وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار– عنابـــة

Faculté des Sciences de L'Ingéniorat
Département d'Informatique

Année 2017-2018

THESE

Présentée en vue de l'obtention
du diplôme de Doctorat en Sciences

Combinaison d'Approches pour les Résumés Automatiques de Textes Arabes

Option
Systèmes d'informations et de connaissance

Par
Mme LAGRINI Samira

Membres de jury

Mr. Mohamed Tahar KIMOUR	Prof.	Université Badji Mokhtar – Annaba	**Président**
Mme Nabiha AZIZI	MCA.	Université Badji Mokhtar – Annaba	**Directrice de Thèse**
Mr. Mohammed REDJIMI	Prof.	Université 20 Août 1955-Skikda	**Co-directeur de Thèse**
Mr. Mohamed BENMOHAMMED	Prof.	Université Abdelhamid Mehri-Constantine2	**Examinateur**
Mr. Nadir FARAH	Prof.	Université Badji Mokhtar – Annaba	**Examinateur**
Mr. Mohammed Chawki BATOUCHE	Prof.	Université Abdelhamid Mehri-Constantine2	**Examinateur**

Remerciements

Je tiens à exprimer ma profonde reconnaissance et mes chaleureux remerciements à toutes les personnes qui m'ont de près ou de loin aidée dans ce long parcours parfois semé de doute et presque toujours d'incertitudes pour atteindre mon objectif et accomplir ce travail de recherche. J'espère que je n'oublierai personne, et je m'excuse par avance auprès de ceux que j'oublie.

Je tiens à remercier vivement ma directrice de thèse Madame Nabiha AZIZI Professeur à l'université de Badji mokhtar pour ses orientations, son soutien, sa générosité sans pareil et son amitié sincère dont elle a fait preuve tout au long de ce travail de recherche. J'apprécie la confiance qu'elle m'a témoignée et ses conseils avisés qui m'ont permis de découvrir les fabuleux plaisirs de la recherche. J'ai beaucoup appris à ses côtés et je lui adresse toute ma gratitude.

Je voudrais remercier beaucoup celui qui a inspiré et dirigé ce travail, Monsieur Mohammed REDJIMI Professeur à l'université de Skikda et mon co-directeur de thèse. Je lui adresse mes chaleureux remerciements pour ses conseils plus qu'avisés, sa sympathie et ses qualités scientifiques exceptionnelles mêlées d'une gentillesse extraordinaire. Je n'oublierai jamais ses encouragements, son écoute, sa disponibilité et son soutien permanent tant au niveau des connaissances qu'au niveau humain. Qu'il trouve ici l'expression de reconnaissance sincère.

Je ne sais honnêtement comment exprimer ma profonde gratitude à ces deux chères personnes…

Je remercie également les membres de jury pour l'intérêt qu'ils ont exprimé pour ce travail : Monsieur KIMOUR Med Tahar Professeur à l'université Badji Mokhtar -Annaba- pour l'honneur qu'il m'a fait, en acceptant de présider ce jury.

Je suis très reconnaissante envers Monsieur FARAH Nadir, Professeur à l'université d'Annaba et directeur de laboratoire LABGED pour avoir accepté d'évaluer mon travail,

Je suis très honorée et très reconnaissante de l'honneur que m'ont fait Monsieur BATOUCHE Mohammed Chawki et Monsieur BENMOHAMMED Mohamed *Professeurs à L'université de Constantine 2, d'examiner et évaluer mon travail.*

Merci à toute l'équipe du laboratoire LABGED (Gestion Electronique de Document), avec laquelle j'ai passé agréables moments.

Je tiens à remercier également tous mes amies et collègues, enseignants, employés et responsables du département d'informatique de l'université de Badji Mokhtar -Annaba.

Comment pourrais-je trouver les mots pour remercier toute ma famille, et surtout mes parents et mon époux Toufik Lounis qui m'ont beaucoup soutenue tout au long de ce travail et qui m'ont vraiment aidée à garder le cap dans des moments difficiles.
Mille fois Merci ….

Samira Lagrini.

Dédicaces

ملخـــص

نظرا إلى النمو الهائل في كمية المعلومات المتاحة إلكترونيا، أصبح الوصول إلى المعلومات المهمة في غضون فترة زمنية معقولة أمرا صعبا جدا إن لم يكن مستحيلا.

في ضل هذا الوضع، يبدو التلخيص الآلي للنصوص حلا جيدا يتوسط فرعين أساسيين: المعالجة الآلية للغات والبحث الآلي عن المعلومات. التلخيص الآلي للنص هو انتاج صيغة موجزة للنص مع الاحتفاظ بالمعلومات المهمة.

تقتصر معظم أنظمة التلخيص الآلي على تلخيص النصوص باللغات اللاتينية (الإنجليزية ,الفرنسية،.. إلخ)، في حين أصبحت الحاجة إلى تطوير أنظمة تلخيص مخصصة للنصوص العربية أكثر الحاحا في السنوات الأخيرة، نظرا لازدياد عدد الوثائق الإلكترونية باللغة العربية. وبالتالي، ففي إطار المعالجة الآلية للغات الطبيعية (TALN) وبالضبط التلخيص الآلي للنصوص العربية يندرج موضوع هذه الأطروحة.

هدفنا هو تحسين جودة الملخصات الآلية للنصوص العربية من خلال إقتراح منهج جديد يأخذ في عين الاعتبار العلاقات البلاغية التي تربط وحدات النص.

يجمع المنهج المقترح بين المعالجة اللغوية البحتة المعتمدة على تحليل الخطاب في اللغة العربية والمعالجة الإحصائية. يستند تحليل الخطاب على نظرية البنية البلاغية (RST).

على عكس المناهج التقليدية القائمة على تحليل الخطاب، فإن منهجنا يعتمد على استغلال العلاقات البلاغية بدلا من البنية البلاغية النص للإنتاج ملخص أولي للنص، والذي سيخضع بعدها الى معالجة احصائية لإنتاج الملخص النهائي.

حتى نتمكن من استخدام العلاقات البلاغية علينا تحديدها أولا وبشكل آلي. لهذ الغرض فقد قمنا بإنشاء نظام لتحديد العلاقات البلاغية العربية الصريحة والضمنية اعتمادا على تقنيات التلقين الآلي.

لتقييم فعالية المنهج المقترح قمنا بإجراء نوعين من التقييم: تقييم آلي باستخدام ROUGE وتقييم يدوي من طرف حكام مؤهلين.

وقد أثبت كلا النوعين من التقييم فعالية النهج المقترح في إنتاج ملخصات متناسقة تغطي أغلبية الجمل المهمة في النصوص الأم.

الكلمات المفتاحية: المعالجة الآلية للغات، التلخيص الآلي للنصوص، اللغة العربية، العلاقات البلاغية، نظرية البنية البلاغية (RST) ، تحليل الخطاب.

Résumé

Vue la croissance exponentielle de la quantité d'information disponible sous format électronique, l'accès a l'information pertinente dans un temps raisonnable est devenu très difficile voire impossible.

Le Résumé Automatique de Texte semble être une bonne solution qui se trouve à la croisée de deux disciplines : traitement automatique de la langue et recherche d'information. Le Résumé Automatique de Textes consiste à produire une représentation courte d'un texte tout en en conservant l'information pertinente.

De nos jours, la plupart des systèmes de résumé automatique traitent des textes en langues indo-européennes (l'anglais, le français, etc.). Le besoin de développer des systèmes de résumé automatique dédiés pour la langue arabe devient de plus en plus incontournable, ces dernières années, vu l'augmentation du nombre de documents électroniques rédigés en langue Arabe. Ainsi, c'est dans le cadre du Traitement Automatique du Langage Naturel (TALN) et plus précisément celui du résumé automatique de textes arabes que s'inscrit le sujet de cette thèse.

Nous nous sommes fixés comme objectif l'amélioration de la qualité des extraits automatiques de textes arabes par la proposition d'une nouvelle approche qui prend en compte les relations rhétoriques reliant les unités de texte.

L'approche proposée combine un traitement purement linguistique basé sur l'analyse de discours arabe avec un traitement statistique. L'analyse de discours est basée sur la théorie de la structure rhétorique (RST). À l'inverse des approches classiques basées sur l'analyse de discours, notre approche s'appuie sur l'exploitation des relations rhétoriques au lieu de la structure rhétorique de texte pour générer un résumé primaire qui va subir un traitement statistique afin de générer le résumé final de texte.

Afin de pouvoir exploiter les relations rhétoriques, une première étape consiste à en faire une identification de façon automatique. A cet effet, nous avons proposé une approche supervisée pour la classification automatique des relations rhétoriques Arabes explicites et implicites. L'implémentation de ce modèle a nécessité l'élaboration manuelle d'un corpus de discours arabe annoté selon le cadre de la théorie de la structure rhétorique, vue la non disponibilité d'une telle ressource en langue Arabe.

Afin de montrer la faisabilité de l'approche proposée, nous avons également effectué deux types d'évaluation : une évaluation automatique utilisant les mesures ROUGE et une évaluation manuelle établie par des juges qualifiés.

Les deux types d'évaluation ont prouvé l'efficacité de l'approche proposée et ont montré sa capacité à générer des résumés cohérents qui couvrent les phrases les plus pertinentes des textes source.

Mots-clés : Traitement automatique du langage naturel (TALN), résumé automatique de textes, la langue Arabe, les relations rhétoriques, la théorie de la structure rhétorique (RST), analyse de discours.

Abstract

Due to the exponential growth of amount of online information, retrieve the relevant information quickly has become very difficult if not impossible.

Automatic Text Summarization seems to be a good solution that lies between two disciplines: automatic language processing and information retrieval. Automatic text Summarization consists of producing a short representation of a text while preserving the relevant information.

Nowadays, most automatic summarization systems handle texts in Indo-European languages (English, French, etc.). The need to develop automatic summarization systems for Arabic language becomes more and more essential, in recent years, due to the increasing number of online documents written in Arabic. The work presented in this thesis concerns automatic Natural Language Processing (TALN) and more precisely automatic Arabic texts summarization.

Our aim is to improve the quality of automatic extracts of Arabic texts by proposing a new approach that takes into account rhetorical relations that link text segments in the text.

The proposed approach combines a discourse analysis based approach and a statistical based approach. The discourse analysis is based on the rhetorical structure theory (RST).

Unlike conventional RST based approaches, our approach rely on rhetorical relations instead of the rhetorical structure of the text to generate a primary summary. Then it will be pass by a second phase where a statistical method is applied to generate the final summary.

In order to be able to use rhetorical relations, a first step is to make an automatic identification is necessary. To this end, we proposed a supervised learning approach for automatic identification of rhetorical relations in Arabic. Since there is no RST annotated corpora in Arabic, we have manually annotated a large corpus following the rhetorical structure theory framework.

To evaluate the performance of our approach, we have performed two types of evaluations: an automatic evaluation using ROUGE measures and a manual evaluation by human judges.

Both evaluations have proved the efficiency of the proposed approach, and have shown that it can generate coherent summaries that cover the most relevant sentences of the source texts.

***Keywords**: Automatic natural language processing (ANLP), Automatic text summarization, Arabic language, Rhetorical relations, Rhetorical structure theory (RST), Discourse analysis.*

Table des Matières

Table des Figures

Liste des Tableaux

Table des Abbreviations

ANSI	American National Standards Institute
ANLP	Automatic Natural Language Processing ANLP
AutoSummENG	AUTOmatic SUMMary Evaluation based on N-gram Graphs
AWN	Arabic WordNet
CRF	Conditional Random Field
DUC	Document Understanding Conference
EASC	Essex Arabic Summaries Corpus
EDU	Elementary Discourse Unit
FFNN	Les Réseaux de Neurones Feed Forward
GA	Les Algorithmes Génétiques
GMM	Les Modèles de Mélanges Gaussiens
HMM	Hidden Markov Model
IDF	Inverse Document Frequency
ISF	Inverse Sentence Frequency
MeMoG	Merged Model Graph
MLP	Multilayer Perceptron
MMR	Maximal Marginal Relevance
NIST	National Institute Of Standards And Technology
NPowER	N-gram graph Powered Evaluation via Regression
PDTB	Penn Discourse Treebank
PNN	Probabilistic neural network
RNN	Recurrent neural network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
ROUGE-L	Longest Common Subsequence
ROUGE-N	N-gram Co-Occurrence Statistics
ROUGE-S	Skip-Bigram Co-Occurrence Statistics
ROUGE-W	Weighted Longest Common Subsequence
RST	Rhetorical Structure Theory
RST-DT	Rhetorical structure theory- Discourse Treebank
SDRT	Segmented Discourse Representation Theory

SVM	Support Vector Machine
TAC	Text Analysis Conference
TALN	Traitement Automatique Du Langage Naturel
TC	Taux de compression
TF	Term Frequency
TF*IDF	Term Frequency*Inverse Document Frequency
TF*ISF	Term Frequency*Inverse Sentence Frequency
TREC	Text Retrieval Conferences
VSM	Vector Space Model

Introduction Générale

Les développements des nouvelles technologies de l'information et de la communication, l'avènement de l'internet et la multiplication des moteurs de recherches mettent à disposition des quantités titanesques d'informations sous diverses formes (visuelles, sonores, multimédias, textuelles). La compression de ces informations ou leur résumé s'avèrent être des moyens nécessaires sinon indispensables pour n'en garder que l'information pertinente pouvant servir de lien pour la recherche de sa globalité en cas de besoin.

Les informations textuelles véhiculent une grande partie des connaissances actuelles sous forme de documents. Les résumés de ces documents ont constitués depuis le début de l'informatique des domaines de recherches et d'investigations investis de façon intensive par les chercheurs. Le résumé d'un texte consiste dans la production d'un second texte comprenant les idées essentielles et nettement plus bref et plus concis.

1. Contexte

Le résumé automatique de texte consiste à en produire une représentation abrégée qui en couvre l'essentiel du contenu à l'aide d'un système informatique. L'idée de concevoir des systèmes de résumés automatiques de textes n'est pas nouvelle, elle remonte aux années soixante [LUHN, 1958 ; BAXE, 1958 ; EDUM, 1969] dans le but de satisfaire les premiers besoin en termes de résumés automatique de documents. Mais ce besoin est devenu encore plus excessif avec l'apparition d'internet et l'accroissement exponentiel du nombre de documents textuels sous formats électroniques ce qui a provoqué une grande croissance et une surcharge colossale d'informations. Cette situation a initié un grand engouement dans ce domaine de recherches. Beaucoup de chercheurs s'étant investis totalement dans cet axe au sein de la communauté du traitement automatique du langage naturel (TALN).

L'enjeu de produire des résumés automatiquement reste important et s'étend même à de nouvelles applications telles que le résumé d'opinion [NISH, 2010 ; MENG, 2012 ; POTT, 2010], de nouvelles de presse [MCKE, 2003], le résumé de fils de discussion (*email threads*) [ULRI, 2009 ; RAMB, 2004 ; CARE, 2007], des emails [CORS, 2004], et des pages web [DIAO, 2006]. Le résumé de documents pour des domaines de spécialité apporte une aide

précieuse pour les experts et les utilisateurs concernés. Plusieurs catégories peuvent être considérées tels que les textes juridiques [FARZ, 2005], biomédicaux [REEV, 2007], …etc.

Les systèmes de résumés automatiques peuvent être classés selon plusieurs critères tels ceux concernant leur entrée (*mono-document ou multi-document),* le style de la sortie *(Indicatif, Informatif*), L'objectif (*Générique, Guidé, ou mis à jour*), le domaine (*dépendant ou indépendant de domaine*) et L'approche employée. Selon ce dernier critère on peut distinguer deux grandes approches dans le domaine du résumé automatique de textes : l'approche par abstraction et l'approche par extraction. Dans l'approche par abstraction, la construction de résumé passe par une un analyse en profondeur du texte, afin de sélectionner les phrases les plus pertinentes, les fusionner et les générer en utilisant des techniques de génération automatique de texte. Le résumé généré par une telle approche contient des paraphrases des phrases du texte source. Dans un cadre pratique, il est extrêmement difficile de construire des résumés automatique fondés sur une approche par abstraction, vue que cela nécessite beaucoup de ressources linguistiques très avancées pour la représentation des textes, la fusion des phrases, la génération automatique de texte…etc. En revanche, dans l'approche par extraction, un sous ensemble de phrases du document source est sélectionné et assemblé pour construire le résumé. Aucun processus de reformulation n'est accompli. L'idée principale sous-jacente à une méthodologie par extraction consiste à identifier les segments les plus pertinents du texte source et à les extraire pour former un résumé. Pour identifier les segments les plus importants, un algorithme de sélection fondé sur des connaissances statistiques ou linguistiques est utilisé. Ainsi, Ces approches de résumé par extraction sont généralement simples à implémenter et ne nécessitent que certains aspects linguistiques [TORR, 2011]. C'est pour cette raison que dans un cadre pratique les recherches se sont principalement concentrées sur le résumé automatique de texte par extraction (*extractive text summarization*) au lieu du résumé de texte par abstraction (*abstractive text summarization*). Le travail de recherche présenté dans cette thèse est axé sur le résumé automatique de textes par extraction et plus précisément le résumé mono-document.

Plusieurs approches ont été développées au cours du dernier demi-siècle afin de produire des extraits automatiques, c'est-à-dire des résumés de textes par extraction. Les premiers travaux se sont appuyés sur des approches statistiques pour extraire les phrases les plus pertinentes. Les approches statistiques prennent comme critère de pertinence une valeur numérique (un score) attribué à un segment textuel ou une phrase qui est calculé par une fonction de score portant sur une ou plusieurs caractéristiques. Ainsi les segments considérés les plus pertinents sont les segments ayant les scores les plus élevés. Ces segments sont alors sélectionnés et assemblés

dans l'ordre de leur apparition dans le document source pour construire le résumé. Plusieurs caractéristiques ont été utilisées telles que la fréquence de termes, la position de la phrase, similarité avec le titre, ...etc.

Avec l'augmentation du nombre de caractéristiques, les techniques de l'apprentissage supervisé ont commencé à être utilisées. Ces techniques sont utilisées pour identifier un ensemble de caractéristiques appropriées et leurs poids applicables. L'objectif des approches basées sur l'apprentissage supervisé est de construire un modèle qui sert à classer les phrases en phrase pertinentes pour le résumé ou non pertinentes. Le problème majeur avec ces approches porte sur leur dépendance aux corpus des phrases étiquetées manuellement. Afin de remédier à ce problème, plusieurs chercheurs se sont orientés vers les techniques de l'apprentissage non supervisé (*le clustering*) pour construire des résumés automatiques. L'idée sous-jacente aux approches basées sur le clustering consiste à regrouper les phrases très similaires au sein d'un même cluster puis de sélectionner dans chaque cluster les phrases ayant les scores les plus élevés. Ces approches ont réussi à éliminer la redondance dans les résumés générés, qui est l'un des problèmes majeurs de résumé automatique de texte.

D'autre approche à base de graphes ont été également proposées dans le cadre de résumé par extraction. Le principe de ces approches consiste à représenter le document sous forme d'un graphe, de telle façon que les nœuds du graphe correspondent aux phrases du document et les arcs correspondent aux similarités entre les phrases. Un algorithme de classement à base de graphe est utilisé pour pondérer et classer les phrases. Les phrases les mieux classées sont ensuite sélectionnées pour construire le résumé.

Toutes ces approches sont de nature numérique, En ce sens qu'elles s'appuient sur des valeurs calculatoires pour juger la pertinence des segments textuels.

Par ailleurs, On considère un autre groupe d'approches de résumé par extraction est de nature linguistique. Ces approches s'appuient sur des marqueurs linguistiques ou sur des critères de nature discursive pour évaluer la pertinence des phrases dans le texte sans faire appel à une quelconque forme d'évaluation quantitative. Plusieurs approches basées sur l'analyse de discours ont été proposées dans la littérature. Ces approches visent à exploiter la structure discursive du texte ainsi que les relations rhétoriques qui existent entre les unités textuelles pour générer des résumés automatiques [KUMA, 2016]. Parmi ces approches nous citons celles qui sont basées sur la théorie de la structure rhétorique (RST) [MANN, 1988] vu que cette théorie a influencé notre travail. Dans le cadre de la RST, un texte cohérent peut être représenté par

une structure hiérarchique de relations rhétoriques de telle façon que les éléments de cette structure soient les segments de texte reliés par des relations rhétoriques. A chaque élément dans cette structure est assigné un statut rhétorique (noyau ou satellite) qui reflète son degré d'importance. Un segment noyau est un segment primordial dans le texte tandis qu'un segment satellite est un segment de support. Les approches basées sur la RST s'appuient sur cette structure afin de sélectionner les segments les plus pertinents. Généralement un algorithme qui pondère chaque élément dans cette structure est appliqué ; les éléments ayant les poids les plus élevés sont sélectionnés pour construire le résumé. Toutes les approches basées sur la RST donnent plus d'importance aux éléments noyaux de la structure rhétorique de texte.

Généralement, les approches de résumé par extraction ont plus ou moins réussi à extraire les éléments pertinents du texte source. Néanmoins la qualité des résumés générés par ces approches reste à l'heure actuelle à améliorer.

2. Motivations et Objectifs

Les recherches dans le domaine du résumé automatique de texte sont très actives. Néanmoins la majorité de ces recherches concernent les langues indo-européennes (l'anglais, le français, etc.). Cependant pour la langue arabe les recherches dans ce domaine ont commencé à apparaitre seulement depuis les années deux milles [DOUZ, 2004] malgré la croissance colossale des documents arabes disponibles sous format électronique d'une part et bien que la langue Arabe soit la langue parlée par plus de 420 million de personne à travers le monde d'autre part.

Par conséquent, les recherches dans le domaine du résumé automatique de texte arabe n'ont pas encore connu de progrès notables. Nous pensons que cela est dû au manque de ressources linguistiques pour la langue arabe ainsi qu'aux caractéristiques linguistiques assez rigoureux de cette langue qui posent des défis majeurs aux chercheurs.

Dans la littérature de résumés automatique de textes arabes, la majorité des travaux rapportés s'appuie sur des techniques numériques pour extraire les éléments les plus pertinents du texte source et néglige complètement la cohérence globale du résumé généré. Cela nous a motivé à explorer ce domaine de recherche et de tenter d'améliorer la qualité des résumés automatiques de texte arabe par la proposition d'une nouvelle approche qui prend en compte les relations rhétorique qui relient les segments de texte.

3. Contributions

Afin d'améliorer la qualité des résumés automatiques de textes arabes, nous proposons une nouvelle approche qui combine une approche basés sur l'analyse de discours et plus précisément sur la théorie de la structure rhétorique avec un traitement statistique. A l'inverse des approches classiques basées sur la RST, notre approche s'appuie sur l'exploitation des relations rhétoriques reliant les unités de discours élémentaires au lieu de la structure rhétorique de texte pour générer un résumé primaire de texte. Ce dernier va ensuite subir un traitement statistique afin de générer le résumé final.

Pour pouvoir exploiter ces relations, il nous a fallu d'abord les identifier de manière automatique. Pour cela, nous avons proposé une approche supervisée pour l'identification automatique des relations rhétoriques arabes explicites et implicites. Vue la non disponibilité de corpus de discours arabe annoté selon le principe de la théorie de la structure rhétorique, il était nécessaire d'envisager la construction et l'annotation manuelle de ce type de corpus.

Ainsi nos principales contributions sont les suivantes :

1. Annotation manuelle d'un corpus arabe selon le principe de la théorie de la structure rhétorique.
2. Proposition d'une approche supervisée pour l'identification automatique des relations rhétoriques arabes.
3. Proposition d'une approche hybride pour le résumé automatique de textes arabes.

4. Organisation de la thèse

Le manuscrit de thèse est organisé en cinq chapitres et comporte une introduction et une conclusion générales.

Dans le premier chapitre, nous allons cerner l'objet de notre étude à savoir le résumé automatique de textes. Nous y présentons les caractéristiques et les types de résumé automatique ainsi que l'architecture globale de tel système. Nous y abordons également les deux grandes approches dans ce domaine ainsi que les méthodes d'évaluation utilisée. Nous achevons ce chapitre par un aperçu sur les applications de résumé automatique de textes.
Le deuxième chapitre est consacré à l'étude détaillée des approches de résumé par extraction. Nous y passons en revue les différents travaux qui ont été menés dans ce domaine. Par la suite,

nous présentons une synthèse des travaux menés dans notre domaine de recherche, à savoir le résumé automatique de textes Arabes. Nous achevons ce chapitre par un petit aperçu sur quelques ressources disponibles pour la langue Arabe.

Le chapitre trois est dédié à l'annotation rhétorique de notre corpus. Nous y présentons en détail les étapes de l'annotation rhétorique en commençant par l'élaboration de la liste des relations rhétoriques arabes, la segmentation des textes en unité de discours élémentaires, l'annotation des relations rhétoriques ainsi que la détermination des statuts rhétoriques des segments. Nous achevons ce chapitre par la présentation des caractéristiques statistiques de notre corpus annoté.

Dans Le chapitre 4 nous présentons notre modèle pour l'identification automatique des relations rhétoriques arabes. Nous y présentons les caractéristiques que nous avons utilisées, les détails des expérimentations ainsi que les résultats d'évaluation obtenus.

Le chapitre 5 est consacré à la présentation de l'approche proposée pour le résumé automatique de textes arabes. Nous y abordons les différentes étapes par lesquelles passe la génération du résumé. Nous allons également présenter les deux méthodes d'évaluation que nous avons effectuée pour montrer la faisabilité de cette approche ainsi que les résultats obtenus. Enfin nous concluons cette thèse en réaffirmant l'intérêt de notre approche et en résumant les principaux points de notre travail. Nous présentons alors les perspectives de recherche de nos travaux futurs.

Chapitre 1

Résumé Automatique de Textes

Sommaire

1. Introduction

Le résumé automatique de documents consiste à produire une représentation abrégée qui en couvre l'essentiel du contenu à l'aide d'un système informatique. Cela réduit, ainsi, le temps de recherche pour y trouver les informations pertinentes.

L'apparition d'Internet et l'accroissance exponentielle de documents textuels sous formats électroniques a provoqué une grande excroissance et surcharge d'informations. Les lecteurs n'ont ni assez de temps ni assez d'énergie pour lire de longs documents textuels où une version plus courte suffirait. Tous les utilisateurs d'ordinateurs, que ce soit des professionnels ou des utilisateurs novices, sont touchés par cette problématique.

Cette situation a dynamisé encore plus fortement les recherches dans le domaine de résumé automatique de textes au sein de la communauté du traitement automatique du langage naturel (TALN). Au cours du dernier demi-siècle, le résumé automatique de textes a été abordé à partir de nombreuses perspectives, dans des domaines différents et en utilisant des paradigmes différents.

Dans ce chapitre, nous allons cerner l'objet de notre étude à savoir le résumé automatique de documents textuels. Nous débuterons d'abord par des définitions formelles de quelques notions basics. Puis une taxonomie des systèmes de résumé automatique selon plusieurs critères sera présentée dans la section 3. Les différents modules qui composent un système de résumé automatique ainsi que les approches existantes pour élaborer un tel système seront décrites dans les sections 4 et 5 respectivement. Nous étudierons ensuite les méthodes d'évaluation de ces systèmes et quelques applications de résumé automatique de textes.

2. Qu'est-ce qu'un résumé automatique ?

Plusieurs définitions concernant la notion de résumé automatique de textes peuvent être trouvées dans la littérature. D'après Radev et al. [RADE, 2002]:

"*A summary is a text that is produced from one or more texts, that conveys important information in the original texts and that is no longer than half of the original text(s) and usually significantly less than that"*.

D'après Saggion et Lapalme [SAGG, 2002]:

« Un résumé est une version condensée d'un document source qui possède un genre et un propos spécifique pour donner au lecteur une idée exacte et concis de contenu de la source»

D'après Torres-Moreno [TORR, 2011]
« Un résumé est un texte généré par un logiciel, cohérent et contenant une partie important des informations pertinentes de la source et dont le taux de compression est inférieur à un tiers de la taille de documents originaux »

D'après ces trois définitions, on peut tirer trois aspects importants qui caractérisent le résumé automatique:

1) Le résumé devrait contenir l'information pertinente.
2) Le résumé devrait être cohérent.
3) Le résumé devrait être court.

Le taux de compression ou ratio de compression est le rapport entre la taille du résumé et celle de document sources (équation 1). Radev et al. [RADE, 2002] ont précisé que la taille du résumé doit être inférieure à la moitié de la source, tandis que Torres-Moreno [TORR, 2011] soutient que la taille du résumé doit être inférieure à un tiers de la taille de la source. De son côté, l'ANSI[1] (*American National standards institute*) recommande des résumés d'une longueur de 250 mots.

$$Taux\ de\ compression = \frac{nombre\ de\ phrase\ de\ résumé}{nombre\ de\ phrases\ de\ document\ source} \quad (1)$$

Différents types de résumés ont vu le jour. Il y a deux raisons principales qui l'expliquent : la première est que l'on dispose de plusieurs types de documents et la deuxième est que les [2]personnes ne cherchent pas le même type de résumé car leurs besoins en synthèse d'information diffèrent.

3. Classification des Systèmes de Résumé Automatique

Différentes taxonomies sont proposées par les auteurs pour la classification des systèmes de résumé automatique. Chaque auteur propose sa propre classification selon ses propres critères. Selon Jones [JONE, 1999], les systèmes de résumé automatique sont classés selon trois critères: source, objectif, et sortie. Tandis que Mani et Maybury [MANI, 1999a] ont proposé une classification en se basant sur le niveau de traitement. Ainsi, compte tenu de toutes

[1]http:// www.ansi.org

[2]

les taxonomies proposées, les systèmes de résumé automatique peuvent être classés selon les critères suivants (voir Figure 1-1) :

- **L'entrée (*input*)** : l'entrée d'un système de résumé automatique peut être un seul document ou un ensemble de documents. Ainsi, un système de résumé automatique peut être **mono-document ou multi-document.** Dans les systèmes multi-document, Le résumé doit contenir les informations les plus pertinentes contenus dans tous les documents, notons que si ces documents sont liés par un sujet commun, il risque d'y avoir de nombreuses informations communes [BLAI, 2008], de ce fait le résumé construit doit alors éviter toutes forme de redondance d'information.
- **Le style de la sortie (*style of output*)**: selon le style de la sortie, on distingue deux classes de système: indicatif et informatif.
  - **Indicatif** : fait intervenir la notion de thématisassions, c'est-à-dire qu'il ne présente au lecteur que les thèmes développés dans le document source [MANI, 2001]. Le résumé indicatif tend à fournir au lecteur les principaux sujets abordés pour qu'il puisse juger s'il doit consulter ou non le texte source.
  - **Informatif** : fournit une information générale sur tous les points essentiels du document source en cherchant à couvrir tous les sujets traités par le document. Il est plus difficile à produire puisqu'il nécessite un processus complexe de compréhension/généralisation de l'information. Les résumes informatifs sont souvent utilisés comme une vue d'ensemble de documents source. Ils sont appropriés pour répondre aux besoins d'un utilisateur qui veut obtenir une vision globale et générale du contenu d'un document. SUMUM [SAGG, 2002a] est l'un des systèmes de résumé automatique qui fournit en sortie des résumés indicatifs et informatifs.
- **L'approche employée**: Les méthodes de production de résumés automatiques de texte peuvent être regroupées en deux familles : extractives et abstractives. On distingue, ainsi, des systèmes de résumé par extraction et des systèmes par abstraction.
  - **Systèmes de résumé par abstraction** ***(Abstarctive text summarization systems)****: dans* ces systèmes, le résumé produit est comparable à ce que fait un être humain, qui lit un texte, le résume et le reformule avec ses propres mots. Un abstrait contient habituellement des paraphrases des phrases du texte d'entrée, et permet un haut niveau de condensation.
  - **Système de résumé par extraction** ***(Extractive text summarization systems)****:* produit des résumés composés des segments extraits du texte source. Aucun processus de génération n'est accompli. L'idée principale sous-jacente à une méthodologie par extraction consiste

à identifier les parties les plus importantes du texte source et à les extraire pour former un *résumé*. Ces méthodes sont fondées sur l'hypothèse qu'il existe, dans tout texte, des unités textuelles saillantes. Elles emploient un algorithme de sélection fondé sur des connaissances statistiques ou linguistiques pour sélectionner une liste d'unités textuelles qui constitueront le résumé (*extrait*).

Par abus de langage, le mot résumé en Français ne fait pas la distinction entre un extrait et un abstrait, et génère, alors, une certaine ambigüité. Par contre en Anglais, il y a peu d'ambigüité entre les deux termes '*summary*' et '*abstract'*. Ainsi les deux termes '*summarization'* et '*abstraction'* désignent respectivement la génération automatique d'un *extrait* et d'un *abstrait.*

- **L'objectif (*purpose)***: selon l'objectif voulu, un système de résumé automatique peut être: générique, Guidé, ou mis à jour.

- ***Générique*** : consiste à produire des résumés tout en préservant les thématiques principales et sans considérer les besoins d'information du lecteur. Cette tâche, qui paraît être simple, pose plusieurs difficultés. Parmi lesquelles, le type de document que l'on veut résumer. en effet, il est plus ou moins facile de résumer des articles de presse et quasiment impossible de résumer des œuvres littéraires [MIHA, 2007]
- **Guidé** : consiste en la production d'un résumé qui satisfait les besoins d'information d'un utilisateur. Ces besoins sont généralement exprimés au moyen d'une requête et doivent permettre au système d'isoler les parties de document concernant une thématique bien précise [BOUD, 2008]. L'objectif est de produire un résumé de document incluant uniquement les passages en rapport direct avec la thématique demandée.
- ***Mis à jour (update summary)*** : dans ce type de système, on suppose que le lecteur a déjà lu les documents et leurs résumés sur un sujet bien spécifié. Le résumé mis à jour se contente donc de montrer seulement les nouveautés importantes, tout en évitant la redondance d'information avec les documents déjà lus par l'utilisateur. Les résumés mis à jour ont été introduits lors de la campagne d'évaluation DUC (*Document Understanding Conference*) en 2007 (cette campagne sera décrite dans la section 5.2), afin d'améliorer la qualité du résumé lorsque l'on dispose de plus d'informations à propos des connaissances et des attentes de l'utilisateur [BOUD, 2008].

- **Domaine:** selon le domaine des documents source, les systèmes de résumé peuvent être classés en deux catégories: dépendant de domaine (*domain sensitive text summarization*) et indépendant de domaine (*general purpose*). Les système dépendants de domaine sont des systèmes qui ne peuvent résumer que les document appartenant à un domaine de spécialité (scientifique, technique, médical, juridique..etc.).

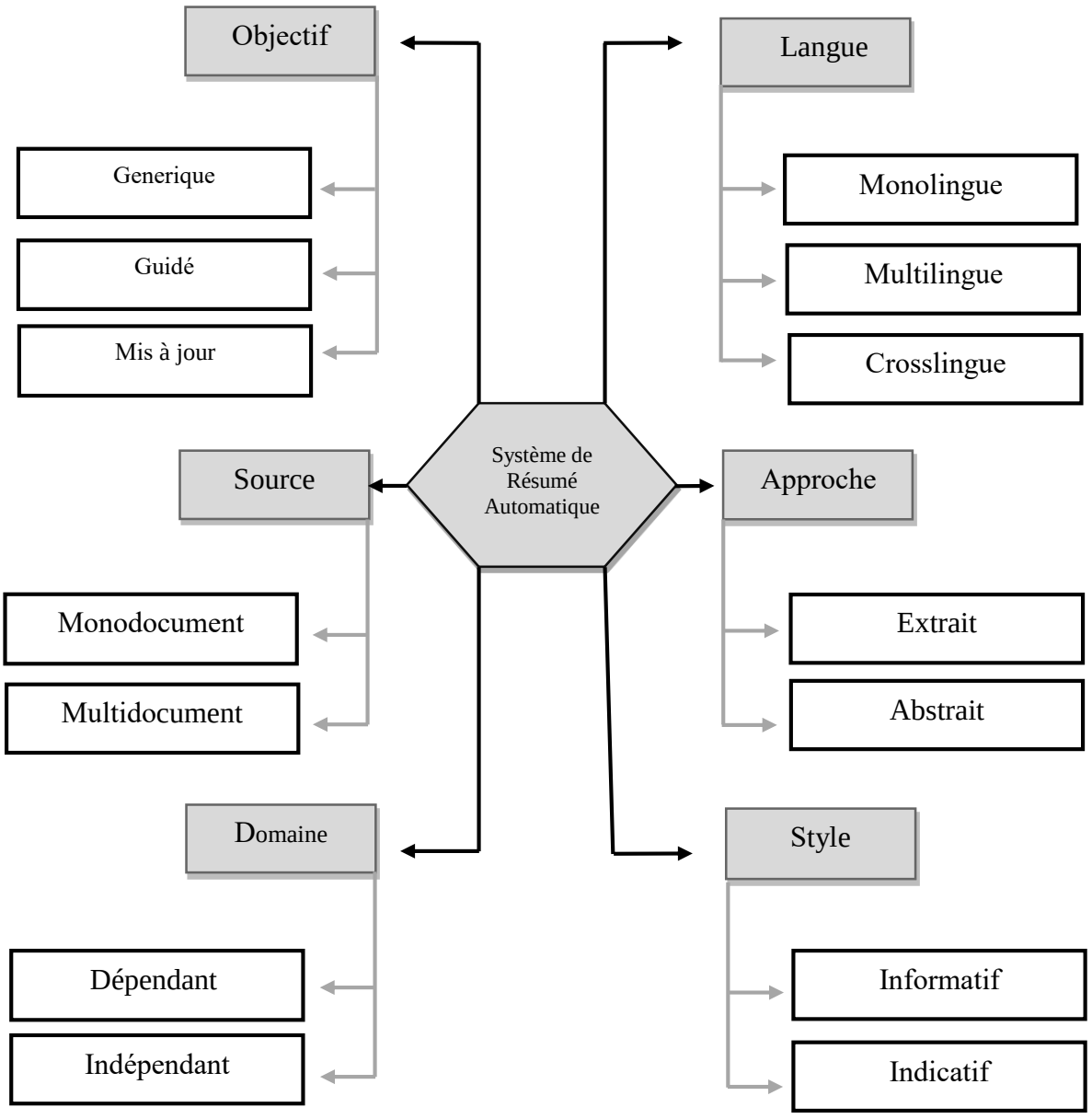

Figure 1-1 : Taxonomie des Systèmes de Résumé Automatique.

- **La langue:** selon la langue du texte source et celle du résumé produit, les systèmes de résumé automatique peuvent être classés en trois catégories:

- **Monolingue:** ce type de système est conçu pour fonctionner avec une seule langue, et génère des résumés dans la langue de du texte source.
- **Multilingue:** le système peut traiter plusieurs langues, mais la langue du résumé produit est toujours la même que celle de documents source. SUMMARIST [HOVY, 1998] et COLUMBIA NEWSBLASTER [EVAN, 2004] sont des exemples de systèmes de résumé multi-document multilingue.
- **Cross-lingue:** avec le développement des techniques de traduction automatique, il est possible de créer des systèmes qui produisent des résumés dans une langue différente de celle du document source. Un exemple d'implémentation de ce type de système est présenté dans [SAGG, 2002b]

4. Architecture Générale d'un système de Résumé Automatique

D'après Bawakid [BAWA, 2011], la génération automatique d'un résumé passe par trois étapes : le prétraitement, l'analyse et la génération de résumé, comme illustré dans la figure 1-2. Ces étapes sont partagées par tous les systèmes de résumé disponibles. Les détails de l'implémentation de ces étapes sont ce qui rend un système différent de l'autre.

