapitres regroupés en deux parties, respectivement : introduction et une conclusion générale

La première thématique porte sur le contexte :

— Dans le chapitre 1 : nous avons présenté l'introduction générale

— Dans le chapitre 2, nous introduisons les concepts de base utilisés le long de cette mémoire, nous exposons les notions de fouille de texte biomédical, les différentes taches dans ce domaine, nous décrivons aussi les différentes techniques pour évaluer les résumés automatiques. Nous pressentons aussi les particularités du texte biomédical,

— Le chapitre 3 présente la formalisation de notre problématique de résumé automatique de texte biomédical, nous exposons un état de l'art de résumé automatique et leurs applications au texte biomédicalet mesures pour développer un système du résumé automatique...

— Dans le chapitre 4, nous présentons une nouvelle méthode de résumé automatique de texte biomédical basée sur la fouille de donnéesNous exposons en détail les formalismes et les différentes étapes de notre

approche. Ce chapitre inclut aussi des exemples d'applications pour simuler notre proposition,

— Dans le chapitre 5 , nous concluons cette mémoire en présentant un bilan général de l'ensemble de nos contributions et en évoquant de nouvelles perspectives de recherche

Chapitre 1
Concepts de base

Introduction

Nous sommes une espèce passionnée par la recherche mais qui a peur de découvrir. Nous répondons à nos peurs par nos croyances, un peu comme ces anciens marins qui refusaient l'idée du voyage, convaincus que chargés de leurs certitudes le monde s'achevait en un abîme sans fin.

La fouille de texte ou la découverte de connaissances à partir de texte pour la première fois mentionnée dans [1] : "agir avec la machine prise en charge l'analyse de texte"

Nous présentons dans ce chapitre introduit les concepts fondamentaux utilisés dans la mémoire. Premièrement, nous abordons le domaine de fouille de texte, nous définissons aussi les concepts fondamentaux du domaine de résumé automatique texte, et les méthodes utilisées dans la littérature résumés.

Fouille de texte

Historique

La fouille de texte, l'extraction de connaissance dans les textes ou le « Textmining » est incluse dans le domaine de l'intelligence artificielle. Elle est apparue durant les années 1990 aux États-Unis comme une nouvelle technique de l'intelligence artificielle . Les premiers algorithmes de fouille de données ont permis de réduire les volumes importants des données qui ont été utilisées dans le domaine du marketing

Définition

Également appelé **traitement automatique du langage**, peut être défini comme étant un ensemble de techniques issues de l'intelligence artificielle, alliant plusieurs domaines : **la linguistique, la sémantique, le langage, les statistiques et l'informatique**. Combinées ensemble, ces techniques permettent d'extraire des données pour recréer de l'information à partir de corpus de textes en les classifiant et les analysants de manière à **établir des tendances**

Fouille de textes, respecte deux étapes principales. La première étape, l'analyse, consiste à **analyser les corpus de textes** de manière à en reconnaître les mots, les phrases, les rôles grammaticaux ainsi que les relations et les sens de ces derniers entre eux. Cette première étape, commune à tous les traitements, ne trouve sa pertinence que lorsqu'elle est couplée à la seconde étape : **l'interprétation de l'analyse**. Cette étape permet de sélectionner des textes en particuliers parmi d'autres

L'extraction d'information

L'Extraction d'Information ou EI (en anglais, Information ExtractionouIE) est une technique de textmining désigne une technologie récente qui vise à extraire et à structurer automatiquement un ensemble d'informations précises apparaissant dans un ou plusieurs documents textuels écrits en langue naturelle, La tâche d'extraction est réalisée grâce au remplissage de formulaires prédéfinis (template).

Les informations extraites par un système d'Extraction d'Information peuvent être consultées par des utilisateurs humains (par exemple via la génération de rapports d'événements), être utilisées pour la génération de résumés (dans la même langue ou dans une langue différente) ou, dans la plupart des cas, alimenter une base de données afin d'être analysées plus tard [2], Parmi les applications de l'extraction d'information dans le domaine biomédical : la détection des interactions entre les protéines et les médicaments, l'analyse des tableaux des expressions ADN, la description des fonctionnalités d'un protéine, la détection des relations entre des gènes et des maladies etc. . . . [3].

Recherche d'informations

Un système de recherche d'information (SRI)est défini par un langage de représentation des documents (qui peut s'appliquer à différents corpus de documents) et des requêtes qui expriment un besoin de l'utilisateur (sous forme de mots-clés par exemple), et une fonction de mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur, c'est-à-dire répondant à son besoin d'information .[4], . Les systèmes de l'IR sont appelés "des moteurs de recherche", par exemple : Google,

Google scholar ouPub-Med [5]

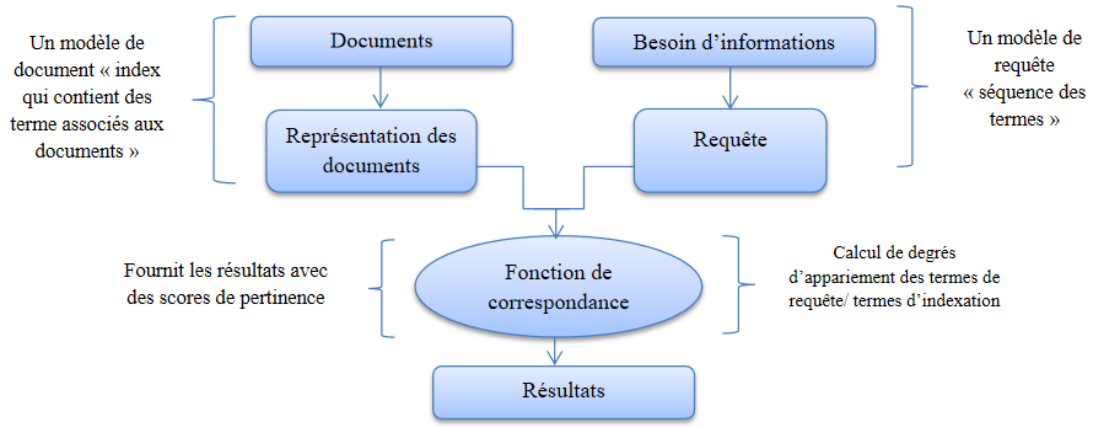

Figure 1 Système de recherche d'information

Traitement automatique de la langue

Le Traitements Automatique des Langues (TAL) est une discipline qui associe étroitement linguistes et informaticiens. Il repose sur la linguistique, les formalismes (représentation de l'information et des connaissances dans des formats interprétables par des machines) et l'informatique. Le TALa pour objectif de développer des logiciels ou des programmes informatiques capables de traiter de façon automatique des données linguistiques. Pour traiter automatiquement ces données, il faut d'abord expliciter les règles de la langue puis les représenter dans des formalismes opératoires et calculables et enfin les implémenter à l'aide de programmes informatiques

