Étude comparative entre trois approches de résumé automatique de documents arabes

Iskandar Keskes[1,2] *Mohamed Mahdi Boudabous*[1] *Mohamed Hédi Maaloul*[1,3]
Lamia Hadrich Belguith[1]

(1) ANLP Research Group, Laboratoire MIRACL, Route de Tunis Km 10, BP 242, Sfax, Tunisie
(2) Laboratoire IRIT, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
(3) Laboratoire LPL, 5 avenue Pasteur, BP 80975, 13604 Aix-en-Provence, France
Keskes@irit.fr, mehdiboudabous@gmail.com
mohamed.maaloul@lpl-aix.fr, l.belguith@fsegs.rnu.tn

RÉSUME

Dans cet article, nous proposons une étude comparative entre trois approches pour le résumé automatique de documents arabes. Ainsi, nous avons proposé trois méthodes pour l'extraction des phrases les plus représentatives d'un document. La première méthode se base sur une approche symbolique, la deuxième repose sur une approche numérique et la troisième se base sur une approche hybride. Ces méthodes sont implémentées respectivement par le système ARSTResume, le système R.I.A et le système HybridResume. Nous présentons, par la suite, les résultats obtenus par les trois systèmes et nous procédons à une étude comparative entre les résultats obtenus afin de souligner les avantages et les limites de chaque méthode. Les résultats de l'évaluation ont montré que l'approche numérique est plus performante que l'approche symbolique au niveau des textes longs. Mais, l'intégration de ces deux approches en une approche hybride aboutit aux résultats les plus performants dans notre corpus de textes.

ABSTRACT

Comparative study of three approaches to automatic summarization of Arabic documents

In this paper, we propose a comparative study between three approaches for automatic summarization of Arabic documents. Thus, we proposed three methods for extracting most representative sentences of a document. The first method is based on a symbolic approach, the second is relied on a numerical approach and the third is based on a hybrid approach. These methods are implemented respectively by the ARSTResume, R.I.A and HybridResume systems. Then, we present the results obtained by the three systems and we conduct a comparative study between the obtained results in order to highlight the advantages and limitations of each method. The evaluation results showed that the numerical approach has better performances than the symbolic approach. But, combining into a hybrid approach achieved the best results for our text corpus.

MOTS-CLES : Résumé automatique, approche symbolique, approche numérique, approche hybride, document arabe.

KEYWORDS: Automatic summarization, symbolic approach, numerical approach, hybrid approach, Arabic document.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 225–238,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Le Traitement Automatique du Langage Naturel (TALN) nous montre que les approches peuvent être convergées pour résoudre le même problème. Chaque approche a ses propres avantages et inconvénients qui peuvent être identifiés par une étude comparative.

Le présent travail présente une étude comparative entre différentes approches de TALN, dans le cadre du résumé automatique de textes.

Ce domaine aide à contribuer à une meilleure compréhension de la façon dont les gens produisent et comprennent la langue, car il peut résoudre les besoins croissants d'information de synthèse dans notre société.

La tâche de résumé semble être intrinsèquement interprétée dans le sens où différentes personnes produisent généralement des résumés très différents pour un texte donné. Ainsi, la qualité des résumés peut être jugée très différemment (Iria et al., 2007).

En matière de résumé automatique, on peut distinguer trois principales approches à savoir, l'approche par compréhension appelée l'approche symbolique, l'approche par extraction appelée l'approche numérique et l'approche qui combine les deux approches précédentes appelée l'approche hybride. L'approche symbolique exploite un savoir purement linguistique, et plus précisément sémantique pour extraire les phrases pertinentes d'un document (Azmi et Al-Thanyyan, 2012). Plusieurs théories entrent dans le cadre de cette approche à savoir : la Théorie de la Structure Rhétorique (RST) (Mann et Thompson, 1988), la Théorie de la Représentation Discursive (DRT) (Kamp, 1981 ; Kamp et Reyle, 1993), la Théorie de la Représentation Discursive Segmentée (SDRT) (Asher, 1993 ; Lascarides et Asher, 1993)… tandis que l'approche numérique repose sur un calcul de poids ou de scores associés à chaque phrase afin d'estimer son degré d'importance dans le texte. On distingue deux grandes techniques à savoir : la technique statistique (mots des titres, position des phrases,…) et la technique d'apprentissage (apprentissage supervisé, apprentissage semi-supervisé et apprentissage non supervisé) (Amini, 2001). L'extrait final contient les unités textuelles qui ont les scores les plus élevés. Concernant l'approche hybride, elle utilise des méthodes linguistiques et numériques pour extraire les phrases du résumé.