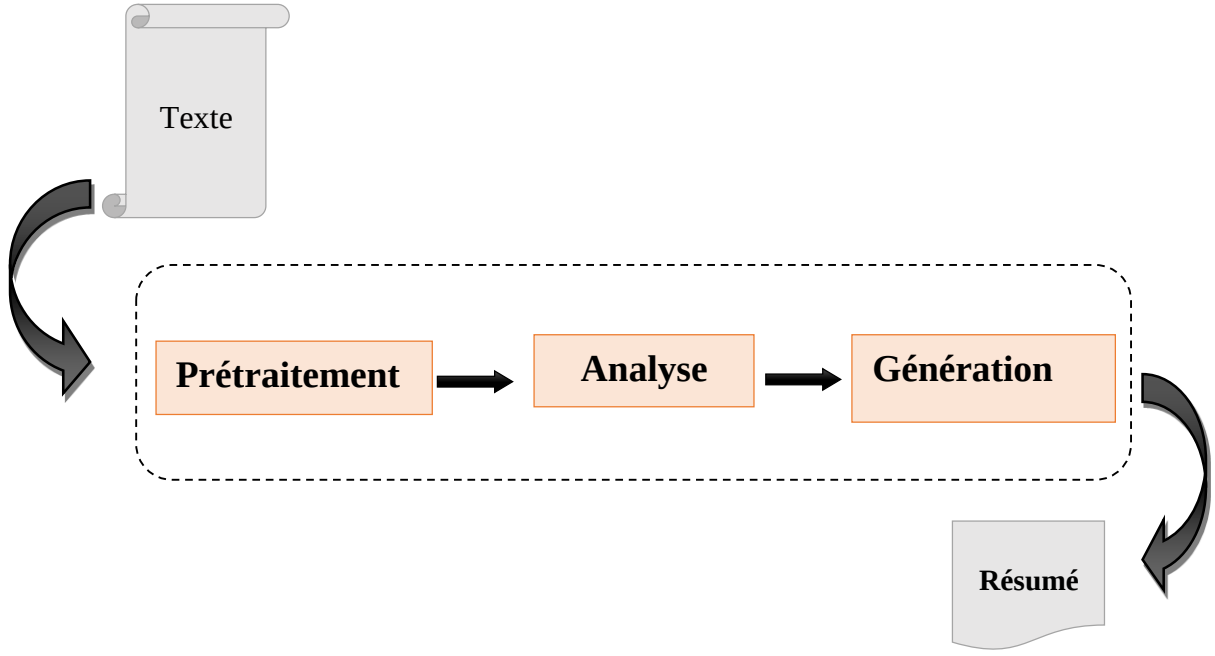

Figure 1-2 : Architecture General d'un Système de Résume Automatique.

4.1 Le prétraitement

La première étape dans tout système de résumé automatique concerne le prétraitement des données d'entrée qui sont : le document à résumer, la requête utilisateur (dans les systèmes guidés) et éventuellement le taux de compression. Le système analyse d'abord ces données, puis il prépare et converti le document en un format acceptable par le module d'analyse. Les symboles inutiles non traitables par le système vont être supprimés à ce niveau. le document est segmenté en phrase en se basant sur quelque marques de ponctuation comme « . » et « ? ». Des informations supplémentaires peuvent être étiquetées à chaque phrase comme sa position dans le document et/ou dans le paragraphe. Dans des systèmes qui emploient des techniques numériques, d'autres traitements peuvent être employés comme le *stemming* qui consiste à convertir chaque mot à sa forme racine ainsi que la suppression des mots vides. Tandis que dans les systèmes qui génèrent des abstraits, cette étape consiste à déterminer une représentation indiquant les sujets abordés et comment ils changent dans le texte [LIOR, 2012].

4.2. L'analyse

A ce niveau, le système doit évaluer et sélectionner l'information pertinente du document Généralement, un score est attribué à chaque phrase du document en fonction de son importance. Ce score est calculé en fonction de certains critères spécifiques comme la position de la phrase, la fréquence des termes..etc. Les phrases ayant les scores les plus élevés sont considérées les plus importantes. Dans les systèmes qui génèrent des abstraits, les techniques de simplification et de compression de phrases sont aussi employées à ce niveau [BAWA, 2011].

4.3. La génération du résumé

La génération automatique de résumé consiste à fusionner et reformuler les phrases/clauses précédemment identifiées. Comme cette étape n'est pas facile à aborder, les approches de résumé par extraction ne se concentrent que sur les deux premières étapes, en extrayant simplement les phrases considérées importantes (les mieux classées) telles qu'elles apparaissent dans les documents [LIOR, 2012], puis les assembler pour produire le résumé. Le nombre de phrases extraites dépend du taux de compression spécifié par l'utilisateur.

5. Description de deux Approches: Extraction vs. Abstraction

Comme cela a déjà été mentionné dans la section 3, les approches de génération automatique de résumé peuvent être regroupées en deux catégories: l'extraction et l'abstraction.

Les approches par *abstraction* sont fondées sur une compréhension profonde du texte et cherchent à produire des résumés de qualité en utilisant des sources de connaissance. Très peu de système ont été créés sous cette optique vu la complexité de la mise en œuvre de ces approches [LIOR, 2012]. En revanche, les approches par *extraction* se contentent de sélectionner les phrases qui semblent importantes en se basant sur plusieurs critères puis de les assembler pour produire le résumé. Ces approches sont généralement simples à implémenter et ne nécessitent que certains aspects linguistiques [TORR, 2011].

Dans ce qui suit, une description globale de ces deux courants de recherche sera présentée. Une étude plus détaillée focalisée sur les approches par extraction fera l'objet du prochain chapitre.

5.1 Approches de résumé par extraction

Les approches de résumé automatique par extraction peuvent être catégorisées en deux groupes: les approches numériques et les approches linguistiques. Nous englobons dans le champ des approches numériques tout ce qui renvoie à des techniques calculatoires sur des valeurs numériques, à savoir les approches statistiques, les approches à base de graphe et les approches à base de l'apprentissage automatique.

Les approches statistiques prennent comme critère de pertinence une valeur numérique attribuée à un segment textuel, calculée par une fonction de score portant sur un ou plusieurs critères [BLAI, 2008], tels que la position de la phrase, les mots-clés, les expressions indicatives…etc. Un segment textuel est alors extrait si son poids est suffisamment élevé par rapport à un seuil ou par rapport au poids des autres segments.

Plusieurs travaux dans la littérature ont adopté des approches statistiques pour produire des résumés automatiques, toutefois nous n'en citerons dans ce chapitre que deux pour illustration : les travaux de Luhn [LUHN, 1958], et ceux de Edmundson [EDMU, 1969] qui sont considérés les premiers travaux portant sur la production de résumé automatique de textes. Luhn a décrit une technique spécifique aux articles scientifiques qui utilise les fréquences des mots dans le document pour pondérer les phrases. Un peu plus tard Edmundson [EDMU, 1969] a étendu ces travaux en tenant compte de la présence des mots provenant de la structure du document (i.e. titres, sous-titres, etc.) et des expressions indices ainsi que la position des phrases. Il est à noter

que l'approche proposée par Edmundson [EDMU, 1969] peut être employée uniquement dans les documents ayant une structure fixe, comme le titre, la section, le paragraphe..etc.
Par ailleurs, d'autres systèmes de résumé automatique sont basés sur les techniques de l'apprentissage automatique. A titre d'exemple Kupiec et al. [KUPI, 1995] ont développé un système de résumé automatique basé sur le classifier bayésien naïf. Lin [LIN, 1999] a utilisé les arbres de décision pour modéliser la problématique d'extraction de phrases. Son système s'est avéré être globalement plus performant.

Le deuxième groupe des approches par extraction regroupe celles de nature linguistique qui s'appuient, pour évaluer la pertinence des segments textuels, sur la présence de marqueurs linguistiques et particulièrement sur certaines de leurs propriétés sémantiques ou discursives, sans essayer de faire appel à une quelconque forme d'évaluation quantitative de la pertinence. Ce type d'approches admet souvent l'hypothèse que certaines marques de surface dans un contexte textuel bien précis permettent d'affecter une valeur sémantique ou rhétorique à la phrase ou la proposition qui les contient, et ainsi de connaître sa pertinence dans le texte afin de l'extraire pour construire le résumé. Certaines de ces approches sont fondées sur la théorie de la structure rhétorique [MANN, 1988] qui vise à exploiter la structure discursive du document pour en produire le résumé.

Il est à noter que la majorité des systèmes qui exploitent ces approches linguistiques, utilise conjointement des techniques statistiques formant ce qu'on appelle des approches hybrides.
L'avantage principal de l'approche par extraction est de ne pas passer par une analyse en profondeur du texte [BLAI, 2008], ce qui permet de produire un résumé de façon plus simple sans également devoir générer du texte automatiquement. Mais l'inconvénient de cette approche porte sur les mauvaises liaisons entre les segments extraits, et ainsi le manque de la cohérence du résumé produit. Néanmoins, cette approche reste actuellement la plus adéquate dans un cadre pratique et applicatif [LIOR, 2012].

5.2 Approches de résumé par abstraction

Les approches de résumé par abstraction considèrent la tâche de résumé automatique comme devant être calquée sur l'activité humaine de production de résumés. La génération d'un abstrait doit ainsi passer par la compréhension du texte pour pouvoir sélectionner les informations pertinentes et générer de nouvelles phrases pour le résumé, en utilisant les ingrédients lexicaux, syntaxiques, sémantiques et rhétoriques du texte original.

Pour le courant de recherche qui a développé ce type de système, le souci central consiste à caractériser les modèles de compréhension et de représentation de connaissances, ainsi que des techniques de génération de textes [BLAI, 2008].

L'abstraction diffère principalement de l'extraction par la génération des résumés ayant un certain degré d'inférence sur des connaissances pas nécessairement présentes dans le document source [BAWA, 2011], ainsi, les abstraits produit sont basés sur des reformulations ou fusion des phrases et souvent en utilisant de nouveaux vocabulaires.

Le système de résumé par abstraction doit construire une représentation du texte afin de pouvoir générer automatiquement un abstrait à partir de celui-ci. Les méthodes de compression automatique des phrases [ZAJI, 2007], la fusion des phrases [BARZ, 2005] ainsi que la génération de langages naturels [RADE, 1998 ; YU, 2007 ; BELZ, 2008] ont été employés pour la génération automatique des abstraits. Pour un approfondissement des techniques de résumé par abstraction nous renvoyons le lecteur à [KHAN, 2014].

Produire un abstrait reste à l'heure actuelle une tâche complexe et ardue. A notre connaissance, il n'existe pas encore d'outils conceptuels et techniques pour pouvoir automatiser la compréhension totale d'un texte. Toutefois quelques systèmes de résumés par abstraction ont vu le jour, nous citons parmi lesquels : La plate-forme FRUMP [DEJO, 1982], conçue pour produire des abstraits des articles de presse. Ainsi, que le système SUSY [FUM, 1982] réalisé pour résumer des textes scientifiques. Ce système propose une séquence de traitements qui passe par la construction de plusieurs représentations de texte jusqu'à la génération finale de l'abstrait. SUSY est particulièrement intéressant. Toutefois, Son ambition fut plus grande que la capacité réelle des moyens conceptuels et techniques employés à l'époque [BLAI, 2008]. Enfin, OPINOSIS [GANE, 2010] est un système à base des graphes conçu pour produire automatiquement des abstraits d'opinions ayant un haut degré de redondance.

6. Evaluation du résumé automatique

Évaluer la qualité d'un résumé est un problème difficile auquel il n'existe que des solutions partielles [LIOR, 2012]. Plusieurs facteurs sont derrière cette problématique. En effet, il n'existe pas un résumé « idéal ou parfait ». Plusieurs résumés peuvent convenir au même document et ils ne sont pas forcément similaires ou convergeant au niveau du contenu. Par conséquent, deux résumés pour le même document peuvent être produits en utilisant un vocabulaire totalement différent. Voire même, la même personne peut résumer le même document de manière différente au fil du temps [LUHN, 1958]. De plus, l'évaluation du résumé

nécessite une comparaison avec des résumés de référence [LIOR, 2012]. Cela implique l'existence de corpus de référence contenant des documents à résumer et leurs résumés de référence. La création de tels corpus de référence est une tâche coûteuse et fastidieuse [EL HA, 2015].

Un autre facteur derrière la complexité de cette tâche est l'évaluation elle-même. L'évaluation manuelle est couteuse en temps et subjective [IMAM, 2013]. Pour remédier à ce problème, des méthodes automatiques et semi-automatiques ont été introduites. Mais le problème avec ces méthodes est qu'elles évaluent le résumé en termes de son contenu informationnel et ne prennent pas en compte la qualité du résumé et sa cohérence.

Dans le reste de cette section, nous allons décrire ces deux courants d'évaluation brièvement, et ainsi passer en revue certaines des méthodes d'évaluation automatique des résumés textuels.

6.1 Méthodes d'évaluation

D'après Jones [JONE, 1995] les méthodes d'évaluation des résumés peuvent être classées en deux catégories : intrinsèques et extrinsèques.

6.1.1. Évaluations extrinsèques

Dans une évaluation extrinsèque, les résumés sont évalués en fonction de leur aptitude à effectuer certaines tâches comme par exemple la classification de documents, la recherche d'information ou l'utilisation des résumés à la place des documents sources dans des systèmes question/réponse. L'objectif est de voir comment ces tâches seront effectuées suivant la qualité du résumé. Par exemple, dans les systèmes question /réponse, Un résumé est considéré de bonne qualité s'il permet à son lecteur de répondre au questionnaire aussi bien que d'autres lecteurs qui ont lu le texte source. Il est à noter que dans ce type d'évaluation, il faut choisir des tâches dont la bonne fonction dépend de la bonne qualité de résumé [BOUD, 2008].

Dans les campagnes d'évaluation de résumé automatique TIPSTER SUMMAC [MANI ,1999]. Deux tâches d'évaluation extrinsèque ont été définies : la tâche *ad hoc* et celle de catégorisation. Dans la tâche de catégorisation, l'évaluation cherche à savoir si un résumé générique peut effectivement présenter suffisamment d'informations pour permettre à un analyste de classer rapidement et correctement un document. Ces méthodes apparaissent dans certains cas plus adéquats pour évaluer les résumés [MANI, 1999].

6.1.2. Évaluations intrinsèques

Ce type évaluation est basé directement sur une analyse de résumés [JONE, 1995]. Les résumés automatiques sont évalués directement en se basant sur leurs propriétés et leurs contenus. Ces méthodes peuvent être divisées d'après El Haj et al. [EL HA, 2011] en deux classes : contenu et forme (voir figure 1-3)

L'évaluation de la forme de résumé peut être réalisée manuellement par un certain nombre de juges qui mettent en valeur la lisibilité, la grammaire et la cohérence des résumés. L'évaluation de contenu mesure la capacité du résumé à identifier les sujets importants du document source. Cela peut être fait automatiquement via une comparaison avec des résumés de références produits par des experts ou par les auteurs de l'article ou encore des phrases extraites par des experts.

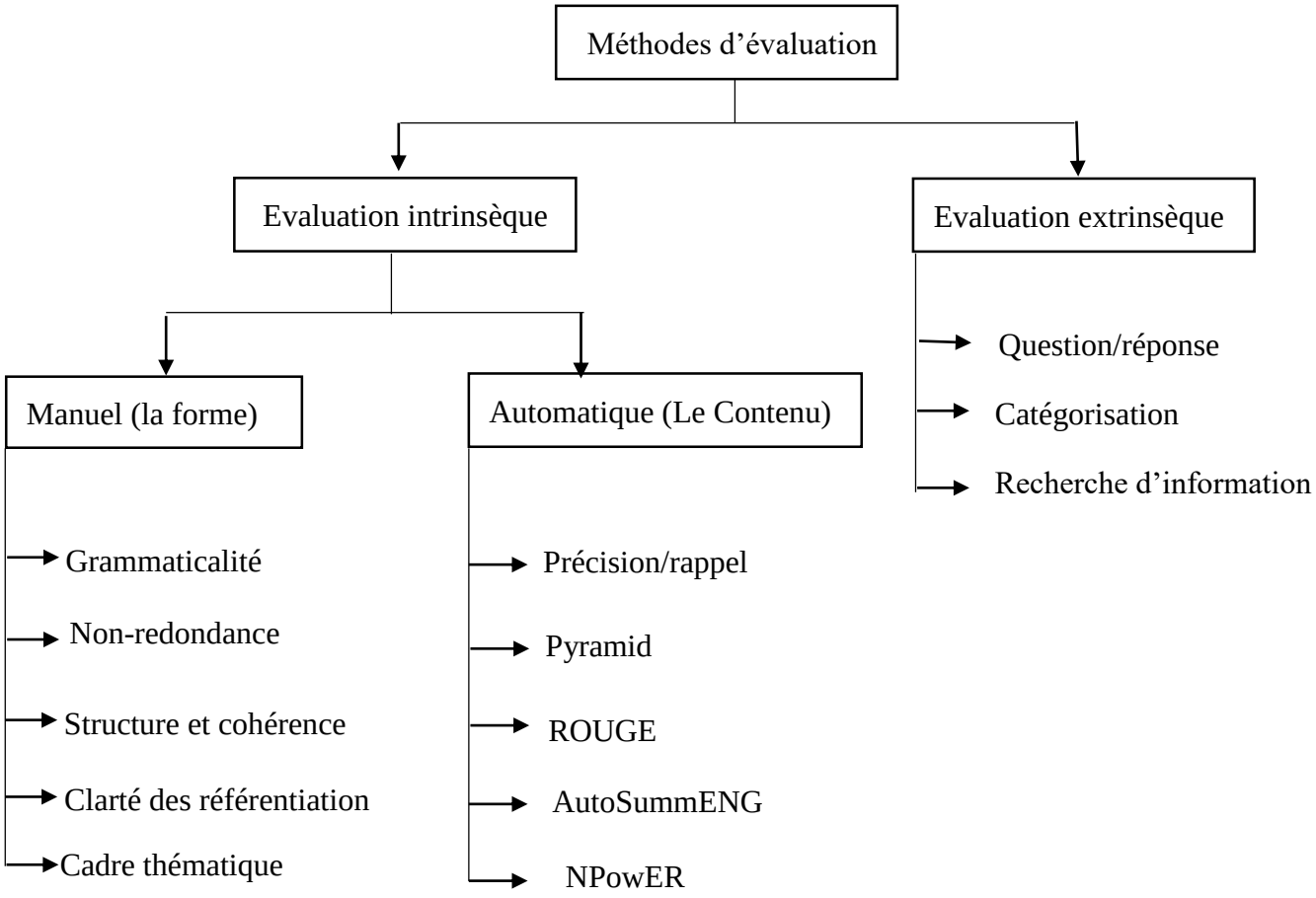

Figure 1-3. Taxonomie Des Méthodes d'évaluation de Résumé Automatique.

6.1.2.1 Méthodes d'évaluation automatique

L'évaluation automatique des résumés automatiques peut se faire au travers des mesures classiques de recherche d'informations telles que la précision et le rappel [NENK, 2006] ou des méthodes semi-automatiques telle que pyramyd [NENK, 2007], ou automatiques tels que ROUGE [LIN, 2004] et AutoSummENG [GIAN, 2011]. Certaines de ces méthodes seront détaillées ci-après.

- **Rappel, précision et F-mesure**

Le rappel et la précision présentent des mesures de similarité classiques de recherche d'information. Ces métriques, mesurent le chevauchement de contenus entre le résumé machine et le résumé de référence. Elles indiquent à quel point un système obtient des performances proches de celles obtenues manuellement [TORR, 2011]. On définit « P » comme étant le nombre de phrases pertinentes extraites par le système ; « NP » le nombre de phrases non pertinentes extraite par le système ; « A » le nombre de phrases pertinentes non extraites par le système et présentes dans le résumé de référence, la précision et le rappel sont calculées comme suit :

$$Rappel = \frac{P}{P+A} \qquad (2)$$

$$précision = \frac{P}{P+NP} \qquad (3)$$

La précision indique l'exactitude du système, autrement dit, combien de bonnes réponses ont été fournies par le système. Tandis que le rappel indique l'exhaustivité, ou combien de bonnes réponses ont été oubliées par le système. Une troisième mesure appelé F-score qui combine la précision et le rappel est généralement calculée. Formellement cette mesure est calculée comme suit :

$$F - score = \frac{2 * précision * rappel}{précision + rappel} \qquad (4)$$

- **ROUGE**

ROUGE (*Recall Oriented Understudy for Gisting Evaluation*) [LIN, 2004] est un ensemble de mesures basées sur le calcul du nombre de n-grammes (séquence de n mots) qui se chevauchent entre le résumé candidat et un ou plusieurs résumés de référence. Il s'agit d'un package qui comprend diverses mesures ROUGE, tels que ROUGE-L (*Longest Common Subsequence*), ROUGE-N (*N-gram Co-Occurrence Statistics*), ROUGE-W (*Weighted Longest Common Subsequence*), et ROUGE-S (*Skip-Bigram Co-Occurrence Statistics*). ROUGE2, ROUGE-S, ROUGE-L sont très utiles pour l'évaluation de résumés mono-document [Lin, 2004]

- **ROUGE-N** : mesure de rappel basée sur le calcul de rappel de n-gramme entre le résumé candidat et les résumés de référence. D'abord, une série de n-grammes (n= 2 ou 3 et rarement 4) est tirée des résumés de référence et du résumé candidat (résumé généré automatiquement). On définit comme *commun (ngram)* le nombre de n-grammes communs entre le résumé candidat et le résumé de référence, *total (ngram)* le nombre total de n-grammes dans les résumés de référence. La formule générale pour ROUGE-N est la suivante :

$$ROUGE_n = \frac{commun(ngram)}{total(ngram)} \qquad (5)$$

Les mesures ROUGE sont Fortement utilisées lors des campagnes DUC et considérées comme des standards par la communauté du fait de leurs fortes corrélations avec les évaluations manuelles.

- **ROUGE-L** : Cette mesure utilise le concept de la plus longue sous-séquence commune (LCS) entre deux séquences de texte. L'intuition est que plus le LCS est long entre deux phrases récapitulatives, plus elles sont similaires. Bien que cette métrique soit plus flexible que la précédente, elle possède l'inconvénient que tous les n-grammes doivent être consécutifs.

- **ROUGE-SU(M) :** Adaptation de ROUGE-2 utilisant des bi-grammes à trous (skip units,(SU)) de taille maximum M et comptabilisant les uni-grammes.

Exemple

Soit l'expression suivante : *'le résumé automatique de document'*

Les unités utilisées lors du découpage de cette expression par les mesures ROUGE sont présentées dans le tableau 1-1.

ROUGE	**Unités utilisées**
ROUGE1	Le- résumé-automatique-de-document
ROUGE2	Le résumé-résumé automatique-automatique de –de document
ROUGE-SU(2)	Le- résumé-automatique-de-document- Le résumé - résumé automatique-automatique de –de document-le automatique- le de- résumé de- résumé document
ROUGE-SU(4)	Le- résumé-automatique-de-document- Le résumé - résumé automatique-automatique de –de document-le automatique- le de- le document-résumé de- résumé document

Tableau 1-1 : Découpages Utilisés dans le Calcul des Mesures ROUGE.

Bien que basée uniquement sur le contenu des résumés (et pas sur la forme), ROUGE souffre de nombreuses lacunes liées à sa dépendance vis-à-vis des unités (n-grammes) utilisées pour le calcul des scores. Les unités multi-mots et les mots relativement peu importants comme « le », « mais », etc. biaisent le nombre de cooccurrences [BOUD, 2008]. En outre, de nombreuses étapes de prétraitement reposant sur des ressources dépendant du langage sont nécessaires auparavant [LIOR, 2012].

Deux nouvelles méthodes ont été proposées afin d'apporter une solution à cette problématique. AutoSummENG (*Évaluation SOMMAIRE automatique basée sur des graphes N-gram*) [GIAN, 2011] a été introduite en tant que méthode d'évaluation indépendante de la langue. L'idée de base derrière cette méthode est de créer d'abord un graphe de n-gramme pour le résumé candidat ainsi que pour les résumés de référence. Ensuite, la moyenne des similarités entre le résumé candidat et chaque résumé de référence est calculée afin d'évaluer le système. Plus tard, la méthode MeMoG (*MergedModel Graph*) [GIAN, 2011] a été introduite Comme variante d'AutoSummENG. MeMoG s'appuie sur un seul graphe représentant les résumés de références pour calculer sa similarité avec le résumé candidat plutôt que d'utiliser tous les graphes de

résumés de référence. Récemment, lors de ACL 2013 *MultiLing Workshop*, la méthode NPowER (*N-gram graph Powered Evaluation via Regression*) [GIAN, 2013] a été ajoutée aux méthodes d'évaluation de résumés automatiques. Cette méthode combine les deux dernières méthodes AutoSummENG et MeMoG. Pour plus de détails sur cette méthode, voir [GIAN, 2013].

6.1.2.2. Evaluation manuelle de résumé automatique (la forme)

Dans le cadre des campagnes d'évaluation DUC (*Document Understanding Conference*), l'évaluation des systèmes participants est faite sur la forme ainsi que sur le contenu. L'évaluation de la forme consiste à attribuer à chaque résumé une note qualitative (de A : très bien à E : très mauvais) selon les critères suivants :

- ***grammaticalité*** : le résumé ne doit pas contenir de phrases clairement agrammaticales (e.g., parties manquantes) qui rendent difficile la lecture du résumé.
- ***non redondance* :** le résumé ne doit pas contenir des informations redondantes.
- ***clarté des références :*** on doit pouvoir identifier à qui ou à quoi les pronoms et les groupes nominaux se réfèrent dans le résumé. Par exemple, le pronom « il » doit signifier quelqu'un dans le contexte du résumé.
- ***cible :*** le résumé ne doit pas contenir d'informations sortant de la thématique désirée.
- ***structure et cohérence :*** le résumé doit être structuré, les phrases doivent être disposées de manière cohérente.

Pour chaque critère linguistique, une note entre 1 et 5 est attribuée : 1 correspond a très mauvais et a 5 à très bon.

6.2. Les Campagnes d'évaluation

Depuis 1998, plusieurs conférences et workshops traitant le domaine de résumé automatique de textes ont vu les jours tels que SUMMAC[3] , DUC[4], TAC[5]. Ces campagnes sont organisées par NIST (*National Institute of Standards and Technology*). Leur but est de promouvoir les progrès réalisés dans le domaine du résumé automatique de textes et surtout de permettre aux chercheurs de participer à des expérimentations de grande envergure tant au point

[3] http://www.nlpir.nist.gov/related_projects/tipster_summac.
[4] http://www.nlpir.nist.gov/projects/duc/index.html.
[5] http://www.nist.gov/tac.

de vue du développement que de l'évaluation de leurs systèmes. Par exemple DUC était une série importante de conférences qui se déroulait chaque année entre 2001 et 2007. DUC a organisé des concours qui comprenaient plusieurs tâches de résumé automatique (voir le tableau 1-2) en fournissant plusieurs corpus de référence (DUC 2001 , DUC 2002, DUC 2004). De nombreux travaux de résumé automatique sur les documents en langue Arabe ont été évalués et comparés à l'aide des corpus de référence DUC [DOUZ, 2004; FATT, 2009; EL HA, 2011a].

En 2008, la conférence DUC a rejoint les conférences TREC "*Text Retrieval Conferences*" pour former une seule conférence TAC "*Text Analysis Conference*".

Plusieurs sessions DUC et TAC se sont tenues jusqu'à présent. Le tableau 1-2 présente les tâches introduites par ces compagnes pour encourager les recherches dans le domaine du résumé automatique de documents.

Conférence	Tâches
DUC 2001 –DUC 2002	- Résumés génériques mono et multi-documents
DUC 2003	- Resume court (headline) et multi-document
DUC 2004	- Résumés Courts, - Résumé multilingues multi-documents - résumé biographiques
DUC 2005- DUC 2007	- Résumés orientés multi documents
TAC 2008 – TAC 2009	- Résumés mise à jour (*Update task summarization*) - Résumé guidé multi-documents
TAC 2010	- Résumés orientés multi-documents (*Guided summarization*) - Evaluation automatic de résumé (*Automatically Evaluating Summaries Of Peers)*
TAC 2011	- Résumé guidé *(Guided summarization),* - Evaluation automatique de résumé - et la nouvelle tâche *MultiLing pilot* pour favoriser l'utilisation d'algorithmes multilingues pour le résumé.
TAC 2014	- résumé de textes biomédicales (*Biomedical summarization*)

Tableau 1-2 : Tâches Introduits par les Conférences DUC/TAC.

Dans le cadre de ces campagnes, l'évaluation des systèmes de résumés est réalisée de manière intrinsèque sur le fond ainsi que sur la forme des résumés produits.

7. Applications de Résumé Automatique de Textes

Avec l'accroissance rapide de nouveaux contenus sur Internet, de nouvelles tâches de résumé automatique ont pu voir le jour, ainsi leurs applications sont devenues multiples, parmi lesquelles nous citons les suivant :

- Améliorer les performances des systèmes de recherche d'information: le résumé automatique de textes peut être utilisé dans les systèmes de recherche d'information et de question réponse comme une étape intermédiaire, afin de réduire la longueur des documents [PERE, 2013]. Cela permet de réduire le temps de recherche et facilite l'accès à l'information désirée. Dans ce contexte, Perea-Ortega et al. [PERE, 2013] ont proposé une approche pour la génération automatique des résumés génériques et géographiques dans le perspective d'améliorer les performances des systèmes de recherche d'information géographique.
- Résumer les nouvelles de presse: Les articles journalistiques ont été au centre de la plupart des systèmes de résumé proposés jusqu'à présent. McKeown et al. [MCKE, 2003] ont proposé un système robuste, *Columbias Newsblaster*, qui produit des résumés qui mettent à jour un utilisateur sur de nouvelles informations à propos d'un événement.
- Résumer de documents spécialisés **(**juridique, biomédical..etc**)**: les documents issues d'un domaine de spécialité posent des difficultés particulières due à leurs structures, ainsi qu'à leur terminologie spécifique. Les algorithmes de résumé automatique doivent ainsi s'adapter pour pouvoir traiter ce genre d'information. Dans ce contexte, Farzindar [FARZ, 2005], a proposé une approche pour générer un résumé très court pour les documents juridiques longs en explorant l'architecture des documents. d'un autre coté Reeve et al. [REEV, 2007] ont proposé une approche hybride pour la génération automatique de résumé de textes biomédicaux.
- Résumer de fils de discussion (email threads) [ULRI, 2009; RAMB, 2004 ; CARE, 2007], résumés des emails [CORS, 2004], et des pages web [DIAO, 2006].
- Résumer les opinions : c'est l'une des tendances actuelles de résumé automatique [NISH, 2010 ; MENG, 2012 ; POTT, 2010]. Le résumé d'opinion est une tâche difficile et un peu différente des résumés de texte classiques. En résumé d'opinion, il ne s'agit pas de résumer les commentaires des clients à propos d'un produit en sélectionnant un sous ensemble de phrases , mais plutôt de rassembler d'une façon cohérente les avis positifs, négatifs ou neutres sur un produit [TORR, 2011].

- Résumé multimédia : C'est un sujet actif dans la communauté scientifique de résumé automatique. Il consiste à combiner plusieurs sources d'information sur internet: documents textes, vidéo, audio, image. En France le projet RPM2[6] (Résumé Pluri-média, Multi-documents et Multi-opinions) lancé en 2007 a pour but la mise au point de méthodes de résumé multi-documents pour les médias texte, audio et vidéo ainsi que pour des contextes pluri-média.

8. Conclusion

Ce chapitre a permis de brosser un portrait global du domaine du résumé automatique de documents textuels. Tout d'abord, des notions de base de résumé automatique ont été présentées, les différente étapes par lesquelles passe l'élaboration automatique de résumé ainsi les différente approches utilisées dans ce domaine ont été clairement étudiées. Nous avons également décrit les méthodes d'évaluation de résumés automatique existantes, tout en mentionnant que ces méthodes sont des solutions partielles permettant de se faire une idée de la qualité de résumé.

Enfin, nous avons exposé quelques applications de résumé automatique de textes.

En un mot, on peut dire que le résumé automatique de textes est un domaine loin d'en être à ses débuts, mais qui présente encore plusieurs défis. Il s'agit d'une technologie ayant le potentiel de soutenir des applications très utiles et intéressantes, mais démontrant encore plusieurs lacunes que ce soit pour les approches de construction qui restent un peu loin de ce que peut faire un humain ou même pour les stratégies d'évaluation.

[6] http://rpm2.org/index.html

Chapitre 2

Approches de Résumé Automatique de Textes

Sommaire

1. Introduction

Au cours du demi-siècle dernier, différentes méthodes et techniques issues du domaine de l'intelligence artificielle et du traitement automatique des langages naturels (TALN) ont été développées pour produire des résumés automatiques de textes. Ces méthodes peuvent être classées en cinq catégories : les approches statistiques basées sur le calcul des scores, les approches à base de l'apprentissage supervisé, les approches à base de l'apprentissage non supervisé (ou *clustering*), les approches à base de l'analyse de discours et les approches à base de l'analyse sémantique.

Ce chapitre a pour objet de rendre compte d'une recherche bibliographique concernant le résumé automatique de textes. Nous présenterons, d'abord, chacune des approches susmentionnées ainsi que les concepts fondamentaux qui s'y rapportent en prêtant une importance particulière à l'approche à base de l'analyse de discours, vu qu'elle a influencé est inspiré notre travail. Nous allons également passer en revue les différents travaux existants dans ce domaine de recherche encore en plein développement. Par la suite, nous ferons une synthèse des travaux menés dans le domaine du résumé automatique de textes Arabes qui sont au centre de la problématique de cette thèse. En dernier lieu, nous achèverons ce chapitre par un petit aperçu sur quelques ressources disponibles pour la langue Arabe.

2. Approches de Résumé Automatique

2.1 Les approches statistiques

Les approches statistiques regroupent les méthodes dont l'objectif est de déterminer les caractéristiques qui reflètent l'importance de la phrase. Ainsi, l'évaluation des pertinences des unités textuelles dépend de certaines mesures statistiques. Dans ce type d'approches, chaque unité reçoit un score basé sur certaines caractéristiques qui peuvent être statistiques telles que les fréquences des termes [LUHN, 1958] ou linguistiques, basées sur l'occurrence de certains marqueurs linguistiques. Les phrases ayant les scores les plus élevés sont assemblées dans l'ordre de leur apparition dans le document source pour construire le résumé. Plusieurs caractéristiques ont été utilisées, parmi lesquelles nous citons les suivantes :

- **Fréquence des termes (term frequency TF):** Cette caractéristique est considérée parmi les premières utilisées dans le domaine du résumé automatique de textes. elle a été proposée par Luhn [LUHN, 1958]. L'idée de Luhn s'articule autour du fait que les mots les plus

fréquents dans un document (à l'exception des mots vides) sont les mots les plus importants dans le sens où ils véhiculent le maximum d'informations. Par conséquent, les phrases qui renferment les mots les plus fréquemment utilisés dans le texte sont considérées comme étant des phrases renfermant le plus d'informations (importantes).

- **Position de la phrase:** Baxendale [BAXE, 1958] a montré que le degré d'importance d'une phrase peut être déterminé par sa position relative dans le texte. Suite à une étude faite sur un corpus de 200 paragraphes, Baxendale a constaté que, dans 85% des paragraphes, la phrase la plus importante se trouve en début de paragraphe (première position), et dans 7% des paragraphes, la phrase la plus importante s'est produite en dernière position.

- **Expressions indicatives (*cue phrase*)**: D'après Edmundson [EDMU, 1969], la pertinence d'une phrase dépend de deux types d'expressions : *bonus et stigma*. L'existence des expressions bonus telles que « l'objectif de cet article », « dans ce rapport, on propose » « en conclusion » signale l'importance de la phrase qui les renferment et par voie de conséquence ; elle augmente son score. Par contre les expressions « stigma » telles que « par exemple » « c'est-à-dire » pénalise le poids de la phrase. Les méthodes basées sur cette caractéristique utilisent généralement des dictionnaires contenant les expressions indicatives extraites à partir des résumés produits manuellement.

- **Similarité avec le titre *(Title similarity):*** les principaux thèmes abordés dans le texte sont généralement véhiculés dans le titre et les sous-titres [DOUZ, 2004]. Par conséquent Les phrases pertinentes sont celles qui contiennent les mots de titre. La similarité ou la ressemblance d'une phrase avec le titre est le chevauchement du vocabulaire entre une phrase donnée et le titre. Les scores attribués aux phrases sont, ainsi, fonction des cooccurrences des mots de titres et de leurs regroupements.
  La similarité avec le titre peut être aussi calculée en utilisant d'autres mesures de similarité telle que la mesure *cosinus* [FATT, 2009].

- **Les termes clés *(keywords):*** Les termes clés sont des mots ou des expressions (multi-mots) représentant les aspects principaux qui sont abordés dans le document. Ainsi, les phrases contenant ces termes sont considérées comme des phrases pertinentes et elles ont une forte chance d'être extraites pour construire le résumé.

- **Fréquence normalisée TF*ISF (*terme frequency-inverse sentence frequency*)**

  TF*ISF [BRAN, 1995] est une version spéciale de la pondération TF*IDF [SALT, 1997]. La pondération TF*IDF prend en considération le corpus auquel appartient le texte à résumer lors du calcul des fréquences des termes. IDF (*inverse document frequency*) suppose que les termes fréquents dans le document et rares dans le corpus sont importants. Cependant dans le cas de résumés mono-document, l'ISF (*inverse sentence frequency*) est utilisé au lieu de l'IDF comme indiqué dans la formule 1. Le poids TF*ISF de chaque terme est donné par la formule 2. Le score assigné à une phrase est la somme des poids TF*ISF de ses termes constituants (Les mots vides sont supprimés à l'avance).

$$ISF = \log(\frac{N}{N(ti)}) \quad (1)$$

$$TF * ISF(ti, S) = TF(ti, S) * ISF(ti) \quad (2)$$

$$TF * ISF(S) = \sum TF * ISF(ti, S) \quad (3)$$

Avec :

N: nombre total de phrases dans le document.

N (t_i) : nombre de phrases qui contiennent le terme (t_i).

TF (t_i, S) : fréquence du terme (t_i) dans la phrase S.

- **Noms propres** : les noms propres sont les noms des personnes ou des lieux. D'après Nobata et al. [NOBA, 2002], la présence des noms propres dans une phrase signale son importance.

- **La cohésion des phrases *(sentence to sentence cohesion)*** : les phrases ayant un haut degré de cohésion avec les autres phrases sont considérées comme étant plus importantes. Pour calculer le score d'une phrase, la similarité de la phrase avec chacune des autres phrases du document est d'abord calculée, la somme des similarités obtenues représente le score de la phrase.

La Figure 2-1 illustre le processus de résumé fondé sur une approche statistique.