Traitement automatique de la langue naturelle

On regroupe sous le vocable detraitement automatique du langage naturel(TALN) l'ensemble des recherches et développements visant à modéliser et reproduire, àl'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication, L'objectif général du (TALN) est d'atteindre une meilleure compréhension du langage naturel [6]. D'autres incluent également l'emploi de techniques simples et durables pour le traitement rapide du texte

Prétraitement de texte

Les données textuelles sont une forme particulière de données complexes. Elles ne sont pas délimitées, structurées et étiquetées sémantiquement de façon explicite. En conséquence ces données nécessitent un traitement préalable. De manière générale, Pour réaliser les différentes tâches de fouille de texte, il est nécessaire de prétraiter les documents textuels et de stocker les données résultantes dans un format structuré pour un traitement ultérieur, l'objectif de ce prétraitement est de minimiser l'espace de recherche. En effet, même si les capacités des ordinateurs évoluent constamment, il n'est malheureusement pas possible de traiter les documents dans leur intégralité. En fonction des différentes thématiques de recherche [7], la plupart des approches de fouille de texte sont basées sur l'idée qu'un document textuel est représenté par un ensemble de mots (appelé sac de mots)

Fin d'obtenir tous les mots utilisés dans un texte brut, un processus de tokénisation est requis, c'est-à-dire qu'un document textuel est divisé en un ensemble de phrases et chaque phrase est divisée en ensemble de mot. L'ensemble de tous les mots obtenus en fusionnant tous les documents d'une collection est appelé un dictionnaire d'une collection de documents. Afin de réduire la taille du dictionnaire, l'ensemble des mots décrivant les documents peuvent être réduits par filtrage, lemmatisation ou racinisation.

1 **Les méthodes de filtrage** : suppriment des mots du dictionnaire. Une méthode de filtrage standard est le filtrage des mots vides. L'idée de filtrage de mots vide consiste à supprimer les mots contenant peu ou pas d'informations sur le contenu, comme des conjonctions, des prépositions (e.g : il, le, la, de, ce, lui, et . . .). De plus, les mots qui apparaissent beaucoup dans un texte peuvent être peu informatifs, et aussi des mots qui apparaissent très rarement sont susceptibles d'être sans intérêt particulier et peuvent être aussi supprimés du dictionnaire

2 Les méthodes de lemmatisation

Transformation d'un mot sous forme fléchie en sa forme canonique exemple : distances → distance, ce processus prend généralement beaucoup de temps et sujettes aux erreurs, dans la pratique, des méthodes de dérivation sont fréquemment appliquées

3 Les méthodes de racinisation : ou (stemming) tentent de construire une forme de mot où on ne pas faire un changement, la racine (stem) ne correspond pas toujours à un vrai mot, mais seulement au fragment d'un mot qui ne change jamais,Identification de la plus petite chaıne de caractères porteusede sensTrèsutilisée en recherche d'information [8], Il a défini un ensemble de règles de production pour transformer de manière itérative les mots en anglais en leurs racines.

Résumé automatique de texte

Définition

L'élaboration de systèmes plus performants passe donc par le détour de recherches fondamentales, en matière notamment decompréhension de texteset degénérationde textes, Résumer un texte peut revenir à assigner une phrase du texte originel au résumé=> classification statistique

En règle générale, un résumé doit être beaucoup plus court que le texte source. Cette caractéristique est définie par un taux de compression, qui

mesure le rapport entre la longueur du résumé et la longueur du texte original en mots ou en phrases.

Facteurs de résumé automatique

La présentation des facteurs qui suit vise donc avant tout à mettre l'accent sur la gamme et la richesse des influences et, partant, des variétés de la résumé Il est commode de distinguer trois classes de facteurs : entrée, objectif et de sortie

Facteurs d'entrée

Pour le résumé multi-documents est plus complexe que pourle résumé mono-document. En effet, dans un document unique, les phrases sont toutes issues de la même structurediscursive, et peuvent être restituées dans l'ordre du document source. En revanche, en multi-documents, lesphrases peuvent être extraites de documents épars, et une structure discursive doit être recomposée, La plupart des techniques utilisées dans les résumés d'un seul document sont également utilisé dans les résumés multi-documents, mais ce dernier présente quelques défis supplémentaire comme par exemple :

_ Le degré de redondance d'informations contenu dans une collection de documents parlant de même sujet

_Le taux de compression sera considérablement moins pour une collection de documents que pour les résumés d'un seul document

— Les problèmes de cohérence, de cohésion et de coréférence dans les résumés multi-documents est beaucoup plus fréquents

Résumé monolingue, multilingue et cross-lingue

Monolingue

Tout d'abord, un dictionnaire monolingue est celui qui utilise une seule langue dans ses entrées, donc qui donne des explications pour les mots vedettes dans une même langue, il existe deux types de dictionnaires monolingues. Le premier est le dictionnaire monolingue classique adressé aux locuteurs natifs. Quant au deuxième, il est plutôt pédagogique et il est destiné aux personnes apprenant une langue étrangère.

Multilingue :

Le principal objectif de la recherche d'information multilingue est de fournir à l'utilisateur qui ne serait pas familier avec une langue

particulière, mais qui serait quand même intéressé à obtenir des documents dans une autre langue ou plusieurs autres langues. Les principales tâches reliées à l a recherche d'information multilingue sont le filtrage, la sélection et le classement de documents qui pourraient être pertinents pour l'usager

cross-lingue

Le résumé automatique cross-lingue serait d'appliquer un système de Traduction automatique (TA) directement sur les sorties d'un système de résumé automatique classique

Facteurs de sortie

Le résumé extractif vise à choisir des parties du document d'origine, comme des phrases entières voire même un paragraphe.

Un résumé abstractif paraphrase le contenu du document original tout en prenant en compte la cohésion et la concision du résumé en sortie [9]

Facteurs d'objectifs

-
RésuméindicatifousignalétiqueIls'agitdesignaleroud'indiquerd'unemanière brèvelesthèmesd'étude.Cerésumédonneuneindicationsurletyped'informati onqu'onpeutespérertrouverdansledocumentmaisilnedonnepasl'information elle – même. L'objectifdecetypederésuméestdefairesavoiràl'usagers'ildoitlireletexteinté gralement,s'ilnedoitenlirequ'unepartieousiledocumentsignalénel'intéressep asdutout.Lerésuméindicatifn'adoncpaspourvocationdedispenserdelalectur edesdocumentspertinents.Ilestunoutildesélection,detri.