Nous proposons dans cet article une étude comparative entre les trois approches (symbolique, numérique et hybride). Cette étude a pour objectif d'évaluer la robustesse de chacune de ces approches ainsi que la mise en relief de leurs avantages et de leurs inconvénients pour le résumé automatique.

La suite de cet article se structure autour de cinq piliers. Le premier pilier présente la méthode symbolique pour le résumé automatique de documents arabes implémentée dans le système ARSTResume. Le deuxième pilier présente la méthode numérique pour le résumé automatique de documents arabes implémentée dans le système R.I.A. Le troisième pilier présente la méthode hybride pour le résumé automatique de documents arabes implémentée dans le système HybridResume. Le quatrième pilier expose le corpus

d'évaluation, l'évaluation de ces trois systèmes et les résultats obtenus. Enfin, le cinquième, montre une étude comparative entre les trois approches.

2 Méthode symbolique proposée

Dans cette section, nous présentons la méthode symbolique que nous proposons pour le résumé automatique de documents arabes, ainsi qu'une description détaillée des différentes étapes de cette méthode (Keskes, 2011).

2.1 Présentation

La méthode symbolique proposée pour le résumé automatique des documents arabes se base principalement sur des techniques d'extraction moyennant des critères linguistiques. Elle repose sur la théorie de la structure rhétorique (RST) (Mann et Thompson, 1988). Il s'agit de détecter les relations sémantiques et les relations intentionnelles qui existent entre les segments d'un document. En effet, l'analyse rhétorique a pour but d'établir les relations et les dépendances ainsi que l'importance relative à des phrases ou propositions les unes par rapport aux autres (Keskes et Maâloul, 2010). Notre méthode se déroule en trois temps. D'abord, le repérage des relations rhétoriques entre les différentes unités minimales du texte dont l'une possède le statut de noyau – qui est le segment de texte primordial pour la cohérence – les autres ayant un statut de noyau ou de satellite, sont des segments optionnels. Ensuite, le dressage et la simplification de l'arbre RST. Enfin, la sélection des phrases noyaux formant le résumé final, selon type de relation rhétorique choisi pour l'extrait.

À l'issue de notre étude du corpus, formé de cent textes en langue arabe annotés par trois linguistes (ces derniers ont sélectionné les phrases pertinentes), nous avons pu repérer des *frames* de relations rhétoriques. Ces *frames* sont des règles rhétoriques formées par des signaux linguistiques. Ces signaux sont principalement des marqueurs linguistiques indépendants d'un domaine particulier pour le repérage des relations rhétoriques (Minel, 2002). Toutefois, ces marqueurs peuvent être répertoriés en deux types : indicateurs déclencheurs et indices complémentaires. Les indicateurs déclencheurs énoncent la présence d'une relation rhétorique. Les indices complémentaires sont recherchés dans un espace défini à partir de l'indicateur (dans le voisinage de l'indicateur). Ils peuvent ainsi agir, dans le contexte, afin de confirmer ou d'infirmer la relation rhétorique énoncée par l'indicateur déclencheur. Ces règles rhétoriques sont appliquées pour construire par la suite l'arbre rhétorique. À partir de notre corpus d'étude, nous avons énuméré vingt relations rhétoriques. La table 1 présente quelques relations :

Liste des relations rhétoriques	Condition / شرط
	Concession / استدراك
	Enumération / تفصيل
	Restriction / استثناء
	Confirmation / توكيد
	Réduction / تقليل
	Joint / ربط

	Evidence / قاعدة
	Négation / نفي

TABLE1 -Exemples de relations rhétoriques

Le *frame* suivant est utilisé pour détecter la relation rhétorique négation:

Nom de relation :	{négation /نفي }
Contrainte sur (1) :	Contient un/des indice(s) complémentaire(s){لكن , لكنني ,لكنهم ,لكنه }
Contrainte sur (2) :	Contient l'indice déclencheur {ليست , ليسوا , ليس , لن , ولم ,لم }
Position de l'indice déclencheur	Milieu
Unité retenue	(1)

TABLE 2 -*Frame* de la relation rhétorique négation

2.2. Description détaillée de la méthode

La mise en œuvre fonctionnelle de notre méthode est représentée par la figure 1. Elle repose sur une segmentation à différents niveaux (titres, sections, paragraphes, phrases) ainsi que sur une recherche basée sur les règles rhétoriques afin de détecter les relations rhétoriques. Ces règles rhétoriques sont utiles pour la construction de l'arbre rhétorique. Enfin, à travers le choix du type de résumé (i.e. résumé indicatif, résumé informatif, …), on précède à la simplification de l'arbre et à la sélection des phrases du résumé.