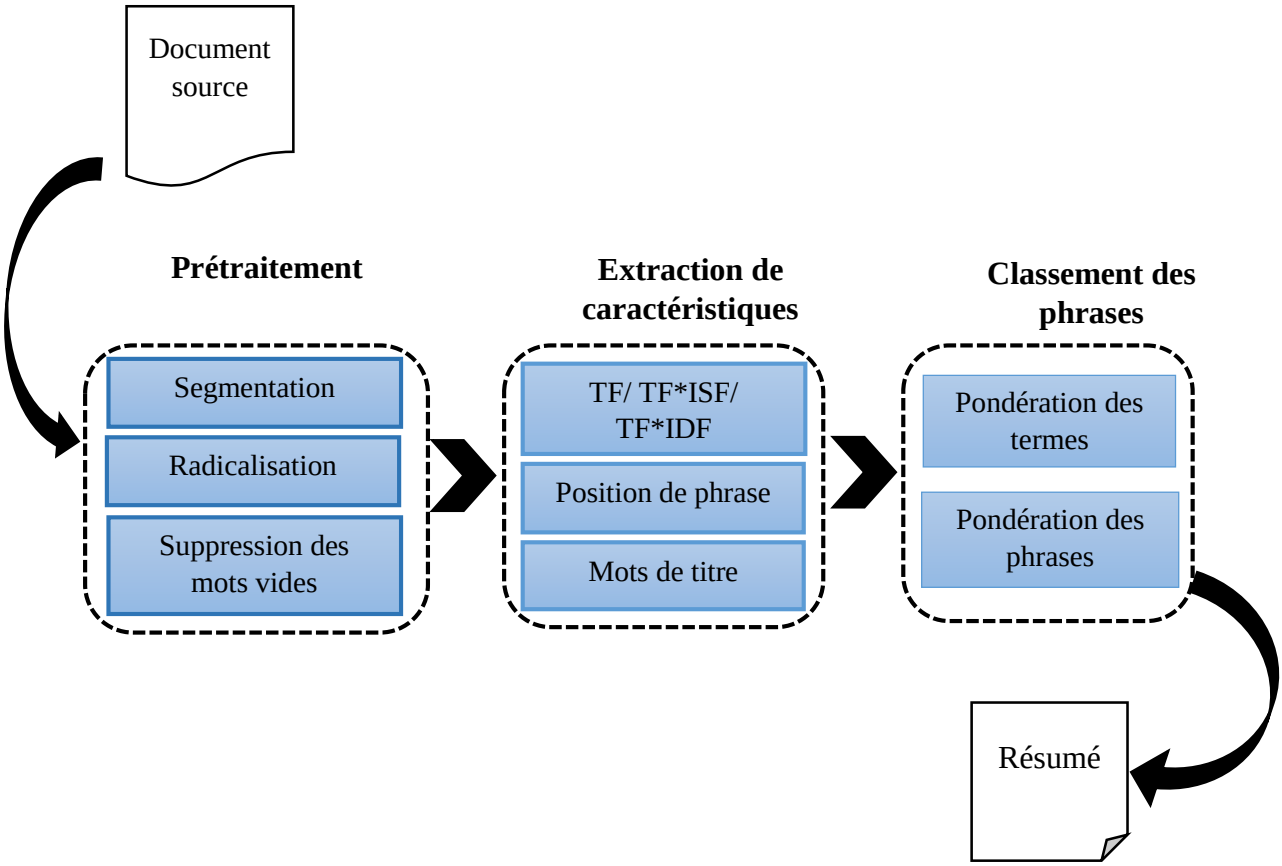

Figure 2-1. Processus de Résumé Automatique fondé sur une approche Statistique.

Plusieurs travaux dans le domaine de résumé automatique se sont appuyés sur des méthodes statistiques pour extraire les phrases pertinentes du texte source [BERGE, 2000; GALA, 2008; KNIG, 2000]. Le travail le plus ancien et peut être le plus cité est celui de Luhn [LUHN, 1958]. Dans son travail, il a d'abord cherché les mots-clés qui sont pour lui les mots les plus fréquents dans le texte. Puis il a sélectionné les phrases avec des mots-clés fortement pondérés pour produire le résumé. Luhn a souligné le besoin de caractéristiques qui reflètent l'importance des phrases dans le texte. Il a proposé plusieurs idées clés qui ont pris de l'importance dans les travaux ultérieurs de résumé automatique. Plus tard, Edomnson [EDOM, 1969] a étendu ces travaux par l'ajout de deux autres caractéristiques : les mots provenant des titres et les expressions indicatives. Edomnson [EDOM, 1969] a combiné ces deux caractéristiques avec celles proposées par Luhn [LUHN, 1958] et Baxendale [BAXE, 1958]. Brandow et al. [BRAN, 1995] ont utilisé la pondération TF*IDF pour sélectionner les phrases importantes. L'intuition

derrière ceci était de capturer les mots qui sont, généralement, moins fréquents. Nobata et al. [NOBA, 2002] ont proposé une approche basée sur les noms propres, fréquence de termes, mots clés, similarités avec le titre et position de la phrase dans le texte. L'approche a été évaluée en utilisant le corpus d'évaluation DUC 2001 et il a été observé que l'approche proposée donne de meilleurs résultats par rapport à d'autres systèmes participant à DUC 2001.

Les techniques statistiques sont des techniques qui ne dépendent pas fortement d'une langue donnée et par conséquent elles sont bien généralisables à diverses langues. En effet, Pour des systèmes utilisant par exemple des vecteurs de fréquences, l'application à plusieurs langues ne pose pas de problème parce que les fréquences sont calculées à partir des formes graphiques qui peuvent être des mots ou des n-grammes. Cela explique l'utilisation de ces techniques dans beaucoup de systèmes de résumés multilingues [BLAI, 2008].

2.2. Approches basées sur l'apprentissage automatique

Dès les années 1990, les techniques de l'apprentissage automatique ont commencé à être utilisées pour produire des résumés automatiques de textes. Plusieurs approches ont été proposées pour aborder cette tâche, allant des techniques de l'apprentissage supervisé, semi-supervisé à non supervisées. Ainsi plusieurs caractéristiques ont été explorées. Dans le reste de cette section, nous allons présenter deux grandes approches : l'approche basée sur l'apprentissage supervisé et celle basée sur l'apprentissage non supervisé *(clustering),* ainsi que différents méthodes proposées sous ces optiques.

2.2.1 Approches basées sur l'apprentissage supervisé

La majorité de méthodes fondées sur l'apprentissage automatique reposent sur l'apprentissage supervisé, dans lequel le processus de résumé automatique est modélisé comme un problème de classification binaire. L'objectif est de construire un modèle (fonction d'appartenance) qui sert à classer les phrases en deux catégories : phrases pertinentes et non pertinentes pour le résumé. Cela nécessite systématiquement un corpus d'entrainement composé de documents sources et de leurs résumés qui vont être considérés comme des références. Chaque phrase dans le corpus d'entrainement est étiquetée pour indiquer si elle fait partie ou non du résumé de référence. Une fois l'apprentissage du système accompli, le modèle construit sera utilisé pour identifier les phrases importantes du résumé lorsqu'un nouveau document est fourni au système. La figure 2-2 illustre le processus de résumé automatique basé sur l'apprentissage supervisé.

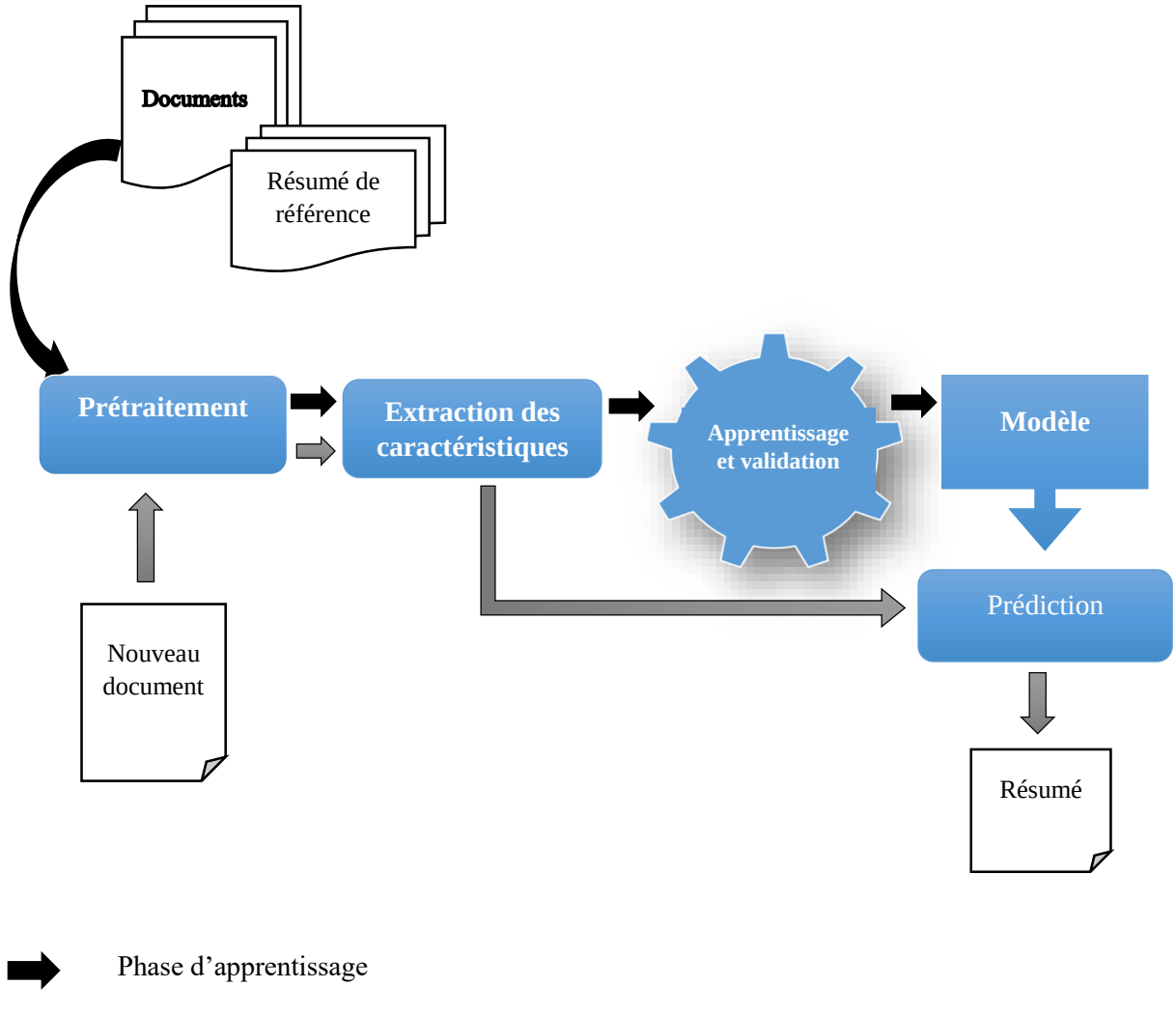

Figure 2-2 : Processus de Résumé Automatique Fondé Sur L'apprentissage Supervisé

Les techniques de l'apprentissage supervisé ont été employées dans plusieurs travaux pour produire des résumés automatiques de textes [LIN, 1999; FATT, 2008; SCHI, 2008]. Au début, certains auteurs s'étaient appuyés sur des méthodes bayésiennes naïves qui supposent l'indépendance des termes [KUPI, 1995; AONE, 1998], tandis que d'autres ont modélisé le problème d'extraction des phrases pertinentes à l'aide des arbres de décision. Ils ont, ainsi, mis fin à cette hypothèse [LIN, 1999].

Par la suite, les auteurs ont commencé a profité des avantages des réseaux de neurones et de leurs capacités d'apprentissage pour générer des extraits automatqiues [KAIK, 2004; YONG, 2006; SVOR, 2007 ; FATT, 2008]. Kaikhah [KAIK, 2004] a utilisé un modèle de réseaux de neurones (*Feed-forward)* à trois couches pour apprendre les caractéristiques des phrases de résumé en utilisant sept descripteurs. Une fois que le réseau a appris les caractéristiques qui représentent au mieux les phrases de résumé, certains descripteurs sont supprimés et d'autres sont combinés. Le modèle élagué est ensuite utilisé pour identifier les phrases de résumé. Dans le même contexte, NetSum, un système de résumé mono-document basé sur un modèle de réseau de neurones a été introduit par Svore et al. [SVOR, 2007]. Le système a démontré une meilleure performance qui dépassait significativement les Baseline définis dans les campagnes d'évaluation DUC 2007. Récemment, Nallapati et al. [NALL, 2017] a utilisé les réseaux de neurones récurrents (RNN) pour créer un système de résumé appelé SummaRuNNer [NALL, 2017]. Le système possède comme avantage la possibilité d'apprendre uniquement sur les résumés de référence générés par l'homme, éliminant le besoin d'étiqueter les phrases de documents d'entrainement.

Contrairement aux méthodes précédentes, qui sont principalement basées sur le choix des descripteurs, Conroy et O'leary [CONR, 2001] a utilisé un modèle de Markov caché (HMM) avec deux états. La motivation derrière l'utilisation des modèles séquentiels est de tenir compte des dépendances locales entre les phrases. Les auteurs ont utilisés trois descripteurs : position de la phrase, nombre de termes dans la phrase ainsi que la probabilité des termes. Ils suggèrent que la probabilité de sélectionner une phrase pour le résumé dépende des phrases déjà sélectionnées. D'autres variantes des HMM ont été aussi utilisées dans [KNIG, 2000 ; KNIG, 2002 ; TURN, 2005].

Il est à noter que ces travaux ont permis le développement de méthodes performantes mais sans garantie que le résumé puisse être exploité. En effet, les phrases ainsi sélectionnées peuvent ne pas être cohérentes. De même, les descripteurs utilisés lors de l'apprentissage peuvent s'avérer ne pas être consistants au travers des différents types de documents.

Les techniques fondées sur l'apprentissage automatique sont généralement assez variables et s'utilisent de plusieurs façons. Cependant, leur inconvénient majeur porte sur leur dépendance aux corpus des phrases étiquetées manuellement. L'étiquetage manuel est très couteux en temps, de plus il est difficile d'étiqueter manuellement des phrases de résumé similaires aux résumés humains.

Un autre inconvénient lié à ces techniques réside dans leur incapacité à maîtriser le processus résumant [BLAI, 2008]. Par cette remarque, il faut entendre le fait que tout traitement effectué, telle que l'estimation de la pertinence ne peut pas être totalement explicable comme cela peut l'être pour un système à base de règles ou de connaissances. L'apprentissage est conditionné par l'objectif d'obtenir les sorties désirées, peu importe les traitements internes du système. Le travail étant alors de paramétrer le système correctement et de ne retenir que la combinaison des descripteurs qui donnent les meilleurs résultats.

2.2.2 Approches basées sur l'apprentissage non supervisé (*clustering*)

Le *clustering* fait référence au regroupement d'instances similaires dans un même groupe (cluster). Dans le cas du résumé automatique, ces instances sont les phrases. Cela peut être fait à l'aide du calcul de similarités entre phrases. Les phrases très similaires sont regroupées dans le même groupe (cluster). Par la suite, les phrases de chaque groupe ayant les scores les plus élevés sont rassemblées pour former le résumé.

Radev et al. [RADE, 2004] ont été les premiers à avoir utilisé les centroïdes du cluster dans leur système de résumé automatique multi-documents, appelé MEAD. Le centroïde a été défini comme un pseudo-document composé de mots ayant des scores TF* IDF supérieur à un seuil prédéfini. Les auteurs ont utilisé ces centroïdes pour identifier les phrases pertinentes de chaque cluster, qui sont les phrases les plus similaires au centroïde, en utilisant la mesure de similarité cosinus. Les phrases les plus pertinentes de chaque cluster ont été sélectionnées et assemblées pour produire le résumé. Wan et Yang [WAN, 2008a] ont proposé une méthode de résumé multi-documents basé sur trois algorithmes de clustering. Les phrases ont été regroupées en utilisant ces trois algorithmes en différents clusters (sous-thèmes). Le nombre de clusters a été défini comme la racine carrée du nombre de phrases dans l'ensemble des documents.

Les algorithmes de *clustering* par Partitionnement ont été utilisés dans plusieurs travaux de résumé automatique [RADE, 2000; KOLL, 2007]. Tandis que Nie et al. [NIE, 2006] ont employé une approche hybride qui ne dépend pas uniquement de la similarité avec le cluster pour la sélection des phrases, mais considère également la similarité avec le contenu global du document. Dans le travail de Aliguliyev [ALIG, 2010], l'accent a été mis sur le renforcement de la diversité des clusters. Pour y parvenir, l'auteur a utilisé un algorithme d'optimisation en rajoutant une opération de mutation adoptée à partir des algorithmes génétiques pour optimiser la similarité intra-cluster et la dissimilarité inter-cluster.

Les méthodes basées sur le *clustering* ont réussi à représenter la diversité et à éliminer la redondance d'information [EL HA, 2012], qui est un des problèmes majeurs de résumé automatique multi-documents. Toutefois un résumé ne peut pas être suffisamment significatif au niveau de son contenu si la pertinence de ses phrases est jugée simplement sur la base des clusters [KUMA, 2016]. En effet, dans ces méthodes, les phrases sont classées en fonction de leur similarité avec le centroïde du cluster qui représente simplement les termes les plus fréquents des documents.

2.3 Approches basées sur la théorie des graphes

Dans les approches à base de graphes, les unités textuelles (phrases) d'un document ainsi que les relations entre elles sont représentées sous forme d'un graphe. Chaque phrase dans le document est représentée par un nœud dans le graphe. Deux nœuds sont reliés par un arc si la similarité entre les deux phrases est supérieure à un seuil prédéfini. Ainsi le poids associé à chaque arc correspond au degré de similarité entre les deux phrases. Cette similarité peut être calculée en utilisant différentes mesures de similarité. Dans la littérature de l'approche à base de graphe, la mesure cosinus est la mesure de similarité la plus utilisée [KUMA, 2016].

Une fois le graphe construit, un algorithme de classement à base de graphe tel que TextRank [MIHA, 2004] est utilisé pour identifier l'importance ou la centralité des phrases. Une phrase est considérée importante ou centrale si elle est fortement liée à de nombreuses autres phrases importantes. Enfin, les phrases les mieux classées sont ensuite sélectionnées pour produire le résumé.

Dans le résumé automatique multi-documents, les liens intra-document ainsi qu'inter-documents sont pris en compte. Certain auteurs ont tenté d'améliorer les performances de leur système en modifiant les algorithmes de classement des phrases. Wan et Yang [WAN, 2006] ont assigné des poids aux liens intra-document différent de ceux attribués aux liens inter-documents. Ainsi, Les auteurs ont donné plus d'importance aux phrases fortement liées par des liens inter-documents.

Plusieurs auteurs se sont s'intéressés à l'approche à base de graphes pour générer des résumés automatiques multi-documents [ERKA, 2004 ; MIHA, 2004 ; WAN, 2008a ; HARI, 2009; WEI, 2010].

Les méthodes issues de ce paradigme ont réussi à démontrer leur capacité à identifier des phrases prestigieuses au travers de multiple documents. Cependant cette approche dépend

fortement de similarité entre phrases pour générer le résumé sans comprendre la relation entre ces phrases [KUMA, 2016].

2.4 Approches basées sur l'analyse de discours

Les approches basées sur l'analyse de discours visent à exploiter la structure discursive de texte ainsi que les relations rhétoriques et intentionnelles qui existent entre les unités textuelles afin de parvenir à générer des extraits [KUMA, 2016]. Elles reposent sur l'idée qu'un texte est défini par sa structure discursive et ses relations rhétoriques dépendantes de la langue [TORR, 2011]. La structure de discours est une représentation inhérente d'un document qui détermine les relations entre les unités textuelles et leur saillance [HIRA, 2015]. La prise en compte de cette structure permet de générer des résumés cohérents[1]. Plusieurs méthodes qui prennent en compte la cohérence globale entre les unités textuelles dans les documents ont été proposées. Parmi les méthodes les plus intéressantes et les plus étudiées, citons celles qui sont basées sur la théorie de la structure rhétorique (RST) [MANN, 1988], vue que cette théorie est à la base de notre propre approche. Les méthodes fondées sur la RST considèrent un document comme un arbre de discours, ainsi les unités textuelles les plus pertinentes sont extraites selon les contraintes de l'arbre [HIRA, 2015].

Afin de mieux comprendre l'approche fondée sur la RST, nous allons d'abord présenter les principes de base de cette théorie ainsi que la manière avec laquelle cette théorie de discours a été utilisée pour générer des résumés automatiques.

2.4.1 La théorie de la structure rhétorique

La théorie de la structure rhétorique (RST) [MANN, 1988] est une théorie descriptive de l'organisation du texte par le bais des relations qui existent entre les unités textuelles [TABO, 2006]. Elle a été créée par Mann et Thompson suite à une analyse de plus de 400 textes. Bien que cette théorie fût initialement destinée à être utilisée dans la génération automatique de textes, elle est devenue par la suite un cadre célèbre pour l'analyse de discours [TABO, 2006].

RST se focalise sur une analyse rhétorique, qui vise à structurer le texte sous une forme hiérarchique des relations rhétoriques qui existent entre unités textuelles. Plus formellement, dans le cadre de la RST, si deux unités de discours élémentaires sont reliées par une relation de discours (appelée aussi relation de cohérence ou relation rhétorique) elles constituent ensemble

[1] *'coherence is defined as the discourse connectivity that focuses on the overall logical structure and links among clauses or sentences in the text' [MANN, 1988]*

un autre élément discursif. L'élément composé peut à son tour participer à une autre relation de discours comme s'il s'agissait d'une unité de discours élémentaire [MANN, 1988]. Sous une analyse complète, un texte cohérent est structuré hiérarchiquement sous forme d'un arbre, appelé structure rhétorique, dans lequel les nœuds supérieurs sont les plus représentatifs du message de l'auteur.

Deux éléments clés sont à définir dans le cadre de la RST : les unités de discours et les relations rhétoriques.

- **Les Unités de Discours**

Les unités de discours élémentaires (EDUs) sont les unités de texte minimales. Elles ne sont pas nécessairement des clauses syntaxiques, où il y'a forcément des indices lexicaux explicites pour indiquer leurs limites [FENG, 2015]. D'après Mann et Thompson [MANN, 1988], RST fournit une façon générale de décrire les relations entre les clauses d'un texte, qu'elles soient signalées ou non grammaticalement ou lexicalement.

Les unités de discours peuvent avoir le statut ' noyau' ou 'satellite', tout dépend de leur importance relative dans le texte. La distinction entre le noyau et le satellite vient de l'observation empirique que le noyau exprime ce qui est essentiel à l'intention de l'auteur tandis que le satellite fournit une information secondaire et que le noyau est compréhensible indépendamment du satellite. Cette relation n'étant pas inversible. [MANN, 1988].

- **Les Relations Rhétoriques**

Dans le cadre de la RST, les relations qui relient deux unités de discours peuvent être mononucléaires ou multi-nucléaires. Les relations mononucléaires relient deux unités de discours ayant des statuts différents : un noyau et un satellite, tandis que les relations multi-nucléaires relient des unités de discours d'importance égale ; en ce sens que les unités liées sont toutes des noyaux. Les auteurs de la RST ont défini 24 relations dont la majorité est mononucléaires (21 relations). Cet ensemble de relations a été étendu par la suite par [CARL, 2003] à 78, ce qui permet un plus haut niveau d'expression. Ces relations sont regroupées en 16 classes en fonction de leur similarité rhétorique.

- **Etapes dans l'analyse rhétorique**

L'analyse rhétorique d'un texte dans le cadre de la RST consiste en trois étapes :

1) Segmentation du texte source;
2) Identification des relations rhétoriques;
3) et construction de la structure rhétorique/discursive du texte.

La première étape de l'analyse rhétorique consiste à segmenter le texte source en unités de discours élémentaires. A cette fin, Marcu [MARC, 1997] a utilisé les signes de ponctuation et un ensemble de marqueurs linguistiques appelés expressions indicatives *(par exemple, bien que,…)* pour identifier les frontières des unités de discours. D'après [FENG, 2015], s'appuyer seulement sur les informations lexicales n'est pas suffisant pour la segmentation de textes en EDUs, ainsi des approches plus sophistiquées sont nécessaires et doivent être considérées.
Une fois le texte segmenté en EDUs, les relations rhétoriques entre les différentes unités de discours doivent être identifiées. Lors de cette étape, il faut être capable de reconnaître non seulement la présence d'une relation, mais aussi son type et sa portée [FENG, 2015]. Les marqueurs linguistiques peuvent être utilisés lors de cette étape pour identifier la relation entre deux unités discursives. Mais, souvent, un marqueur linguistique peut être un indicateur de plus d'une relation. Par exemple, le marqueur « mais » peut indiquer la relation 'CONTRASTE' aussi bien que 'CONCESSION' ou 'ANTITHESE'. La désambiguïsation des marqueurs linguistiques est plus difficile pour l'identification des relations rhétoriques que pour la segmentation, qui consiste seulement à distinguer si un marqueur linguistique est un marqueur de discours ou non. L'identification des relations rhétoriques et les différentes méthodes proposées à propos ce sujet feront l'objet du chapitre 4 de cette thèse.
La dernière étape dans l'analyse rhétorique consiste à relier les unités de discours (EDU et unités de discours plus larges) par le bais des relations rhétoriques pour construire la structure rhétorique de texte.

Considérons à titre d'exemple le fragment de texte de l'exemple 1 composé de quatre EDUS (e1, e2, e3, e4) et sa structure rhétorique présentée dans la Figure 2.3. Nous avons utilisé la notation introduite par [MANN, 1988], dans laquelle les lignes horizontales indiquent les unités de discours et les lignes verticales les noyaux. Les satellites sont reliés aux noyaux par des flèches incurvées.

Exemple

[*Le ministre a déclaré :]* $_{e1}$ *[l'état cherchait à négocier avec les syndicats]* $_{e2}$ *[pour mettre fin aux grèves répétées]*$_{e3}$ *[et trouver des solutions fonctionnelle aux problèmes posés]*$_{e4}$*.*

Sur la figure 2-3, les deux EDUs *e3* et *e4* sont reliées par une relation multi-nucléaire *joint,* formant une unité de discours plus grande (*e3-e4*). Cette dernière est reliée à *e2* par la relation mononucléaire *but*, tel que *e2* est plus centrale (indiquée par une ligne verticale) que l'unité de

discours (*e3-e4*). Enfin l'EDU *e1* est reliée à l'unité de discours (*e2-e4*) par la relation mononucléaire *attribution* pour former la structure rhétorique du fragment textuel.

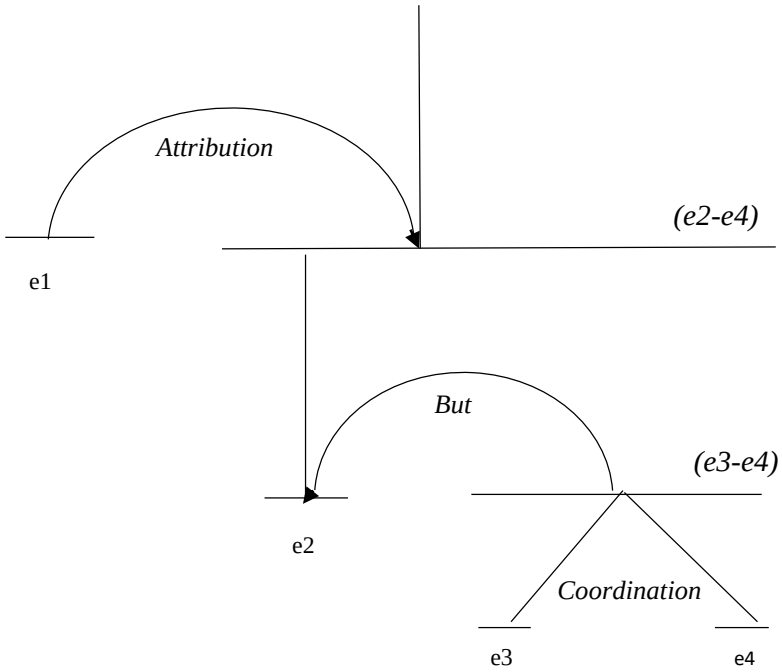

Figure 2-3 : Structure Rhétorique de Texte.

De nombreuses ressources ont été créées suivant le principe de la RST. Le RST *Discours Treebank* (RST-DT) [CARL, 2003] est le premier corpus anglais annoté dans le cadre de la RST. Le corpus est composé de 385 articles : 347 articles pour l'apprentissage et 38 articles pour le test. RST-DT [CARL, 2003] a été largement utilisé comme un corpus de référence standard dans les recherches de l'analyse de discours car il fournit un guide systématique de plusieurs concepts dans le développement original de la RST ; y est compris la définition des EDUs ainsi que les relations rhétoriques. Des corpus similaires ont été également créés pour d'autres langues telles que l'espagnol [DA C, 2011], le néerlandais [VAN, 2011], le portugais [PARD, 2004] et l'allemand [STED, 2014].

2.4.2 Méthodes de résumé basées sur la théorie de la structure rhétorique

De nombreuses méthodes de résumé mono-document basées sur la RST ont été proposées. Toutes ces méthodes donnent plus d'importances aux segments nucléaires des relations.

De plus, chaque méthode utilise des critères différents pour sélectionner les éléments pertinents, ou en d'autres termes pour sélectionner les satellites à éliminer vu que ces derniers fournissent une information secondaire. Généralement le texte source est segmenté en unités discursives. Une fois la structure rhétorique de texte créée, un algorithme qui pondère et ordonne chaque élément de cette structure est appliqué. Les segments les mieux classés (ayant les poids les plus élevés) sont sélectionnés pour produire le résumé tout en tenant en compte du taux de compression. Typiquement, le score associé à un segment est en fonction de son occurrence dans l'arbre. Il faut noter que toutes les méthodes proposées emploient un ordre partiel lors du classement des segments, cela veut dire que certains segments peuvent avoir le même ordre parce qu'ils ont le même score.

Ono et al. [ONO, 1994] ont proposé l'une des méthodes les plus simples de résumé automatique basées sur la RST. Dans leur algorithme, la racine de l'arbre RST est attribuée un score équivalent au nombre de niveaux hiérarchiques de l'arbre. Puis en traversant l'arbre depuis la racine vers les nœuds terminaux, à chaque niveau est associé le score de son niveau supérieur. Chaque fois qu'on part par un satellite, le score de niveau est diminué de un. Ensuite les segments sont classés (ordre partiel) selon leur score final. Cette méthode a été évaluée manuellement en vérifiant le nombre de phrases pertinentes incluses dans le résumé. Les auteurs ont rapporté que dans les meilleurs des cas, 51% des phrases pertinentes ont été sélectionnées et que dans 74% des cas, le résumé produit contient la phrase la plus importante de chaque document.

Pour O'Donnell [O'DON, 1997], l'importance des relations rhétoriques a été prise en compte pour classer les segments. L'auteur a assigné un facteur d'importance entre 0 et 1 pour chaque relation. Ce facteur a été utilisé dans la pondération des segments satellites. Plus tard, Marcu [MARC, 2000a] a proposé l'utilisation des ensembles de promotion pour déterminer les segments les plus importants de l'arbre. L'ensemble de promotion pour chaque nœud dans l'arbre RST est construit des éléments nucléaires de ses nœuds fils nucléaires. L'ensemble de promotion d'un nœud terminal (feuille) est composé du nœud lui-même. Ces ensembles sont ensuite utilisés pour le classement des segments. Dans ses expériences, Marcu a rapporté que sa méthode a été évaluée en utilisant seulement 5 documents textuels. Les résultats ont révélé que la méthode permettait de déterminer les unités textuelles les plus importantes avec des taux de rappel et de précision qui avoisine les 70%. Un peu plus tard, Uzeda [UZED, 2010] a étendu cette méthode par l'intégration des scores basés sur les mots clés issus de GistSumm [PARD, 2003].

Récemment Hirao et al. [HIRA, 2015] ont proposé une méthode qui utilise des arbres de discours de dépendance au lieu d'utiliser l'arbre RST. Les arbres de discours utilisés sont obtenus en transformant les arbres RST en arbre de dépendances qui expriment explicitement les relations de dépendance entre les unités textuelles. Produire un résumé avec cette méthode revient à chercher le sous arbre optimal à partir de l'arbre du discours.

2.5 Approches basées sur l'analyse sémantique

L'analyse et les corrélations sémantiques entre les phrases ont été également exploitées pour produire des résumés automatiques [HOVY, 1998; BLAI, 2006; BAWA, 2008]. L'analyse sémantique consiste à représenter le texte par ses entités textuelles et à modéliser les relations entre ces entités [BAWA, 2011]. Les entités textuelles peuvent être des termes simples, des séquences de mots (n-grammes), ou des phrases. Les relations entre ces entités peuvent prendre la forme de similarités, d'implications textuelles (*textual entailements*), ou de relations de cooccurrence [MCKE, 1999].

Dans [BAWA ,2008] la similarité entre la requête et les phrases du document est utilisée pour pondérer les phrases. Chaque phrase a été attribué un score de pertinence qui dépend de sa similarité sémantique avec la requête utilisateur. Ce score est ensuite utilisé pour sélectionner les phrases les plus pertinentes pour construire le résumé.

La similarité sémantique est calculée en se basant sur la méthode de chevauchement des mots telle qu'elle est implémenté dans [HOVY, 1998]. Dans sa version originale, la méthode de chevauchement des termes applique son analyse sur les termes explicitement mentionnés dans les documents et ne considère pas la relation de synonymie entre les mots. Par exemple, si un document contient les termes «maman» et «mère» chaque mot serait traité indépendamment de l'autre. Pour remédier à ce problème les auteurs ont utilisé WordNet [FELL, 1998].

WordNet [FELL, 1998] est une base de données lexicale pour la langue anglaise. WordNet regroupe les mots en ensembles de synonymes appelés synsets et fournit des définitions générales courtes appelées gloses. Il détermine, ainsi, les différentes relations sémantiques entre ces ensembles de synonymes.

Les mesures de similarités sémantiques basées sur WordNet ont été introduites par [JIAN, 1997] et [LIN, 1998]. Ces mesures ont été largement utilisées dans plusieurs systèmes de résumé automatique [HOVY, 1998 ; BAWA, 2008 ; WANG, 2008].

Barzilay et Elhadad [BARZ, 1999] ont proposé une méthode de résumé automatique basée sur les chaînes lexicales. Les chaines lexicales sont des séquences des mots liées par des relations lexico-sémantiques [TORR, 2011]. Les auteurs ont utilisé des algorithmes pour calculer les chaînes lexicales dans le texte source. Ces algorithmes utilisent WordNet pour trouver des mots et des chaînes liés sémantiquement. D'autre méthodes de résumés basées sur les chaines lexicales ont également été proposées dans [DORA, 2004 ; POUR, 2012].

L'implication textuelle (*Textual Entailment*) est une autre forme de relation qui peut être définie entre les entités textuelles. L'implication textuelle a été introduite comme un cadre général pour la modélisation de la variabilité sémantique dans plusieurs tâches de traitement du langage naturel [LIOR, 2008]. Une relation d'implication consiste à déterminer si le sens d'une phrase peut être déduit d'une autre. Formellement, on dit : Le texte T→H Si le sens de H peut être inféré du sens de T.

Ce type de relations permet de trouver les phrases fortement liées dans le document. L'implication textuelle a été utilisée par un certain nombre de chercheurs dans le résumé automatique de texte afin d'améliorer la qualité des résumés générés et d'éliminer la redondance. Lioret et al. [LIOR, 2008] ont montré l'importance des relations d'implication textuelle dans le résumé automatique, et que l'utilisation de ce type de relations conjointement avec une approche de résumé conduit à une amélioration significative des performances globale du système.

3. Le résumé automatique de textes Arabes

Les recherches sur le résumé automatique de textes arabes ont commencé à apparaître seulement depuis les années deux milles et par conséquent, elles n'ont pas encore progressé aussi vite que celles effectuées pour d'autres langues tel que l'anglais. Douzidia et al. [DOUZ, 2004] étaient parmi les premiers chercheurs à s'intéresser à ce domaine lorsqu'ils ont développé LAKHAS, le premier système de résumé automatique de textes Arabes. LAKHAS génère des résumés très courts d'environ dix mots afin de satisfaire les conditions proposé par DUC 2004. Les auteurs ont proposé une méthode statistique pour développer LAKHAS. Pour s'assurer que le résumé final soit très court et respecte la taille requise par DUC 2004, les auteurs ont utilisé quatre méthodes de réduction de phrases afin de compresser les phrases du résumé. Il est à noter que le résumé de texte Arabe n'était pas inclus dans les tâches de DUC-2004. Pour surmonter ce problème, les auteurs ont traduit le corpus anglais de DUC-2004 en Arabe en utilisant un outil de traduction automatique. Pour pouvoir évaluer les résumés, ils ont traduit

encore une fois les résumés générés de l'Arabe vers l'Anglais. Tous les systèmes de la compétition ont été évalués en utilisant les mesures ROUGE. LAKHAS a été classé entre la 5ème et la 6ème place parmi les systèmes concurrents. D'autre travaux allant dans la même direction de LAKHAS ont été également proposés, nous citons à tire d'exemple les travaux de Alotaiby et al. [ALOT, 2012].

Les techniques de l'apprentissage automatique ont attiré l'attention de plusieurs chercheurs travaillant dans le domaine du résumé automatique de textes Arabes [SOBH, 2006; FATTAH, 2009 ; BOUD, 2010 ; BELK, 2015]. Sobh et al. [SOBH, 2006] ont combiné les algorithmes génétiques avec le classifeur bayésien naïf pour construire leur système de résumé automatique. Plus tard, Fattah et Ren [FATT, 2009], ont étudié l'utilisation de plusieurs méthodes de classification y compris: les réseaux de neurones probabilistes (PNN), les algorithmes génétiques (GA), les modèles de mélanges Gaussiens (GMM), les réseaux de neurones Feed Forward (FFNN) et la régression linéaire (MR) pour la tâche de résumé automatique de textes Arabes. Les machines à vecteurs support (SVM) ont été aussi utilisées dans [BOUD, 2010]. Un peu plus tard, Belkebir, et al. [BELK, 2015] ont proposé une méthode supervisée en utilisant AdaBoost pour produire des extraits Arabes.

Bien que les techniques de clustering ont démontré un grand succès dans le résumé automatique multi-documents dans les langues latines, à notre connaissance, peu de travaux ont appliqué ces technique dans le résumé automatique multi-documents en langue arabe [EL HA, 2011, OUFA, 2014, NOOR, 2014].

Dans [EL HA, 2011], les auteurs ont examiné l'utilisation du clustering dans le résumé multi-documents pour éliminer la redondance. Deux expériences ont été effectuées. Dans la première expérience, quelques phrases ont été sélectionnées de manière aléatoire en tant que centroïdes initiaux, puis toutes les phrases ont été attribuées aux groupes (*clusters*) les plus proches en fonction de leur similarité aux centroides. Pour produire le résumé, les auteurs ont utilisé deux méthodes de sélection : Dans la première méthode, la première phrase de chaque groupe est sélectionnée, tandis que dans la seconde, toutes les phrases du plus grand cluster sont sélectionnées. Pour la deuxième expérience, la première phrase de chaque document et la phrase qui lui est la plus similaire sont sélectionnées avant l'étape de clustering qui est effectué en employant l'algorithme K-means. Toutes les étapes ultérieures sont similaires à la première expérience. Pour l'évaluation, les auteurs ont utilisé l'ensemble de données DUC-2002 et une version parallèle traduite en Arabe. Les résultats d'évaluation en termes de ROUGE-1 ont démontré que les modèles proposés atteignent les meilleures performances. Dans le même

contexte, Oufaida et al. [OUFA, 2014] ont proposé une méthode de résumé mono et multi-documents basée sur le clustering et une méthode d'analyse discriminante adaptée. Fejer et al. [FEJE, 2014] ont proposé un modèle basé sur une approche hybride de clustering (partitionnement et k-means). Le modèle regroupe des documents arabes en plusieurs clusters, puis il extrait les phrases-clés de chaque cluster pour construire le résumé. Le modèle a obtenu de bons résultats pour le résumé mono et multi-documents, mais aucune comparaison avec d'autres systèmes n'a été effectuée.