-
RésuméinformatifouanalytiqueCetypederésumésertvéritablementàinforme rl'utilisateursurlecontenududocumentoriginalanalyséetpeutremplacerlalect ureinextensodecelui-
ci.Laproblématiquedurésuméinformatifestdouble.Ils'agitdecomprendreceq uin'estpasinformationdansuntexteetdeconnaîtrelebesoindel'utilisateurfinal. Ildoitdoncsesituerenfonctiondusavoirdel'usagerfinal(c'estsouventunspécia liste)etdesonbesoin.Cettedistinctionestsouventétendueàuntroisièmetypede résumé,

Lerésumécritique/évaluatif.Ilévaluelaproblématiquedudocumentsource,en exprimantl'opiniondurésumeurconcernantlaqualitédutravaildel'auteur

Les méthodes de résumé automatique ne peuvent pas produire de tels résumés en raison de l'impossibilité d'encoder ce type d'expertise à ce jour

Le résumé indicatif, le résumé informatif, le résumé générique, le résumé orienté et le résumé de mise à jour :

Le résumé indicatif a pour objectif d'aider le lecteur à agir sur sa décision à consulter ou pas un document, en lui indiquant les thématiques abordées et développées dans le document source, sans considérer les détails. Le résumé informatif a pour objectif principal de renseigner le lecteur sur les principales informations quantitatives et qualitatives, il est considéré comme une version abrégée, conservant l'organisation générale du document source. Le résumé générique résume le document sans prendre en compte les besoins en information des utilisateurs, par contre, le résumé orienté a pour objectif de ne résumer que les informations qui répondent à une requête de l'utilisateur. Le résumé de mise à jour se contente de fournir un résumé sous l'hypomémoire que l'utilisateur a déjà des connaissances sur la thématique et qui n'a besoin que des nouveautés importantes, tout en évitant la redondance de l'information

Résumé général, spécifié à un domaine

Un système de résumé de texte automatique général est un système génère des résumés pour n'importe quel type de document sans prise en considération les caractéristique du document traité

Les systèmes de résumé de texte spécifiques à un domaine prisent en considération les spécificités de chaque document [10]

Caractéristiques des résumés

Un résumé doit être doté d'un certain nombre de caractéristiques : la concision, la couverture, la fidélité, la cohésion et la cohérence.

La concision

Est à relier directement au taux de réduction qui est le rapport entre la longueur du texte source et celle du résumé. D'une manière générale, le taux de réduction est proportionnel au caractère restrictif du (ou des) critère(s) employé(s) pour générer le résumé .Par exemple, un critère très restrictif du genre "ne retenir que les expressions conclusives et de résultats" produira des résumés plutôt courts. À l'inverse, un critère plus vague et beaucoup moins restrictif du genre "faire un résumé informatif" conduira dans la plupart des cas à des résumés plus longs. Il est aussi

important de retenir qu'un processus de résumé faisant intervenir la reformulation permet de générer des résumés considérablement plus courts que ceux produits sans processus de reformulation

Couverture

La couverture est en quelque sorte le rapport entre le nombre de thèmes ou d'éléments présents dans le texte source et ceux présents dans le résumé. La nature des éléments est fonction du type de résumé considéré : pour un résumé indicatif, on retiendra uniquement les thèmes a bordés ; dans un résumé résultat, on s'intéressera principalement aux expressions conclusive set aux résultats. Dans le cas d'un résumé informatif, c'est-à-dire, une réduction "à l'identique " du texte, la couverture est plus difficile à déterminer.

Fidélité

La fidélité est aussi un critère important pour caractériser un résumé. Elle représente la relation de similarité objective existant entre le résumé et le texte source. C'est en quelque sorte une mesure de la qualité globale du résumé. La notion de fidélité intègre, comme composante, la couverture. En règle générale, un résumé ayant une couverture correcte sera assez fidèle au texte source.

Cohésion et cohérence

Les deux derniers critères que nous exposerons comme définissant un résumé sont intimement liés à la notion de texte elle-même. Il s'agit de la cohésion et de cohérence.

La cohésion peut être vue comme le résultat de l'application de mécanismes visant à maintenir une unité référentielle (mécanisme d'anaphore) et argumentative (emploi des connecteurs).

La cohérence quant à elle découle plus de la bonne application des mécanismes rhétoriques et thématiques(suivi du thème), qui rendent un texte intelligible.

Évaluation de résumé automatique

Les évaluations du résumé de texte sont de deuxtypes : extrinsèque et intrinsèque

Dans une évaluation extrinsèque, les résumés sont évalués dans le contexte d'une tâche spécifique réalisée par un humain ou une machine.

Dans l'évaluation intrinsèque, les résumés sont évalués par rapport à une référence ou modèle idéal

Évaluation de qualité de texte : Les mesures basées sur la qualité de texte contrôlent les ses aspects linguistiques, par exemple :

- Grammaire
- Non-redondance
- Clarté de la référence

Évaluation de contenu

Mesures de Co-sélection :

Pour les extraits de phrases (les résumés extractifs), elle est souvent mesurée par la Co-sélection c'est-à-dire combien de phrases idéales dans le texte ont été sélectionné dans le résumé. La précision, le rappel et la F-mesure sont les principaux paramètres d'évaluation de la Co-sélection

— **Précision :** est le nombre de phrases qui existent dans le résumé généré par le système et le résumé idéal divisé par le nombre de phrases dans le résumé du système.

— **Rappel :** est le nombre de phrases survenant à la fois dans le système et le résumé idéal divisé par le nombre de phrases dans le résumé idéal.