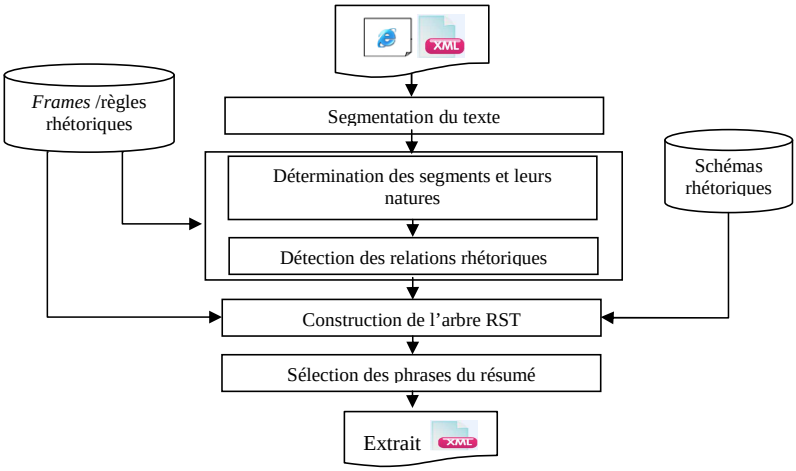

FIGURE1-Principales étapes de la méthode symbolique

2.1.1 Segmentation du document source

La segmentation du document est une étape nécessaire pour la tâche du résumé automatique. Cette étape consiste à hiérarchiser et à structurer le texte source en différentes unités (titres, sections, paragraphes et phrases).

Signalons, à ce niveau de traitement, une grande difficulté. En effet, La segmentation des textes en langue arabe ne peut pas se reposer uniquement sur la ponctuation puisqu'elle utilise certaines particules telles que " و " (waw) et "ف" (fā) et certains mots connecteurs pour séparer entre les phrases (Belguith et al., 2005).

Pour notre corpus constitué de textes en format HTML, nous utilisons une segmentation basée sur les signes de ponctuation et sur un ensemble de balises HTML. Cette étape de segmentation fournit en sortie un texte en format XML enrichi avec des balises encadrant les titres : <عنوان>…</عنوان>, les sections : <جزء>…</جزء>, les paragraphes : <فقرة>…</فقرة > et les phrases : <جملة>…</جملة>.

La deuxième étape de la segmentation est la segmentation des phrases en unités minimales, en utilisant les indicateurs principaux des règles rhétoriques, afin de descendre à un niveau plus bas dans l'analyse et de mieux dégager les relations. Ces dernières sont encadrées par les balises <قطاع>…</قطاع> (Tofiloski et al., 2009).

2.1.2 Application des règles rhétoriques

L'application des règles rhétoriques à un double but : déterminer la nature des segments (noyau ou satellite) et détecter les relations rhétoriques entre ces segments.

2.1.2.1 Détermination du segment Noyau et Satellite

Cette étape consiste à repérer les indicateurs principaux dans les phrases déjà segmentées et à préciser leurs positions dans l'unité minimale afin d'appliquer les règles rhétoriques, en cherchant les indices complémentaires.

Dans cette étape, nous allons donner, pour chaque unité minimale, un statut qui indique l'importance de cette unité par rapport à la phrase ou pour lui donner plus d'importance par rapport à une autre unité minimale. Le statut peut être un noyau ou un satellite.

Le noyau est un segment de texte qui comporte une information très pertinente. C'est un élément essentiel pour comprendre l'intention de l'auteur. Lorsqu'on élimine le noyau, nous ne pouvons pas comprendre le sens de la phrase. De même, un satellite est un segment de texte, mais qui comporte une information moins pertinente que le noyau. Donc, le noyau est un segment de texte primordial pour la cohérence et le satellite est un segment optionnel.