Par ailleurs, très peu de travaux ont été rapportés sur le résumé automatique de textes Arabes à base de graphes [AL TA, 2014 ; ALAM, 2017]. Récemment Alami et al. [ALAM, 2017] ont proposé une méthode de résumé mono-document basée sur les graphs pondérés. Chaque paire de phrases (représentée par des nœuds dans le graphe) est reliée par deux arcs qui représentent les mesures de similarité sémantique et statistiques. La mesure de similarité statistique s'appuie sur la méthode de chevauchement de termes entre les deux phrases, tandis que la mesure sémantique est basée sur des informations sémantiques extraites de la base lexicale Arabic WordNet (AWN).

Le score associé à chaque phrase est fonction de son score généré en appliquant *l'algorithme page Rank* et un ensemble de caractéristiques statistiques tel que TF-ISF ainsi que la position de la phrase.

Les phrases les mieux classés sont sélectionnées ensuite pour construire le résumé. Pour la redondance et la diversité de l'information les auteurs ont utilisé la méthode de pertinence marginale maximale (MMR). Les résultats des expérimentations sur le corpus EASC [EL HA, 2010] sont très satisfaisants.

La base lexicale AWN a été également utilisée par Imam et al. [IMAM, 2013] pour enrichir la requête utilisateur dans son système de résumé automatique appelé OSSAD. De façon générale, on peut dire que l'exploitation des relations sémantiques dans le résume automatique de documents arabes est encore à ses débuts. A notre connaissance, le seul travail qui a utilisé les relations d'implication textuelle dans le résumé Arabe est celui de Al-Khawaldeh et Samawi [AL KH, 2015]. Dans leur système appelé LCEAS, les auteurs ont essayé d'aborder le problème de la redondance dans les extraits arabes en s'appuyant sur les relations d'implication textuelle. L'algorithme d'implication textuel déjà suggéré dans [TATA, 2009] a été amélioré afin de l'adapter à la langue Arabe. La cohésion lexicale a été appliquée pour distinguer les phrases pertinentes. Par conséquent, les informations inutiles ont été supprimées avant d'appliquer l'algorithme d'implication textuelle.

De même, les recherches dans le domaine de l'analyse du discours Arabe non plus progressé vu le manque de ressources ainsi que les particularités linguistiques de la langue Arabe qui posent des défis sérieux aux chercheurs. Néanmoins, quelques efforts ont été déployés ces dernières années afin de créer des systèmes de résumé de textes arabes basés sur la RST [AZMI, 2012 ; IBRA , 2013, MAAL, 2012].

Dans [AZMI, 2012] les auteurs ont proposé une méthode hybride à deux phases. Dans la première phase, une analyse rhétorique de texte est effectuée afin de générer sa structure rhétorique à base duquel le résumé primaire est généré. Dans la deuxième phase, chaque phrase dans le résumé primaire reçoit un score basé sur quelques caractéristiques statistiques et linguistiques. Les phrases ayant les scores les plus élevés sont sélectionnées par la suite pour produire le résumé final tout en tenant compte du ratio de compression fixé par l'utilisateur. Dans le même contexte, Ibrahim et al. [IBRA, 2013] ont proposé un modèle qui combine la RST avec le modèle de l'espace vectoriel (VSM). Le modèle découvre d'abord les paragraphes les plus importants en se basant sur des critères sémantiques liée à la RST, puis ces paragraphes sont classés en se basant sur le modèle de l'espace vectoriel (VSM). Les auteurs ont montré que le modèle hybride combinant la RST et le VSM est capable de prendre les avantages des deux. Un résultat similaire a été aussi obtenu par Maaloul [MAAL, 2012] dans l'implémentation de son système appelé *Allkhas al Ali* (L.A.E) qui combine la RST avec les machine à vecteurs support (SVM). Après avoir effectué une analyse rhétorique de texte à résumer, L'algorithme SVM est utilisé pour décider si deux unités textuelles reliées par la relation "Autres" sont pertinentes ou non pour l'extrait final. Cette relation est assignée lorsque la relation rhétorique entre deux unités textuelles n'est pas reconnue. Les résultats obtenus confirment que l'application d'une approche hybride pourrait améliorer les performances du système de résumé. A part leurs utilisation dans le résumé automatique, la RST a été aussi utilisé récemment dans le système de question/ réponse [AZMI, 2017]

La théorie des représentations discursives segmentées (SDRT) a été également utilisée dans [KESK, 2015] pour générer des résumés automatiques de textes Arabes. Les auteurs ont proposé une approche sémantique pour analyser le discours arabe suivant le cadre de la SDRT. Ils ont montré aussi comment la structure du discours peut être utilisée pour produire des résumés indicatifs de documents Arabes.

Enfin, Nous finissons cette section par le tableau 2-1 qui présente un résumé des travaux Arabes étudiés dans un ordre chronologique.

Travail de Recherche	Année	Approche	Type	Méthode d'évaluation	Corpus
[DOUZ, 2004]	2004	statistique	Mono-document	ROUGE	DUC 2004
[SOBH, 2006]	2006	Classifieur bayésien et AG	Mono- document	Précision Rappel f-score	Corpus de l'auteur
[FATT, 2009]	2009	GA, MR, FFNN, PNN, GMM	Mono- document	ROUGE1 Rappel précision	Corpus de l'auteur
[BOUD, 2010]	2010	SVM	Mono- document	Précision Rappel f-score	Corpus de l'auteur
[EL HA, 2011]	2011	Clustering (K-means)	Multi-document	ROUGE-1, Rappel, précision	DUC 2002 corpus and Arabic translated version
[ALOT, 2012]	2012	statistique	Mono-document	ROUGE	Arabic Gigaword
[AZMI, 2012]	2012	RST	Mono- document	Précision, rappel, F-score, Mesures ROUGE	Corpus de l'auteur
[IBRA, 2013]	2013	RST et VSM	Mono- document	Rappel, précision, F-score	Corpus de l'auteur
[IMAM, 2013]	2013	Arbre de décision (C4.5) + AWN	Mono-document	ROUGE-L	Corpus de l'auteur + EASC

[OUFA, 2014]	2014	Clustering+ MRMR	Mono et multi-documents	ROUGE-1 ROUGE-2	EASC, TAC2011 MultiLing Pilot
[NOOR, 2014]	2014	Clustering	Mono et multi-documents	ROUGE	EASC
[AL TA, 2014]	2014	GRAPH	Mono document	Précision Rappel f-score	EASC
[BELK, 2015]	2015	Adaboost	Mono-document	Précision Rappel F-score	Corpus de l'auteur
[KESK, 2015]	2015	SDRT	Mono-document	Précision Rappel F-score	Corpus de l'auteur, Arabic Treebank (ATB v3.2 part3)
[AL KH, 2015]	2015	Cohésion lexical, Implication textuelle	Mono et multi document	ROUGE-2 ROUGE-L ROUGE-W ROUGE-S AutoSummEng	Corpus de l'auteur, EASC
[ALAM, 2017]	2017	Graphe	Mono-document	ROUGE	EASC

Tableau 2-1: Synthèse des Travaux Etudié dans L'ordre Chronologique [LAGR, 2018]

Il est à noter que, nous ne pouvons pas comparer les résultats obtenus par ces travaux, car ces systèmes ne sont pas évalués sur le même corpus et en utilisant la même méthode d'évaluation. Comme nous pouvons le voir dans le tableau 2-1, dans la majorité des travaux les auteurs ont utilisé leur propre corpus pour évaluer leurs systèmes. Cela est dû essentiellement au manque de corpus de référence standard pour le résumé automatqiue de textes Arabes pendant plusieurs années.

4. Corpus disponibles pour le résumé automatique en langue Arabe

Plusieurs efforts ont été déployés pour remédier au problème de manque de corpus de référence standard pour le résumé automatique de textes Arabes. El Haj et al. [EL HA, 2010] ont créé un corpus de résumé mono-document appelé EASC (*Essex Arabic Summaries Corpus*). Le corpus est composé de 153 articles collectés de Wikipedia et de deux journaux arabes. Chaque document dans EASC est associé à cinq résumés de référence créés manuellement par différentes personnes. Comme ce corpus ne permet pas aux chercheurs de comparer leurs résultats avec d'autres systèmes qui ont produit de bons résultats lors de campagne d'évaluation tel que DUC, les auteurs se sont orienté en 2011vers la construction d'un autre corpus par la traduction automatique de corpus de résumé multi-documents DUC 2002 de l'anglais vers l'Arabe [EL HA, 2012]. DUC 2002 a été traduit phrases par phrases en utilisant Google traduction afin de créer une version parallèle, ce qui permet de comparer les résultats avec des systèmes de résumés non seulement Arabe mais aussi Anglais.

Le tableau 2-2 présente quelque corpus disponibles pour le résumé automatique de textes Arabes. Pour plus de détails sur la construction de ces corpus, Nous renvoyons le lecteur à [LI, 2013 ; EL HA, 2012 ; EL HA, 2015].

Nom de corpus	Nombre de documents	Nombre de résumés
EASC	153	765 résumés manuels (5 résumés par document)
DUC 2002 (traduit en Arabe)	567	118
TAC 2011 MultiLing (Arabe)	100	30 résumés : (3 résumés manuels pour chaque 10 document)
TAC 2013 MultiLing pilot (Arabe)	150	3 résumés manuels pour chaque 15 document

Tableau 2-2 : Quelques Corpus disponibles pour le Résumé Automatique de textes Arabes.

5. Conclusion

Dans ce chapitre, nous avons exploré les différentes approches proposées pour produire des extraits automatiques. Nous avons vu que plusieurs méthodes et techniques ont été utilisées pour évaluer la pertinence des segments textuelles. Les méthodes numériques sont généralement assez simples à implémenter, mais elles restent toujours limitées et confrontées aux barrières de la langue puisque ces méthodes ne peuvent être testées qu'avec la manipulation de valeurs numériques liées à des propriétés d'objets à la surface de textes [BLAI, 2008]. Les méthodes linguistiques basées sur l'exploitation en profondeur de la structure linguistique et discursive de textes semblent prometteuses pour faire avancer qualitativement les systèmes de résumés automatiques. Néanmoins, ces méthodes nécessitent beaucoup de ressources linguistiques avancées ce qui représente un véritable obstacle devant les chercheurs travaillant dans le domaine de résumés automatique de textes Arabes.

Les méthodes hybrides qui combinent à la fois des techniques linguistiques et statistiques donnent de meilleurs résultats, mais avec ces méthodes aussi la cohérence des résumés produits reste toujours à améliorer. Le manque de cohérence des résumés produits est dû essentiellement à la négligence des relations rhétoriques qui relient les phrases de textes source. Pour remédier à cette lacune, nous proposons une méthode hybride basée principalement sur l'exploitation des relations rhétoriques reliant les unités de discours élémentaires. Cette approche sera présentée en détail dans le cinquième chapitre de cette thèse.

Chapitre 3

Contribution 1

Annotation Rhétorique d'un Corpus Arabe

Sommaire

1. Introduction

Comme déjà mentionné dans le chapitre 2, notre approche de résumé automatique tient compte principalement de l'exploitation des relations rhétoriques reliant les unités de discours élémentaires dans le texte. Afin de pouvoir exploiter ces relations, une première étape consiste, d'abord, à en faire une identification façon automatique. Pour cela nous avons proposé une approche supervisée pour l'identification automatique des relations rhétoriques Arabe (cf. Chapitre 4). Notre approche a nécessité un corpus de discours arabe annoté selon le cadre de la théorie de la structure rhétorique. Vue la non disponibilité d'une telle ressource pour la langue arabe, il était nécessaire d'envisager la construction manuelle de ce type de corpus.

Dans ce chapitre nous allons focaliser sur l'annotation manuelle de notre corpus de discours arabe selon le principe de la théorie de la structure rhétorique (RST). Ce processus commence par la segmentation du texte en unité de discours élémentaire (EDUs). Les EDUs sont principalement des propositions verbales ou nominales. L'étape suivante consiste à l'annotation des relations rhétoriques qui existent entre deux EDUs adjacentes. A ce niveau, nous nous sommes limités à l'annotation des relations intra-phrases, c'est à dire les relations entre les EDUS adjacents dans une même phrase. Pour accomplir cette tâche, nous avons d'abord élaboré une liste des relations rhétoriques arabes en nous appuyant sur trois piliers : la rhétorique Arabe, les relations RST déjà définies dans des campagnes d'annotation antérieures et l'analyse de corpus.

2. La construction du corpus

Pour construire notre corpus nous avons sélectionné une grande collection d'articles extraits d'un grand corpus Arabe non étiqueté disponible en ligne pour exploration appelé *Arabic Corpus*[1]. Ce corpus est composé de plusieurs ressources réparties en cinq catégories: articles de presse, littérature moderne, non-fiction, familier égyptien et le prémoderne. Le corpus formé d'articles de presse contient environ 135.360.804 mots tirés de plusieurs journaux[2] Arabes publiés entre 1996 et 2010.

[1] *http://arabicorpus.byu.edu*

[2] *Les journaux sont : Al-Hayat de Londres, Al-Ahram d'Egypte, Al-Watan du Koweït, Al-Thawra de Syrie, Al-Ghad de Jordanie Al-Masri Al-Yawm ,* Shuruq d'Egypte, *At-Tajdid de Moroc.*

Nous avons sélectionné au hasard 200 articles de presse de ce corpus vu qu'il couvre plusieurs thèmes. Ces articles sont d'abord convertis sous format textuel puis annotés manuellement par deux experts. Les documents de notre corpus ne sont pas trop longs (environ 22 phrases par document) mais les phrases sont longues et syntaxiquement complexes. L'exemple suivant présente une phrase extraite de notre corpus.

' والآن، وحيث فشلت القيادة في اثبات مصدقيتها وانتمائها الى الديموقراطية، باتت الحاجة ملحّة لرؤية الحركة بقامتها الفعلية، ومؤهلاتها الحقيقية ، وكشف المستور من عيوبها، بعيداً عن هالات التمجيد وصراع التبجح الفارغة.'

'Et maintenant que la gouvernance n'a pas réussi à prouver sa crédibilité et son caractère démocratique, il est urgent de voir le mouvement dans sa dimension effective, ses vraies qualifications et la détection cachée de ses défauts, loin des halos de glorification et des luttes d'arrogance vides.'

3. Etapes de l'annotation rhétorique

L'annotation de notre corpus est basée sur la théorie de la structure rhétorique (RST) [MANN, 1988]. Ce choix est motivé par le fait que la RST permet de générer des annotations riches qui capturent non seulement les relations rhétoriques, intentionnelles et sémantiques qui relient les segments textuels ainsi elle fournit aussi une information sur l'importance relative des segments mis en jeu par ces relations dans le texte.

L'annotation de notre corpus implique trois tâches principales :

1. Déterminer les relations rhétoriques intra-phrase dans la langue Arabe et les organiser dans une nouvelle hiérarchie,
2. Elaborer le manuel d'annotation,
3. Annoter manuellement tous les documents du corpus.

Usuellement, dans le cadre de la RST, l'annotation rhétorique d'un texte passe par trois étapes :

1. Segmentation du texte en unités de discours élémentaires (EDUS) ;
2. Identification des relations qui existent entre les unités de discours ;
3. Construction de la structure rhétorique de texte (arbre RST).

Comme nous nous focalisons principalement sur l'annotation des relations rhétoriques intra-phrases, la troisième étape, à savoir, la construction de la structure rhétorique de texte a été ignorée dans notre processus d'annotation.

Ainsi, Pour accomplir les trois tâches susmentionnées, nous avons sélectionné 50 documents pour la détermination des relations rhétoriques arabes et 15 autres documents pour mesurer le degré d'accord entre annotateurs.
Dans les sections qui suivent, nous allons détailler chacune des tâches impliquées dans notre travail d'annotation.

3.1 Détermination des relations rhétoriques arabes

3.1.1. Méthodologie suivie

Dans le domaine de l'analyse de discours, il n'y a pas un consensus standard ni sur le nombre de relations rhétoriques ni sur leur taxonomie. Ainsi, Chaque théorie de discours définie sa propre liste de relations en se basant sur ses propres critères. Dans le cadre de la RST, les relations rhétoriques sont définies en se basant sur des critères fonctionnels et sémantiques et non sur des signaux morphologiques ou syntaxiques [TABO, 2006]. Ce principe nous a amené à suivre deux axes en parallèle dans notre recherche : étudier les relations RST définies dans les campagnes d'annotation antérieures et mener une recherche intensive dans le domaine de la rhétorique arabe (علم البلاغة) afin d'appréhender ses sens rhétoriques. L'objectif fondamental étant de déterminer si les relations RST définies dans les campagnes d'annotation antérieures peuvent être adaptées pour tenir compte des particularités de la langue Arabe. D'autre part, essayer de découvrir d'autres relations rhétoriques spécifiques pour la langue arabe.

Nous avons choisi de commencer par l'ensemble des relations de base définies dans le RST-DT [CARL, 2003]. Cet ensemble comporte 53 relations reparties en 16 catégories. Le tableau 3-1 présente ces relations avec leur taxonomie.

Parmi ces relations, nous nous somme focalisé sur les relations intra-phrases c'est-à-dire celles qui relient généralement les unités de discours adjacentes au sein de la même phrase. Les relations inter-phrases qui joignent généralement des phrases ou des paragraphes tels que *problème-solution, Résumé, arrière- plan, ..etc*. sont écartés.

Class	Relation
Attribution	*Attribution, Attribution-negative*
Background	*Background, Circumstance*
Cause	*Cause, Result, Consequence*
Comparison	*Comparison, Preference, Analogy, Proportion*
Condition	*Condition, Hypothetical, Contingency, Otherwise*
Contrast	*Contrast, Concession, Antithesis*
Elaboration	*Elaboration-additional, Elaboration-general-specific, Elaboration-part-whole, Elaboration-process-step, Elaboration-object-attribute, Elaboration-set-member, Example, Definition*
Enablement	*Purpose, Enablement*
Evaluation	*Evaluation, Interpretation, Conclusion, Comment*
Explanation	*Evidence, Explanation-argumentative, Reason*
Joint	*List, Disjunction*
Manner-Means	*Manner, Means*
Topic-Comment	*Problem-solution, Question-answer, Statement-response, Topic-comment, Comment-topic, Rhetorical-question* .
Summary	*Summary, Restatement.*
Temporal	*Temporal-before, Temporal-after, Temporal-same-time, Sequence, Inverted-sequence*
Topic Change	*Topic-shift, Topic-drift*

Tableau 3-1 : Liste des relations de Base dans RST-DT [CARL, 2003]

Les relations intra-phrases sélectionnées sont ensuite raffinées en s'appuyant sur une analyse de corpus et les sens rhétoriques arabes fournis par des études antérieures sur la rhétorique arabe [ABDU, 2006 ; AL JA, 2012]. Pour accomplir cette tâche, nous avons fait

intervenir trois professeurs en linguistique Arabe. Nous leur avons fourni tout d'abord un aperçu général sur le principe de la RST, ainsi qu'une définition précise pour chaque relation telle qu'elle est définie dans le manuel d'annotation de Carlson et al. [CARL, 2001], puis nous avons leur demandé d'analyser comment chacune de ces relations est instanciée au sein de notre corpus et quel est son sens rhétorique Arabe correspondant. Il en résulte quatre situations qui peuvent se produire:

1. La relation a un sens rhétorique arabe qui lui correspond et elle est instanciée dans notre corpus. Dans ce cas, cette relation est sélectionnée pour être rajoutée à l'ensemble des relations rhétoriques Arabe, et nos experts devraient déterminer comment cette relation est signalée dans le corpus explicitement (déterminer ses marqueurs de discours possibles).
2. La relation n'a pas un sens rhétorique Arabe, mais elle est instanciée dans notre corpus. Dans ce cas, cette relation est sélectionnée et elle est définie en se basant sur son rôle fonctionnel et sémantique instancié dans le corpus. C'est le cas par exemple de la relation « *كيفية/ manière*»
3. Quelques relations de l'ensemble des relations de la RST-DT correspondent à un seul sens rhétorique Arabe. Dans ce cas, ces relations doivent être généralisées et une nouvelle relation qui correspond au sens rhétorique doit être créée et rajoutée à l'ensemble des relations rhétoriques Arabe. Par exemple la relation زمني/temporel a été créée comme une généralisation des relations : Temporelle-avant, Temporelle-après et Temporale-même temps.
4. La relation n'a pas un sens rhétorique Arabe correspondant et elle n'est pas instanciée dans le corpus. Dans ce cas, cette relation est supprimée. C'est le cas de la relation « *enablement* » par exemple.

Ensuite, nos experts ont été sollicités pour rajouter des relations rhétoriques définies dans la rhétorique Arabe qui sont instanciées dans notre corpus mais dont la sémantique ne correspond à aucune relation dans la RST-DT. Cela a abouti à l'ajout de deux relations : تشبيه /analogie et اضراب /rectification.

Enfin et étant donné qu'il n'y a pas de consensus standard sur la taxonomie des relations rhétoriques, nous avons demandé à nos experts de classer les relations rhétoriques Arabes obtenues en fonction de leurs rôles fonctionnels. Cette étape a abouti à une nouvelle taxonomie de 7 classes: السببية / Causal, المقارنة / Comparaison, العطف / coordination, تفصيل / Elaboration,

التوضيح / Explication, الإسناد / Attribution, الشرط / Conditionnel, avec un total de 20 relations tel qu'illustré dans la figure 3-1.

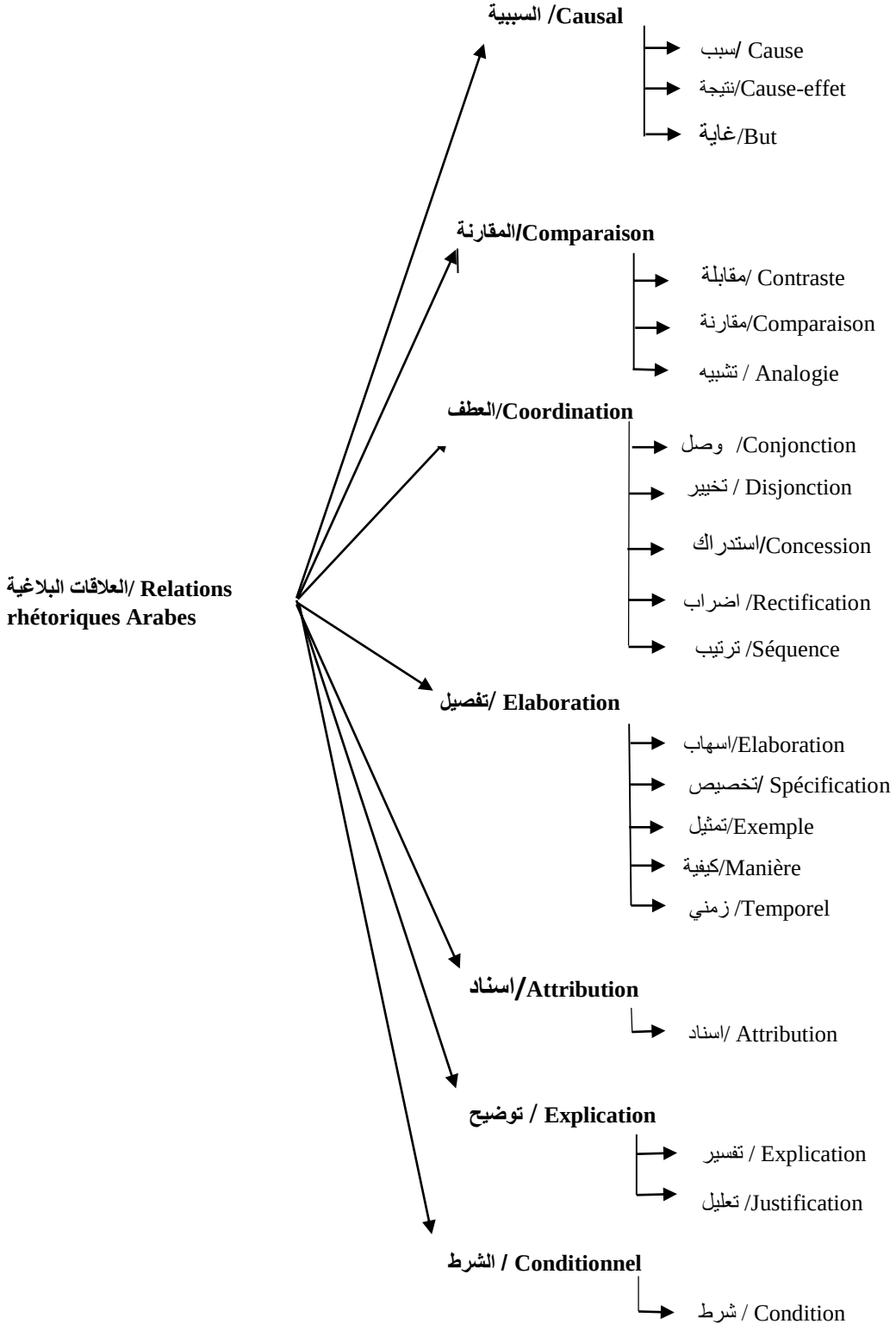

Figure 3-1 : Taxonomie des Relations Rhétoriques Arabes [LAGR, 2018a].

3.1.2. Définition des relations rhétoriques Arabes

Dans cette section ; nous allons présenter une brève définition de chacune des relations rhétoriques arabes. Le nom de chaque relation est suivi de son type: mononucléaire ou multi-nucléaire. Rappelons qu'une relation mononucléaire relie un EDU noyau et un autre satellite. Tandis qu'une relation multi-nucléaire ne joint que des segments noyaux.

Les notations utilisées sont les suivantes :

- R.N-S (x, y) : signifie que les EDU X et Y sont reliés par la relation mononucléaire R. EDU X est le noyau et EDU Y est le satellite.
- R.S-N (x, y) signifie que les EDU X et Y sont reliés par la relation mononucléaire R. EDU X est le satellite et EDU Y est le noyau.
- R.N-N (X , Y) signifie que les deux EDUs X et Y sont reliés par la relation multi-nucléaire R.

1. سبب/cause (mononucléaire)

Définition : Dans la relation ' **سبب /cause'**, La situation présentée dans le satellite est la cause de la situation présentée dans le noyau. C'est-à-dire, le satellite spécifie pourquoi un évènement s'est produit dans le noyau. (cf. exemple 1)

(1) [هنالك عشرات آلاف الوحدات السكنية الخالية،] 1 [بسبب ارتفاع اثمانها، التي لا تتناسب ودخل المواطنين.] 2

(1)[Il y a des dizaines de milliers de logements vacants,] 1 [en raison de leurs prix élevés, qui ne correspondent pas aux revenus des citoyens.] 2

(2 ,1) S-N. سبب/Cause (1)

2. نتيجة/ cause-effet (multi-nucléaire)

Définition : Dans cette relation la situation présente dans un EDU est le résultat de la situation présente dans l'autre. Autrement dit, C'est une relation causale dans laquelle le premier EDU représente la cause et l'autre représente l'effet (cf. exemple 2). Les deux EDUs sont d'importances égales.

(2) [إن ازدياد حجم الطلب وقلة العروض]1 [يؤدي الى زيادة نشاط سوق الأسهم،]2

(2)[L'augmentation de la demande et la diminution de l'offre]1[entraine une augmentation dans la bourse des actions]2

(2,1) N-N نتيجة/cause –effet (2)

3. غاية/ But (mononucléaire)

Définition : Dans cette relation, la situation présentée dans le satellite est le but ou la motivation derrière la situation présentée dans le noyau (cf. exemple 3).

(3) [عدد كبير من الأزواج اليوم يختارون الصمت او المسايرة]1 [من اجل تجب نوبات الغضب لدى الزوجات]2

(3) [un grand nombre d'hommes choisissent aujourd'hui le silence ou l'appariement] 1[afin d'éviter les crises de colères de leurs femmes] 2

(2,1) S-N. غاية/But (3)

4. مقابلة / Contraste (multi-nucléaire)

Définition : dans la relation ' مقابلة/CONTRASTE', deux noyaux se contredisent. Dans la langue Arabe, cette relation est signalée par la présence de plus d'un mot dans le premier EDU et leurs antonymes dans le second EDU (cf. exemple 4).

(4) [ضحكت قليلا]1 [و بكيت كثيرا]2

(4) [Elle a peu ri]1 [et beaucoup pleuré]2

(2,1) N-N. مقابلة / Contraste(4)

5. مقارنة/Comparaison (multi-nucléaire):

Définition **:** Elle est équivalente à la relation '*comparaison*' dans la RST-DT. Cette relation indique que certaines entités dans les deux segments reliées par la relation ' مقارنة / comparaison ' sont similaires, différentes, supérieures, inferieurs…etc. (voir exemple 5)

(5) [بلغت نسبة التخلف العقلي في الاحياء المتوسطة 3.3%]1 [بينما وصلت هذه النسبة الي 7.1% في الاحياء الفقيرة]2

(5)[Le taux de retard mental dans les quartiers moyens est de 3,3%]1[alors que ce taux atteint 7,1% dans les quartiers pauvres]2

(2,1)N-N.مقارنة/Comparaison(5)

6. تشبيه / Analogie (mononucléaire)

Définition :' فن التشبيه '/ l'art de l'analogie dans la rhétorique Arabe est un mode esthétique du discours dont le but pragmatique est de rapprocher deux significations les unes des autres et de comparer une entité donnée à une autre partageant les mêmes caractéristiques. Il est utilisé généralement pour réaliser la fonction de parabole (cf. exemple 6). On trouvera plus de détails sur ce mode de discours et ses différents types dans la rhétorique arabe dans [Abdul-Raof, 2006].

(6) [بكل وحشية فقد استل خنجره و تقدم نحوهم] 1 [وكأنه وحش ثائر لا يمد للإنسانية بصلة]2

(6) [Il empoignât son poignard avec une grande férocité et il s'avança vers eux]1 [comme si c'était une bête féroce n'ayant rien d'un être humain]2

(2,1) S-N.تشبيه/Analogie(6)

7. وصل/ Conjonction (multi-nucléaire)

Définition : C'est une relation de coordination multi-nucléaire dont les constituants sont listés en parallèle. Ces constituants sont généralement reliés par la conjonction de coordination ' و ' ainsi que par d'autre marqueurs linguistiques tels que كما , الى جانب, زيادة على ذلك…etc. comme illustré par l'exemple 7.

(7) [على المسلم ان يحرص دائما على افشاء السلام] 1 [و الصلاة بالليل و الناس نيام] 2

(7) [Un musulman se doit toujours d'être désireux de divulguer la paix]1[et prier la nuit alors que les gens sont endormis] 2

(2,1)N-N.وصل/Conjonction (7)

8. تخيير/Disjonction (multi-nucléaire)

Définition : c'est une relation multi-nucléaire dont les constituants peuvent être listés comme alternatives. Généralement, les deux EDUs concernés par cette relation sont reliés par la conjonction de coordination 'أو ' comme dans l'exemple 8.

(8) [تطوير الاقتصاد الوطني يعتمد على تطوير القطاع الصناعي] 1 [أو تنمية القطاع الزراعي أو كليهما]2

(8) [Le développement de l'économie nationale dépend du développement du secteur industriel]1[ou du développement du secteur agricole ou des deux] 2

(2,1)N-N. تخيير/Disjonction (8)

9. اضراب/ Rectification (multi-nucléaire)

Définition : Dans cette relation, la situation présentée dans le satellite indique un changement d'attitude ou une rectification dans la situation présente dans le noyau. Les deux constituants de cette relation sont généralement séparés par la conjonction 'بل' tel qu'illustré par l'exemple 9.

(9) [ان ضعف القطاع الصناعي «ليس قدرا» لا راد له،]1 [بل هو نتيجة عوامل يمكن التغلب عليها.]2

(9) [La faiblesse du secteur industriel n'est pas une fatalité incontournable,] 1 [mais c'est le résultat de facteurs qui peuvent être surmontés.]2

(2,1)N-N.اضراب/Rectification (9)

10- **ترتيب /séquence (multi-nucléaire)**

***Définition* :** C'est une relation multi-nucléaire qui impose un ordre chronologique entre les événements présents dans ses constituants (cf. exemple 10)

(10) [فقد نشأ في فلسطين] 1 [ثم تعلم في القاهرة] 2 [ودرس فيها]3

(10) [Il a grandi en Palestine]1[puis il a étudié au Caire]2[et il y a enseigné.]3

(2,1)N-N.ترتيب/séquence (10)
(3,2).N-N. وصل/Conjonction

11.استدراك /Concession (mononucléaire)

***Définition* :** dans cette relation, l'information présentée dans le satellite corrige une information mentionnée dans le noyau. Les deux constituants de cette relation sont généralement reliés par la conjonction 'لكن'. (cf. exemple 11)

(11) [وما ظلمناهم] 1 [ولكن كانوا أنفسهم يظلمون]2

(11) [Nous ne les avions pas lésés]1[mais ils se sont faits eux-mêmes du tort]2

(2,1)S-N.استدراك/ Concession(11)

12. اسهاب/Elaboration

Définition : c'est une relation mononucléaire très commune. Elle est équivalente à 'ELABORATION-ADDITIONAL' dans RST-DT. Dans cette relation, le satellite donne plus de détails ou ajoute des informations supplémentaires sur la situation présentée dans le noyau.

(12) [تشهد البيوت في عصرنا حالات صعبة،]1[حيث تتضارب الأمزجة بين أفراد العائلة فيتنافر الأخوة ....]2

(12) Les familles connaissent des situations difficiles, à notre époques,]1[où les humeurs contradictoires entre les membres de la famille mènent à des discordes entre frères]2

(2,1)S-N.اسهاب/Elaboration (12)

13 . تخصيص/ Spécification

Définition : C'est une relation dont la sémantique est proche de la relation 'اسهاب'. Elle est équivalente à la relation '**ELABORATION-PART-WHOLE'** dans RST-DT**.** Dans cette relation, le satellite spécifie ou élabore sur une partie du noyau. Elle est généralement signalée par des marqueurs linguistiques tels que خاصة/ surtout, بالأخص/ en particulier, على وجه الخصوص/en particulier.., comme dans l'exemple 13.

(13) [تسعى الدولة الى التوسع في الصادرات الى الخارج] 1[خصوصا القطن بتصدير نحو 5.1 مليون قنطار]2

(13) [Le pays cherche à développer les exportations vers l'étranger] 1[, en particulier les exportations de coton d'environ 5,1 millions de quintaux]2

(2,1)S-N.تخصيص/Spécification (13)

14. تمثيل/ Exemple (mononucléaire)

Définition : Dans cette relation, le satellite fournit un exemple en ce qui concerne les informations présentées dans le noyau. Elle est introduit généralement par على سبيل المثال/ à titre d'exemple, مثل/ tel que.., comme dans l'exemple 14.

(14) [اسقطت بعض البلدان اجزاء من ديونها على الأردن] 1 [مثل الولايات المتحدة والمانيا]2

(14) [Certains pays ont dissous une partie des dettes de la Jordanie]1[comme les États-Unis et l'Allemagne]2

(14)تمثيل.N-S(2,1)

15. كيفية /manière (mononucléaire)

Définition : équivalente à la relation '*manner*' dans RST-DT. Le satellite d'une telle relation spécifie la manière pour accomplir une tâche mentionnée dans le noyau. Il devrait nous informer comment une chose est faite ou doit être faite. En ce sens que le satellite réponde à la question «par quels moyens?» ou «comment?». Question pouvant être posée à propos d'une information présente dans le noyau. Cette relation est souvent signalée par des marqueurs tels que بواسطة/ par, باستعمال/ en utilisant, عن طريق/ par le biais de .. , comme dans l'exemple 15.

(15) [و قد نجح في السيطرة على المعارضة 1] [عن طريق استخدام وحدات من الجيش تتولى تنفيذ أوامره] 2

(15) […et il a réussi à maitriser l'opposition]1[en utilisant des unités de l'armée pour exécuter ses ordres]2

(2,1)S-N.كيفية/Mannière (15)

16. زمني/ Temporel (mononucléaire)

Définition : cette relation impose un ordre temporel entre les évènements. Elle est généralement signalée par des adverbes de temps tels que قبل/avant, بعد/ après…etc. Parfois le satellite de cette relation indique le temps d'un évènement présent dans le noyau comme dans l'exemple 16.

(16) [عشية الاحتفال بعيد الأضحى المبارك] 1 [تم اطلاق صراح العديد من المساجين] 2

(16) [La veille de la célébration de l'Aid El Adha,]1[Beaucoup de prisonniers ont été libérés.]2

(2,1)N-S. زمني / temporel(16)

17. اسناد/ Attribution (mononucléaire)

Définition : Cette relation concerne les instances du discours direct rapporté. Le noyau de cette relation est le message rapporté. Le satellite est la source de l'attribution appelé ' السند' qui est la clause qui contient le verbe de la parole tel que 'قال' /dire, ' أعلن'/annoncer, ' أكد'

/confirmer… etc, ou une phrase qui commence par 'حسب /selon' tel qu'illustré par les exemples 17 et 18.

(17) [اعلن الوزير] 1 [ان أسعار المواد الغذائية ستشهد ارتفاعا طفيفا]2

(17) [Le ministre annonce]1[que les prix des produits alimentaires vont connaitre de légères fluctuations]2

(18) [حسب التصريحات التي أدلى بها الرئيس]1 [فان موعد الانتخابات التشريعية سيكون نهاية الشهر المقبل]2

(18) [Selon les déclarations du président,]1[le rendez-vous électoral des législatives est fixé à la fin du mois d'avril prochain]2

(2,1)N-S.اسناد/Attribution (18)(17)

18. تفسير/ Explication(mononucléaire)

***Définition* :** dans cette relation, le satellite fournit une explication à propos d'une situation présente dans le noyau. Cette relation est généralement indiquée par les marqueurs ' أي ', ' بمعنى ' /c'est-à-dire , comme dans l'exemple 19.

(19) [العلم يتقدم نافيا ما سبقه،]1 [بمعنى أن آخر ما ينجزه العلم هو الأكثر صحة]2

(19) [La science avance dans la négation de ce qui précède] 1 [dans le sens où ce sont les dernières découvertes scientifiques qui ont le plus de crédibilité]2

(2,1)S-N. تفسير/Explication (19)

19. تعليل/ Justification (mononucléaire)

***Définition* :** le satellite d'une relation 'تعليل' justifie une situation mentionné dans le noyau, elle est souvent marqué par le marqueur 'لأن'. Autrement dit; le rôle de cette relation répond à la question «pourquoi?» qui peut être posée à propos d'une information présente dans le noyau comme dans l'exemple 20.