— **F-mesure** : est la moyenne entre la précision et le rappel.

$$F = 2 \cdot \frac{(\text{précision} \cdot \text{rappel})}{(\text{précision} + \text{rappel})}$$

Mesures basées sur le contenu : les mesures basées sur le contenu comparent les mots d'une phrase plutôt que la phrase entière. Leurs avantages est qu'ils peuvent comparer des résumés générés par un système avec des résumés générés par des humains contenant des nouvelles phrases écrites. Alors que les mesures de Co-sélection ne peuvent pas le faire, les mesures de similitude basées sur le contenu peuvent le faire

Statistiques de Co occurrence N-gram (ROUGE)

ROUGE (Recall-OrientedUnderstanding pour Gisting Evaluation) a été utilisé comme méthode d'évaluation automatique universelle pour la première fois dans les conférences DUC en 2003. ROUGE est une famille de mesures qui reposent sur le calcul de la similarité entre les n-grammes, les séquences de mots, et les mots assemblés entre un résumé de système

et des résumés de références. La méthode ROUGE contient cinq variantes : ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, et ROUGE-SU

Le texte biomédical présente certaines propriétés uniques qui doivent être prise en compte dans le développement d'un système de résumé automatique. Premièrement, comme nous avons annoncé précédemment, les informations médicales apparaissent sous forme de différents types de documents, Chaque type de document présente des caractéristiques à prendre en compte dans le processus de résumé. Nous nous concentrons sur les articles scientifiques, qui sont principalement composés de texte, Les articles Biomédicaux suivent souvent la structure IMRaD (Introduction, Méthode, Résultats et discussion),

Deuxièmement, les particularités de la terminologie rendent difficile le traitement automatique des informations biomédicales, le premier défi est le problème des synonymes, les homonymes.

Conclusion

Nous avons présenté dans ce chapitre les principales notions et concepts utilisé tout en long dans cette mémoire, nous avons défini le domaine de la fouille de texte, nous avons donné quelques taches dans ce domaine en se focalisant sur le domaine choisi qui est domaine de résumé automatique de texte biomédical , dont lequel nous avons défini les caractéristiques des résumés et les différents facteurs qui contribuent dans le processus de développent des méthodes de résumé automatique de résumés, les différentes méthodes connues pour évaluer les performances des systèmes de résumé automatique existant dans la littérature.

Dans le chapitre qui suit nous dressons un état de l'art sur les approches de résumé de texte.

Chapitre 02 :
L’état de l’art

Introduction

Le résumé automatique se propose de faire une extraction de l'information jugée importante d'un texte d'entrée pour construire, à partir de cette information, un nouveau texte de sortie, condensé. Ce nouveau texte permet d'éviter la lecture en entier du document source. Nous présentons dans ce chapitre l'état de l'art du résumé automatique.

Les Premiers travaux

Le travail de Luhn [Luhn 1958] est considéré comme le premier travail dans le domaine de résumé automatique. La plupart des systèmes de résumé actuels s'inspirent de ce travail. Luhn, dans son travail, utilise la fréquence de mots ainsi que la position des mots dans la phrase, pour calculer l'importance d'une phrase. Dans ce qui suit, nous allons présenter les étapes de résumé comme proposé par Luhn, dont le prétraitement et le calcul des scores de phrases. Ces étapes sont, ensuite, reprises et étendues dans les systèmes de RA statistiques.

Dès le début des années 1960, H.P. Edmundson et les autres participants du projet TRW (Thompson RamoWoolridge Inc.) [Edmundson 1963] ont proposé un nouveau système de résumé automatique où celui-ci combinait plusieurs critères pour évaluer la pertinence des phrases à extraire. Il préfigurait alors de nombreux systèmes à venir qui reprendraient par la suite le même principe général.

LexRank est un autre système de résumé multi-document bien connu qui utilise un graph pour déterminer les phrases les plus importants. Tout d'abord, le corpus est représenté comme un graphe pondéré non orienté où les nœuds représentent des vecteurs de phrases de valeurs TF-IDF, et les liens sont étiquetés par les similarités cosinus entre eux. Seulement les liens avec des similarités supérieurs à un seuil prédéfini sont. Ensuite, l'algorithme PageRank est appliqué pour scorer les phrases. Enfin. Les phrases les plus scorés sont sélectionnées et une technique d'anti-redondance est utilisée pour construire un résumé informatif et avec une redondance minimale d'informations. Un algorithme très similaire au LexRank est Textrank mais pour un seul document à la fois et il génère un ensemble de motsclés ou d'expressions clés. TextRank utilise une représentation des unités textuelles (mots, phrases) comme LexRank mais les liens sont étiquetés par co-occurrentes ou par le chevauchement de mots respectivement

Synmémoire automatique de textes basée sur des techniques statistiques

Plusieurs études ont examiné l'application de techniques superficielles pour la notation de phrases pour la synmémoire de texte. Ces techniques sont assez simples à mettre en œuvre, nécessitent peu ou pas de ressources linguistiques et ont un faible temps de traitement informatique (FERREIRA et al. 2013) mis en œuvre quinze méthodes de notation élaborées dans la littérature au fil des ans. Afin de sélectionner les phrases les plus pertinentes, on utilise des méthodes de notation des mots, des phrases et des graphiques. Dans la notation des mots, les scores sont donnés les mots et les mots avec les scores les plus élevés sont les plus importants. Parmi les méthodes de notation des mots, mentionnons la fréquence des mots (LUHN, 1958), la TF/IDF et la similitude lexicale (MURDOCK, 2007).dans la notion des phrases, des notes sont attribuées à chaque phrase individuelle des documents. Certaines des méthodes de notion des phrases sont la position de la phrase (FATTAH ; REN, 2009 ; BARRERA ; VERMA ,2012) et la centralité de la phrase (FATTAH ; REN, 2009).dans la notion graphique, les score sont calculés en modélisant les liens entre les phares comme un bord entre les graphes des nœuds de phrase.

(MENDOZA et al. 2014) a proposé une méthode de synmémoire générique extractive pour des documents uniques en utilisant des opérateurs génériques et en guidant la recherche locale.Cette méthode utilise un algorithme mimétique qui a combiné la recherche basée sur la population d'évolution à une stratégie de recherche locale guidée.La tâche de synmémoire est traitée comme un problème d'optimisation binaire.Peu de fonctionnalités indépendantes du domaine et de la langue sont utilisées pour rechercher les phrases importantes à partir des documents tels que la position de la phrase, la ressemblance de la phrase avec le titre, la longueur de la phrase, la cohésion et la couverture

Synmémoire automatique du texte à partir de graphiques

(BARALIS et al. 2013) a proposé GRAPHSUM, un outil de synmémoire graphique à usage général pour la synmémoire de plusieurs documents.Cette approche explore et utilise une technique d'apprentissage automatique (règles d'association) pour découvrir des corrélations entre plusieurs termes et ne dépend pas des modèles avancés basés sur la sémantique.Après le prétraitement, la collecte des documents est organisée comme un ensemble de données transactionnelles afin que

l'exploration des règles d'association puisse être effectuée sur ces derniers.Ensuite, les ensembles d'éléments fréquemment présents qui présentent des corrélations élevées entre les termes sont extraits de l'ensemble de données transactionnelles et un graphique de corrélation est généré à partir de ces termes, ce qui aidera davantage à sélectionner des phrases importantes pour le résumé[11].Donc, l'identification de ces chaines lexicales est très importante pour déterminer les unités cohésives qui expriment le sujet global d'un texte. Ils ont aussi créé un cadre pour construire les chaines lexicales d'un texte qui se compose de 3 étapes :

1 Sélectionner un ensemble de mots candidats.

2. Pour chaque mot candidat, trouvez une chaîne appropriée pour lui parmi les membres des chaînes en fonction de son sémantique.