2.1.2.2 Détection des relations rhétoriques

Cette étape consiste à chercher les indices complémentaires de validation au voisinage de l'indicateur principal, c'est-à-dire le segment qui contient l'indicateur principal et le segment qui le précède. C'est l'indicateur principal qui signale la relation rhétorique

entre ces deux segments et c'est le rôle des indices complémentaires de confirmer ou non cette relation et de valider aussi le statut des deux segments.

Cette technique nous permet une analyse plus profonde, en tenant compte de la spécificité de la langue arabe sachant qu'on a des relations qui peuvent donner des sens proches comme les relations "حصر" et "استثناء" et aussi "تفسير" et "تفصيل".

2.1.3 Construction de l'arbre RST

Une fois l'étape de détection du type des unités minimales et des différentes relations rhétoriques existantes est achevée, nous ajoutons à notre technique les schémas rhétoriques (Mann et Thompson, 1988) afin de spécifier la composition structurale du texte et construire l'arbre RST.

Ces schémas rhétoriques décrivent l'organisation structurelle d'un texte, quelque soit le niveau hiérarchique de ce dernier. Ils permettent de lier un noyau et un satellite, deux ou plusieurs noyaux entre eux, et un noyau avec plusieurs satellites (Marcu, 1999).

Ainsi, les schémas rhétoriques se présentent sous la forme de cinq modèles de schémas (figure 2) qui peuvent être utilisés récursivement pour décrire des textes de taille arbitraire.

Généralement, le schéma le plus utilisé est celui liant un satellite unique à un noyau unique représenté dans la figure 3.

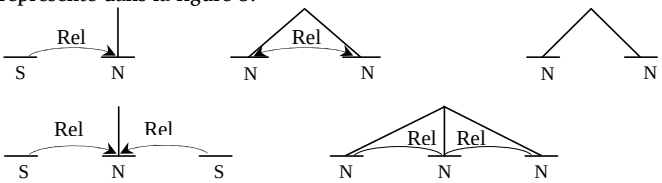

FIGURE 2 -Schéma rhétorique de base de la technique RST (Mann et al., 1988)

En plus des schémas rhétoriques, nous avons utilisé d'autres règles que nous avons dégagées suite à une étude empirique. Ces règles ont été validées par des linguistes. Elles permettent de traiter aussi les cas où nous n'avons pas de relations entre les phrases et assurent ainsi le maximum de couverture de texte que possible.

Afin de déterminer l'arbre RST le plus approprié pour le texte, nous avons essayé d'étudier le texte et la manière dont l'auteur l'a écrit. En effet, les auteurs veulent principalement donner un message aux lecteurs. Ce message est mentionné comme plusieurs faits; cependant, l'étude que nous avons faite sur le corpus prouve que les auteurs tendent à mentionner ces faits dans l'ordre, et chaque fait est suivi par des rapports qui le soutiennent. À travers cette étude empirique, nous avons pu dégager des règles de construction d'arbre RST représentées dans la table 3 :

Règles	Si (Indice principal est au début de la phrase) alors la relation détectée relie cette phrase avec la phrase précédente.
	Si (Indice principal est à la fin de la phrase) alors le segment qui contient cet indice est le seul qui contribue à la définition de la relation.
	Si (on a une ou plusieurs phrases qui n'admettent pas de relation entre elles) et (l'indice principal qui les suit est au début de la phrase) alors La relation relie toutes les phrases qui précèdent cet indice avec la phrase où il se trouve.

TABLE3 –Exemple de règles de construction d'arbres

Prenons par exemple la première règle, elle exprime le fait que s'il existe un marqueur principal, qui déclenche une relation rhétorique, situé au début de la phrase, alors cette relation relie entre le segment qui contient le marqueur principal et la phrase qui la précède. Car, sémantiquement, cette relation doit être subordonnante ou coordonnante de la relation rhétorique qu'elle précède et non pas le segment qu'il précède (Keskes et al., 2010b).

2.1.4 Sélection des phrases du résumé

Une fois l'arbre généré, nous allons faire l'élagage (simplification de l'arbre) selon le type de résumé indicatif ou selon les relations choisies par l'utilisateur tout en tenant compte des segments noyaux.

Tous les noyaux ne sont pas d'égale importance. En effet, l'étape de sélection des unités minimales importantes (noyaux), profite des relations entre les structures de discours pour décider du degré de leur importance. L'extrait final affiche les unités noyaux retenues après la simplification de l'arbre RST.