(20) [اعتزل الفن مؤخرا]1 [لأنه كان يعاني من مرض عضال]2

(20) [Il s'est récemment retiré de la scène artistique] [1] parce qu'il souffrait d'une maladie grave]2

(2,1)S-N. تعليل/ justification(20)

20. شرط/ Condition (multi-nucléaire)

Définition **:** c'est une relation multi-nucléaire dont la vérité de la proposition de l'un de ses constituants (جواب الشرط) dépend de l'accomplissement de la condition mentionnée dans son autre constituant tel qu'illustré par l'exemple 21. Dans la langue ; arabe cette relation est signalée par des marqueurs suivants { ان, اذا , لو , من,مهما, كيفما , أينما , اين,انى, ايان , أي , حيثما, اذا , كلما , لولا.}

La relation de condition peut aussi être marquée par la préposition 'حتى' précédée par une négation marquée par 'لن' comme dans l'exemple 22.

(21) [اذا استمر الوضع على هذا الحال] 1 [فسيخسر الفريق المباراة]2

(21) *[Si cette situation perdure]*$_1$*[, l'équipe va perdre le match]*$_2$

(22) [لن تنالوا البر] 1 [حتى تنفقوا مما تحبون]2

(22) [*Vous ne recevrez pas la miséricorde]*$_1$*[avant de dépenser de ce que vous aimez]*$_2$

(2,1)N-N.شرط/ Condition (21)

3.2 Elaboration du manuel d'annotation

Après avoir défini la liste des relations rhétoriques utilisées pour l'annotation de notre corpus, nous avons procédé à l'élaboration du manuel d'annotation. Ce manuel couvre toutes les instructions nécessaires à l'annotation de notre corpus.

Deux points cruciaux seront abordés dans le reste de cette section: la segmentation des textes en unités de discours élémentaires (EDUs) ainsi que comment déterminer la nucléarité des segments liés par une relation.

3.2.1 Segmentation des textes en unité de discours élémentaires

Afin de segmenter les textes en unités de discours élémentaires (EDUs), il était nécessaire d'établir un ensemble de règles et de principes pour guider le processus de segmentation.

Certaines de nos règles sont inspirées du manuel de référence de Carlson et al [CARL, 2001] utilisé pour annoter le RST-DT [CARL, 2003]. Ce manuel fournit aux annotateurs des instructions d'annotations détaillées qui reflètent une analyse trop méticuleuse à certains égards. Les règles inspirées de ce manuel sont ajustées pour prendre en compte les particularités de la langue arabe. Nous avons également ajouté des règles spécifiques pour la langue Arabe pour traiter certains cas particuliers tels que 'المفعول لاجله'/accusative de but,' 'المصدر' / la construction infinitive, la coordination,…etc.

Dans notre manuel, chaque règle est accompagnée d'exemples qui illustrent les cas nécessitant une segmentation ainsi que ceux qui n'en nécessitent pas.

Dans cette section nous allons présenter d'abord les principes de base puis les règles de segmentation accompagnée d'exemples illustratifs.

A. Principes de base

1) Chaque texte doit être d'abord segmenté en phrases. Une phrase est considérée comme un passage textuel qui se termine par un '.' (cf. exemple 23).

(23) وبأسلوب يدل على التواضع افتتح مقاله بالقول :ان هذه، على أية حال، هي صيغتي، مهما تكن قيمتها، واذا لم تكن ذات فائدة لأحد فلتنتهي في سلة المهملات .

(23) *D'une manière humble, il a entamé son article en disant: Ceci est, cependant, ma voie, quelle que soit sa valeur et si elle n'est utile à personne, elle finira à la poubelle.*

2) Chaque phrase doit être ensuite segmentée en EDUs délimités par des crochets en se basant sur les règles de segmentations (voir section B) .

2) Un EDU peut être une proposition nominale (جملة اسمية) qui commence par un marqueur de discours (cf. exemple 24) ou verbale (جملة فعلية) comme illustré dans l'exemple 25 . Les compléments de verbes ou de sujets ne sont pas traités comme EDUs (i.e. le verbe ne doit être séparé ni de son sujet ni de son complément. Une exception est faite pour les verbes de paroles).

(24) [نظرا لسوء احوال الطقس]

(24) [*En raison de mauvaises conditions météorologiques*]

(25) [الغيت كل الرحلات الجوية]

(25) [*Tous les vols ont été annulés*]

3) Les EDUs ne peuvent pas se chevaucher mais juxtaposés.

B. Les règles de segmentation

1) Les marqueurs de discours

- Toute proposition qui commence par un marqueur de discours fort doit être segmentée et traitée comme un EDU séparé (cf. exemple 26). Les marqueurs de discours sont des expressions ou signes linguistiques tels que' لكي' */pour,' بسبب '/ a cause de, ' يؤدي '/ conduit à,' من اجل '/pour...etc.*

(26)[علينا أن نلوذ بأفضل ما لدينا من معارف علمية وهندسية] [من اجل الوصول بالبشرية إلى بر الأمان.]

(26) [*Nous devons utiliser nos meilleures connaissances scientifiques et techniques] [pour mettre l'humanité en sécurité]*

Nous avons établi une liste des marqueurs de discours qui justifie la segmentation d'une phrase en EDUs. Nos marqueurs sont classés en deux catégories : forts et faibles.

Les marqueurs forts tels que ' *من اجل' /pour,' بسبب'/ a cause , ' نظرا' /vue que,' في حين'/ tandis que*,..etc, indiquent toujours le début d'un EDU (cf. exemple 27).

(27) [تم الغاء كل الرحلات الجوية][نظرا لسوء الحوال الطقس]

(27) [*En raison de mauvaises conditions météorologiques*] [*Tous les vols ont été annulés*]

Tandis que Les marqueurs faibles tels que {' *حتى' ,' بينما' , le connecteur و ,..etc* } sont ambigus et leur présence ne signale pas forcement le début d'un EDU. Considérons, par exemple, le marqueur ' حتى ' dans les exemples 28 et 29 respectivement. Dans l'exemple 28 ce connecteur ne signale pas un point de segmentation par contre dans l'exemple 29, le même marqueur indique le début d'un EDU et justifie la segmentation de la phrase.

(28) [استهدفت حملات التوعية كل الفئات حتى الشيوخ]

(28) [*les campagnes de sensibilisation ont ciblées toute les catégories de la société] [, même les vieux*]

(29) [توزيع مشاريع الاستثمار على المحافظات ركن يجب له الأولوية] [حتى يتسنى ايجاد فرص عمل للشباب المؤهل الذي يعاني من بطالة]

(29) [*La distribution de projets d'investissement aux provinces devrait être une priorité] [afin de créer des opportunités d'emploi pour les jeunes qualifiés qui souffrent du chômage]*

2) Verbes de parole ou verbe introductifs (*Attribution verbs*)

Les verbes de parole sont des verbes généralement utilisés pour rapporter les discours directs et indirects tels que : *' قال' / dire, ' اعلن' / annoncer,' صرح' / déclarer ...etc.*

Nous nous sommes intéressés juste au discours direct et plus précisément quand le message rapporté commence par le connecteur '' ان ' ou précédé par ' :' .
Dans ce cas, la phrase doit être segmentée en deux EDUs : Le premier EDU concerne la proposition qui contient le verbe introductif ou qui commence par 'حسب/' et le deuxième EDU contient le message rapporté (cf. exemple 30).

(30) [وكان الوزير الأول للبلاد قد اعلن] [ان الدولة تدرس فرض عقوبات إضافية على التجار المخالفين للقانون]

(30) [*Le premier ministre du pays a déclaré que] [l'Etat envisage d'imposer des sanctions supplémentaires contre les commerçants qui enfreignent la loi*]

Dans les autres cas, la phrase est traitée comme une seule unité (cf. exemple 31)

(31) [اعلن الوزير عن تعديلات جوهرية على السياسات الاقتصادية السابقة،]

(31) [*Le ministre a annoncé des changements substantiels dans les politiques économiques précédentes*]

3) La coordination

- Les phrases et les propositions coordonnées sont segmentées et divisés en EDUs. Les phrases coordonnées sont généralement séparées par une virgule ou un point-virgule plus une conjonction de coordination. En langue arabe, la coordination est exprimée à l'aide de six conjonctions :{ و, او, ف, ثم, حتى, ام, بل, لكن, لا }. Ces conjonctions sont fortement ambigus ce qui rend leur traitement une tâche difficile. Par exemple, le connecteur '(و) / et' peut avoir plusieurs sens et plusieurs rôles autres que la coordination [KHAL, 2011].

Les conjonctions de coordination en arabe peuvent relier des mots simples, des propositions ou des phrases. Notre traitement de segmentation concerne les phrases et les propositions coordonnées. Ainsi deux phrases séparées par une conjonction de coordination sont segmentées et traitées comme deux EDUs séparées (cf. exemple 32).

(32) [ادانت السلطات المعنية اعمال التخريب] [وقامت بفرض إجراءات ردعية صارمة ضد المشتبه فيهم]

(32) [*Les autorités concernées ont condamné les actes de vandalisme,] [et elles ont imposé des mesures de dissuasion strictes contre les suspects*]

- Deux propositions coordonnées sont segmentées tel qu'illustré par l'exemple 33.

(33) [امْ يَقُولُونَ بِهِ جِنَّةٌ] [بَلْ جاءَهُمْ بِالْحَقِّ[3]]

(33) [Ou diront-ils Il est fou][Au contraire, c'est la vérité qu'il leur a apportée]

- Les mots coordonnés ne sont pas traité comme des EDUs (cf. exemple 34)

(34) [يا أَيُّهَا الَّذِينَ آمَنُواْ أَنفِقُواْ مِمَّا رَزَقْنَاكُم] [مِّن قَبْلِ أَن يَأْتِيَ يَوْمٌ لاَّ بَيْعٌ فِيهِ وَلاَ خُلَّةٌ وَلاَ شَفَاعَةٌ[4]]

(34) [*Ô les croyants! Dépenser de ce que Nous vous avons attribué][avant que vienne le jour où il n'y aura ni rançon ni amitié ni intercession]*

- les verbes coordonnés ne sont pas marqués comme des EDUs séparés (cf. exemple 35)

[3] Verset 70 de Sourate Al-Muminune-Le Saint Coran.
[4] Verset 187 de Sourate al-Baqarah - Le Saint Coran.

(35) [قام و شمر عن ساعديه]

(35) *[Il s'est levé et a retroussé ses manches]*

4) les expressions temporelles

- Une proposition qui commence par un adverbe de temps (ظرف زمان) tel que :{ يومئذ,حينئذ , اثناء , خلال, بعد,قبل,حيث ,اذ ..etc} est segmentée et traitée comme un EDU (cf. exemple 36)

(36) [فقد نصره الله] [اذ اخرجه الذين كفروا]

(36) [*Dieu l'a fait triompher] [alors que les mécréants l'ont chassé]*

- Une proposition qui commence ou se termine par une expression temporelle est aussi marquée comme un EDU séparé (cf. exemple 37).

(37) [عشية الاحتفال براس السنة اهجريه العام الفارط] [شهدت البلدة عدة اعمال انتحارية]

(37)[*La veille de la célébration du nouvel an de l'hégire de l'année passée][la ville a connu plusieurs opérations suicidaires]*

- Si l'expression temporelle ou l'adverbe de temps sépare le verbe de son sujet ou de son complément, alors l'expression temporelle dans ce cas n'indique pas un point de segmentation (cf. exemple 38).

(38) [استقبل, صباح اليوم, وزير الشباب و الرياضة الفريق الوطني]

(38) *[Ce matin, le ministre de la jeunesse et du sport a reçu l'équipe nationale]*

- Une expression temporelle composée d'une combinaison des adverbes de temps tels que {عشية, يوم, صبيحة, ساعة , سنة …etc} est segmentée en EDUs séparées si elle précède ou suit une phrase verbale tel que dans l'exemple 39.

(39) [صبيحة يوم الاثنين 13 يناير 2014,] [تم اطلاق صراح المعتقلين السياسيين المحكوم عليهم بالإعدام.]

(39) [Dans la matinée du lundi 13 janvier 2014]$_1$[les prisonniers politiques condamnés à mort ont été libérés]$_2$

5) Propositions relatives

- Les propositions relatives ne doivent pas être segmentées en EDU séparé (cf. exemple 40). En langue arabe une proposition relative (الجملة الموصولة) commence toujours par un pronom relatif tel que { الذي, اللذان, الذين , التي ,من,اللائي } . Nous avons décidé de ne pas segmenter les propositions relatives parce que dans la majorité des contextes, nous n'avons pas trouvé des raisons suffisant pour segmenter ces phrases du point de vue discursif.

(40) [لم يعرف قطاع السياحة التطور الذي كان متوقعا]

(40) [Le secteur de tourisme n'as pas connu le développement attendu]

6) Le conditionnel

- La phrase conditionnelle doit être toujours segmentée en deux EDUs séparés. le premier EDU est composé de particule de condition (أداة الشرط) ainsi la proposition de condition (جملة الشرط) et le deuxième est composé de la réponse (جملة جواب الشرط) comme illustre l'exemple 41.

(41) [كلما ازداد العالم بشاعة] [ازدادت حاجتنا الى الجمال]

(41) [Plus le monde est laide,][plus nous avons besoin de beauté]

La phrases conditionnelle en langue arabe est composé de trois éléments : une particule de condition (أداة الشرط) tel que : {لولا , كلما, اذا, حيثما, أي, ايان, انى, اين, أينما, كيفما, مهما, من, لو, ما,اذ,ان } et deux propositions : la première exprime la condition appelée (جملة الشرط) et l'autre exprime la réponse appelée جملة جواب الشرط .

7) المفعول لأجله/ Accusative de but

- Une proposition qui commence par un accusatif de but est segmentée et marqué comme un EDU séparé tel que dans l'exemple 42.

'المفعول لاجله' est une construction infinitive (مصدر) dans le cas accusative (منصوب) utilisé pour spécifier le but, le motif ou la raison derrière une action (il signale la relation غرض).Par exemple le mot 'ابتغاء' dans l'exemple (42) justifié pourquoi ils déposent leurs biens.

*(42) [ينفقون اموالهم] [**ابتغاء** مرضاة الله]*

(42) [Ils dépensent leurs biens] [cherchant l'agrément d'Allah]

'المفعول لاجله' peut être aussi précédé par une des prépositions suivants :{ 'لام التعليل ' et 'من' } comme illustre les exemples (43) et (44).

*(43) [تواصل الدولة المفاوضات مع النقابات] [**للحد** من الإضرابات المتكررة]*

(43) L'état poursuit les négociations avec les syndicats pour mettre fin aux grèves répétées

*(44) [لَوْ أَنْزَلْنَا هَذَا الْقُرْآنَ عَلَى جَبَلٍ] [لَرَأَيْتَهُ خَاشِعاً مُتَصَدِّعاً] [**مِنْ خَشْيَةِ** اللَّهِ]*

(44) Si Nous avions fait descendre ce Coran sur une montage, tu l'aurais vu s'humilier et se fendre par crainte d'Allah.

- Si l'accusatif de but est précédé par un pronom démonstratif tel que هذا, ذلك alors le pronom doit être inclue comme une partie de EDU tel qu'il est illustré par l'exemple (45).

(45) [اقامة الطاقم الذي سيكون سريا حتى ما بعد المباراة] [**وذلك تجنبا** لحدوث أي اتصالات من أي مسؤول .]

(45) [*La résidence de l'équipe sera secrète][afin d'éviter toute intervention d'un responsable quelconque]*

8) 'Almasdar /المصدر '

Dans certain cas,' **المصدر** /almasdar' préfixé par la préposition « بـ » indique un point de segmentation s'il apparait dans une proposition telle que dans l'exemple 46. Nous avons décidé de marqué la proposition qui commence par une telle construction comme un EDU séparé parce que nous avons trouvé qu'une telle proposition explique généralement la façon dont quelque chose (raconté dans la première proposition) est fait, c'est-à-dire elle répond à la question comment ? Ou de quelle manière ? Donc il signale la relation manière.

(46) [تسعى الدولة لتعزيز الامن في البلاد] [**بـاتخاذ** اجراءات ردعية قاسية]

(46) [L'Etat cherche à renforcer la sécurité dans le pays][en prenant des mesures dissuasives sévères]

Après une analyse approfondie de notre corpus, nous avons remarqué que cette relation peut être signalé aussi par d'autres expressions tel que { بواسطة ,عن طريق , من خلال }. Ainsi tout proposition qui commence par ces expressions est segmentée et marqué comme un EDU séparé tel qu'il est illustré par les exemples (47) et (48).

(47) [وتحاول خدمة المجتمع بطريقتها الخاصة] [من خلال العمل التطوعي في المشاريع الخيرية لمصلحة الايتام]

(47) *[Et elle essaie de servir la société à sa manière] [à travers le travail bénévole dans des projets caritatifs au profit des orphelins]*

(48) [ونجح اخناتون في السيطرة على معارضة الكهنة] [عن طريق استخدام وحدات من الجيش تتولى تنفيذ أوامره]

(48) [*Akhenaton a réussi à contrôler les prêtres de l'opposition*] [*en utilisant des unités de l'armée pour exécuter ses ordres*]

1) Les signes de Ponctuation

- Une proposition ou une phrase mise entre parenthèses est traitée comme un EDU séparé si elle n'apparait pas à l'intérieur d'une phrase (cf. l'exemple 49). Autrement dit, si elle sépare le verbe de son complément ou de son sujet, elle ne doit pas être traitée comme un EDU séparé (cf. exemple 33).

(49) تتضمن بقية الرسالة اقتراحات لحل قضايا محددة][(مماثلة الى حد ما لمسودة الاتفاق التي اعدت في اواخر 1995)

(49) [*Le reste de la lettre contient des suggestions pour résoudre des problèmes spécifiques] [(quelque peu semblable au projet d'accord préparé à la fin de 1995)]*

(50) [وكانت الغاية المعلنة (الحقوق القومية للاكراد) كافيةً لدفع كل شيء الى الصمت والتواطؤ.]

(50) [*L'objectif déclaré (les droits nationaux des Kurdes) était suffisant pour pousser tout au silence et à la complicité*.]

- Une proposition ou une phrase suivi par « : » est traitée comme un EDU séparé (cf. exemple 51)

(51) [فاندهش و قال:][ان هذا امر لا يصدق]

(51) [Il a été surpris et a dit:] [Ceci est incroyable]

- Une expression mise entre apostrophes n'est pas marquée comme un EDU.
- La virgule (,), le point- virgule (;) et le point d'exclamation (!) n'indiquent pas un point de segmentation.

Les signes de ponctuation dans les textes arabes ne sont pas utilisés d'une façon régulière et parfois ils sont rarement utilisés. De ce fait, les phrases dans le discours arabe sont longues et trop complexes. C'est pour cela qu'il n'était pas possible de compter beaucoup sur ces signes pour nous guider dans la segmentation du discours, comme c'est le cas pour certaines langues telles que la langue anglaise.

3.2.2 Détermination du statut rhétorique

Dans la cadre de la théorie de la structure rhétorique, les unités de discours faisant partie d'une relation rhétorique sont caractérisées par un statut rhétorique qui reflète leur degré d'importance dans le texte. Une relation mononucléaire relie deux unités de discours : un noyau qui représente l'information la plus saillante dans la relation, et un satellite qui fournit une information secondaire, dite aussi information de support, à propos du noyau dont il dépend.
L'unité noyau d'une relation mononucléaire est compréhensible indépendamment du satellite mais pas vice versa. Un satellite peut ne pas donner un sens en isolation de son noyau. A cet effet, pour distinguer entre un noyau et un satellite, nous avons choisi d'appliquer le test de la suppression défini dans [CARL, 2001]. Ce test consiste à supprimer à chaque fois un des constituants de la relation et à analyser le résultat. Lorsqu'un segment satellite est supprimé, le segment qui reste, qui est le noyau, peut effectuer la même fonction dans le texte, c'est-à-dire que le texte reste encore cohérent. Mais lorsque le segment supprimé est un noyau, le segment qui reste peut ne pas avoir un sens et le texte perd sa cohérence.
Déterminer la nucléarité des segments constituants une relation est une étape qui a été faite simultanément avec l'annotation des relations rhétoriques.

3.3. Annotation du corpus

3.3.1. Processus d'annotation

Deux professeurs en langue arabe ont été invités à annoter notre corpus. Nous leur avons fourni un guide d'annotation dans lequel nous avons détaillé les principes et les règles de la segmentation des textes en EDUs, la liste des relations rhétorique qui doit être utilisé pour l'annotation de notre corpus, une définition claire pour chaque relation rhétorique accompagnée par une liste des marqueurs possibles et des exemples illustratifs. Ainsi que le principe de la nuclearité dans la RST et le principe de la détermination du statut rhétorique de deux segments liés par une relation rhétorique y est également détaillé.

La tâche attribuée aux annotateurs consiste d'abord à segmenter le texte en EDUs, puis annoter les relations rhétoriques qui relient deux EDUs adjacents dans la même phrase, ainsi qu'assigner un statut rhétorique à chaque EDU faisant partie d'une relation.

Pour accomplir cette tâche, Les annotateurs sont d'abord passés par une phase d'entrainement, durant laquelle ils se sont entrainés à annoter 20 documents de corpus en se basant sur le manuel d'annotation que nous leur avons fourni. Tout au long de cette phase, les annotateurs ont été encouragés à discuter les points de désaccord et à apporter des modifications sur le manuel lorsque cela était nécessaire.

Ensuite, pour estimer les degrés d'accord entre annotateurs, nous avons sélectionné 15 documents de notre corpus, et nous avons demandé à chacun de nos annotateurs de les annoter. Autrement dit, une double annotation sur 15 documents a été effectuée. Le degré d'accord entre annotateurs a été estimé en utilisant la mesure kappa de cohen [COHE, 1960] pour les deux tâches : identification des relations rhétoriques et assignation de nucléarité ou de statut rhétorique. Concernant l'identification des relations, nous avons obtenu un kappa =0, 72 sur les 20 relations et 0, 89 pour l'assignation du statut rhétorique. Ce qui peut être interprété comme étant un très bon accord.

Finalement, et après avoir discuté tous les points de désaccord, les experts ont entamé l'annotation de corpus. Cette étape a duré presque 4 mois.

3.3.2 Détails statistiques du corpus annoté

Notre corpus annoté est composé de 140 documents avec des tailles presque équivalentes. Chaque document contient en moyenne 21 phrases. Le nombre total des EDUs est de 7639 avec une moyenne de 54.56 EDUs par document. Chaque document a été annoté en utilisant l'ensemble des relations extraites de notre corpus enrichi avec une notation de nucléarité qui

détermine le statu rhétorique des segments faisant partie de la relation, cela a abouti au total à 23 relations fines telles que : N-S.سبب/ cause, S-N. سبب/cause, N-N. مقابلة/contraste, S-N.غاية/but… Etc. Le nombre total de relations annotées est de 3387, la répartition de ces relations dans notre corpus est présentée dans le tableau 2-2.

Classes de relations	Relation rhétorique	Relations rhétoriques fines	Pourcentage (%)
السببية Causal	سبب/Cause	S-N.سبب/cause	7.26
		N-S.سبب/cause	4.80
	نتيجة/Cause-effect	N-N. نتيجة /Cause-effet	1,02
	غاية /but	S-N. غاية/ but	6.78
		Total	**18.76**
المقارنة Comparaison	مقابلة/Contraste	N-N.غاية/Contrast	3.2
	مقارنة/Comparaison	N-N. مقارنة/Comparaison	9.72
	تشبيه/Analogie	S-N. تشبيه/ Analogie	3.97
		Total	**16.89**
العطف Coordination	وصل/ Conjonction	N-N. وصل/Conjonction	8.64
	تخيير/ Disjonction	N-N. تخيير/Disjonction	0.98
	اضراب/ Rectification	N-N. اضراب/Rectification	2.09
	ترتيب/ sequence	N-N. ترتيب/séquence	0.18
	استدراك/ concession	S-N. استدراك/ Concession	2.17
		Total	**14.06**
تفصيل Elaboration	اسهاب/Elaboration	S-N. اسهاب/Elaboration	10.08
	تخصيص/ spécification	S-N. تخصيص/Spécification	4.8
	مثال/ Exemple	S-N. مثال/Exemple	1.42
	كيفية/ Manière	S-N.كيفية/Manière	5.19
	زمني/Temporel	N-S. زمني/Temporel	4.16
		S-N. زمني/Temporel	4.26
		Total	**28.91**
توضيح Explication	تفسير/Explication	S-N.تفسير/Explication	3.53
	تعليل/Justification	S-N.تعليل/justification	4.91
		Total	**8.44**
اسناد	اسناد/Attribution	N-S.اسناد/Attribution	7.53

Attribution		S-N.اسناد/Attribution	1.09
	Total		**8.62**
الشرط/Conditionnel	شرط/Condition	N-N.شرط/Condition	**4.12**

Tableau 3-2 : Distribution des Relations Rhétoriques dans Notre Corpus Annoté

D'après les statistiques présentées dans le tableau 3-2, nous pouvons constater que les relations les plus fréquentes dans notre corpus sont : اسهاب/ Elaboration (10.08%) et وصل/ conjonction (6.84%). Les relations les moins fréquentes sont : تخيير/ disjonction (0.98%) et ترتيب/ séquence (0.18%). Il est à noter, aussi, que dans notre corpus la majorité des relations sont explicites (77%), c'est-à-dire, signalées par des marqueurs de discours tels que لذا,لكي, لكن, بل, من اجل , بسبب … etc. L'annotation des relations explicites est plus facile par rapport aux relations implicites vue l'absence des indices précis qui signalent cette dernière.

4. Conclusion

Dans ce chapitre, nous avons présenté notre corpus arabe annoté selon le cadre de la théorie de la structure rhétorique. Nous nous sommes limités à l'annotation des relations rhétoriques intra-phrases, c'est à dire annoter les relations entre les unités de discours élémentaires adjacents au sein de la même phrase. Pour effectuer ce travail, il était nécessaire d'abord d'élaborer la liste des relations rhétoriques arabes qui va être utilisée pour l'annotation. L'élaboration de cette liste a nécessité une étude de la rhétorique arabe ainsi qu'une analyse approfondie du corpus. Notre travail d'annotation a impliqué aussi la spécification des statuts rhétoriques des segments faisant partie d'une relation, cela signifie que notre corpus est annoté avec des relations plus fines qui spécifient aussi le niveau d'importance des segments reliés.

Notre corpus annoté nous a servi à construire un modèle à base de l'apprentissage supervisé pour l'identification automatique des relations rhétoriques arabes, dont les détails feront l'objet du prochain chapitre.

Chapitre 4

Contribution 2

Identification Automatique des Relations Rhétoriques Arabes

Sommaire

1. Introduction

Un texte cohérent n'est pas une séquence d'unités textuelles indépendante, mais plutôt un texte cohérent à une structure rhétorique qui relie ses unités pour exprimer un sens [JOTY, 2015]. Les relations rhétorique définissent la nature de la relation qui relies ces unités textuelles et contribue ainsi à interpréter et à créer la structure rhétorique de texte.

Les relations rhétoriques (nommées aussi relations de discours ou relations de cohérence) peuvent être explicites ou implicites. Les relations explicites sont des relations signalées par des marqueurs de discours. Ainsi, ces marqueurs jouent un rôle central dans l'identification automatique de ce type de relations. Par exemple, la présence du marqueur « لأن /parce que » signale fortement la relation 'تعليل'/*justification*. Cependant, en l'absence de ces marqueurs, comme dans «عليك القيادة بحذر. المنعرج خطير » la relation est dite implicite et l'identification automatique de telles relations reste, à l'heure actuelle, un défi sérieux.

L'identification automatique des relations rhétorique est une étape cruciale dans l'analyse du discours et elle s'est avérée très utile aussi dans plusieurs types d'applications dans le domaine du traitement automatique du langage naturel (TALN) telles que les systèmes question réponse [AZMI, 2017 ; SADE, 2016], et l'analyse d'opinion [SOMA, 2009]. Elle a suscité un grand intérêt dans la littérature au sein des différents cadres théoriques (la rhétorique Structure Théorie (RST) [MANN, 1988], le modèle *Penn Discourse Treebank* (PDTB) [PRAS, 2008], le modèle *Graphbank* [WOLF, 2005] et La théorie des représentations discursives segmentées (SDRT) [ASHE, 2003]). Ainsi, plusieurs approches ont été proposées pour aborder ce problème allant des approches supervisées, semi supervisées à non supervisées [LIN, 2009; PITL, 2009; LOUI, 2010; PARD, 2012; BIRA, 2013; MIHA, 2016; LI, 2017].

Dans ce chapitre, nous allons présenter notre approche pour l'identification automatique des relations rhétorique Arabes définie dans le cadre de la théorie de la structure rhétorique. Nous nous somme limité à l'identification automatique des relations intra-phrases, c'est-à-dire les relations qui relient deux unités de discours élémentaires (EDU) adjacents au sein de la même phrase.

D'abord nous allons présenter les travaux connexes effectués dans ce domaine, puis le modèle proposé ainsi les différents expérimentations menées.

2. Travaux Connexes

Nous présentons dans cette section, un aperçu des principales approches computationnelles proposées pour l'identification automatique des relations rhétoriques explicites et implicites dans le cadre de la RST ainsi que le modèle PDTB.

Marcu et Echihabi [MARC, 2002] ont présenté la première approche d'apprentissage non supervisé pour identifier quatre classes de relations RST : *Contraste, Explication-Preuve, Condition et Elaboration*. Ils ont été les premiers à utiliser les paires de mots pour identifier les relations rhétoriques. Saito et al. [SAIT, 2006] ont étendu ce travail en combinant les paires de mots avec d'autres caractéristiques pour identifier les relations rhétoriques implicites en japonais. D'autres auteurs ont également proposé des approches semi-supervisées qui exploitaient à la fois les données étiquetées et celles non étiquetées. Hernault et al. [HERN, 2010a] ont proposé une méthode basée sur la cooccurrence des mots observées dans les données non étiquetées. Les auteurs ont utilisé un ensemble de caractéristiques parmi lesquelles les paires de mots et les règles de productions. Ils ont montré que leur méthode améliorait significativement la précision de la classification pour les relations peu fréquentes.

La publication des corpus de discours annotés a ouvert des opportunités d'aborder ce domaine de recherche en s'appuyant sur les techniques de l'apprentissage supervisé. En utilisant le RST-DT [CARL, 2003], Soricut et Marcu [SORI, 2003] ont présenté un analyseur rhétorique des phrases basé sur des caractéristiques lexicales et syntaxiques extraites de l'arbre syntaxique des phrases. Hernault et al. [HERN, 2010b] ont présenté HILDA, un analyseur de discours entièrement mis en œuvre sur la base des machines à vecteurs de support (SVM). Pour l'identification des relations, les auteurs ont utilisé un classifieur SVM multi-classe. Plusieurs caractéristiques lexicales et syntaxiques ont été prises en compte, y compris des caractéristiques structurelles, lexicales et organisationnelles. Feng et Hirst [FENG, 2012] ont étendu ce travail en y intégrant d'autres caractéristiques linguistiques telles que les règles de production pour refléter la dépendance entre les différentes relations de discours et plusieurs caractéristiques contextuelles. Les auteurs ont pris une décision importante concernant les relations rhétoriques. Cette décision consiste en la discrimination entre les relations intra-phrases et inter-phrases, qui impliquait la spécification des caractéristiques différentes pour chaque niveau. Cela a été également pris en compte par Joty et al. [JOTY, 2015] lors de l'implantation de leur analyseur de discours CODRA à l'aide des champs conditionnels aléatoires (CRF).

Un peu plus tard, le PDTB [PRAS, 2008] a été présenté comme un grand corpus de discours annoté. Le corpus fournit une annotation plus claire et plus exhaustive des relations implicites

et une plateforme précieuse pour les chercheurs leur permettant de développer des systèmes centrés sur le discours. Le premier travail abordant l'identification des relations implicites en utilisant le corpus PDTB était celui entrepris par Pilter et al. [PILT, 2009]. L'auteur a utilisé différentes caractéristiques linguistiques telles que la modalité, les classes de verbes et la polarité. Lin et al. [LIN, 2009] ont proposé un modèle qui permet de classifier les relations implicites de deuxième niveau dans le PDTB en utilisant plusieurs groupes de caractéristiques. Récemment, et vus les résultats remarquables obtenus par les modèles d'apprentissage profond (*deep learning*) dans le traitement automatique du langage naturel [SOCH, 2013; KIM, 2014], beaucoup de chercheurs se sont orientés vers l'utilisation des modèles de réseaux de neurones profonds et leurs méthodes de représentation pour l'identification des relations rhétoriques implicites [ZHAN ,2015 ; JI, 2015 ; LIU ,2016].

Pour la langue Arabe, très peu de travaux ont vu le jour dans ce domaine. Alsaif et Markert [ALSA, 2011] ont proposé des algorithmes supervisés pour identifier les marqueurs de discours et les relations explicites qui existent entre les EDUs adjacentes dans le modèle PDTB. Les auteurs ont utilisé quelques caractéristiques déjà utilisées pour l'identification des relations implicites en Anglais. Ce travail a été étendu, plus tard par Keskes et al. [KESK, 2014] qui ont abordé à la fois l'identification des relations explicites et implicites entre des unités adjacentes et non adjacentes dans le cadre de la SDRT. Les auteurs ont utilisé la majorité des caractéristiques utilisées par Alsaif et Markert [ALSA, 2011] combinées à d'autres inspirées de travaux antérieurs sur l'identification des relations rhétoriques en langue anglaise. Les auteurs ont utilisé un corpus de discours Arabe annoté manuellement dans le cadre de la SDRT.

3. Le modèle proposé

Pour identifier automatiquement les relations rhétoriques arabes définies dans le cadre de la théorie de la structure rhétorique (voir chapitre 3), nous proposons une approche supervisée basée sur les réseaux de neurones artificiels et plus précisément le perceptron multicouche (MLP).

Nos instances sont composées de paires d'EDUs qui sont reliées par une relation rhétorique dans notre corpus annoté (cf. chapitre 3). Pour effectuer un apprentissage supervisé sur ce corpus annoté, nous avons créé un vecteur de caractéristiques pour chaque instance de relation i.e. pour chaque couple d'EDUs étiqueté entre elles par une relation fine. L'exemple 1 ci-dessous présente une phrase extraite de notre corpus composé de deux instances et de deux relations rhétoriques : N-N.تخيير et S-N.تفسير, telle que l'instance de la relation N-N.تخيير est le couple d'EDU(2,1) et l'instance de la relation S-N.تفسير est le couple d'EDU (3,2).

Dans ce cas d'exemple, on doit créer deux vecteurs de caractéristiques qui correspondent à ces deux instances.

(1) [ينبغي ان يستند الحل الى حد ما على فكرة التسامح بدافع الخوف]1 تخيير.N-N [او الاحساس بخطر شيء ما يكون اسوأ بكثير،]2 تفسير.S-N [اي بحرب ضارية يمكن ان تلحق بكلا الطرفين اذى يتعذر اصلاحه]3.

(2,1) N-N.تخيير /Disjonction

(3,2) S-N.تفسير/ Explication

Nous avons exploré et évalué plusieurs groupes de caractéristiques. Ces groupes seront décrits en détail dans le reste de cette section.

- ## Caractéristiques utilisés

Nous avons proposé dix groupes de caractéristiques, parmi lesquels trois nouveaux groupes nommés Accusatif de but/ المفعول لأجله, connecteur spécifique, et le nombre de mots antonymes. Tandis que Al-Masdar/المصدر est inspiré de [ALSA, 2011], et les six autres groupes sont inspirés de travaux antérieurs qui ont prouvé leurs efficacité dans l'identification des relations rhétoriques explicites et implicites [MARC, 2000; SUBB , 2009; HUAN, 2011; KESK , 2014]. Dans ce qui suit, nous allons exposer chaque groupe de caractéristiques en l'accompagnant d'exemples illustratifs. Le noyau et le satellite dans ces exemples sont indiqués respectivement par les lettres majuscules [S] et [N] entre crochets au début de chaque EDU.

F1) **Les marqueurs de discours**

Pour encoder les informations liées aux marqueurs de discours, nous avons utilisé un dictionnaire des marqueurs de discours construit manuellement durant le processus d'annotation. Dans ce dictionnaire nous avons associé à chaque marqueur :

- un numéro spécifique,
- une classe, qui correspond au numéro de la relation qu'il peut signaler
- un Type: non ambigu / ambigu. Les marqueurs non ambigus sont ceux qui signalent une seule relation de discours telle que ' بالمقابل/ *en revanche'*, ' من أجل /*pour*', ' لأن /*parce que*'. Tandis que les marqueurs ambigus peuvent signaler plusieurs relations rhétoriques. Par exemple le marqueur ' حتى / *jusqu'à*' peut signaler la relation 'غاية/ *but*'

(cf. exemple 1), ou la relation ‘ شرط /*condition*’ (cf. exemple 2) ainsi que d’autres relation (signalons que ce marqueur est assez particulier en langue arabe).

Ainsi pour chaque EDU, on associe trois caractéristiques numériques :
- la première caractéristique encode le numéro de marqueur s’il existe sinon il prend la valeur zéro (0).
- la deuxième encode sa classe.
- la troisième encode sa position (1 : début de EDU, 2 milieu et 3 : fin de EDU).

(1)[N]أَتَعَلَّمُ اللغة العَرَبِيَّةَ[S]حتى أَعْرِف ثَقَافَتِهَا وَحَضَارةَ أَهْلِهَا

(1) [N] *J'apprends l'Arabe [S] pour connaître sa culture et la civilisation de son peuple*.

(2) [N] لن يلغى الاضراب[N] حتى تستجيب الدولة لمطالب النقابات

(2) [N] *La grève ne sera pas annulée [N] jusqu'à ce que l'État réponde aux demandes des syndicats*

F2) Verbe introductif

Nous avons utilisé deux caractéristiques binaires pour vérifier la présence de verbes introductifs dans chaque EDU en utilisant un lexique construit manuellement composé de 41 verbes introductifs Arabes tels que ‘ كشف / *exposer*’,’ أعلن / *annoncer*’, ‘ قال / *dire* ‘, ‘ أكد / *confirmer’*. La présence de verbes introductifs est un indicateur fort qui signale la relation «اسناد /Attribution » (cf.example.3).

(3) [S]قال الوزير :[N] ان أسعار النفط في انخفاض مستمر

(3) *[S] Le ministre a déclaré: [N] que le prix du pétrole continue de baisser.*

F3) connecteur spécifique

Nous avons utilisé une caractéristique binaire pour vérifier la présence de connecteur ‘أن /*que* ‘ au début de chaque EDU.

Le connecteur ‘أن/*que*’ est généralement utilisé avec les verbes introductifs dans le discours rapporté, donc sa présence au début de la deuxième EDU avec le verbe introductif dans la première signalent fortement la relation « اسناد / *attribution* » (cf. l'exemple.3).

F4) Al-masdar (المصدر)

Al-masdar est une catégorie nominale indiquant des événements sans précision du temps de l’action tel que ‘ اعتذار / *excuse*’, ‘ استعمال / *utilisation’*. Dans la langue arabe, la présence de ' al-masdar/ المصدر ’ préfixé par les particules de justification { لام التعليل (لـ)/*pour* et باء السببية/ بـ /par } signalent généralement la relations ‘ غاية / *but*’ (cf. exemple 4), ou ‘كيفية /*manière*’ (cf. exemple 5).