3. Si elle est trouvé, insérer le mot dans la chaîne et mettez, sinon créer un nouvelle chaine qui contient le mot correspondant.

Afin de créer des résumés, les chaines ont été scorées en fonction des fréquences des membres qu'elles les contiennent, et les phrases du résumé ont été sélectionné à partir de la plus longue chaine[12]

Ces synsets sont ensuite liés les uns aux autres via des relations différentes pour former des relations sémantiques entre des concepts. Une amélioration de la méthode de Morris et al [13]proposée par Barzilay et al pour tacler le problème du désambiguation des membres de chaines lexicales en créant différentes chaines lexicales pour les synsets de chaque entrée.

Conclusion

Dans ce chapitre nous avons présenté un état de l'art sur les différents approches proposées dans le domaine du résumé de texte biomédical, les premiers travaux ont été génériques, elles ont été ensuite substituées par des méthodes sémantiques, des méthodes très récentes et prometteuses utilisent les nouveaux techniques du plongement de mots. Dans le chapitre qui suit, nous allons présenter nos contributions.

Chapitre 03 : implémentation d’une nouvelle approche de résumé de texte biomédicale

Introduction

Après avoir vu les différents travaux sur le résumé automatique, en mettant l'accent sur les méthodes statistiques. Celles-ci sont des méthodes simples, faciles à manipuler, et rapides si on les compare aux méthodes linguistiques l'objectif majeur de nos propositions est d'améliorer ou de proposer des nouvelles techniques, nous proposons dans ce chapitre une méthode de résumé de texte biomédical qui améliore la qualité des résumés en utilisent les techniques de fouille de données, nous commençons par une description générale puis une définition formelle de notre approche proposée.

Méthode

Les systèmes de résumé automatiques, statistiques, linguistiques, ou hybrides ont tous besoins d'une phase de prétraitement pour rendre le texte d'entrée conforme à la phase de traitement, nous utilisent, ainsi, un module de post-traitement qui s'occupe, généralement, de la présentation du résumé au lecteur, En examinant les systèmes de résumé automatique, on peut dire que le module de traitement est celui qui définit la différence entre tous ces systèmes, Notre système ne fait pas l'exception par rapport aux autres systèmes de résumé automatique, il comporte aussi les trois modules classiques qui sont le prétraitement, le traitement, et le post-traitement

Prétraitement du document

Pour préparer le document d'entrée aux tâches suivantes, notre méthode nécessite plusieurs étapes de prétraitement Les points suivants illustrent les titres de ces techniques :

1. Suppression des sections non pertinentes : les sections non importantes pour inclusion dans le résumé sont supprimées : les intérêts concurrents, les remerciements, les références, les titres, images, figures, tableaux et titres
2. Diviser le texte en phrases : Dans cette étape, nous avons divisé le texte en un ensemble de phrases, le document est représenté comme un ensemble de phrases notées $D = S_1, S_2, \ldots, S_n$
3. Tokenisation des phrases : chaque phrase $s \in D$ est exprimée comme un ensemble de tokens, noté par $S = w_1, w_2, , w_k$

Extraction des mots clés (top mot)

1. choisir un document dans le domaine biomédicale

2. Supprime les sections non pertinentes
3. après le faire les deux étapes précèdent nous entrent le document dans un algorithme De notre création qui extrait les mots de ce document, il est séparé avec l'espace, après il est calculer le nombre d'occurrence de chaque mot dans le document c.-à-d. chaque mot dans le document il y'a un score appelé score-mot
4. le résultat de l'algorithme précèdent donne une liste des mots avec le score, nous avons copié la liste dans l'Excel pour filtrer
5. en utilise les trois méthodesde prétraitement : méthode de filtrage de lemmatisation et la méthode de racinisation,suppriméles motsvide :les mots qui n'a pas de sens dans le texte (prépositions, pronoms, etc.), et nous avons supprimé les mots répété, supprimé les mots de même sens et diffèrent dans l'écriture

Après ces étapes nous nous trouvons extrait les tops mots biomédicale dans un sujet spécifique

Exemple

mots	score
diagnose	12
mammograph	12
predicted	12
mammograms	13
recent	13
studies	13
image	14
feature	15
data	18
accuracy	19
dataset	27
mass	27
model	27
classifcation	30
use	32
breast	35
cancer	36
detection	39

Figure 2 Les mots de score supérieur à 11

Représentation du document

Le résultat de l'étape précédente est un ensemble de concepts, dont le document traité consiste en un ensemble de phrases $D = S_1, S_2, \ldots, S_n$, et chaque phrase contient un ensemble de concepts $S = C_1, C_2, \ldots, C_k$,nous utilisons la matrices pour représenter les phrase verticalementDS_i, et les

mots clés horizontalementDS_j, d'autre part une représentation vertical les poids des phrasePH_i.