La simplification de l'arbre, prendra en considération la liste des relations retenues par l'utilisateur. Au cas où ce dernier ne précise aucun choix, le système détermine automatiquement les relations retenues pour le type de résumé indicatif. En effet, la réduction de l'arbre RST se fait par la suppression de tous les descendants qui viennent d'une relation rhétorique non choisie par l'utilisateur (Keskes et al., 2010a).

Cette méthode proposée a été implémentée dans le système ARSTResume.

3 Méthode numérique proposée

Dans cette section nous présentons la méthode numérique proposée pour le résumé automatique de documents arabes, ainsi qu'une description détaillée des différentes étapes de cette méthode.

3.1 Présentation

La méthode numérique pour le résumé automatique, d'articles de journaux en langue arabe, se base sur une technique d'apprentissage. Plus précisément, elle est basée sur la technique d'apprentissage semi-supervisé, qui se compose de deux phases à savoir :

La phase d'apprentissage qui permet au système d'apprendre à extraire les phrases du résumé. Cette phase se compose de deux étapes, une étape de segmentation et d'annotation, et une étape d'apprentissage.

La deuxième phase est la phase d'utilisation qui permet aux utilisateurs de résumer un nouveau document. Cette phase est composée de deux étapes, une étape de segmentation et d'annotation et une étape de classification (Boudabous et al., 2010).

Les différentes phases de notre méthode sont illustrées dans la figure 3.

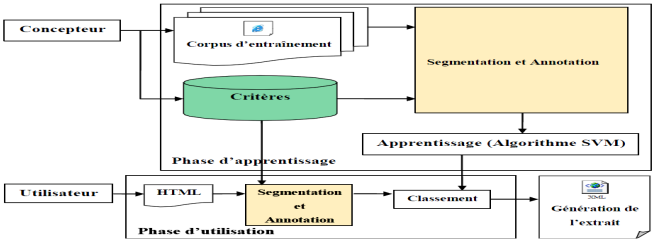

FIGURE 3 -Principales étapes de la méthode numérique

3.2 Description détaillée de la méthode

3.2.1 Phase d'apprentissage

La phase d'apprentissage nécessite l'utilisation d'un corpus d'entraînement ainsi qu'une base de critères d'extraction.

Le corpus d'entraînement est constitué de cent documents étiquetés (textes sources et leurs résumés) en format HTML (au moyen de trois pages par document). Les résumés de référence sont faits par trois experts humains afin d'apprendre au système comment produire des résumés similaires à ceux des experts humains de manière automatique.

Les critères d'extraction sont utilisés pour annoter les phrases des documents constituant notre corpus d'entraînement.

Nous avons classé les critères dans deux classes : les critères positionnels et les critères lexicaux. Ces derniers associent un score normalisé à chaque phrase, par contre les critères positionnels classent les phrases selon leurs postions dans le texte, présentés dans la table 4.

Critères positionnels	Position_ ph_texte	Classe la phrase selon sa position dans le texte : 1 si la phrase est dans le premier tiers du texte, 2 si elle est dans le deuxième tiers et 3 autrement.
	Position_ ph_sec	Classe la phrase selon sa position dans la section : 1 si la phrase est dans le premier tiers du texte, 2 si elle est dans le deuxième tiers et 3 autrement.
Critères lexicaux	Nb_mot_titre	Calcule le nombre d'apparition des mots du titre dans la phrase.
	Nb_exp_bonus	Calcule le nombre d'expressions bonus dans la phrase.
	Tf*Idf	Calcule le score tf*idf de la phrase.

TABLE 4 -Critères d'extraction

- **Segmentation et annotation du corpus**

Cette étape aboutit à la construction d'un vecteur d'extraction pour chaque unité du texte. L'ensemble des vecteurs d'extraction forme un fichier d'entrée pour l'étape d'apprentissage. La sous étape segmentation a pour but de découper le texte en unités minimales. Nous avons adopté la même segmentation utilisée dans la méthode symbolique. Concernant la sous étape d'annotation, l'acte annotatif consiste à donner une valeur ou un jugement à un segment du texte en se référant aux critères d'extraction. Cette étape a pour but d'annoter chaque segment du texte selon les différents critères d'extraction présentés précédemment. Chaque phrase de la collection est décrite par vecteur d'extraction, où la valeur donnée d'un critère correspond à la valeur d'analyse de la phrase selon ce critère.