Pour chaque EDU, une caractéristique binaire est utilisée pour vérifier si le premier mot de chaque EDU contient al-masdar/المصدر.

Une autre caractéristique numérique est ajoutée pour encoder la particule préfixé a cette construction.

(4) [N] تواصل الدولة المفاوضات مع النقابات [S] للحد من الإضرابات المتكررة

(4) [*N] L'Etat poursuit les négociations avec les syndicats [S]afin de mettre fin aux grèves répétées.*

(5) [N] تسعى الدولة لتطوير اقتصادها [S] بالاعتماد على مصادر متنوعة

(5) *L'État cherche à développer son économie en s'appuyant sur diverses sources.*

F5) Les signes de ponctuation

Nous avons utilisé trois caractéristiques binaires pour tester la présence de certaines signes de ponctuations (deux points (:), virgule (,) et les parenthèses ()). Ces caractéristiques peuvent nous aider à identifier certaines relations rhétoriques, telles que ‘ اسناد /*attribution*’ et ‘ تفصيل / *élaboration*’.

F6) Accusatif de but/ المفعول لأجله

Nous avons utilisé une caractéristique binaire pour indiquer si le premier ou le deuxième mot de chaque EDU est un accusatif de but /’ 'المفعول لأجله’.

Comme indiqué dans le chapitre 3, ‘ المفعول لأجله / l’Accusatif de but’ est un nom indéfini dans le cas accusatif manṣūb (منصوب) utilisé pour spécifier le but, le motif ou la raison derrière une

action. Par exemple, l'accusatif de but « تجنبا / pour éviter », dans l'exemple (6) précise la raison derrière la réunion organisée hier. Ainsi, il signale fortement la relation ' غاية / but '.

(6) عقد اجتماع عاجل مع اللاعبين مساء امس [S]تجنبا لاحتجاجاتهم.

6) [N] *Hier, une réunion urgente avec les joueurs est organisée [S] pour éviter leurs protestations.*

F7) les entités nommées

Les entités nommées sont les noms des personnes, des endroits et des organisations. Nous avons utilisé une caractéristique binaire pour indiquer si chaque EDU contient des entités nommées. Une telle information pourrait nous aider à reconnaitre des relations rhétoriques telles que ' اسناد / *Attribution*',' مقارنة / *comparaison*', et ' تفصيل/ *élaboration*'.

F8) La longueur

C'est une caractéristique numérique qui indique quelle EDU est plus longue que l'autre (la longueur en termes de mots). Dans certaines relations comme' تفسير / *explication*' ou' اسهاب / *élaboration*', la deuxième EDU est généralement plus longue que la première. C'est pour cela nous pensons que cette information peut être utile pour notre tâche.

F9) Données numériques

Nous utilisons deux caractéristiques binaires pour indiquer si chaque EDU contient des données numériques telles que les chiffres. Cela nous permet d'identifier certaines relations rhétoriques telles que ' مقارنة / *comparaison*' et ' زمني / *temporel*'.

F10) Nombre de mots antonymes

L'information sémantique a une grande importance dans l'identification des relations rhétoriques, en particulier pour la langue arabe où l'identification de certaines relations de discours implicites repose principalement sur les relations sémantiques entre les mots. Par exemple : la relation ' مقابلة / *contraste*' en arabe (cf.example.7) est signalée par la présence de plus d'un mot dans la première EDU et leurs antonymes dans la deuxième. C'est pourquoi, nous avons utilisé une caractéristique numérique qui calcule le nombre de mots des deux EDUs qui sont antonymes.

(7) فَلْيَضْحَكُوا قَلِيلًا [N] وَلْيَبْكُوا كَثِيرًا[1]. [N]

(7) *Alors laissez-les rire un peu [N] et pleurer plus.*

4. Expérimentations

Pour effectuer nos expérimentations, nous avons choisi les 18 relations les plus fréquentes dans notre corpus annoté générant une base de 3318 instances. Les fréquences de ces relations sont présentées dans le tableau 4-1.

Relations fusionnées	Relations rhétoriques fines	Fréquence
السببية Causal	S-N.سبب/cause	246
	N-S.سبب/cause	163
	N-N. غاية/ but	230
المقارنة Comparaison	N-N.غاية/Contrast	106
	N-N. مقارنة/Comparaison	231
	S-N. تشبيه/ Analogie	134
العطف Coordination	N-N. وصل/Conjonction	324
	S-N. استدراك/ Concession	113
	S-N. اضراب/ Rectification	72
تفصيل Elaboration	S-N. اسهاب/Elaboration	341
	S-N. تخصيص/Spécification	162
	S-N.كيفية/Manière	175
	N-S. زمني/Temporal	141
	S-N. زمني/Temporal	155
توضيح Explication	S-N.تفسير/Explication	120
	S-N.تعليل/Justification	166
اسناد/Attribution	N-S.اسناد/Attribution	295
الشرط/Conditionnel	N-N.شرط/Condition	**144**

Tableau 4-1: Fréquence des Relations dans le corpus d'apprentissage.

[1] Verset 82 de Surat Al – Tawbah – le Saint-Coran.

Nous avons utilisé un ensemble de ressources parmi lesquelles : l'environnement d'apprentissage Weka[2], l'analyseur morphologique Arabe [BOUD, 2017], la base lexicale WordNet pour la langue Arabe (AWN) [BLAC, 2006] ainsi qu'un dictionnaire des noms propres arabe « *Gazetteer* », nommé ANERgazet [BENA, 2007] pour vérifier la présence des entités nommées dans les EDUs.

Pour implémenter notre modèle, nous avons utilisé le classifieur perceptron multi couche implémenté dans l'environnement WEKA. Nous avons utilisé le perceptron multicouche avec une seule couche cachée vue quelle nous a donné les meilleurs résultats. Pour toutes les expérimentations nous avons utilisé la validation croisée (k=10).

Nous rapportons dans ce qui suit les résultats des expérimentations menées pour l'identification des relations fines (18 relations) ainsi que sur leurs classes (sept (07) relations fusionnées). En fait, nous avons généré deux modèles en utilisant les mêmes caractéristiques : le premier pour identifier les relations fines et le second pour identifier les classes de ces relations, appelées dorénavant les relations fusionnées.

Nos modèles sont comparés à deux modèles de référence (*baselines*) :

- Le premier modèle est la classe majoritaire (algorithme zeroR dans weka). Dans ce modèle, la relation prédite pour toutes les instances est la relation ' S-N.اسهاب/ *Elaboration*' pour la classification des relations fines et ' تفصيل / *Elaboration*' pour la classification des relations fusionnée.
- Le deuxième modèle de référence est le modèle basé sur les marqueurs de discours (F1).

5. Résultats et Analyse

Pour l'évaluation de performances, nous avons adopté pour chaque relation rhétorique (fine ou fusionnée), les métriques d'évaluation communément utilisées : la Précision, le Rappel et le F-score. Pour évaluer tout le système, nous avons utilisé les mesures d'exactitude et la macro F-score.

Formellement pour chaque classe i, la précision, le rappel et le F-score sont calculées comme suit :

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (1)$$

[2] https://sourceforge.net/projects/weka/

$$\text{Rappel} = \frac{\text{TP}}{\text{TP}+FN} \quad (2)$$

$$\text{F} - \text{score} = \frac{2*\text{Precision}*\text{Rappel}}{\text{Precision}+\text{Rappel}} \quad (3)$$

Avec:

TP : nombre d'instances de la classe i correctement classée

FP : nombre d'instances de la classe i mal prédites

FN : nombre d'instances n'appartenant pas à la classe i prédites par le système (mal classées)

La macro F-score est calculée comme suit

$$macro\ F - score\ = 1/k \sum_{i=1}^{k} F - score \quad (4)$$

K étant le nombre de classes.

$$exactitude = \frac{\text{nombre d'instances correctement classé}}{nombe\ total\ des\ instances} \quad (5)$$

5.1 Résultats globaux

Le tableau 4-2 présente les résultats globaux pour la classification des relations fines ainsi que pour les relations fusionnées. Comme nous pouvons le voir, les performances de nos modèles sont nettement meilleures que les deux modèles de référence pour les deux niveaux de classification. Pour la classification des relations fines, notre modèle surpasse le premier modèle (B1) et le deuxième modèle (B2) en termes d'exactitude par 62.46 % et 6.56% respectivement. De même, pour la classification des relations fusionnées, la performance de notre modèle est supérieure à celle du deuxième modèle de référence (B2) en termes de F-score de 5,05%. Ce qui démontre que l'intégration de nouvelles caractéristiques a permis d'enrichir la représentation des relations rhétoriques Arabes d'une manière globale.

	Classification des relations fines		Classification des relations fusionnés	
	F-score %	Exactitude %	F-score %	Exactitude %
BI	-	20.93	-	51.03
B2	70.82	74.12	81.69	85.54
Modèle proposé	78.49	83.39	86.74	91.03

Tableau 4-2 : Résultats Globaux pour les Deux Niveaux de Classification.

Pour les deux niveaux de classification, nous pouvons également noter que le deuxième modèle de référence (B2) fournit de très bons résultats par rapport à (B1).

Afin d'évaluer l'efficacité de chaque groupe de caractéristiques sur la classification des relations fines, nous avons mené plusieurs expérimentations dans lesquelles nous avons construit neuf classifieurs individuels. Chaque modèle a été construit en utilisant un ensemble de caractéristiques différentes. Nous avons d'abord utilisé uniquement la caractéristique F2 (verbes introductifs) et F1 (marqueurs de discours), puis nous avons appliqué une stratégie gloutonne, intégrant à chaque fois une nouvelle caractéristique. Le choix de cette dernière peut être assuré d'une manière aléatoire ou itérative. Dans notre cas, l'ajout de caractéristiques est fait itérativement selon l'ordre d'apparition de caractéristiques (dans la section 3). La Figure 4-1 présente les résultats en termes de F-score de ces expérimentations.

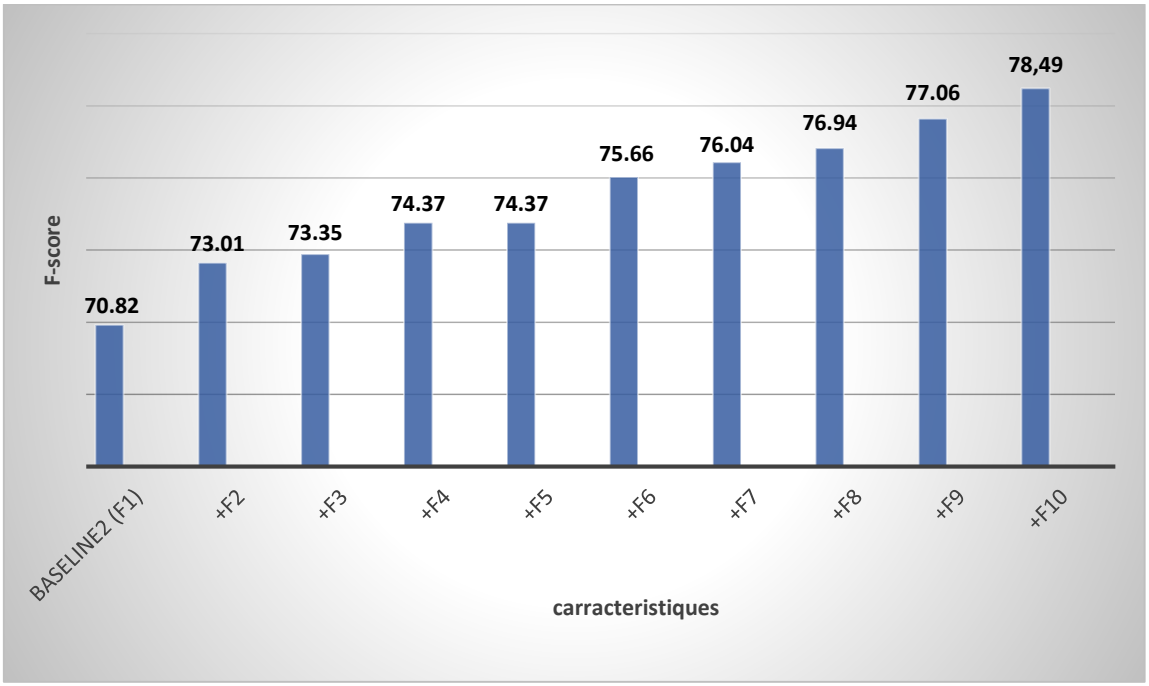

Figure 4-1 : Effet de Différents Caractéristique sur la Classification des Relations Fines.

Lorsque nous analysons l'impact de chaque groupe de caractéristiques, nous pouvons constater que la performance globale du système est notablement affectée par les marqueurs de discours (F1) et les verbes introductifs (F2). En effet, en ajoutant (F2) à (F1), la macro F-score a été améliorée par 2.19% par rapport à (B2) qui est basé sur (F1). L'ajout du connecteur spécifique (F3) a amélioré le F-score de 0,34%, tandis que (F6) a eu un impact important vue que la macro F-score a été amélioré de 1,29%. Il est à remarquer, cependant, que les signes de ponctuation (F5) n'ont aucun impact sur la performance globale d système.

En revanche, L'ajout des F7 (entités nommées) et F8 (la longueur) conduit à une amélioration en termes de F-score. L'ajout de F9 (données numériques) conduit à une amélioration de 0.12% tandis que l'ajout de F10 conduit à une amélioration estimée à 1.43%.

Afin de prouver l'efficacité des trois nouvelles caractéristiques (F3 , F6 , F10) et savoir si l'une de ces caractéristiques peut être omise en gardant le même degré de performance, nous avons effectué une deuxième série d'expérimentations dans laquelle nous avons d'abord omis à chaque fois une seule caractéristique puis nous avons omis les trois caractéristiques à la fois. La figure 4-2 présente les résultats de ces expérimentations.

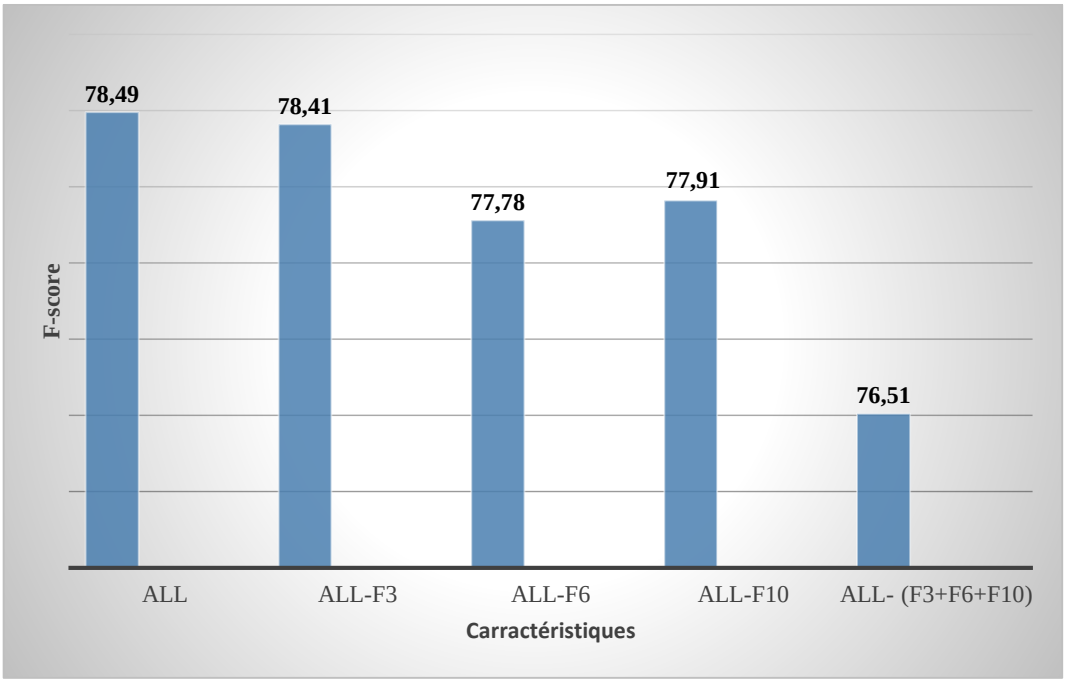

Figure 4-2 : Performance du notre modèle sans les trois nouvelles caractéristiques.

L'analyse des résultats obtenus, montre que l'efficacité de chaque caractéristique est réduite lorsque tous les autres groupes de caractéristiques sont pris en compte. Par exemple, supprimer (F3) (connecteur spécifique) conduit à une dégradation marginale en termes de F-score estimée à 0,08%. Ainsi supprimer (F6) (accusatif de but) et (F10) (nombre de mots antonymes) conduit à une dégradation des F- scores estimés à 0,29% et 0,58% respectivement. Tandis que la suppression des trois groupes de caractéristiques à la fois conduit à une dégradation en termes de f-score estimé à 1,98%.

Après avoir évalué l'efficacité de chaque groupe de caractéristiques, nous avons ensuite évalué la performance de notre modèle pour prédire les relations rhétoriques implicites. Les résultats ont montré que la performance de notre modèle en termes d'exactitude pour prédire des relations rhétoriques implicites est inférieure de 18,05% à sa capacité de prédire les relations explicites. Cela est dû à la couverture partielle des relations rhétoriques implicites dans notre corpus annoté (21%). Par conséquent, il est nécessaire d'avoir plus d'instances de relations rhétoriques implicites pour améliorer la performance de notre modèle.

5.2 Classification des Relations Fines

Le tableau 4-4 présente les résultats détaillés en termes de F-score pour la classification des relations fines. Les colonnes 3 à 9 montrent les résultats obtenus par l'ajout progressif de différents groupes de caractéristiques à B2 (modèle basé sur les marqueurs de discours). La dernière colonne montre les résultats obtenus par l'utilisation de tous les groupes de caractéristiques.

L'analyse des résultats obtenus nous permet de déduire que les caractéristiques utilisées ont une influence différente sur la prédiction des différentes relations rhétoriques. Par exemple, l'ajout de (F2) a fortement influencé la performance de la relation 'N-S.اسناد/ Attribution', vue que son F-score correspondant a été augmenté de 15% par rapport à (F1). Tandis que l'ajout de F3 (connecteurs spécifique) et F4 (المصدر/Al-masdar) a augmenté les performances de la relation ' S-N.كيفية / *manière*' de 11% et ' S-N.غاية/ *but*' de 4% par rapport à F1 + F2. Ils ont également augmenté le F-score de ' S-N.تعليل / *justification*' de 3%. D'un autre côté, l'ajout de F6 (accusatif de but) a fortement amélioré la performance de la relation
'S-N.غاية /*but*' vue que le F-score correspondant à cette relation a été augmenté de 7%.

Relation rhétoriques	(F1)	+F2	+F3	+F4	+F5	+F6	+F7	+F8	+F9	+F10
N-S.سبب	66.32	62.32	63.47	65.18	65.18	**69.95**	69.83	70.03	71.10	**71.13**
S-N.سبب	78.06	78.33	78.13	78.21	78.21	82.29	82.17	82.19	82.21	**82.22**
S-N.غاية	72.63	72.54	**72.54**	**77.12**	77.11	**84.04**	84.04	84.25	84.22	**84.19**
N-N.مقابلة	45.28	46.17	46.08	46.22	46.22	46.18	46.22	48.12	47.71	**61.48**
N-N.مقارنة	78.25	79.02	79.01	79.13	79.13	79.41	**79.66**	80.12	**86.23**	**88.27**
N-N.تشبيه	56.45	56.45	56.79	55.11	55.11	56.22	58.03	58.11	58.02	**58.00**
N-N.وصل	76.23	76.28	76.28	76.19	76.17	76.31	76.31	76.22	76.22	**76.27**
S-N.استدراك	73.32	73.27	73.22	74.01	74.01	74.12	75.08	75.19	75.22	**75.31**
S-N.اضراب	84.36	84.36	84.36	84.28	84.28	84.28	85.12	85.32	85.31	**85.31**
S-N.اسهاب	78.15	78.65	78.06	78.66	78.66	79.28	**81.02**	82.17	82.22	**82.19**
S-N.تخصيص	85.79	85.82	85.83	85.84	85.84	85.87	85.88	86.25	86.03	**86.09**
S-N.كيفية	69.25	69.22	**69.24**	**80.72**	80.72	80.49	81.38	81.22	81.21	**81.33**
N-S.زمني	75.97	75.97	75.94	75.95	75.95	75.96	75.96	75.98	**83.39**	**83.46**
S-N.زمني	47.23	74.23	74.18	74.29	74.30	75.08	75.08	75.18	82.18	82.26
S-N.تفسير	79.35	79.39	78.96	76.49	76.48	**79.88**	**80.12**	82.82	82.82	**82.76**
S-N.تعليل	73.23	73.14	75.22	76.88	76.88	**78.21**	78.18	78.21	78.21	**78.03**
N-S.اسناد	79.95	**94.21**	**98.21**	98.21	98.21	98.23	98.23	98.25	98.25	**98.28**
N-N.شرط	54.94	54.94	54.92	56.24	56.24	56.24	56.47	56.41	56.40	**56.41**
Macro F-score	70.82	73.01	73.35	74.37	74.37	75.66	76.04	76.94	77.06	**78.49**

Tableau 4-3 : Performance du Modèle Pour Chaque Relation en Termes de F-score

Contrairement aux caractéristiques précédentes, F5 (les signes de ponctuation) n'ont pas eu un impact sur la prédiction de toutes les relations. Nous pouvons expliquer cela par le fait que les signes de ponctuation ne sont pas utilisés régulièrement dans les textes arabes, voire que parfois ils sont rarement utilisés, c'est pour cela qu'ils n'ont pas eu d'influence sur la prédiction de toutes les relations.

En revanche, F7 (les entités nommées) et F8 (la longueur) ont amélioré les performances des relations ‘N-N. مقارنة / *comparaison*’, ‘S-N. اسهاب /*élaboration*’ et ‘S-N. تفسير / *explication*’, alors que F9 (donnés numériques) a eu un bon impact uniquement sur les relations ‘S-N.زمني/*temporel*’ et ‘N-S.زمني/*temporel*’.

En ce qui concerne les autres relations, il est à noter que la relation ‘N-N.مقابلة /contraste’ a atteint sa meilleure performance lors de l'ajout de (F10) (c'est-à-dire le nombre de mots antonymes). Ceci est en cohérence avec la définition de cette relation en Arabe qui tient quand il y a plus d'un mot dans le premier segment de texte et ses antonymes correspondants dans le deuxième segment. L'ajout de cette caractéristique a clairement discriminé entre les relations ‘N-N. مقابلة /*contraste*’ et ‘N-N. مقارنة/*comparaison*’ et a augmenté leurs performances respectives de 14% et 3%.

En général, nous pouvons dire que chaque groupe de caractéristiques a un effet différent sur la prédiction de différentes relations. Certaines caractéristiques sont cruciales pour prédire certaines relations et en même temps, elles ont un léger impact voire même un impact négatif sur les autres. Les marqueurs de discours (F1) sont très utiles pour les relations rhétoriques explicites, alors que les verbes introductifs (F2), l’accusative de but (F6), les entités nommées (F7) et les mots antonymes (F10) sont plus utiles pour les relations implicites.

L'analyse des erreurs à ce niveau a montré que notre modèle a échoué à discriminer entre les relations ‘S-N. سبب / *Cause*’ et ‘S-N. اسهاب / *Elaboration*’ quand ils sont implicitement signalés.

5.3 Classification des Relations Fusionnées

Le tableau 4-4 présente les résultats détaillés de la classification des relations fusionnées en termes de précision, rappel et F-score. La dernière ligne présente la précision moyenne, le rappel moyen et la macro F-score.

D’après les résultats présentés, on déduit que pour la classification des relations fusionnées, notre modèle a obtenu de très bons résultats avec une macro F-score = 86,74% et une macro précision = 91,05%. On remarque, aussi, que la meilleure performance a été obtenue pour la classe ‘ اسناد / Attribution’ avec un F-score = 98,32%, c’est-à-dire que c’est la classe la plus reconnue, tandis que la classe ‘الشرط/conditionnel’ est la classe la moins reconnue (F-score = 67,71%).

Relation	Precision%	Rappel%	F-score %
السببية/Causal	90.03	94.10	92.02
المقارنة/Comparison	87.48	93.05	90.17
العطف/Joint	82.24	94.86	88.10
اسهاب/Elaboration	85.71	96.28	90.68
اسناد/Attribution	99.27	97.39	98.32
توضيح/Explanation	81.05	79.41	80.22
الشرط/Conditional	66.32	69.18	67.71
la moyenne	**84.58**	**89.18**	**86.74**

Tableau 4-4 : Performance de Modèle pour la Classification des Relations Fusionnées

L'analyse des erreurs à ce niveau a montré que la majorité des erreurs sont entre la classe ' تفصيل / élaboration' et ' شرط / conditionnel'. Cela est dû à la distribution de ces relations dans notre corpus. Comme la relation ' تفصيل / élaboration' est la relation la plus dominante, le système a eu tendance à classer les relations ambiguës comme ' تفصيل / *élaboration*', ce qui explique le haut rappel et la faible précision de cette relation.

6. Conclusion

Dans ce chapitre, nous avons présenté notre approche pour l'identification automatique des relations rhétorique implicites et explicites Arabe définies dans le cadre de la RST. Nous nous sommes limités à l'identification des relations intra-phrases qui relient deux unités de discours au sein de la même phrase. Pour apprendre automatiquement ces relations, nous avons exploré plusieurs groupes de caractéristiques qui ont déjà prouvé leurs efficacités dans les travaux antérieurs sur l'identification automatique des relations rhétoriques. Nous avons également contribué en introduisant trois nouvelles caractéristiques nommées : 'المفعول لاجله */l'accusatif de but*', connecteur spécifique et le nombre de mots antonymes. Pour implémenter notre modèle nous avons utilisé le perceptron multicouches (MLP) qui nous donné de meilleurs performances par rapport à d'autres classifieurs déjà testés.

Les résultats expérimentaux ont montré que notre modèle a obtenu de très bonnes performances pour la classification des relations fines ainsi que pour leurs classes, et qu'il surpasse significativement tous les modèles de références (*baslines*).

Les résultats obtenus ont prouvé aussi l'efficacité de nos nouvelles caractéristiques et leurs impacts sur l'identification des relations rhétoriques.

Enfin, il est à noter que le modèle présenté dans ce chapitre est utilisé pour identifier les relations rhétoriques entre les unités de textes adjacents, qui est une étape primordiale dans notre approche de résumé automatique de textes arabes. Les détails de cette approche feront l'objet de prochain chapitre de cette thèse.

Contribution 3

Chapitre 5

Nouvelle Approche Pour le Résumé Automatique de Textes Arabes

Sommaire

1. Introduction

Dans la littérature, plusieurs méthodes concernant le résumé automatique de textes arabes ont été proposées pour évaluer la pertinence des segments textuels afin de générer des extraits automatiques. Bien que ces méthodes aient plus ou moins réussi à extraire les contenus les plus pertinents des textes sources, la cohérence des résumés générés par ces méthodes demeure un point qui est toujours sujet à des améliorations possibles. Le manque de cohérence dans les résumés automatiques est dû à la négligence des relations rhétoriques qui relient les segments constituants le texte source. En fait, un texte n'est pas un ensemble de phrases indépendantes mais plutôt une séquence de phrases liées par des relations rhétoriques et sémantiques pour exprimer un sens.

Ce chapitre a pour objet de présenter notre approche pour le résumé automatique de textes Arabes. L'approche proposée tente de remédier au problème de la cohérence dans les résumés automatiques de textes arabes par l'utilisation conjointe d'une méthode purement linguistique avec une méthode statistique.

Nous allons présenter d'abord l'approche proposée ainsi que les principales étapes puis nous passerons à l'évaluation de cette approche et à la discussion des résultats obtenus.

2. Approche proposée

Afin d'améliorer la qualité des extraits automatiques de textes Arabes, nous proposons une méthode hybride basée sur l'utilisation conjointe d'une méthode linguistique basée sur l'analyse de discours et plus précisément sur la théorie de la structure rhétorique avec une méthode statistique. Notre objectif est d'exploiter à la fois les relations rhétoriques qui existent entre les unités textuelles afin de générer des extraits cohérents. Ainsi la génération d'un extrait par une méthode hybride passe par deux phases principales. Dans la première phase, une analyse rhétorique du texte est effectuée afin de mettre en évidence les relations rhétoriques qui relient les unités élémentaires du texte. Ces relations constituent l'élément de base sur lequel un résumé primaire du texte source est généré.

Dans la deuxième phase, chaque phrase dans le résumé primaire est assigné un score en se basant sur certaines caractéristiques. Les phrases ayant les scores les plus élevés sont ensuite sélectionnées pour produire le résumé final tout en prenant en compte le ratio de compression fourni par l'utilisateur.

Ce procédé permet de coupler les informations issues d'une analyse profonde de texte avec celles issues d'un traitement statistique. Ce qui permet de profiter d'avantage des deux aspects dans l'amélioration des qualités des extraits Arabes.

Une des originalités de l'approche proposée réside dans sa capacité à identifier les passages pertinents d'un texte en s'appuyant principalement sur l'exploitation des relations rhétoriques entre les unités de textes élémentaires au lieu de la structure rhétorique du texte.

Dans la section suivante nous allons présenter les étapes par lesquelles passe la génération automatique du résumé obtenu par l'approche hybride proposée.

3. Etapes de génération du résumé

Comme nous l'avons déjà mentionné, la construction d'un résume passe par deux phases : la phase de l'analyse rhétorique et la phase de traitement statistique. Chacune de ces phases est composée de trois étapes principales. La Figure 5-1 présente le processus général du résumé automatique par l'approche proposée.

3.1. La phase de l'analyse rhétorique

C'est une phase primordiale dans notre approche. Cette phase est composée des étapes suivantes : segmentation du texte source, identification des relations rhétoriques, et compression des phrases.

3.1.1 Segmentation du document source

La segmentation est une étape incontournable dans tout système de résumé automatique. Cette étape consiste à décomposer le texte en unités textuelles plus petites en se basant généralement sur les signes de ponctuation. Comme notre approche s'appuie principalement sur une analyse rhétorique de texte, notre méthode de segmentation prend en compte les marqueurs de discours arabes et quelques signes de ponctuation. Ainsi, dans notre méthode, le texte est d'abord segmenté en phrases, de telle sorte qu'une phrase représente un passage textuel délimité par (.), puis chaque phrase est segmentée en unités de discours élémentaires (EDUs) en se basant sur un ensemble des règles de segmentation. Ces règles s'appuient principalement sur les marqueurs de discours Arabe (les marqueurs de discours définis lors de l'annotation de notre corpus annoté décrit dans le chapitre 3). De ce fait une phrase peut ne pas être segmentée en EDUs comme elle peut être segmentée en plusieurs EDUs ; tout dépend de la présence des marqueurs de discours.

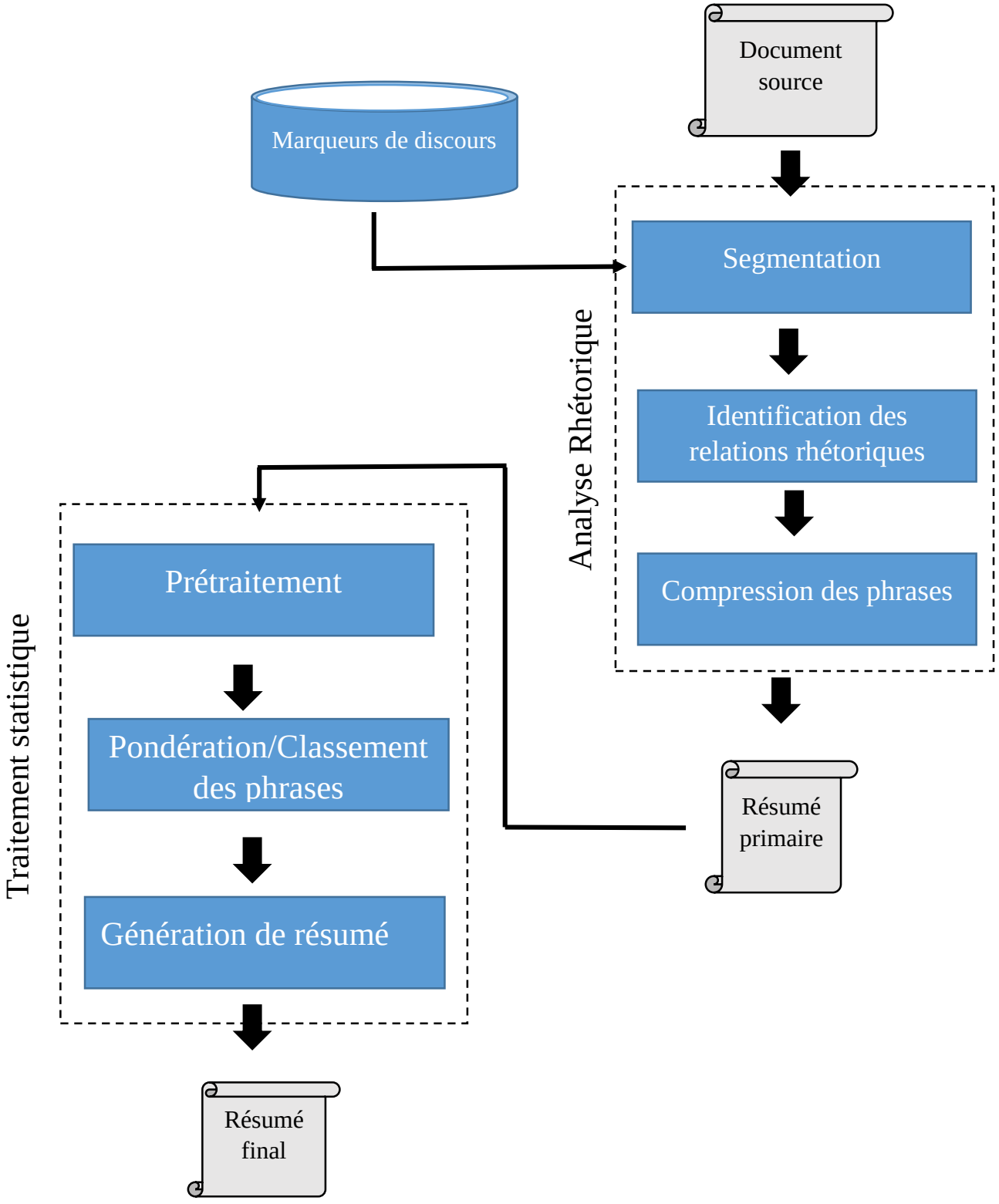

Figure 5-1: Etapes Principales de L'approche Proposée

Pour accomplir cette tâche, nous avons implémenté un segmenteur rhétorique à base de règles. Ce segmenteur reçoit en entrée un texte (sous format textuel) et fournit en sortie un texte segmenté en phrase et en EDUs. Le Tableau 5-1 présente un exemple du texte segmenté. Les EDUs sont séparés par le signe '#'.

N° de la phrase	Phrases segmenté en EDUs
1	والحقيقة ان قطاع غزة كان يقض مضاجع السلطات الاسرائيلية منذ زمن بعيد ، #اذ وقبل اندلاع الأعمال الفدائية بزمن، كان هناك مناضلون فلسطينيون ينطلقون من ذلك القطاع وعبره للقيام بعمليات قاسية ضد قوات الاحتلال الاسرائيلية.|
2	وكان الوضع يصل الى لحظات توتر قصوى، عند بدايات 1955،# حيث اندلعت أعمال عنف ضد القوات الاسرائيلية #وكذلك ضد المنشآت التابعة للامم المتحدة. |
3	ولقد دفع ذلك التوتر اسرائيل الى شن غارات قاتلة على القطاع، #مما ولد رد فعل قاس لدى السلطات المصرية بزعامة جمال عبدالناصر الذي، كان يحاول ان يهدئ الأوضاع، #مراعاة لخاطر الاميركيين من جهة، #وتأجيلاً للانفجار المحتمل بين مصر واسرائيل من جهة ثانية .|
4	ومن هنا حين احتلت القوات الاسرائيلية غزة #في العام 1956 #خيل للكثيرين انها لن تنسحب منها بعد ذلك، #على رغم الضغوط الدولية. |
5	ولقد حاولت فئات نيابية كثيرة، ومنها مجموعات من حزب العمل الحاكم، نفسه،ان تطرح الثقة في الحكومة، #لكن هذا كله لم يوهن من عزيمة حكومة بن غوريون التي لم تبد أي اهتمام حتى بالتظاهرات التي نظمتها المعارضة في الشارع# داعية الى الابقاء على احتلال قطاع غزة |
6	غير ان بن غوريون لم يأبه بكل تلك الاعتراضات،# بل واصل سياسته وانسحب وسلم الامم المتحدة مكان جيشهم| .

Tableau 5-1 : Exemple d'un Texte Segmenté en EDUs.

3.1.2. Identification des relations rhétoriques

Cette étape est cruciale dans notre processus de résumé. Elle consiste à prédire la relation rhétorique fine qui relie deux unités de discours élémentaires adjacents. Comme nous l'avons déjà mentionné dans le chapitre 3, les relations fines sont les relations enrichies par une notation de nucléarité (S-N, N-S, N-N) qui spécifie le statut rhétorique des segments reliés.

A titre d'exemple, considérons la phrase suivante composée de deux EDUs (séparées par #).

(1) غير ان بن غوريون لم يأبه بكل تلك الاعتراضات، # بل واصل سياسته وسلم الامم المتحدة مكان جيشهم| .

(1) Mais Ben Gourion ne se souciait pas de toutes ces objections, # mais il a poursuivi sa politique et remis aux Nations Unies, le lieu de leur armée.

Ces deux segments sont reliés par la relation fine '***S-N.اضراب/ rectification'*** tels que le premier segment est un noyau (dénoté par N) et le deuxième segment est un satellite (dénoté par S). Le segment satellite est un segment optionnel, tandis que le segment noyau est un segment primordial dans le texte.
La notation de nucléarité attachée au nom de la relation rhétorique est très intéressante dans notre travail, vu qu'elle nous fournit une information sur l'importance relative des segments faisant partie d'une relation rhétorique.

Afin d'identifier automatiquement les relations rhétorique fines entre les pairs de segments adjacents, nous avons employé un modèle à base de perceptron multicouche dont l'implémentation a déjà été détaillée dans le chapitre 4 de cette thèse. Ainsi, pour prédire les relations rhétoriques qui relient les segments d'un texte, nous avons procédé comme suit :

- D'abord, un vecteur de caractéristiques a été extrait pour chaque couple d'EDUs adjacents au sein de la même phrase. Nous avons utilisées les mêmes caractéristiques déjà détaillées dans le chapitre 4.
- Ensuite, l'ensemble des vecteurs de caractéristiques est transmis au modèle afin qu'il prédise les relations fines correspondantes. Le tableau 5-2 présente un exemple de relations prédites par notre modèle pour le texte présenté dans le tableau 5-1.