Présentation des mots clés

Nous proposons une fonction qui calcule le poids de chaque phrase PH_i , avant le calcule il faut sauvegarder la liste des mots avec le score dans un fichier .txt

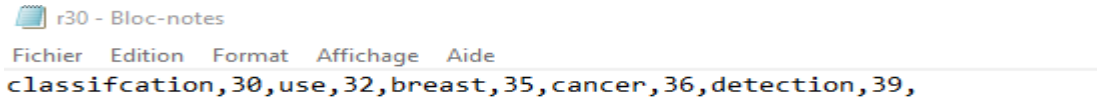

Figure 3Présentation de fichier txt des mots clés

Avec une fonction on a séparé les mots et le score de chaque mot

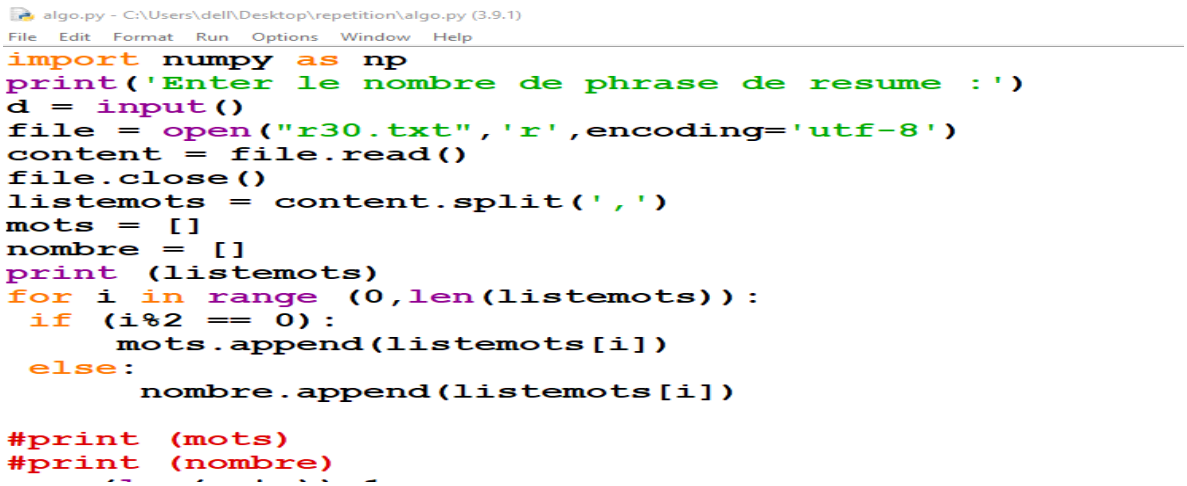

Figure 4 Fonction de séparation des mots et ces scores

Vérification l'existence des mots clés

avec une fonction on a vérifié l'existence des mots clés dans chaque phrase, s'il existe on a modifié dans la case avec DS_{ij}=1, s'il n'existe pas modifie dans la case avec DS_{ij}=0,

```
for a in range (0,m):
    for j in range (0,len(listemotcommuns)):
        if (mots[a] == listemotcommuns[j]):
            A[i,a]= 1
        if ((A[i,a]== 0)&(mots[a] !=  listemotcommuns[j])):
            A[i,a]= 0
listemotcommuns = []
```

Figure 5 Fonction de vérifier l'existence des mots

Calculer le poids des phrase

Pour calculer le poids de chaque phrase on a créé une fonction pour le calculer il faut tester si la case = 1 on a ajouté le score de ce mot a le poids de cette phrase, on à faire ça avec tous les mots jusqu'à terminer le teste de tous les mots clés, et faire le même travaille avec tous les phrase.

```
enfin = []
somme = 0

for i in range (0,n):
    for j in range (0,m):
        a = int (nombre[j])
        b =A[i,j]
        somme = (a * b)+ somme
    enfin.append(somme)
    somme = 0
#print (enfin)
```

Figure 6 Fonction de calculer le poids des phrases

A la fin, il produit pour nous un vecteur de poids de chaque phrase ce vecteur il triée par ordre descendent on donne le droit par l'utilisateur pour choisir le nombre de phrase du résumé a affiché les phrase affiché ordonné Selon la valeur la plus élevée à la valeur la plus faible

```
IDLE Shell 3.9.1
File  Edit  Shell  Debug  Options  Window  Help
Python 3.9.1 (tags/v3.9.1:1e5d33e, Dec  7 2020, 16:33:24) [MSC v.1928 32 bit (In
tel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
============== RESTART: C:\Users\dell\Desktop\repetition\algo.py ==============
Enter le nombre de phrase de resume :
3
 That is why CNN has become the most widely used method for image interpretation
 tasks in many domains such as the breast cancer detection and classification
 Then, the recent published work in the breast cancer detection and classificati
on from 2017 to 2019 are compared with respective to different aspects including
the used datasets, data preprocessing and data augmentation in section three fol
lowed by the main building blocks they build on the cancer classification and de
tection

This section lists the most common datasets containing mammograms used for the b
reast cancer detection and classification
>>>
```

Figure 7 L'affichage de résumé

Processus d'expérimentation

Le but de cette section est d'évaluer les performances de notre méthode, avec un programme de nos créations

On a obtenir un résumé d'un expert de même documentCe que nous étudions.

Nous avons entre le résumé d'expert et le résumé de l'algorithme précèdent

Dans l'algorithme qui calculer les déférentes métriques d'evaluation rappelle précision et f-mesure

```
#calcule rappel
print('le  rappel =  (nbr de mots communs /nbr des mots de resume de systemme)
rappel= nbrphrase/sumsys
print (rappel)
#calcule precision
print('le  precision =  (nbr de mots communs /nbr des mots de resume manuel)  ')
precision= nbrphrase/summan
print (precision)
#calcule F_mesure
print('le F_mesure = 2*[(precision *rappel)/(precision + rappel)] ')
f = rappel*precision
b = (rappel+precision)
print (2*(f/b))
```


Figure 8 Les métriques d'evaluation

Résultats et discussions

Résultats

Pour évaluer l'impact du nombre de phrase dans le résumé et le nombre des mots clés, sur les performances de notre synthétiseur, nous avons réalisé un ensemble d'expériences sur un petit corpus qui contient 20 pages pour observer l'effet dunombre des phrases sur les déférents ensembles des mots clés , nous étudions les metrique d'evaluation chacun a la fois

La table 1 représente les résultats rappel pour les différents nombre de phrase et les différents scores des mots clés

P : représente les paragraphes

R : représente les mots clés

		p8	p9	p10
Rappel	r1	0.470588	0.469649	0.492668622
	r20	0.508475	0.494253	0.482649842
	r30	0.512712	0.527559	0.512544803

Tableau 1 Résultats rappel

La table 2 représente les résultats précision pour les différents nombre de phrase et les différents scores des mots clés

		p8	p9	p10
précision	r1	0.931507	1	1
	r20	0.821918	0.883562	1
	r30	0.828767	0.917808	0.979452055

Tableau 2 Résultats précision

La table 3 représente les résultats f-mesure pour les différents nombre de phrase et les différents scores des mots clés

		p8	p9	p10
F-mesure	r1	0.625287	0.640523	0.660117878
	r20	0.628272	0.633907	0.65106383
	r30	0.633508	0.67	0.672941176

Tableau 3 Résultats f-mesure

Notrerésumé reste plus performant que le résumé d'expert en terme de rappel,Plus le score de motsclés est élevé il est Augmente la valeurde rappelle, même observation dans l'évolution de nombre de phrase du résumé, en terme de précision il le dépasse dans la plupart des cas

Le F-mesure il est combine entre la précision et le rappelle, 2*[(précision *rappel)/ (précision + rappel)] ont Remarqué Plus le score de mots clés et le nombre de phrases est élevé, plus la valeur de f-mesure il est augmenter

Discussion

Les figures 3.1, 3.2 et 3.4 montrent l'exécution de la méthode, lorsque nous varions le score des mots clés et le nombre de phrase du résumé en observant leurs effets sur la qualité des résumés générés en termes de rappel, de précision et de f-mesure.