- **Étape d'apprentissage**

L'algorithme d'apprentissage utilisé est l'algorithme SVM (Machines à Vecteurs de Support). Le choix de cet algorithme se justifie par sa robustesse de classification binaire, sa vitesse d'exécution et son adaptation aux problèmes non linéairement séparables. Cet algorithme génère une seule règle d'extraction appelée équation de l'hyperplan qui sépare les phrases pertinentes des phrases non pertinentes. Ainsi, l'algorithme d'apprentissage élimine les critères qui sont inutiles pour la phase d'apprentissage.

3.2.2 Phase d'utilisation

Cette phase permet à l'utilisateur du système de bénéficier des résultats de la phase d'apprentissage pour résumer un nouveau document. Les étapes par lesquelles passe le

texte à résumer sont : l'étape de segmentation et d'annotation, et l'étape de classement. L'étape de classification prend comme entrées les vecteurs d'extraction générés par l'étape de segmentation et d'annotation et l'équation de l'hyperplan générée par la phase d'apprentissage. L'équation de l'hyperplan est utilisée pour calculer le score de chaque phrase en se basant sur les vecteurs d'extraction. Cette méthode a été implémentée dans le système Résumeur Intelligent Arabe (R.I.A) (Boudabous et al., 2010).

4 Méthode hybride proposée

Dans cette section, nous proposons une méthode hybride pour le résumé automatique. Elle consiste à coupler la méthode linguistique et la méthode numérique.

4.1 Présentation

La méthode hybride, pour le résumé automatique des documents arabes, consiste à combiner la méthode symbolique basée sur la RST et la méthode numérique à base d'apprentissage. La figure 4 illustre le principe de cette méthode.

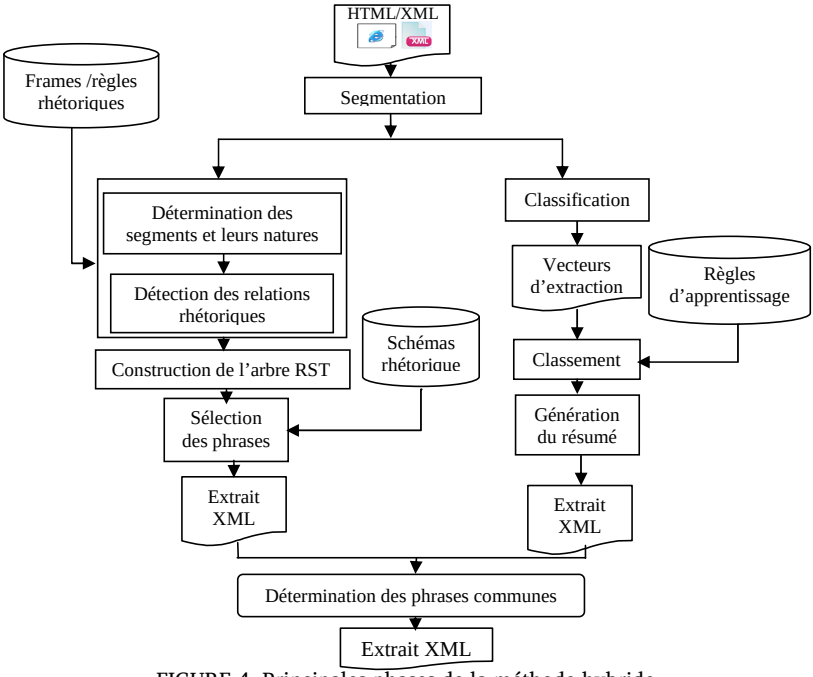

FIGURE 4 -Principales phases de la méthode hybride

4.2 Description détaillée de la méthode hybride

La méthode hybride que nous proposons se base sur la méthode symbolique et la méthode numérique, qui ont en commun le corpus d'étude et l'étape de segmentation des textes. Ces deux méthodes sont exécutées simultanément (en parallèle) comme nous l'avons décrit ci-dessus (section 2.2 et 3.2), puis, nous avons ajouté une étape de combinaison des résultats des deux méthodes.

L'étape de combinaison consiste à sélectionner les phrases communes des deux résumés générés par la méthode symbolique et la méthode numérique. Cette combinaison permet d'avoir un seul résumé pour chaque texte qui contient les phrases sélectionnées à la fois par la méthode symbolique et par la méthode numérique.

L'implémentation de cette méthode est basée sur l'intégration des deux systèmes ARSTRésume et R.I.A., à laquelle nous avons ajouté l'étape de combinaison. Le système développé s'appelle HybridResume.