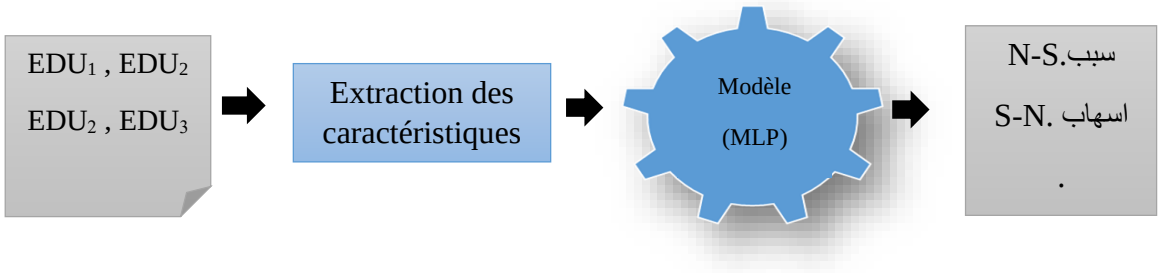

Figure 5-2 : Processus d'identification des Relations Rhétoriques.

Paire d'EDUs	Relation Rhétorique prédite
(والحقيقة ان قطاع غزة كان يقض مضاجع السلطات الاسرائيلية منذ زمن بعيد ، #اذ وقبل اندلاع الأعمال الفدائية بزمن، كان هناك مناضلون فلسطينيون ينطلقون من ذلك القطاع وعبره للقيام بعمليات قاسية ضد قوات الاحتلال الاسرائيلية)	اسهاب .S-N
(وكان الوضع يصل الى لحظات توتر قصوى، عند بدايات 1955،# حيث اندلعت أعمال عنف ضد القوات الاسرائيلية)	اسهاب .S-N
(اندلعت أعمال عنف ضد القوات الاسرائيلية #وكذلك ضد المنشآت التابعة للامم المتحدة.)	وصل.N-N
(ولقد دفع ذلك التوتر اسرائيل الى شن غارات قاتلة على القطاع، #مما ولد رد فعل قاس لدى السلطات المصرية بزعامة جمال عبدالناصر الذي كان يحاول ان يهدئ الأوضاع)	نتيجة.N-N
(.. الذي كان يحاول ان يهدئ الأوضاع#مراعاة لخاطر الاميركيين من جهة	غاية.S-N
(مراعاة لخاطر الاميركيين من جهة، #وتأجيلاً للانفجار المحتمل بين مصر واسرائيل من جهة ثانية)	وصل.N-N
(ومن هنا حين احتلت القوات الاسرائيلية غزة #في العام 1956)	زمني.S-N
في العام 1956 #خيل للكثيرين انها لن تنسحب منها بعد ذلك، على رغم الضغوط الدولية.	زمني.N-S
(ولقد حاولت فئات نيابية كثيرة، ومنها مجموعات من حزب العمل الحاكم، نفسه،ان تطرح الثقة في الحكومة، #لكن هذا كله لم يوهن من عزيمة حكومة بن غوريون التي لم تبد أي اهتمام حتى بالتظاهرات التي نظمتها المعارضة في الشارع)	استدراك.N-N
(لكن هذا كله لم يوهن من عزيمة حكومة بن غوريون التي لم تبد أي اهتمام حتى بالتظاهرات التي نظمتها المعارضة في الشارع# داعية الى الابقاء على احتلال قطاع غزة)	غاية.S-N
(غير ان دايان ومائير وبن غوريون لم يأبهوا بكل تلك الاعتراضات،# بل واصلوا سياستهم وانسحبوا وسلموا الأمم المتحدة مكان جيشهم)	اضراب.S-N

Tableau 5-2 : Exemple de Relations Rhétoriques Prédites.

3.1.3. Compression des phrases

Une fois les relations rhétoriques entre chaque couple d'EDUs identifiées, nous avons procédé à la compression des phrases par l'élimination des EDUs satellites tout en prenant en compte la cohérence globale de la phrase. Nous entendons par cette dernière remarque qu'un EDU ou un segment satellite est supprimé sous certaines contraintes qui garantissent que sa suppression n'altère pas le sens global de la phrase. Pour accomplir cette tâche, nous avons implémenté un algorithme dont le pseudo code est présenté dans le tableau 5-3. L'algorithme prend en entrée une phrase 'S' et fourni en sortie sa version compressé.

Procedure compression (S)

Suppr-set={} ;
Si N>=2 alors
{ i=1 ;
Tant que (i<N) faire
{ x=i, y=x+1
Si (Relation (EDU_x, EDU_y) ∈ RX alors ajouter EDU_x, EDU_y à suppr_Set
Sinon
{ Si (Relation (EDU_x, EDU_y)= R.S-N) Alors Ajouter EDU_x à suppr_Set
Sinon
Si Relation (EDU_x, EDU_y) = R.N-S Alors Ajouter EDU_y à suppr_Set
Sinon
Si (relation (EDU_x, EDU_y) =R.N-N) && (EDU_x ∈ suppr-set) alors
Ajouter EDU_y à suppr set
} i++
}
Si suppr-set ≠ {} alors
Pour i=1 à N faire {
If EDU_i ∈ suppr-set alors supprimer EDU_i de S
}
Return S
}

Tableau 5-3 : Pseudo Code pour l'algorithme de Compression.

Avec :

N : le nombre d'EDUs dans la phrase S, suppr_set : l'ensemble des EDUs à supprimer, Relation={Relation1 (EDU1 ,EDU2),…Relation2 (EDUn-1,EDUn)} la liste des relations prédite par notre modèle entre chaque paire d'EDUs adjacents dans la phrase 'S', R : relation rhétorique , RX : ensemble des relations rhétoriques non prises en compte={N-S.إسناد, N-N.مقابلة, S-N.تشبيه}

Il est à noter que, la compression des phrases s'appuie non seulement sur le statut rhétorique de ses constituants EDUs, mais aussi sur la relation rhétorique qui relie ses segments. Comme nous pouvons le voir dans le pseudo code ; nous avons établi une liste de relations (notée RX) qui ne sont pas utiles pour la tâche de résumé. Autrement dit, les relations dont les constituants EDUs ne sont pas vraiment importantes pour le résumé. Cette liste inclut les relations suivantes :{ N-S.اسناد, N-N.مقابلة, S-N.تشبيه}. Ainsi, la présence de telles relations implique la suppression de ses constituants EDUs même si ces constituants sont des noyaux.

Pour mieux appréhender la compression des phrases par l'application de notre algorithme, considérons à titre d'exemple la phrase suivante :

(1) ولقد دفع ذلك التوتر هذا البلد الى شن غارات قاتلة على جيرانه(1) مما ولد رد فعل قاس لدى السلطات العسكرية بزعامة ضباطها الذين، كانوا يحاولون تهدئة الأوضاع، (2) مراعاة لخاطر الاميركيين من جهة، (3) وتأجيلاً للانفجار المحتمل بين بلدهم وهذا البلد من جهة ثانية (4).

(2) Cette tension a conduit le pays à lancé des raids meurtriers sur ses voisins (1) provoquant une réaction brutale de la part des autorités militaires menées par ses officiers qui tentaient de calmer la situation, (3) en prenant en considération le risque des Américains d'un côté (4) et afin de retarder l'explosion probable entre eux et ce pays. (5)

Cette phrase est composée de quatre EDUs, reliées par les trois relations rhétoriques suivantes :

R(1,2)= N-N.نتيجة

R(2,3)=S-N. غاية

R(3,4)=N-N.وصل

Lors de l'application de notre algorithme présenté dans le tableau 4-3 sur cette phrase, l'EDU 3 est supprimée d'abord parce que c'est un segment satellite de EDU 2, puis l'EDU 4 est aussi supprimée bien que ce soit un segment noyau mais ce EDU est reliée par une relation multi-nucléaire à un EDU figurant dans la liste des EDUs qui doivent être supprimées.
Ainsi la phrase compressée après l'application de l'algorithme est la suivante :

(3) ولقد دفع ذلك التوتر هذا البلد الى شن غارات قاتلة على جيرانه(1) مما ولد رد فعل قاس لدى السلطات العسكرية بزعامة ضباطها الذين، كانوا يحاولون تهدئة الأوضاع.

(3) Cette tension a conduit le pays à lancé des raids meurtriers sur ses voisins (1) provoquant une réaction brutale de la part des autorités militaires menées par ses officiers (2).

L'application de l'algorithme de compression sur toutes les phrases du document génère un texte compressé et cohérent dont le nombre de phrases est le même que celui du texte source, mais ses phrases sont beaucoup plus abrégées de celles du texte source. Nous considérons ce texte comme un résumé primaire. La figure 5-3 présente un exemple de résumé primaire généré pour le fragment textuel présenté dans le tableau 5-1.

1. والحقيقة ان قطاع غزة كان يقض مضاجع السلطات الاسرائيلية منذ زمن بعيد
2. وكان الوضع يصل الى لحظات توتر قصوى، عند بدايات 1955.
3. ولقد دفع ذلك التوتر اسرائيل الى شن غارات قاتلة على القطاع، مما ولد رد فعل قاس لدى السلطات المصرية بزعامة جمال عبدالناصر الذي كان يحاول ان يهدئ الأوضاع.
4. ومن هنا حين احتلت القوات الاسرائيلية غزة خيل للكثيرين انها لن تنسحب منها بعد ذلك،
5. ولقد حاولت فئات نيابية كثيرة، ومنها مجموعات من حزب العمل الحاكم، نفسه، ان تطرح الثقة في الحكومة، لكن هذا كله لم يوهن من عزيمة حكومة بن غوريون التي لم تبد أي اهتمام حتى بالتظاهرات التي نظمتها المعارضة في الشارع.
6. غير ان بن غوريون لم يأبه بكل تلك الاعتراضات.

Figure 5-3. Un exemple du résumé primaire.

3.2. La phase de traitement statistique

Dans cette phase, le résumé primaire généré par la première phase va subir un traitement statistique. L'objectif de cette phase est de réduire le nombre de phrases dans le résumé primaire

en sélectionnant les phrases les plus pertinentes afin de construire le résumé final. L'accomplissement de cette tâche passe par les étapes suivantes : le prétraitement du résumé primaire, le classement des phrases et enfin la génération de résumé final. Chacune de ces étapes sera détaillée dans le reste de cette section.

3.2.1. Prétraitement du résumé primaire

Le prétraitement consiste en trois tâches principales : *tokenization*, suppression des mots vides et radicalisation ou *stemming*.

- ***Tokenization* :** Comme le résumé primaire est déjà segmenté en phrases, cette étape consiste à segmenter chaque phrase en mots distincts après avoir effectué un nettoyage du texte en enlevant les signes de ponctuation et les chiffres. Ainsi chaque phrase dans le résumé primaire est transformée en une liste des mots appelés *tokens*..
- **Suppression des mots vides**

L'objectif de cette étape est de supprimer les mots 'outils' qui sont du point de vue linguistique des mots vides de sens tels que les conjonctions, les prépositions...etc. Pour accomplir cette tâche, nous avons élaboré une liste contenant 128 mots vides arabes.

- **Radicalisation /*Stemming***

Le *stemming* est une technique morphologique qui consiste à chercher la racine lexicale (radical) des mots en langue naturelle et ceci par l'élimination des affixes (suffixes et préfixes) qui leurs sont attachées. En d'autres termes, chercher à regrouper sous un même identifiant des mots dans la racine est commune.

Dans notre travail nous avons utilisé un lemmatiseur pour la langue Arabe disponible en ligne[1] nommé '**JAVA Arabic Stemmer'.**

3.2.2. Pondération et classement des phrases

Apres l'étape de prétraitement, chaque phrase est assigné un score en se basant sur certaines caractéristiques visant à évaluer sa pertinence. Dans la littérature de résumé automatique de textes, plusieurs caractéristiques ont été adoptées y compris les termes clés, les expressions indicatives, la position de la phrase, les cohésions entre phrases...etc. (voir chapitre 2).

Dans notre approche, nous avons pris en compte les trois caractéristiques suivantes : position de la phrase, longueur de la phrases et similarité avec le titre. Ces caractéristiques ont été utilisées avec succès dans plusieurs travaux rapportés dans la littérature de résumé automatique

[1] https://sourceforge.net/projects/arabicstemmer/

de textes Arabe [AL RA, 2018 ; AL AB, 2017 ; DOUZ, 2004 ; FATT, 2009 ; SOBH, 2006]. Chacune de ces caractéristiques est présentées dans le reste de cette section.

A. Caractéristiques utilisées

1. similarité avec le titre

Comme le titre couvre généralement le thème principal abordé dans le texte, les mots de titre sont des mots pertinents et peuvent être considérés comme des termes clés. De ce fait, les phrases qui contiennent les mots du titre sont des phrases pertinentes et doivent être incluses dans le résumé final. Le score de similarité avec le titre attribué à chaque phrase est fonction des cooccurrences des mots de titre dans la phrase. Ce score est calculé en utilisant la formule suivante :

$$similarité(Si, T) = \frac{\text{Nombre de mos de titre dans la phrase (Si)}}{\text{Nombre de mot de titre}} \qquad (1)$$

Il est à noter que pour estimer correctement la similarité lexicale des phrases avec le titre, les mêmes traitements (tokenization , élimination des mots vides et stemming) ont été appliqués sur le titre au préalable.

2. Position de la phrase

La position d'une phrase dans le texte peut être un bon indicateur qui reflète son degré d'importance. Généralement les phrases les plus liées au thème du document sont soit situées au début du document, c'est-à-dire dans les premières positions, soit à la fin de ce document. C'est pour cela que nous considérons que la première et la dernière phrase dans le résumé primaire sont très importantes et doivent être incluses dans le résumé final. C'est pour cela qu'à ces deux phrases sont assignés des poids plus élevés qu'au reste des phrases. Ainsi, le score de la position assigné à une phrase est calculé comme suit :

$$pos\ (Si) = \begin{cases} 1 & si\ i = 1\ ou\ i = N \\ 0.5 & sinon \end{cases} \qquad (2)$$

Avec :

i : le numéro d'ordre de la phrase

N : le nombre total des phrases.

3. Longueur de la phrase

Un poids est assigné à chaque phrase en se basant sur sa longueur (la longueur en termes de mots). Comme la majorité des phrases dans le résume primaire ne contient que les éléments noyaux du texte source, c'est à dire les éléments primordiaux dans le texte, on peut dire que les phrases les plus longues sont celles qui sont les plus susceptibles à contenir plus d'informations pertinentes. Selon ce principe, les phrases très courtes ont une très faible chance d'être incluses dans le résumé final vue qu'elles couvrent moins d'informations pertinentes. Ainsi un score portant sur la longueur est assigné à chaque phrase et est calculé comme suit :

$$longuer(Si) = \frac{N(Si)}{N(Sl)} \qquad (3)$$

Avec :

N(Si) : nombre de mots dans la phrase Si

N(SL) : nombre de mots dans la phrase la plus longue dans le résumé primaire.

B. Classement des phrases

Le classement des phrases dans le résumé primaire est basé sur les scores finaux qui leurs sont attribués. Le score final de chaque phrase est une combinaison linéaire de ses scores attribués pour chaque caractéristique. Il est calculé en utilisant la formule suivante :

$$Score\ (Si) = similarité(Si, T) + longuer(Si) + pos\ (Si) \qquad (4)$$

Le score final d'une phrase représente son degré de pertinence dans le texte (le résumé primaire). Ainsi, les phrases sont classées dans un ordre décroissant en fonction de leurs scores finaux. Les phrases ayant les scores les plus élevés sont sélectionnées pour produire le résumé final dans l'étape suivante.

3.2.3. Génération du résumé final

Cette étape consiste à sélectionner les phrases les mieux classées pour produire le résumé final. Les phrases sélectionnées sont assemblées et arrangées selon leur ordre d'apparition dans le résumé primaire pour assurer la lisibilité du résumé. Le nombre de phrases extraites pour produire le résumé final dépend du taux de compression spécifié par l'utilisateur.
Il faut noter que la génération automatique des résumés nécessite des technique de fusion et de reformulation des phrases, mais comme cela est pratiquement difficile à maintenir et nécessite beaucoup de ressources linguistiques avancées, les approches de résumés par extraction se limitent à l'assemblage et à l'ordonnancement des phrases sélectionnées. La figure 4 présente le résumé final généré avec un taux de compression de 50% pour le fragment textuel présenté dans le tableau 1.

والحقيقة ان قطاع غزة كان يقض مضاجع السلطات الاسرائيلية منذ زمن بعيد
ولقد دفع ذلك التوتر اسرائيل الى شن غارات قاتلة على القطاع، مما ولد رد فعل قاس لدى السلطات المصرية بزعامة جمال عبد الناصر الذي كان يحاول ان يهدئ الأوضاع.
ولقد حاولت فئات نيابية كثيرة، ومنها مجموعات من حزب العمل الحاكم، نفسه، ان تطرح الثقة في الحكومة، لكن هذا كله لم يوهن من عزيمة حكومة بن غوريون التي لم تبد أي اهتمام حتى بالتظاهرات التي نظمتها المعارضة في الشارع.

Figue 5-4 : Résumé Final du texte.

4. Evaluation de l'approche proposée

L'approche proposée a été implémenté en java. Plusieurs ressources ont été utilisées parmi lesquelles nous citons l'analyseur morphologique [BOUD, 2017], notre classifieur des relations rhétoriques Arabes (voir chapitre 4), l'environnement d'apprentissage WEKA, …etc

Pour évaluer les performances de notre système nous nous sommes appuyés sur une évaluation intrinsèque. Comme nous l'avons déjà mentionné dans le chapitre 1, les méthodes intrinsèques cherchent à évaluer les résumés automatiques en se basant sur leurs formes et leurs contenus. L'évaluation de la forme peut être faite manuellement par un certain nombre de juges

qui mettent en valeur la lisibilité, la cohérence et la grammaire des résumés. Tandis que l'évaluation des contenus mesure la capacité du résumé à identifier les phrases pertinentes du document source, cela peut être fait automatiquement via une comparaison avec des résumés de référence produits par des experts.

Pour évaluer notre système. Nous avons effectué deux types d'évaluations : l'une est manuelle faite par des humains qualifiés et l'autre est automatique. Ainsi, notre évaluation porte sur deux ensembles de données : une collection d'articles issue de *Arabic corpus* pour l'évaluation manuelle et le corpus EASC pour l'évaluation automatique.

4.1 Evaluation automatique

L'évaluation automatique d'un système de résumé automatique de textes implique une comparaison des résumés générés par le système avec des résumés de référence produits par des experts. Pour cela, nous avons utilisé le corpus EASC créé par El Haj et al. [El HA, 2010] (EASC). Ce corpus est composé de 153 documents extraits de Wikipédia et de deux journaux arabes : ALRai et ALwatan. Le corpus couvre dix thèmes différents : art et musique, éducation, environnement, finance, santé, politique, religion, science et technologie, sport et tourisme

Pour Chaque document dans EASC, cinq résumés de référence produits par des personnes Arabes sont disponibles. C'est-à-dire que le corpus est composé au total de 153 articles et 765 résumés de référence dont la taille ne dépasse pas 50% de la taille du texte source. Le corpus est disponible en ligne avec deux encodages : UTF-8 et ISO-Arabic.

La majorité des documents dans EASC sont issus de Wikipédia (106 articles). Comme nous nous focalisons dans notre travail sur le résumé automatique des articles de presse, nous avons sélectionné les quarante-sept (47) articles de presse issus du journal Alwatan et Alrai pour l'évaluation de notre système.

Comme les articles dans EASC n'ont pas de titres, nous avons demandé à deux experts de lire attentivement chaque article dans la collection sélectionnée et de lui associer un titre adéquat. Un titre n'est attribué à un article que s'il y a un accord total entre les deux experts.

Pour l'évaluation de notre système, nous avons utilisé les deux métriques de ROUGE [LIN, 2004] : ROUGE-1 et ROUGE-2. Nous avons utilisé la version 2.0 de ROUGE qui peut être utilisée pour plusieurs langues y compris la langue Arabe.

Dans le reste de cette section nous allons d'abord présenter un aperçu général sur les métriques d'évaluation que nous avons utilisé, puis nous passons à la discussion des résultats obtenus.

4.1.1 Métriques d'évaluation

ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [LIN, 2004], est un ensemble de métriques utilisé pour l'évaluation des systèmes de résumé automatique de textes. ROUGE est basé sur une comparaison de résumé machine avec un ou plusieurs résumé de référence. C'est un package qui comprend les cinq mesures suivantes : ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S et ROUGE-SU (voir chapitre 1).

ROUGE-N calcule le nombre de N-grammes (N mots successifs) communs entre le résumé machine et un ou plusieurs résumés de référence. Différentes variantes de N-grammes peuvent être utilisées tels les uni-grammes (un seul mot), les bi-grammes (deux mots successifs), les trigrammes (trois mots successifs)...etc. Par exemple ; ROUGE-1 calcule le nombre des uni-grammes cooccurrentes entre le résumé généré par le système et les résumés de référence, ROUGE-2 calcule le nombre des bi-grammes communs entre le résumé machine et les résumés de référence. Pour chacune de ces métriques, la précision, le rappel et le F-score sont calculés afin de fournir une information complète sur le système.

- Le Rappel dans le contexte de ROUGE indique l'exhaustivité du système. Autrement dit, combien de N-grammes du résumé de référence ont été récupérés par le système. Le rappel est calculé en utilisant la formule suivante.

$$Rappel = \frac{nombre\ de\ N-gramme\ commun}{nombre\ de\ N-gramme\ dans\ le\ résume\ de\ référence} \qquad (1)$$

- La précision indique l'exactitude du système. C'est-à-dire combien de phrases dans le résumé généré par le système sont pertinentes. La précision est calculée à l'aide de la formule suivante :

$$precision = \frac{nombre\ de\ n-gramme\ commun}{nombre\ de\ n-gramme\ dans\ \ dans\ le\ résumé\ machine} \qquad (2)$$

Plus la précision est élevée plus le système est apte à exclure les phrases non pertinentes. Une précision égale à 1 signifie que toutes les phrases extraites par le système sont pertinentes.

- F-score : combine la précision et le rappel, cette mesure est calculée comme suit :

$$F - score = \frac{2 * precision * rapell}{precision + rapell} \qquad (3)$$

Afin de mieux appréhender l'évaluation automatique des résumés en utilisant les métriques ROUGE-1 et ROUGE-2, considérons à titre d'exemple le résumé généré par le système ' S' et le résumé de référence ' R' tels que :

S : دخل الولد القسم

R :دخل الولد المدرسة مبتسما

Les valeurs de la précision et le rappel en utilisant ROUGE-1 et ROUGE-2 sont présentées dans les colonnes 5 et 6 respectivement du tableau 5-3. La colonne 4 présente les uni-grammes (respectivement les bi-grammes) communs aux résumés générés par le système (S) et le résumé de référence (R).

	S	**R**	$S \cap R$	**Précision**	**rappel**
ROUGE-1	دخل-الولد-القسم	-دخل- الولد- المدرسة- مبتسما	دخل- الولد	2/3	2/4
ROUGE-2	دخل الولد- الولد القسم	دخل الولد- الولد المدرسة- المدرسة مبتسما	دخل الولد	1/2	1/3

Tableau 5-4 : Exemple d'application de ROUGE-1 et ROUGE-2

4.1.2. Résultats et analyse

Afin d'évaluer notre système, nous avons effectué deux expérimentations : dans la première expérimentation, chaque article dans la collection des articles de presse du corpus EASC a été résumé automatiquement par notre système de résumé automatique avec un taux de

compression TC= 50%. Ce qui fait que nous avions au total 47 résumés machines et 235 résumés de référence. Chaque résumé généré par notre système a été évalué para port à 5 résumé de référence dans EASC en utilisant les métriques ROUGE-1 et ROUGE-2. Figure 5-5 présente les résultats d'évaluation obtenus.

Par la suite ; nous avons effectué une deuxième expérimentation similaire à la première mais en utilisant un taux de compression =30%, c'est-à-dire que les 47 articles ont été résumé par notre système avec un ratio de compression TC=30%. Figure 5-6 présente les résultats obtenus.

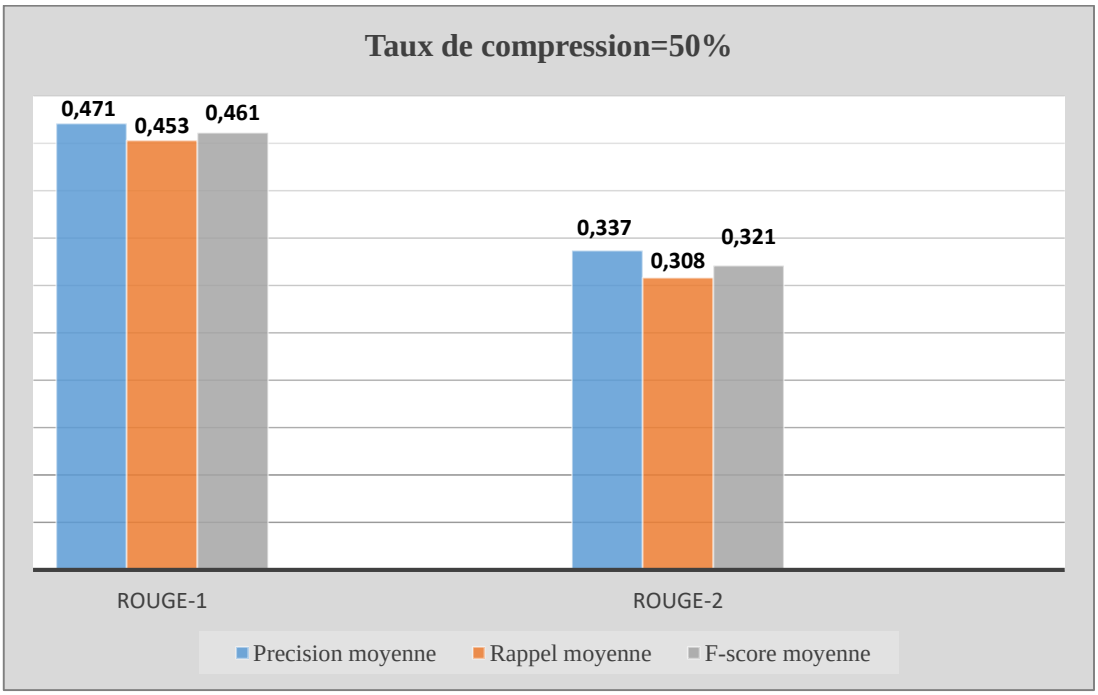

Figure 5-5 : Résultats d'évaluation pour un TC=50%.

En analysant les résultats présentés Figure 5-6, nous pouvons remarquer que notre système atteint de très bonnes performances en utilisant les deux métriques ROUGE-1 et Rouge-2. En utilisant ROUGE-1, le F-score moyen du système =0.453 et pour ROUGE-2, le F-score moyenne= 0.32, ce qui indique un haut degré de performance. Nous pouvons noter aussi que la précision moyenne de notre système pour les deux métriques est très élevée (précision moyenne= 0.471 en utilisant ROUGE-1 et 0.337 en utilisant ROUGE-2) cela signifie que notre système est vraiment apte à exclure les phrases non pertinentes.

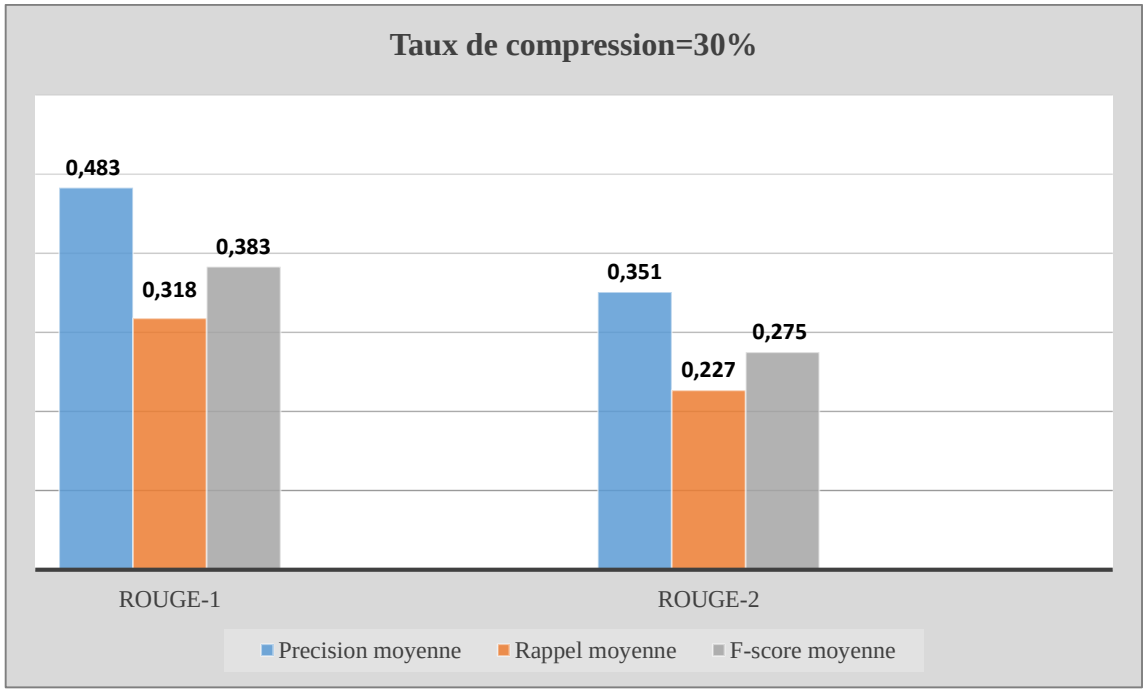

Figure 5-6 : Résultats d'évaluation pour un TC=30%

En analysant les deux figures en parallèle, on peut noter que l'augmentation du taux de compression du texte source (Tc=30%) a conduit à une dégradation en termes de rappel moyen. Cela est justifié par le fait qu'en augmentant le taux de compression (TC =30%) le nombre de phrases à inclure dans le résumé diminué et par conséquent le nombre de phrases pertinentes dans le résumé va systématiquement diminuer ce qui peut être interprété par une dégradation en terme de rappel moyen du système. Par ailleurs, on diminuant le taux de compression (TC =50%), le nombre de phrases à inclure dans le résumé est augmenté, ainsi le nombre de phrase pertinente est augmenté ce qui conduit à une augmentation en termes de rappel moyen.

Nous pouvons noter aussi que la diminution du taux de compression a augmenté la précision du système. Une justification possible est que plus le nombre de phrases à inclure dans le résumé est réduit plus le nombre de phrases non pertinentes est minime, vue que le système va sélectionner juste les phrases les mieux classées, c'est à dire les plus pertinentes du document source et cela va augmenter la précision du système

D'après les valeurs du F-score moyen de notre système, on peut dire qu'avec un taux de compression =50%, notre système est capable d'identifier la majorité des phrase pertinentes dans les résumés de référence et avec une grande précision.

4.2. Evaluation manuelle

4.2.1 Démarche suivie

Apres avoir effectué une évaluation quantitative de notre système, nous avons procédé à une évaluation qualitative faite par des juges humains afin de confirmer ou non les tendances qui s'étaient dégagées lors de la première évaluation. Pour cela, nous avons sélectionné au hasard 30 articles de presse de '*Arabic corpus'*. Comme nous l'avons déjà mentionné dans le chapitre 3, *Arabic corpus* est un grand corpus Arabe non étiqueté disponible en ligne pour exploration. Les articles sélectionnés sont de tailles différentes et couvrent plusieurs thèmes dont la majorité est politique. Ces articles ont été résumés par notre système avec un taux de compression égale à 50%. C'est-à-dire que le nombre de phrases dans le résumé final correspond à la moitié du nombre de phrases dans le texte source. Il est à noter que ce taux est considéré comme un taux de compression élevé, vu que les phrases dans le résumé final sont fortement compressées par rapport à celles dans le texte source.

Quinze étudiants de master ont participé à cette évaluation. Nous avons affecté à chaque étudiant deux textes avec leurs résumés automatiques, et nous leur avons demandé d'attribuer une note qualitative à chaque résumé après avoir lu le texte source correspondant.

Les notes qualitatives que nous avons utilisées sont les suivantes :

- Très mauvais : le résumé ne contient que des phrases non pertinentes.
- Mauvais : le résumé contient quelque phrases pertinentes mais la majorité est non pertinente.
- Acceptable : toutes les phrases dans le résumé sont pertinentes mais le résumé ne couvre pas toutes les phrases pertinentes du document.
- bon : la majorité des phrases dans le résumé sont pertinentes mais le résumé est peu cohérent.
- Très bon : le résumé est cohérent et couvre la majorité des phrases pertinentes du texte source.

Chaque étudiant dispose de deux heures pour noter soigneusement les deux résumés que nous lui avons affectés. Nous pensons que cette durée est largement suffisante vue que la taille maximale des textes source ne dépasse pas quarante phrases.

4.2.2 Résultats et analyse

Figue 5-7 présente les résultats de l'évaluation manuelle. Dès la première observation nous pouvons constater qu'à aucun résumé n'est attribuée la note 'très mauvais', qui est assigné lorsque le résumé ne contient aucune phrase pertinente. Cela signifie que tous les résumés générés par notre système contiennent au moins quelques phrases pertinentes. D'un autre côté, nous pouvons remarquer que neuf résumés son notés comme étant des bon résumés, c'est-à-dire qu'ils contiennent la majorité des phrases pertinentes mais qu'ils sont peu cohérents, tandis que treize résumés sont notés comme de très bons résumés, c'est-à-dire qu'ils sont des résumés cohérents et qu'ils contiennent la majorité des phrases pertinentes. Cela démontre que l'approche proposée est capable non seulement de capter les éléments pertinents du texte source mais aussi d'assurer la cohérence globale du résumé généré. Cette cohérence est due essentiellement à la prise en considération des relations rhétoriques fines qui relient les unités des textes lors de l'évaluation de la pertinence des phrases. Les relations rhétoriques fines que nous avons utilisées sont définies dans le cadre de la théorie de la structure rhétorique, c'est à dire que nos relations indiquent non seulement la relation rhétorique qui se tient entre deux segments de texte mais aussi le statut rhétorique de ces segments dans le texte. Ainsi la prise en compte de telles informations nous a permis de capter les segments les plus pertinentes (les noyaux) des textes et d'assurer la cohérence du résumé final.

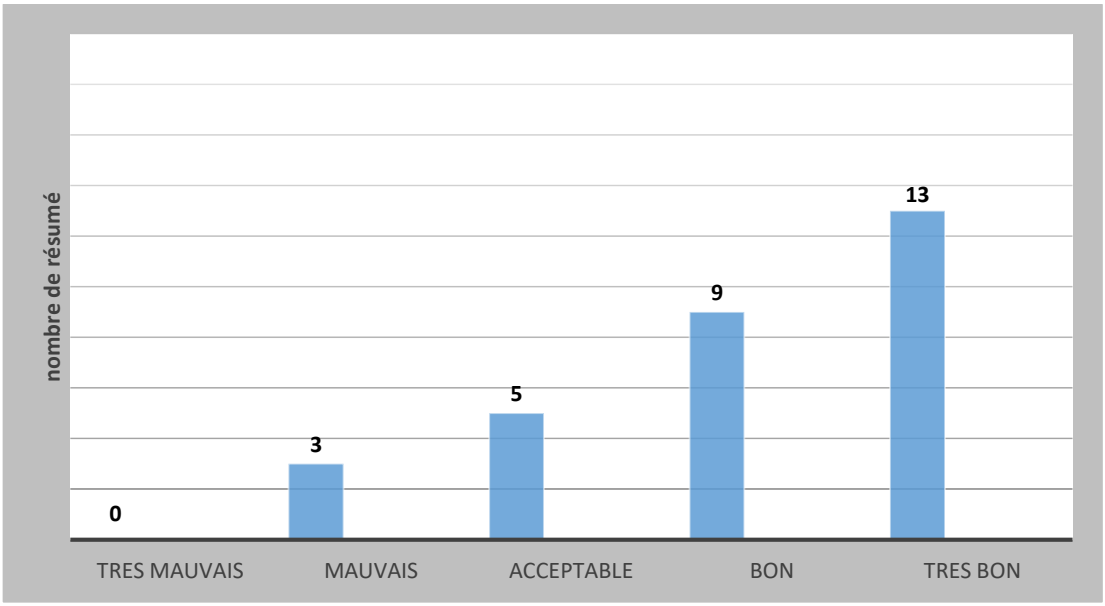

Figure 5-7 : Résultats de L'évaluation Manuelle.

Pour les autres documents notés mauvais (3 résumés) et acceptables (5 résumés), nous pensons que cela est dû aux erreurs dans la phase de segmentation. En fait, la segmentation d'un texte en EDU est une tache extrêmement difficile à cause de l'ambiguïté des marqueurs de discours dans la langue Arabe. Des erreurs commises à ce niveau vont influencer négativement l'identification des relations rhétoriques ce qui conduit à une mauvaise évaluation de la pertinence des phrases.

5. Conclusion

Dans ce chapitre nous avons présenté une nouvelle approche pour le résumé automatique de texte arabe. L'approche proposée combine un traitement purement linguistique basé sur une analyse rhétorique de texte avec un traitement statistique.

L'objectif d l'analyse rhétorique est de compresser les phrases tout en ne gardant que les éléments primordiaux. La compression de phrases est basée sur l'exploitation des relations rhétoriques intra-phrases définies dans le cadre de la théorie de la structure rhétorique

Le traitement statistique a pour objectif de réduire le nombre de phrases compressées issues de l'analyse rhétorique. La sélection des phrases les plus pertinentes est basée sur trois caractéristiques : la position de la phrase, la similarité avec le titre ainsi que la longueur de la phrase.

L'évaluation de l'approche proposée a permis de prouver son efficacité non seulement dans l'extraction des éléments les plus pertinents mais aussi dans la génération de résumés cohérents. En fait la négligence des relations rhétoriques et sémantiques entre les phrases est la cause principale du manque de cohérence dans les résumés automatiques. La prise en compte de ces relations nous a permis de générer des résumés cohérents et plus représentatifs du texte source.

Conclusion Générale et Perspectives

La problématique abordée dans cette thèse concerne le domaine de résumés automatique de textes en langue Arabe. Nous nous sommes fixés comme objectif majeure celui d'améliorer la qualité des extraits automatiques en langue arabe par la proposition d'une nouvelle approche qui exploite au mieux les relations rhétoriques reliant les segments textuels et de confirmer que la prise en compte de ces relations va conduire à la génération de résumés cohérents. Nos objectifs ont été finalement atteints en grande partie.

Notre approche s'appuie principalement sur une analyse rhétorique afin d'éliminer les segments inutiles dans une phrase. Puis un traitement statistique est appliqué afin de réduire le nombre de phrases compressés. Les résultats de l'évaluation automatique sont très satisfaisants ; nous avons obtenu des F-scores moyens de 0.46 et 0.32 en termes de ROUGE-1 et ROUGE -2 respectivement. Nous rappelons que dans le domaine des résumés automatiques de texte, de telles performances sont très bien perçues.

Pour les deux métriques d'évaluation ci-dessus, la précision moyenne de notre système est supérieure au rappel moyen. Cela signifie que la capacité de notre système à éliminer les segments inutiles est supérieure à sa capacité à sélectionner tous les segments pertinents. Ce qui signifie que les caractéristiques statistiques que nous avons utilisé pour juger de la pertinence des phrases dans le résumé primaire ne sont pas suffisantes. Il faut, ainsi, soit augmenter le nombre de caractéristiques ou tout simplement s'appuyer sur un autre traitement plus poussé dans la deuxième phase. Nous avons également effectué une évaluation manuelle de notre système, les résultats de cette évaluation ont bien confirmé les tendances dégagées par l'évaluation automatique.