Tout d'abord, nous observons selon les scores et nombre du phrase, que la valeur de score supérieure à 30 a un impact crucial sur la qualité des résumés générés par rapport à score inferieur a 30, Plus le nombre de

phrase est élevéPlus l'augmentation de rappel est élevée il est dépassé la valeur de 0.5 , au contraire de la valeur du précision , lorsque nous appliquons des score des mots clés très bas supérieure à 1 , nous obtenons une valeur très élevé , mais le changement du nombre des phrase joue le rôle majeure etPlus le nombre de phrase est élevéPlus l'augmentation de précision est élevée il est dépassé la valeur de 0.8

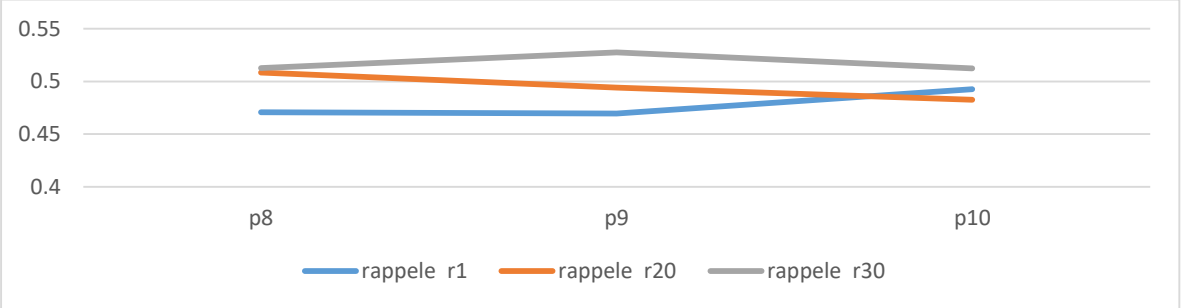

Figure 9 Rappel

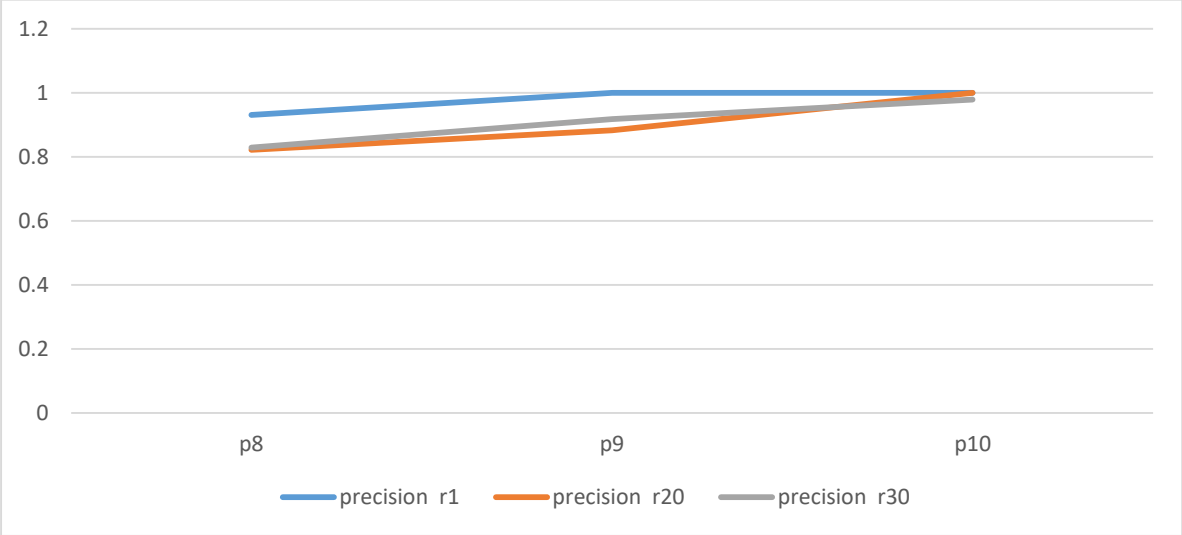

Figure 10 Précision

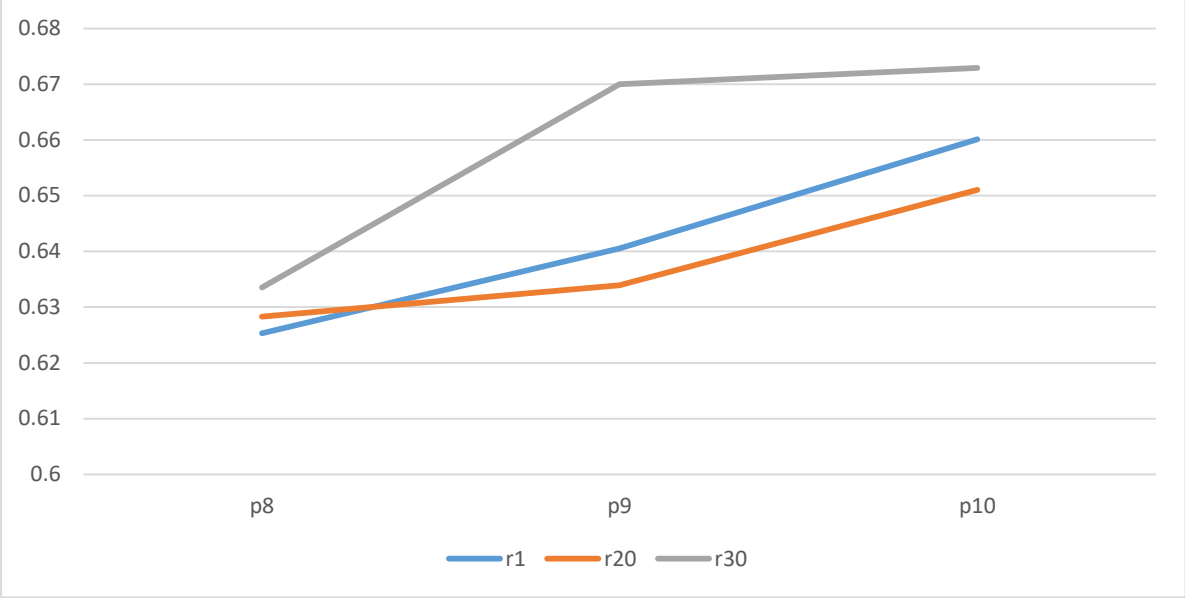

Figure 11 f-mesure

F-mesure il est combine avec les métriques précèdent on a vu que la valeur de score des mots clés plus élevé la valeur de f mesure il est augmenté même dans le nombre de phrase du résumé le plus élevé marquer des valeurs dépassé 0.6, f-mesure montrer une proportion direct entre le score des mots clés et le nombre de phrase du résumé