5 Évaluation

Le corpus d'évaluation est formé de cent articles de presse, en langue arabe, rapatriés du journal Dar El Hayet[1] sans restriction quant a leurs contenu, taille, domaine et auteur. Ainsi, nous avons procédé à l'évaluation de la performance et de la pertinence des résumés générés par les trois systèmes, à l'aide d'une étude comparative qui mettra en jeu les résultats générés par les systèmes avec ceux réalisés par trois experts humain.

Nous avons utilisé le même corpus d'évaluation pour évaluer les trois systèmes (ARSTResume, R.I.A et HybridResume). Notons que ces trois systèmes ont utilisé le même module de segmentation pour avoir le même ensemble de phrases à traiter.

Nous avons procédé à trois expérimentations pour évaluer les trois systèmes. Chaque expérimentation compare les résumés de nos systèmes avec un résumé de l'expert. Le tableau suivant présente la moyenne de rappel, de précision et de f-mesure pour chacun des trois systèmes par rapport aux trois experts.

	ARSTResume			R.I.A.			HybridResume		
	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure	Rappel	Précision	F-mesure
Expert 1	0.52	0.58	0.52	0.59	0.62	0.6	0.52	0.66	0.63
Expert 2	0.39	0.62	0.46	0.53	0.7	0.6	0.58	0.74	0.7
Expert 3	0.5	0.59	0.51	0.63	0.7	0.66	0.6	0.79	0.71
Moyenne	0.47	0.6	0.5	0.58	0.67	0.62	0.57	0.73	0.68

TABLE 5 –Résultats d'évaluation des trois systèmes

[1] Source : http://www.daralhayat.com

Nous remarquons que l'approche numérique est plus performante que l'approche symbolique et qu'HybridResume surclasse l'approche numérique sur ce corpus, et ce pour les 3 types de mesures effectuées.

6 Discussion des résultats obtenus

Suite à l'évaluation des trois systèmes, nous avons obtenu comme valeurs moyennes de rappel, de précision et de F-Mesure respectivement : 47%, 60% et 50% pour le système ARSTRésume, 58%, 67% et 62% pour le système R.I.A et 57%, 73% et 68% pour le système HybridResume. Nous remarquons, que ces mesures différentes d'un système à un autre et d'un expert à l'autre. Cela se justifie par le fait que chaque système à sa propre méthode, et que le résumé avec lequel nous faisons la comparaison dépend du jugement vis-à-vis du domaine d'intérêt de l'expert.

En comparant les mesures des trois systèmes simultanément, nous avons remarqué que le système HybridResume présente toujours les mesures les plus élevées. Voyons d'où cela provient en comparant les deux systèmes ARSTRésume et R.I.A.

En examinant ses mesures calculées sur le corpus d'évaluation pour chacun des deux systèmes, ARSTRésueme et R.I.A, nous avons remarqué que plus le texte est long, plus le système ARSTRésume présente les mesures de rappel et de précision les plus élevées. En effet, cette déduction se justifie par le fait que plus le texte est long, plus il contient de marqueurs linguistiques et de relations rhétoriques. Par conséquent, le système ARSTRésume fait le maximum de couverture pour générer un extrait semblable à celui réalisé par l'expert humain.

A contrario, le système R.I.A., présente ses mesures de rappel et de précision, les plus élevées lorsque le texte est court, car, plus le texte est long, plus nous avons un calcul complexe qui diminue la performance du système.

HybridResume se comporte mieux en moyenne sur un corpus de texte bien distribué entre textes longs et courts, ce qui justifie ses meilleures performances.

7 Conclusion

L'étude, que nous avons présentée, s'inscrit dans le cadre des travaux de recherche effectués sur les résumés automatiques de documents arabes. Dans ce contexte, nous avons présenté trois méthodes différentes de résumé automatique (i.e. une méthode symbolique, une méthode numérique et une méthode hybride). Nous avons implémenté ces trois méthodes respectivement dans les trois systèmes ARSTResume, R.I.A et HybridResume.

Ces trois systèmes ont été évalués sur un même corpus d'évaluation composé de cent textes résumés par trois experts. L'évaluation, a montré que le système R.I.A produit des résultats meilleurs que ceux produits par le système ARSTResume. En effet, les mesures de précision sont respectivement de 60% et 67% pour les systèmes ARSTResume et R.I.A. La performance relative au système R.I.A par rapport au système ARSTResume s'explique

par la difficulté de l'analyse linguistique. En effet, l'absence de relations rhétoriques, la présence des mots ambigus et le manque d'informations morphologiques ont une influence négative sur les valeurs de rappel et de précision. Toutefois, le système HybridResume, qui implémente une méthode hybride, donne les meilleurs résultats (73% de précision).