Tout au long de ce travail, nous avons pu ressentir que le manque de ressources en langue Arabe constitue un obstacle qui ralentit la progression des recherches dans ce domaine. En fait, à l'heure actuelle et à notre connaissance, il n'existe pas de corpus de discours arabe annoté selon le principe de la théorie de la structure rhétorique quoi que cette théorie soit apparue en 1988 et qu'elle a prouvé beaucoup de succès non seulement pour le résumés automatique de textes mais dans plusieurs domaines tels que la traduction automatique, la génération automatique de textes…etc. Ainsi, plusieurs corpus annotés dans le cadre de cette théorie ont été développés dans plusieurs langues autres que l'arabe, ce qui a permis effectivement la

progression des recherches dans le domaine de l'analyse de discours ainsi que ses applications pour ces langues.

L'absence de telles ressources en langue arabe a constitué un véritable obstacle pour notre recherche. Ceci nous a amené à élaborer manuellement ce type de corpus, ce qui représente en soi une tâche assez fastidieuse qui a nécessité beaucoup d'efforts, quoi que dans notre processus d'annotation nous nous sommes limités juste à l'annotation des relations intra-phrases.

Le manque de ressource en langue Arabe ne concerne pas seulement les corpus de discours annotés, il concerne aussi les outils de base du traitement automatique des textes arabes tels que les segmenteurs de texte, les plateformes dédiées au traitement automatiques de la langue arabe… et autres outils.

S'appuyer sur l'analyse de discours arabe pour résumer les textes Arabes était extrêmement difficile vus les défis susmentionnés mais en contrepartie ; cette technique est très avantageuse vue qu'elle permet de générer des extraits avec un haut degré de cohérence.

Notre travail comporte certaines limites ouvrant la voie à plusieurs axes de recherche. Ainsi, parmi les perspectives futures qui peuvent être dégagées, nous citons les suivantes qui nous semblent les plus importantes :

- Exploiter les relations rhétoriques inter-phrases pour réduire le nombre de phrases au lieu de s'appuyer sur des calculs statistiques. Les relations rhétoriques interphases sont les relations qui relient les phrases dans le texte. Identifier ces relations pourrait permettre d'éliminer les phrases ou même les paragraphes non pertinents. En effet, une phrase ou un paragraphe qui est une élaboration d'une information présente dans une autre phrase qui la précède peut être supprimée sans altérer le sens et la cohérence globale du texte. A cet effet, nous pensons que l'exploitation des relations rhétoriques intra-phrases et interphases va nous servir à générer des résumés plus cohérents et qui couvrent l'information pertinente du texte source.
- Afin de remédier au problème de la redondance dans les résumés automatiques de texte Arabe, nous envisageons également d'exploiter les relations d'implication textuelle (*textual entailment*) entre les phrases. En fait une phase 'S' qui implique une autre phrase 'P' signifie que le sens de la phrase 'P' peut être inféré du sens de la phrase 'S'. Par conséquent l'existence de la phrase 'P' avec la phrase 'S' dans le résumé va engendrer une redondance. L'exploitation des relations d'implication textuelle nous semble un axe de recherche très promoteur.

- Focaliser notre recherche sur un domaine de spécialité tel que le résumé de textes biomédicaux ou de textes juridiques en langue Arabe. De tels outils fourniraient une aide intéressante pour la prise de décision des experts du domaine car ils auraient accès à l'information pertinente de façon très rapide.
- Un autre axe de recherche, encore plus prometteur, concerne l'utilisation de l'ingénierie des connaissances dans la production de résumés de textes arabes. Il faudrait, ainsi, constituer des ontologies de domaine ainsi que des outils pour exploiter la sémantique des textes arabes et en déduire des résumés en s'appuyant cette fois-ci sur la sémantique des phrases, voire même des paragraphes. Actuellement, beaucoup de recherches convergent vers l'exploitation rationnelle du web sémantique. Le résumé de texte arabe sémantique annonce une ère nouvelle de recherche et d'applications dans ce domaine.

  En ce qui nous concerne, nous pensons orienter notre recherche future vers cet objectif, qui est en passe de s'imposer par la force des choses et par le développement d'objets intelligents.

Références Bibliographiques

[ABDU, 2006] Abdul- Raof, H. (2006). 'Arabic Rhetoric. A Pragmatic Analysis'. Routledge, Taylor & Francis Group, LONDON and NEW YORK, ISBN10: 0-415-38609-8.

[AL AB, 2017] Al-Abdallah, R. Z., Al-Taani, A. T. (2017). 'Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm'. *Procedia Computer Science*, *117*, pp. 30-37.

[AL JA, 2012] Al Jarim, A. Amin, M. (2012). 'Al-Balagha Al-Wadiha'. *Dar Al-Maaref, LONDON*, ISBN13:9789770276617.

[AL KH, 2015] Al Khawaldeh, F., Samawi, V. (2015). 'Lexical cohesion and entailment based segmentation for Arabic text summarization (LCEAS)'. *The World of Computer Science and Information Technology Journal (WSCIT), 5*(3), pp.51-60.

[AL RA , 2018] Al-Radaideh, Q. A., Bataineh, D. Q. (2018).' A Hybrid Approach for Arabic Text Summarization Using Domain Knowledge and Genetic Algorithms'. *Cognitive Computation*, pp.1-19.

[AL TA, 2014] Al Taani, A. T., Al-Omour, M. M. (2014). 'An extractive graph-based Arabic text summarization approach'. *The International Arab Conference on Information Technology, Jordan.*

[ALAM, 2017] Alami, N., El Adlouni, Y., En-nahnahi, N., et Meknassi, M. (2017).'Using Statistical and Semantic Analysis for Arabic Text Summarization'. *International Conference on Information Technology and Communication Systems*, Springer, Cham. pp. 35-50.

[ALIG, 2010] Aliguliyev, R. M. (2010). 'Clustering Techniques and Discrete Particle Swarm Optimization Algorithm for Multi-Document Summarization'. *Computational Intelligence*, *26*(4), pp. 420-448.

[ALOT, 2012] Alotaiby, F., Foda, S., Alkharashi, I. (2012). 'New approaches to automatic headline generation for Arabic documents'. *Journal of Engineering and Computer Innovations*, *3*(1), pp.11-25.

[ALSA, 2011] Alsaif, A., Markert, K. (2011). 'Modelling discourse relations for Arabic'. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics. pp. 736-747.

[AONE, 1998] Aone, C., Okurowski, M. E., Gorlinsky, J. (1998). 'Trainable, scalable summarization using robust NLP and machine learning'. *Proceedings of the 17th international conference on Computational linguistics*, Montreal, Quebec, Canada, *Vol 1*, pp. 62-66.

[ASHE, 2003] Asher, N., Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

[AZMI, 2012] Azmi, A. M., Al-Thanyyan, S. (2012). 'A text summarizer for Arabic'. *Computer Speech & Language*, *26*(4), pp. 260-273.

[AZMI, 2017] Azmi, A. M., Alshenaifi, N. A. (2017). 'Lemaza: An Arabic why-question answering system'. *Natural Language Engineering*, *23*(6), pp. 877-903.

[BARZ, 1999] Barzilay, R., Elhadad, M. (1999). 'Using lexical chains for text summarization'. Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 10–17.

[BARZ, 2005] Barzilay, R., McKeown, K. R. (2005). 'Sentence fusion for multidocument news summarization'. *Computational Linguistics*, *31*(3), pp. 297-328.

[BAWA, 2008] Bawakid, A., Oussalah, M. (2008). 'A Semantic Summarization System: University of Birmingham at TAC 2008'. *Proceedings of the First Text Analysis Conference (TAC2008).*

[BAWA, 2011] Bawakid, A. (2011). 'Automatic documents summarization using ontology based methodologies'. Phd. Thesis Birmingham university.

[BAXE, 1958] Baxendale, P. B. (1958). 'Machine-made index for technical literature—an experiment'. *IBM Journal of Research and Development*, *2*(4), pp. 354-361.

[BELK, 2015] Belkebir R., Guessoum A. (2015). 'A Supervised Approach to Arabic Text Summarization Using AdaBoost'. In: Rocha A., Correia A., Costanzo S., Reis L. (eds) New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing, vol 353. Springer, Cham

[BELZ, 2008] Belz, A. (2008). 'Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models'. *Natural Language Engineering*, *14*(4), pp. 431-455.

[BENA, 2007] Benajiba Y., Rosso P., BenedíRuiz J.M. (2007) ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2007. Lecture Notes in Computer Science, vol 4394. Springer, Berlin, Heidelberg.

[BERG, 2000] Berger, A., Mittal, V. O. (2000). 'Query-relevant summarization using FAQs'. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Hong Kong. pp. 294-301.

[BIRA, 2013] Biran, O., McKeown, K. (2013). 'Aggregated word pair features for implicit discourse relation disambiguation'. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 69-73.

[BLAC, 2006] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., Fellbaum, C. (2006). 'Introducing the Arabic WordNet project'. *Proceedings of the third international WordNet conference*, pp. 295-300.

[BLAI, 2008] Blais, A. (2008).'Résumé automatique de textes scientifiques et construction de fiches de synthèse catégorisées : approche linguistique par annotations sémantiques et réalisation informatique'. *Thèse de doctorat, Université Paris IV-Sorbonne, France.*

[BOUD, 2008] Boudin, F. (2008). 'Exploration d'approches statistiques pour le résumé automatique de texte '. *Thèse de doctorat, Université d'Avignon.*

[BOUD, 2010] Boudabous M.M., Maaloul M.H., Belguith L.H. (2010). 'Digital Learning for Summarizing Arabic Documents'. In: Loftsson H., Rögnvaldsson E., Helgadóttir S. (eds) Advances in Natural Language Processing. NLP 2010. Lecture Notes in Computer Science, vol 6233. Springer, Berlin, Heidelberg.

[BOUD, 2017] Boudchiche, M., Mazroui, A., Bebah, M. O. A. O., Lakhouaja, A., Boudlal, A. (2017). 'AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer'. *Journal of King Saud University-Computer and Information Sciences, 29*(2), pp. 141-146.

[BRAN, 1995] Brandow, R., Mitze, K., Rau, L. F. (1995).'Automatic condensation of electronic publications by sentence selection'. *Information Processing & Management, 31*(5), pp. 675-685.

[CARE, 2007] Carenini, G., Ng, R. T., Zhou, X. (2007). 'Summarizing email conversations with clue words'. *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada. pp. 91-100.

[CARL, 2001] Carlson, L., Marcu, D. (2001). 'Discourse tagging reference manual'. *Technical Report. University of Southern California. Information Sciences Institute*.

[CARL, 2003] Carlson L., Marcu D., Okurowski M.E. (2003). 'Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory'. In: van Kuppevelt J., Smith R.W. (eds) Current and New Directions in Discourse and Dialogue. Text, Speech and Language Technology, vol 22. Springer, Dordrecht.

[COHE, 1960] Cohen, J. (1960). 'A coefficient of agreement for nominal scales'. *Educational and psychological measurement*, *20*(1), pp. 37-46.

[CORS, 2004] Corston-Oliver, S., Ringger, E., Gamon, M., Campbell, R. (2004). 'Task-focused summarization of email'. ACL-04 Workshop: Text Summarization Branches Out, pp. 43–50.

[DA CU, 2011] Da Cunha, I., Torres-Moreno, J. M., Sierra, G. (2011). 'On the development of the RST Spanish Treebank'. *Proceedings of the 5th Linguistic Annotation Workshop*, Portland, Oregon. pp. 1-10.

[DEJO, 1982] DeJong, G. (1982). 'An overview of the FRUMP system'. *Strategies for natural language processing*, *113*, pp. 149-176.

[DIAO, 2006] Diao, Q., Shan, J. (2006). 'A new web page summarization method'. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, Washington, USA. pp. 639-640.

[DORA, 2004] Doran, W., Stokes, N., Carthy, J., Dunnion, J. (2004). 'Comparing lexical chain-based summarisation approaches using an extrinsic evaluation'. Proceedings of the 5th International conference on Intelligent Text Processing and Computational Linguistics CICLing2004, MIT Press. pp. 112–117.

[DOUZ, 2004] Douzidia, F. S., Lapalme, G. (2004). 'Lakhas, an Arabic summarization system'. *Proceedings of DUC 2004*.

[EDMU, 1969] Edmundson, H. P. (1969). 'New methods in automatic extracting'. *Journal of the ACM (JACM)*, *16*(2), pp. 264-285.

[EL HA, 2010] El Haj, M., Kruschwitz, U., Fox, C. (2010).'Using Mechanical Turk to Create a Corpus of Arabic Summaries. *Editors & Workshop Chairs*, pp. 36-39.

[EL HA, 2011] El Haj, M., Kruschwitz, U., Fox, C. (2011). 'Exploring clustering for multi-document arabic summarisation'. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H.(eds.) AIRS 2011. LNCS, vol. 7097, Springer, Heidelberg. pp. 550–561.

[EL HA, 2011a] El Haj, M., Kruschwitz, U., Fox, C. (2011). 'Multi-document Arabic text summarisation'. *Computer Science and Electronic Engineering Conference (CEEC)*, IEEE. pp. 40-44.

[EL HA, 2012] El Haj, M. (2012). 'Multi-document Arabic text summarisation*'*. Phd. Thesis, *University of Essex, Britain*.

[EL HA, 2015]: El Haj, M., Kruschwitz, U., Fox, C. (2015). 'Creating language resources for under-resourced languages: methodologies, and experiments with Arabic'. *Language Resources and Evaluation*, *49*(3), pp. 549-580.

[ERKA, 2004] Erkan, G., Radev, D. R. (2004). 'Lexrank: Graph-based lexical centrality as salience in text summarization'. *Journal of artificial intelligence research*, *22*, pp. 457-479.

[EVAN, 2004] Evans, D. K., Klavans, J. L., & McKeown, K. R. (2004). 'Columbia newsblaster: Multilingual news summarization on the web'. *Demonstration Papers at HLT-NAACL 2004*, Boston, Massachusetts. pp. 1-4.

[FARZ, 2005] Farzindar, A. (2005). 'Résumé automatique de textes juridiques'. Thèse de doctorat, Université Paris-Sorbonne, France.

[FATT, 2008] Fattah, M. A., Ren, F. (2008). 'Probabilistic neural network based text summarization'. *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE'08. International Conference on* pp. 1-6. IEEE.

[FATT, 2009] Fattah, M. A., Ren, F. (2009). 'GA, MR, FFNN, PNN and GMM based models for automatic text summarization'. *Computer Speech & Language*, *23*(1), pp.126-144.

[FEJE, 2014] Fejer, H. N., Omar, N. (2014). 'Automatic Arabic text summarization using clustering and keyphrase extraction'. *International Conference on Information Technology and Multimedia (ICIMU)*, IEEE. *Putrajaya, Malaysia.* pp. 293-298.

[FELL, 1998] Fellbaum, C. (Ed.). (1998). '*WordNet: An Electronic Lexical Database'*. MIT Press.

[FENG, 2012] Feng, V. W., Hirst, G. (2012). 'Text-level discourse parsing with rich linguistic features'. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, Jeju Island, Korea. pp. 60-68.

[FENG, 2015] Feng, W. V. (2015). 'RST-style discourse parsing and its applications in discourse analysis'. Phd. Thesis *Université of Toronto, Canada.*

[FUM, 1982]: Fum, D., Guida, G., Tasso, C. (1982). 'Forward and backward reasoning in automatic abstracting'. *COLING '82 Proceedings of the 9th conference on Computational linguistics - Volume 1*, Prague, Czechoslovakia. pp. 83-88,

[GALA, 2008] Galanis, D., Malakasiotis, P. (2008). 'AUEB at TAC 2008'. *Proceeding of the Text Analysis Conference*. (TAC).

[GANE, 2010] Ganesan, K., Zhai, C., Han, J. (2010). 'Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions'. *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China. pp. 340-348.

[GIAN, 2011] Giannakopoulos G, Karkaletsis V (2011). 'Autosummeng and memog in evaluating guided summaries'. *Proceedings of the text analysis conference (TAC), MultiLing Summarisation Pilot*, Maryland, USA

[GIAN, 2013] Giannakopoulos, G. (2013). 'Multi-document multilingual summarization and evaluation tracks in ACL 2013 Multiling workshop'. *Proceeding of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, Sofia, Bulgaria. pp. 20-28.

[GOLD, 2006] Blair-Goldensohn, S., McKeown, K. (2006). 'Integrating rhetorical-semantic relation models for query-focused summarization'. *Proceeding of the Document Understanding Conference, DUC-2006, New York, USA*.

[HARI, 2009] Hariharan, S., Srinivasan, R. (2009). 'Studies on Graph Based Approaches for Singleand Multi Document Summarizations'. *International Journal of Computer Theory and Engineering, 1*(5), 519.

[HERN, 2010a] Hernault, H., Bollegala, D., Ishizuka, M. (2010). 'A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension'. *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, Massachusetts. pp. 399-409.

[HERN, 2010b] Hernault, H., Prendinger, H., Ishizuka, M. (2010). 'HILDA: A discourse parser using support vector machine classification'. *Dialogue & Discourse*, *1*(3), pp. 1-33.

[HIRA, 2015] Hirao, T., Nishino, M., Yoshida, Y., Suzuki, J., Yasuda, N., Nagata, M. (2015). 'Summarizing a document by trimming the discourse tree'. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *23*(11), pp. 2081-2092.

[HOVY, 1998] Hovy, E., & Lin, C. Y. (1998). 'Automated text summarization and the SUMMARIST system'. *TIPSTER '98 Proceedings of a workshop ,* Baltimore, Maryland. pp. 197-214.

[HUAN, 2011] Huang, H. H., & Chen, H. H. (2011). 'Chinese discourse relation recognition'. *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand. pp. 1442-1446.

[IBRA, 2013] Ibrahim, A., Elghazaly, T., Gheith, M. (2013). 'A novel Arabic text summarization model based on rhetorical structure theory and vector space model'. *International Journal of Computational Linguistics and Natural Language Processing*, *2*(8), pp. 480-485.

[IMAM, 2012] Imam, I., Nounou, N., Hamouda, A., Khalek, H. A. A. (2013). 'An ontology-based summarization system for Arabic documents (ossad)'. *International Journal of Computer Applications*, *74*(17), pp. 38-43.

[JI, 2015] Ji, Y., Eisenstein, J., (2015). 'One vector is not enough: Entity-augmented distributed semantics for discourse relations'. *Transactions of the Association for Computational Linguistics*, 3, pp. 329–344.

[JIAN, 1997] Jiang, J., Conrath, D. W. (1997). 'Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy'. *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan.

[JONE, 1995] Jones, K. S., Galliers, J. R. (1995). '*Evaluating natural language processing systems: An analysis and review'*. LNAI, Springer Science & Business Media.

[JONE, 1999] Jones, K. S. (1999). 'Automatic summarizing: factors and directions'. *Advances in automatic text summarization*, MIT Press. pp.1-12.

[JOTY, 2015] Joty, S., Carenini, G., & Ng, R. T. (2015). 'CODRA: A novel discriminative framework for rhetorical analysis'. *Computational Linguistics*, *41*(3),pp. 385-435.

[KAIK, 2004] Kaikhah, K. (2004). 'Automatic text summarization with neural networks'. *Proceedings of the 2nd International IEEE Conference* on *Intelligent Systems*, IEEE. varna, Bulgaria, pp. 40-44.

[KESK, 2014] Keskes, I., Zitoune, F. B., Belguith, L. H. (2014). 'Learning explicit and implicit arabic discourse relations'. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), pp. 398-416.

[KESK, 2015] Keskes, I. (2015). '*Discourse analysis of Arabic documents and application to automatic summarization'.* Thèse de doctorat, Université de Toulouse III-Paul Sabatier, France.

[KHAL, 2011] Khalifa, I., Feki, Z., Farawila, A. (2011). 'Arabic discourse segmentation based on rhetorical methods'. *11*(1), pp.10-15.

[KHAN, 2014] Khan, A., Salim, N. (2014). 'A review on abstractive summarization methods'. *Journal of Theoretical and Applied Information Technology*, *59*(1), pp. 64-72.

[KIM, 2014] Kim, Y. 'Convolutional Neural Networks for Sentence Classification'. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. pp. 1746–1751.

[KNIG, 2000] Knight, K., Marcu, D. (2000). 'Statistics-based summarization-step one: Sentence compression'. *Proceedings of* American Association for Artificial Intelligence, vol 2000, pp.703-710.

[KNIG, 2002] Knight, K., Marcu, D. (2002). 'Summarization beyond sentence extraction: A probabilistic approach to sentence compression'. *Artificial Intelligence*, *139*(1), pp. 91-107.

[KOLL, 2007] Kolla, M., Vechtomova, O., Clarke, C. L. (2007). 'Comparison of models based on summaries or documents towards extraction of update summaries'. *Proceedings of DUC 2007*.

[KUMA, 2016] Kumar, Y. J., Goh, O. S., Halizah, B., Ngo, H. C., Puspalata, C. (2016). 'A review on automatic text summarization approaches'. Journal of Computer Science, *12*(4), pp.178-190.

[KUPI, 1995] Kupiec, J., Pedersen, J., Chen, F. (1995). 'A trainable document summarizer'. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, Washington, USA, pp. 68-73

[LAGR, 2018] Lagrini, S., Redjimi M., Azizi N. (2018). 'A Survey of Extractive Arabic Text Summarization Approaches'. In: Lachkar A., Bouzoubaa K., Mazroui A., Hamdani A., Lekhouaja A. (eds) Arabic Language Processing: From Theory to Practice. Communications in Computer and Information Science, vol 782, Springer, Cham.

[LAGR, 2018a] Lagrini, S., Azizi, N., Redjimi, M., Al Dwairi, M. (2018). 'Automatic Identification of Rhetorical Relations Among Intra-sentence Discourse Segments in Arabic'. *International Journal of Intelligent Systems Technologies and Applications- Indesciences publisher. (Article in press).*

[LI, 2013] Li, L., Forascu, C., El-Haj, M., Giannakopoulos, G. (2013). 'Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian'. *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, Sofia, Bulgaria, , pp. 1-12

[LI, 2017] Li, H., Zhang, J., Zong, C. (2017). 'Implicit discourse relations recognition for english and Chinese with multiview modeling and effective representation learning'. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *16*(3).

[LIN, 1998] Lin, D. (1998). 'An information-theoretic definition of similarity'. *Proceedings of the 15th International Conference on Machine Learning (Icml)*, Vol. 98, pp. 296-304.

[LIN, 1999] Lin, C. Y. (1999). 'Training a selection function for extraction'. *Proceedings of the eighth international conference on Information and knowledge management*, Kansas City, Missouri, USA, pp. 55-62.

[LIN, 2004] LIN, C. Y. (2004). 'ROUGE: A Package for Automatic Evaluation of Summaries'. *Proceedings of Workshop on Text Summarization Branches Out*, Barcelona, Spain.

[LIN, 2009] Lin, Z., Kan, M. Y., Ng, H. T. (2009). 'Recognizing implicit discourse relations in the Penn Discourse Treebank'. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, Singapore. pp. 343-351

[LIOR, 2008] Lioret, E., Ferrández, O., Munoz, R., Palomar, M. (2008). 'A Text Summarization Approach under the Influence of Textual Entailment'. *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)*, pp. 22-31.

[LIOR, 2012] Lloret, E., & Palomar, M. (2012). 'Text summarisation in progress: a literature review'. *Artificial Intelligence Review*, *37*(1), pp. 1-41.

[LIU, 2016] Liu, Y., Li, S., Zhang, X., Sui, Z. (2016). 'Implicit Discourse Relation Classification via Multi-Task Neural Networks'. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pp. 2750-2756.

[LOUI, 2010] Louis, A., Joshi, A., Prasad, R., Nenkova, A. (2010). 'Using entity features to classify implicit discourse relations'. *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL '10)*, Tokyo, Japan, pp. 59-62.

[LUHN, 1958] Luhn, H. P. (1958). 'The automatic creation of literature abstracts'. *IBM Journal of research and development*, *2*(2), pp. 159-165.

[MAAL, 2012] Maaloul, M. H. (2012). *Approche hybride pour le résumé automatique de textes. Application à la langue arabe*'. Thèse de doctorat, Université de Provence-Aix-Marseille I, France.

[MANI, 1999a] Mani, I., Maybury, M. T. (1999). *Advances in automatic text summarization*. MIT press.

[MANI, 1999b] Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B. (1999). The TIPSTER SUMMAC text summarization evaluation. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, Bergen, Norway, pp. 77-85.

[MANI, 2001] Mani, I. (2001). '*Automatic summarization*'. John Benjamins Publishing.

[MANN, 1988] Mann, W. C., Thompson, S. A. (1988). 'Rhetorical structure theory: Toward a functional theory of text organization'. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*(3), pp. 243-281.

[MARC, 1997] Marcu, D. (1997). 'The rhetorical parsing of natural language texts'. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, Madrid, Spain. pp. 96-103.

[MARC, 2000] Marcu, D. (2000). 'The rhetorical parsing of unrestricted texts: A surface-based approach'. *Computational linguistics*, 26(3), pp. 395-448.

[MARC, 2000a] Marcu, D. (2000). '*The theory and practice of discourse parsing and summarization'*. MIT press.

[MARC, 2000b] Marcu, D., Carlson, L., Watanabe, M. (2000). 'The automatic translation of discourse structures'. Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (NAACL 2000), Seattle, Washington, pp. 9-17.

[MARC, 2002] Marcu, D., Echihabi, A. (2002). 'An unsupervised approach to recognizing discourse relations'. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (ACL '02), Philadelphia, Pennsylvania, pp. 368-375.

[MCKE, 1999] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., Eskin, E. (1999). 'Towards multidocument summarization by reformulation: Progress and prospects'. *Proceedings of American Association for Artificial Intelligence AAAI- 999.*

[MCKE, 2003] McKeown, K., Barzilay, R., Chen, J., Elson, D., Evans, D., Klavans, J., Sigelman, S. (2003). 'Columbia's newsblaster: new features and future directions'. *Proceedings of HLT-NAACL 2003 Demonstrations*, pp. 15-16. Edmonton, Canada.

[MENG, 2012] Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., Wang, H. (2012). 'Entity-centric topic-oriented opinion summarization in twitter'. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*, Beijing, China. pp. 379-387.

[MIHA, 2004] Mihalcea, R., Tarau, P. (2004). 'Graph-based ranking algorithms for sentence extraction, applied to text summarization'. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (ACLdemo '04), 20*. Barcelona, Spain.

[MIHA, 2007] Mihalcea, R., Ceylan, H. (2007). 'Explorations in automatic book summarization'. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, pp. 380–389.

[MIHA, 2016] Mihaylov, T., Frank, A. (2016). 'Discourse relation sense classification using cross-argument semantic similarity based on word embeddings'. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 100–107.

[MINE, 2002] Minel, J. L. (2002). '*filtrage semantique de textes problemes, conception et realisation d'une plate-forme informatique'.* Thèse de doctorat, Université Paris-Sorbonne-Paris IV, France.

[NALL, 2017] Nallapati, R., Zhai, F., Zhou, B. (2016). 'SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents'. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pp.3075-3081.

[NENK,2006] Nenkova, A. (2006). 'Summarization evaluation for text and speech: issues and approaches'. *Proceedings of the Ninth International Conference on Spoken Language Processing (ICSLP*), Pittsburgh, Pennsylvania, , pp. 1527-1530.

[NENK,2007] Nenkova, A., Passonneau, R., McKeown, K. (2007). 'The pyramid method: Incorporating human content selection variation in summarization evaluation'. *ACM Transactions on Speech and Language Processing (TSLP), 4*(2).

[NIE, 2006] Nie, Y., Ji, D., Yang, L., Niu, Z., He, T. (2006). 'Multi-document summarization using a clustering-based hybrid strategy'. In: Ng H.T., Leong MK., Kan MY., Ji D. (eds) Information Retrieval Technology. AIRS 2006. Lecture Notes in Computer Science, vol 4182. Springer, Berlin, Heidelberg.

[NISH, 2010] Nishikawa, H., Hasegawa, T., Matsuo, Y., Kikui, G. (2010). 'Opinion summarization with integer linear programming formulation for sentence extraction and ordering'. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, Beijing, China, pp. 910-918.

[NOBA, 2002] Nobata, C., Sekine, S., Isahara, H., Grishman, R. (2002). 'Summarization System Integrated with Named Entity Tagging and IE pattern Discovery'. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Spain

[NOOR, 2014] Noori Fejer, H., Omar, N.(2014). ' Automatic Arabic text summarization using clustering and keyphrase extraction'. *Procedings of International Conference on Information Technology and Multimedia (ICIMU),* pp. 293–298.

[O'DON, 1997] O'Donnell, M. (1997). 'Variable-length on-line document generation'. *Proceedings of the 6th European Workshop on Natural Language Generation* pp. 82-91.

[ONO, 1994] Ono, K., Sumita, K., & Miike, S. (1994). 'Abstract generation based on rhetorical structure extraction'. *Proceedings of the 15th conference on Computational linguistics (*COLING '94 *), Vol 1*, Kyoto, Japan, , pp. 344-348.

[OUFA, 2014] Oufaida, H., Nouali, O., Blache, P. (2014). 'Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization'. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), pp. 450-461.

[PARD, 2003] Pardo T.A.S., Rino L.H.M., Nunes M..G.V. (2003). 'GistSumm: A Summarization Tool Based on a New Extractive Method'. In: Mamede N.J., Trancoso I., Baptista J., das Graças Volpe Nunes M. (eds) Computational Processing of the Portuguese Language. PROPOR 2003. Lecture Notes in Computer Science, vol 2721. Springer, Berlin, Heidelberg.

[PARD, 2004] Pardo T.A.S., das Graças Volpe Nunes M., Rino L.H.M. (2004). 'DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese'. In: Bazzan A.L.C., Labidi S. (eds) Advances in Artificial Intelligence – SBIA 2004. SBIA 2004. Lecture Notes in Computer Science, vol 3171. Springer, Berlin, Heidelberg.

[PERE, 2013] Perea-Ortega, J. M., Lloret, E., Ureña-López, L. A., Palomar, M. (2013). 'Application of text summarization techniques to the geographical information retrieval task'. *Expert systems with applications*, *40*(8), pp. 2966-2974.

[PILT, 2009] Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., Joshi, A. K., 2008, Easily identifiable discourse relations. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, pp.87-90.

[POTT, 2010] Potthast M., Becker S. (2010). 'Opinion Summarization of Web Comments'. In: Gurrin C. et al. (eds) Advances in Information Retrieval. ECIR 2010. Lecture Notes in Computer Science, vol 5993. Springer, Berlin, Heidelberg.

[POUR, 2012] Pourvali, M., Abadeh, M. S. (2012). 'Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet and Wikipedia Knowledge Base'. *International Journal of Computer Science Issues*, 9(1), 3.

[PRAS, 2008] Prasad A., Miltsakaki R., Dinesh E., Lee N., Joshi A., and Webber B. (2008). 'The Penn discourse treebank'. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

[RADE, 1998] Radev, D. R., McKeown, K. R. (1998). 'Generating natural language summaries from multiple on-line sources'. *Computational Linguistics*, *24*(3), pp. 470-500.

[RADE, 2000] Radev, D. R., Jing, H., Budzikowska, M. (2000). 'Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies'. *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization (NAACL-ANLP-AutoSum '00), Vol 4,* Seattle, Washington, pp. 21-30.

[RADE, 2002] Radev, D. R., Hovy, E., McKeown, K. (2002). 'Introduction to the special issue on summarization'. *Computational linguistics*, *28*(4), pp. 399-408.

[RADE, 2004] Radev, D. R., Jing, H., Styś, M., Tam, D. (2004). 'Centroid-based summarization of multiple documents'. *Information Processing & Management*, *40*(6), pp. 919-938.

[RAMB, 2004] Rambow, O., Shrestha, L., Chen, J., Lauridsen, C. (2004). 'Summarizing email threads'. *Proceedings of HLT-NAACL 2004: Short Papers*, Boston, Massachusetts, pp. 105-108.

[REEV, 2007] Reeve, L. H., Han, H., & Brooks, A. D. (2007). 'The use of domain-specific concepts in biomedical text summarization'. *Information Processing & Management*, *43*(6), pp.1765-1776.

[SADE, 2016] Sadek, J., Meziane, F. (2016). 'Extracting Arabic causal relations using linguistic patterns'. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *15*(3), 14.

[SAGG, 2002] Saggion, H., Lapalme, G. (2002). 'Generating indicative-informative summaries with sumum'. *Computational linguistics*, *28*(4), pp.497-526.

[SAGG, 2002b] Saggion, H., Radev, D., Teufel, S., Lam, W., Strassel, S. M. (2002). 'Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment'. *Ann Arbor*, *1001*(48), pp.109-1092.

[SAIT, 2006] Saito, M., Yamamoto, K., Sekine, S. (2006). 'Using phrasal patterns to identify discourse relations'. *Proceedings of the Human Language Technology*

Conference of the NAACL, Companion Volume: Short Papers, New York, pp. 133-136.

[SALT, 1997] Salton, G., Singhal, A., Mitra, M., Buckley, C. (1997). 'Automatic text structuring and summarization'. *Information Processing & Management*, *33*(2), pp.193-207.

[SCHI, 2008] Schilder, F., Kondadadi, R. (2008). 'FastSum: fast and accurate query-based multi-document summarization'. *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*, Columbus, Ohio, pp. 205-208.

[SOBH, 2006] Sobh, I., Darwish, N., Fayek, M. (2006). 'An optimized dual classification system for Arabic extractive generic text summarization'. In *In the 7th Conference on Language Engineering,pp. pp. 149–154.*

[SOCH, 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). 'Recursive deep models for semantic compositionality over a sentiment treebank'. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA *pp. 1631–1642.*

[SOMA, 2009] Somasundaran, S., Namata, G., Wiebe, J., Getoor, L. (2009). 'Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification'. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Vol 1*, Singapore, pp. 170-179.

[SORI, 2003] Soricut, R., Marcu, D. (2003). 'Sentence level discourse parsing using syntactic and lexical information'. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol 1* , Edmonton, Canada, pp. 149-156.

[STED, 2014] Stede, M., Neumann, A. (2014). 'Potsdam Commentary Corpus 2.0: Annotation for Discourse Research'. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014),* Iceland.

[SUBB, 2009] Subba, R., Di Eugenio, B. (2009). 'An effective discourse parser that uses rich linguistic information'. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, pp. 566-574.

[SVOR,2007] Svore, K., Vanderwende, L., Burges, C. (2007). 'Enhancing single-document summarization by combining RankNet and third-party sources'. *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, Prague, , pp. 448–457.

[TABO, 2006] Taboada, M., & Mann, W. C. (2006). 'Rhetorical structure theory: Looking back and moving ahead'. *Discourse studies*, *8*(3), pp. 423-459.

[TATA, 2009] Tatar, D., Mihis, A., Lupsa, D., Tamaianu-Morita, E. (2009). 'Entailment-based linear segmentation in summarization'. *International Journal of Software Engineering and Knowledge Engineering*, *19*(08), pp.1023-1038.

[TORR, 2011] Juan-Manuel, T. M. (2011). *Résumé automatique de documents: une approche statistique*. Lavoisier.

[TURN, 2005] Turner, J., Charniak, E. (2005). Supervised and unsupervised learning for sentence compression. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Michigan, pp. 290-297.

[ULRI, 2009] Ulrich, J., Carenini, G., Murray, G., Ng, R. T. (2009). 'Regression-Based Summarization of Email Conversations'. Proceedings of the Third International ICWSM Conference, pp. 334-337.

[UZÊD, 2010] Uzêda, V. R., Pardo, T. A. S., Nunes, M. D. G. V. (2010). 'A comprehensive comparative evaluation of RST-based summarization methods'. *ACM Transactions on Speech and Language Processing (TSLP)*, *6*(4), 4.

[VAN, 2011] Van Der Vliet, N. , Berzlánovich, I., Bouma, G., Egg, M., Redeker, G. (2011). 'Building a discourse-annotated Dutch text corpus'. *Proceeding of the workshop"beyond semantics: corpus based investigation of pragmaticand discourse phenomena*", Gottingen, Germany, pp. 157-171.

[WAN, 2006] Wan, X., Yang, J. (2006). 'Improved affinity graph based multi-document summarization'. *Proceedings of the human language technology conference of the NAACL, Companion volume: Short papers*, New York, pp. 181-184.

[WAN, 2008a] Wan, X., Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, Singapore, pp. 299-306

[WANG, 2008b] Wang, D., Li, T., Zhu, S., Ding, C. (2008). 'Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization'. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, Singapore, pp. 307-314.

[WEI, 2010] Wei, F., Li, W., Lu, Q., He, Y. (2010). 'A document-sensitive graph model for multi-document summarization'. *Knowledge and information systems*, *22*(2), pp. 245-259.

[WOLF, 2005] Wolf, F., Gibson, E. (2005). 'Representing discourse coherence: A corpus-based study'. *Computational Linguistics*, *31*(2), pp. 249-287.

[YONG, 2006] Yong, S. P., Abidin, A. I., Chen, Y. Y. (2006). 'A neural-based text summarization system'. *WIT Transactions on Information and Communication Technologies*, *37*.

[YU, 2007] Yu, J., Reiter, E., Hunter, J., Mellish, C. (2007). 'Choosing the content of textual summaries of large time-series data sets'. *Natural Language Engineering*, *13*(1), pp. 25-49.

[ZAJI, 2007] Zajic, D., Dorr, B. J., Lin, J., Schwartz, R. (2007). 'Multi-candidate reduction: Sentence compression as a tool for document summarization tasks'. *Information Processing & Management*, *43*(6), pp. 1549-1570.

[ZHAN, 2015] Zhang, B., Su, J., Xiong, D., Lu, Y., Duan, H., Yao, J. (2015). 'Shallow convolutional neural network for implicit discourse relation recognition'. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 2230–2235.

Productions Scientifiques

Publications Internationales

1- Samira Lagrini, Mohammed Redjimi and Nabiha Azizi. 'Automatic Arabic Text Summarization Approaches'. *International Journal of Computer Applications* 164(5):31-37, April 2017.

doi> 10.5120/ijca2017913628

2- Samira Lagrini, Nabiha Azizi, Mohammed Redjimi and Monther Al Dwairi . 'Automatic Identification of Rhetorical Relations Among Intra-sentence Discourse Segments in Arabic'. *International Journal of Intelligent Systems Technologies and Applications- Indesciences publisher* – (in press).

3- Samira Lagrini, Nabiha Azizi, Mohammed Redjimi and Monther Al Dwairi . 'Toward an automatic summarization of Arabic text depending on rhetorical relations'. International Journal of Reasoning-based Intelligent Systems - *Indesciences publisher* (in press)

Communications Internationales

1- Lagrini S., Redjimi M., Azizi N. Extractive Arabic Text Summarization Approaches. *The 6th International Conference on Arabic Language Processing 2017* October 11th - 12th 2017, Fez, Morocco.

Chapitres de Livres

1- Lagrini S., Redjimi M., Azizi N. (2018) A Survey of Extractive Arabic Text Summarization Approaches. In: Lachkar A., Bouzoubaa K., Mazroui A., Hamdani A., Lekhouaja A. (eds) Arabic Language Processing: From Theory to Practice. Communications in Computer and Information Science, vol 782. Springer, Cham. https://doi.org/10.1007/978-3-319-73500-9_12