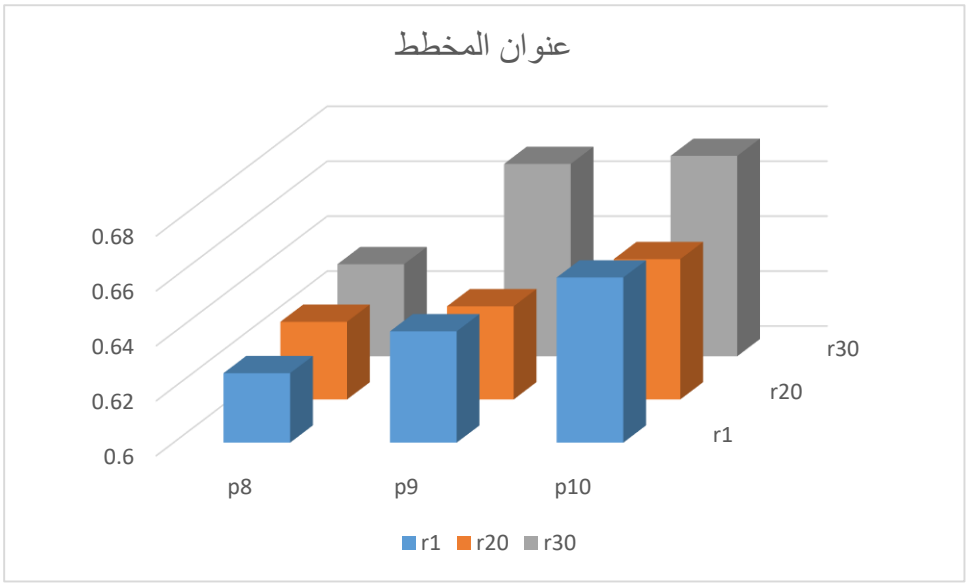

Figure 12 Métriques d'evaluation

Conclusion

dans ce chapitre nous avons présenté notre approche de résumé de texte biomédical mono-document, notre approche utilise une représentation conceptuelle de texte source, elle généré un model basé sur les concepts

biomédicaux , nous avons extrait les mots clés après on a représenté les phrase sous forme un ensemble des mots clés en a calculer le score si le mots il existe en a ajouter a le poids de chaque phrase, à la fin en a obtenir un vecteur de poids de chaque phrase, en afficher le résumer selon les poids des phrase a ordre décroissant .

Nous avons évalué notre travaille avec les métriques d’evaluation, nous avons prouvé que notre approche il est valide Il manque quelques ajustementspour bien travailler.

Conclusion

L’explosion des données textuelles biomédicales dans le web a donné naissance à des nouveaux problèmes de recherche, dans ce contexte nous nous intéressons à l’application de la technique de résumé textuel à la littérature biomédical afin de permettre un accès plus facile et fiable aux scientifiques et aux chercheurs dans ce domaine aux informations désirées en condensant le texte source et en préservant leurs idées principales.

Dans ce cadre, nous avons développé un système de résumé automatique en présentant le texte source sous forme de concepts biomédicaux, nous avons implémenté une nouvelle approche de résumé de texte biomédicale, Utilisant techniques de la fouille de données l’extraction des mots clés

[1] Ronen FELDMAN etIdo DAGAN. « Knowledge Discovery in Textual Databases (KDT) ». In : International Conference on Knowledge Discovery and Data Mining (KDD) June (1995), p. 112-117. ISSN : 1098-6596. arXiv : arXiv:1011.1669v3. URL :http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf.

[2]Extraction d'Information et modélisation de connaissances à partir de Notes de Communication Orale https://tel.archives-ouvertes.fr/tel-00109400/document

[3] Christian BLASCHKE, Lynette HIRSCHMAN et Alfonso VALENCIA. «
Information extraction in molecular biology. » In : Briefings in bioinformatics 3.2 (2002),

[4] MR. ABDELKRIM BOURAMOUL, Thèse pour obtenir le grade de Docteur en Sciences : RECHERCHE D'INFORMATION CONTEXTUELLE ET SEMANTIQUE SUR LE WEB, Année Universitaire : 2010/2011

[5]Ramzan TALIB et al. « Text Mining : Techniques , Applications and Issues ». In : 7.11 (2016), p. 414-418.

[6] Yves KODRATOFF. « Knowledge discovery in texts : A definition, and applications ». In :ficial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, 1999

[7] M. Abdenour MOKRANE, Thèse pour obtenir le grade de Docteur en Sciences:Représentation de collections de documents textuels : application à la caractérisation thématique, Le 17 Novembre 2006,

[8]LadislavGallayetMarián'imko. Utilizing vector models for automatic text lemmatization. In International Conference on Current Trends in Theory and Practice of Informatics, pages 532543. Springer, 2016. (Cité en page 10.)

[9] Carlos Gonzalez-Gallardo. AutomaticMultilingual Multimedia Summarization and Information Retrieval : Résumé automatique multimédia et multilingue et Recherche d'information. Avignon, 2019. (Cité en pages 12, 13 et 14.)

[10]stergos AFANTENOS, Vangelis KARKALETSIS et Panagiotis STAMATOPOULOS. « Summarization from medical documents : A survey ». In : Artificial Intelligence in Medicine 33.2 (2005),

[11]Rinaldo Lima, Towards Coherent Single-Document AutomaticText Summarization: An Integer LinearProgramming-based Approach, 2016, page 25-26

[13] J MORRIS et G HIRST. « Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text ». In : Computational Linguistics 17.1 (1991), p. 21-48