Suite à cette étude comparative, Nous avons conclu que l'approche numérique est plus robuste que l'approche symbolique, lorsque le texte est court et que l'approche symbolique est plus robuste lorsque le texte est long. Par conséquent, nous trouvons que la combinaison de ces deux approches en une approche hybride donne de meilleurs résultats.

Comme perspective, nous envisageons d'introduire une analyse morphologique pour la méthode symbolique en vue de mieux repérer les relations rhétoriques et d'améliorer les performances des systèmes.

8 Bibliographie

AMINI M.R.(2001). Apprentissage Automatique et Recherche d'information: Application à l'extraction d'information de surface et au résumé de texte. Thèse de doctorat, université Paris-6 France.

ASHER N.(1993). Reference to Abstract Objects in Discourse. Kluwer Academic Publishers, Netherlands.

Azmi A.M. et Al-Thanyyan S.(2012). A Text Summarizer for Arabic. *Computer Speech & Language. ISSN :0885-2308.*

BELGUITH H.L., BACCOUR L. et MOURAD G.(2005). Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. 12ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2005), Dourdan, France, 6-10 juin 2005, pp 451–456.

BOUDABOUS M.M., MAALOUL, M.H. et BELGUITH H. L.(2010). Digital Learning for Summarizing ARABIC Documents . IceTAL, Islande.

IRAKY K., ZAKAREYA A. et FARAWILA A.(2011). Arabic Discourse Segmentation Based on Rhetorical Methods. International Journal of Electric & Computer Sciences IJECS-IJENS Vol: 11 No: 01.

IRIA C., SILVIA F., PATRICIA v., VIVALDI J., SANJUAN E. et TORRES-MORENO J. M.(2007). A new hybrid summarizer based on Vector Space Model, Statistical Physics and Linguistics. Lecture Notes in Computer Science 4827. 872-882. ISSN 0302-9743.

KAMP H. et REYEL U.(1993), From Discourse To Logic , Dordrecht Kluwer.

KAMP H.(1981). Evénements, représentations discursives et référence temporelle. Langages, p 34-64.

KESKES I.(2011). Résumé automatique de textes arabes basé sur une approche symbolique. Editeur : EUE. ISBN-13 : 978-3841780232

KESKES I. et MAALOUL M. H.(2010). Résumé automatique de documents arabes basé sur la technique RST . Conférence international de Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (TALN /RECITAL 2010), 12ème edition, Montréal – Canada.

KESKES I., MAALOUL M. H. et BELGUITH L. H.(2010) ,(a). التلخيص الآلي للنصوص العربية اعتمادا على نظرية البنية البلاغية . International Computing Conference in Arabic, 6ème édition, Hammamet – Tunisie, prix du Best Paper.

KESKES I., MAÂLOUL M. H., BELGUITH L. H. et BLACHE P.(2010), (b). Automatic summarization of Arabic texts based on RST technique. International Conference on Enterprise Information Systems, 12ème edition, Madeira – Portugal.

LASCARIDES A. et ASHER N.(1993), Temporal Interpretation, Discourse Relations, and Commonsense Entailment , Linguistics and Philosophy, 16(5).

MAÂLOUL M. H.(2007). Al Lakas El'eli / الآلي اللخاص : Un système de résumé automatique de documents arabes . IBIMA.

MANN W. C. et THOMPSON S. A.(1988). Rhetorical structure theory: Toward a functional theory of text organization . Text, 8(3), p 243 – 281.

MARCU D.(1999). Discourse trees are good indicator of importance in text, Advances in Automatic Text Summarization. p123 – 136.

MINEL J.L.(2002). Filtrage sémantique : du résumé automatique à la fouille de textes. Hermès Science Publications, Paris.

MOURAD G.(1999). La segmentation de textes par l'étude de la ponctuation. CIDE'99, Document Electronique Dynamique, p 155 – 171, Damas, Syrie.

NICOLAS U., AMINI M.R. et GALLINARI P.(2005). Résumé automatique de texte avec un algorithme d'ordonnancement . CORIA.

TOFILOSKI M., BROOKE J. et TABOADA M.(2009). A Syntactic and Lexical-Based Discourse Segmenter. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics.