RECITAL 2010, Montréal, 19-23 juillet 2010

Résumé automatique de documents arabes basé sur la technique RST

Mohamed Hédi Maâloul[1] Iskandar keskes[2]

(1)*Laboratoire LPL, 5 avenue Pasteur - BP 80975, 13604 Aix-en-Provence, France*
mohamed.maaloul@lpl-aix.fr

(2) *Laboratoire MIRACL, Route de Tunis Km 10, BP 242, 3021 – Sfax, Tunisie*
iskandarkeskes@gmail.com

Résumé Dans cet article, nous nous intéressons au résumé automatique de textes arabes. Nous commençons par présenter une étude analytique réalisée sur un corpus de travail qui nous a permis de déduire, suite à des observations empiriques, un ensemble de relations et de frames (règles ou patrons) rhétoriques; ensuite nous présentons notre méthode de production de résumés pour les textes arabes. La méthode que nous proposons se base sur la Théorie de la Structure Rhétorique (RST) (Mann et al., 1988) et utilise des connaissances purement linguistiques. Le principe de notre proposition s'appuie sur trois piliers. Le premier pilier est le repérage des relations rhétoriques entres les différentes unités minimales du texte dont l'une possède le statut de noyau – segment de texte primordial pour la cohérence – et l'autre a le statut noyau ou satellite – segment optionnel. Le deuxième pilier est le dressage et la simplification de l'arbre RST. Le troisième pilier est la sélection des phrases noyaux formant le résumé final, qui tiennent en compte le type de relation rhétoriques choisi pour l'extrait.

Abstract In this paper, we focus on automatic summarization of Arabic texts. We start by presenting an analytical study carried out on a study corpus which enabled us to deduce, following empirical observations, a set of relations and rhetorical frames; then we present our proposed method to produce summaries for Arabic texts. This method is based bases on the **R**hetorical **S**tructure **T**heory (RST) technique (Mann and Al., 1988) and uses purely linguistic knowledge. The principle of the proposed method is based on three pillars. The first pillar is the location of the rhetorical relations between the various minimal units of the text of which one has the status of nucleus - text segment necessary to maintain coherence - and the other has the status of nucleus or satellite - optional segment. The second pillar is the representation and the simplification of RST-tree that is considered most descriptive. The third pillar is the selection of the nucleus sentences forming the final summary, which hold in account the type of rhetorical relations chosen.

Mots-clés : Théorie de la Structure Rhétorique, Relations rhétoriques, Marqueurs linguistiques, Résumé automatique de textes arabes.

Keywords: Rhetorical Structure Theory, Rhetorical relations, Linguistic markers, Automatic summarization of Arabic texts.

1 Introduction

Dans le contexte actuel, nous sommes confrontés à un défi qui consiste à gérer la masse gigantesque des documents textuels électroniques disponibles facilement à travers les réseaux et les supports informatiques. La nécessité d'offrir des outils informatiques de visualisation rapide des textes, afin que l'utilisateur puisse évaluer la pertinence d'un document vis-à-vis de l'information recherchée, devient incontournable. Le résumé automatique fournit une solution qui permet d'extraire utilement les informations intéressantes en vue d'une réutilisation profitable. En effet, le résumé permet d'aider le lecteur à décider si le document source contient l'information recherchée ou pas. Il se peut aussi que le lecteur n'ait pas besoin de lire la totalité du document source, simplement parce que l'information recherchée existe dans le résumé (Maâloul, 2007).

Remarquons que plusieurs tendances se sont manifestées dans le domaine du résumé automatique. En effet, plusieurs approches ont été explorées en symbolique (basées sur l'analyse du discours et de sa structure) et en numérique (basées sur un calcul statistique, probabiliste ou même sur l'apprentissage) (Amini et al, 2002).

Par ailleurs, la plupart des systèmes de résumé automatique traitent des textes de langue latine comme 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 arabe (Maâloul, 2007).

Ainsi, les réalisations dans le domaine de résumé automatique sont reparties généralement selon la/les approches utilisées. On distingue principalement trois approches: numérique, symbolique et hybride. Notre contribution se situe dans le contexte d'approche symbolique et nous proposons un outil de résumé de textes arabes basé sur une technique purement symbolique : la technique RST (Mann et al., 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 comme but d'établir les relations et les dépendances ainsi que l'importance relative des phrases ou propositions les unes par rapport aux autres.

Soulignons que notre méthode aborde la question des *besoins des utilisateurs* étant donné qu'une information n'est pas importante en soi, mais doit correspondre aux besoins d'un utilisateur.

Le présent article se structure autour de trois volets. Le premier volet présente le corpus de travail et l'analyse linguistique qui a été menée sur ce corpus pour la recherche des relations rhétoriques ainsi que l'organisation des marqueurs linguistiques. Le deuxième volet expose la méthode proposée afin de présenter le texte sous forme d'un arbre hiérarchique des relations rhétoriques. Le troisième volet présente le procédé de génération du résumé. Ce procédé se réalise par la sélection des phrases correspondant aux unités textuelles contenant une relation rhétorique choisie par l'utilisateur.

2 Analyse linguistique du corpus d'étude

La réalisation d'un résumeur automatique nécessite au préalable une analyse linguistique du corpus. Le but essentiel de cette analyse est le repérage des unités linguistiques de surface qui représentent des *marqueurs linguistiques* déclencheurs de recherche ainsi *leurs marqueurs de validation* correspondants. Ces *marqueurs linguistiques* sont indépendants d'un domaine particulier et sont organisés dans des *relations rhétoriques.* Afin d'illustrer les particularités de notre domaine d'étude, nous allons présenter les différentes étapes de notre étude effectuée sur un corpus d'articles de presse.

2.1 Présentation du corpus

C'est à partir du Web que nous avons construit notre corpus de textes en langue arabe. Nous avons choisi comme genre textuel les articles de presse[1]. Ces articles sont de type HTML avec un codage UTF-8. Ils ont été rapatriés sans restriction quant à leur contenu et leur volume. Nous estimons en effet que plus le corpus est varié, plus il sera représentatif et contiendra le plus important nombre de marqueurs linguistiques.

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

2.2 Etude du corpus et repérage des relations rhétoriques

Le but de l'étude du corpus consiste à repérer des *frames* (règles ou patrons) des relations rhétoriques. Les frames sont des règles rhétoriques formées par des signaux linguistiques et des heuristiques observés qui sont principalement des marqueurs indépendants d'un domaine particulier mais qui ont des valeurs importantes dans un article de presse (Alrahabi, 2006).

Ces règles rhétoriques sont appliquées pour construire par la suite des différentes structures de l'arbre rhétorique.

Les marqueurs forment les *frames* d'une relation rhétorique et ont un double rôle : premièrement de lier deux unités minimales [2] adjacentes, dont l'un possède le statut de *noyau* – segment de texte primordial pour la cohérence – et l'autre a le statut *noyau* ou *satellite* – segment optionnel (Christophe, 2001) ; deuxièmement les types des relations qui les relient.

Nous avons commencé notre étude analytique par l'analyse sémantique des textes du corpus. Cette étude nous a permis de repérer une vingtaine de relations rhétoriques formées par un ensemble de *frames* rhétoriques. Un *frame* rhétorique est constitué de *marqueurs linguistiques.*

Toutefois, ces marqueurs peuvent être répertoriés en deux types : *indicateurs déclencheurs* et *indices complémentaires* (Minel, 2002). Les indicateurs déclencheurs énoncent des concepts importants qui sont pertinents pour la tâche de résumé automatique. 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.

L'exemple suivant illustre une phrase repérée dans l'un des articles du corpus :

(1)لكن ألبير قصيري **لم** يكن نزيل غرفته في ذلك الفندق فقط، **بل** كان أحد وجوه الشارع وبعض مقاهيها الشهيرة، (2) **لا سيما** مقهى «فلور» الذي كان يقضي فيه ساعات وحيداً أو مع أشخاص عابرين.

(1)Mais Albert Kasiry **n**'était **pas** seulement résidant de sa chambre dans cet hôtel, **mais** il était l'un des gents connus dans la rue et de certains de ses cafés célèbres, (2) **ainsi** le café «Flor» où il passe quelques heures tout seul ou avec des personnes passagers.

Cette phrase contient une relation de spécification / تخصيص entre la première unité minimale (1) et la deuxième unité minimale (2).

Une relation de spécification / تخصيص a généralement pour rôle de détailler ce qui est indiqué et de confirmer le sens et de le clarifier.

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

Nom de relation :	{Spécification / تخصيص}
Contrainte sur (1) :	contient un/des indice(s) complémentaire(s) {mais / بل, ne pas /لم, non /لا, etc.}
Contrainte sur (2)	contient l'indice déclencheur {ainsi / لا سيما}
Position de l'indicateur déclencheur	Milieu
Unité minimale retenue	(2)

A l'issue de notre étude du corpus, nous avons énuméré les relations rhétoriques suivantes :

2. Les auteurs de la RST définissent les unités minimales comme des unités fonctionnellement indépendantes : elles correspondent généralement aux propositions.

Liste des relations rhétoriques	
	Condition / شرط
	Concession / استدراك
	Enumération / تفصيل
	Restriction / استثناء
	Confirmation / توكيد
	Réduction / تقليل
	Joint / ربط
	Evidence / قاعدة
	Négation / نفي
	Exemplification / تمثيل
	Explication /تفسير
	Classement /ترتيب
	Conclusion / استنتاج
	Affirmation / جزم
	Définition / تعريف
	Pondération / ترجيح
	Possibilité / إمكان
	Restriction / حصر
	Spécification / تخصيص

Tableau 1 : Liste des relations rhétoriques

Notons que quelques unes de ces relations rhétoriques sont communes avec celles trouvées par Waleed (Mathkour et al., 2008).

2.3 Organisation des frames rhétoriques en relations

Il s'agit dans cette phase de construire les frames rhétoriques formés par des marqueurs (*indicateurs déclencheur et indices complémentaires*) et de les classer selon les relations rhétoriques. Ainsi nous aurons, dans une relation rhétorique, une liste de patrons linguistiques formés d'un ensemble d'unités linguistiques dont les catégories sont parfois hétérogènes (noms, verbes, connecteurs, mots outils ou grammaticaux, etc.) mais qui remplissent toujours les mêmes fonctions sémantiques discursives.

Voici quelques exemples de frames répartis selon les relations rhétoriques:

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

Nom de relation :	{confirmation /توكيد}
Contrainte sur (1) :	contient un/des indice(s) complémentaire(s) {prenant/ إذ, وإذ}
Contrainte sur (2)	contient l'indice déclencheur {على رغم, رغم, أنه, فإن, إنها, إن, لقد , لئن}
Position de l'indicateur déclencheur	Début
Unité minimale retenue	(2)

3 Méthode proposée

La méthode proposée pour le résumé automatique des articles de presse en langue arabe se base principalement sur des techniques d'extraction moyennant des *critères linguistiques*.

Notre étude du corpus a montré que certains types d'unités linguistiques importantes sont généralement retenus pour résumer un article de presse et que ces unités linguistiques peuvent être repérées en utilisant des *frames* ou règles rhétoriques. Nous avons répertorié ces frames rhétoriques en classes de relations rhétoriques.

Notre méthode mobilise ces ressources linguistiques et utilise automatiquement les marqueurs linguistiques pour mieux focaliser la recherche et le repérage des informations pertinentes dans un texte. Cette étape de repérage permet d'attribuer des étiquettes rhétoriques aux différentes unités du texte source.

Notons que notre proposition cible les besoins potentiels d'un utilisateur. En effet, une information n'est pas importante en soi, mais doit correspondre aux besoins d'un utilisateur. Nous offrons alors, à l'utilisateur, la possibilité de construire ses propres itinéraires à travers le texte et ce en choisissant les relations rhétoriques qui l'intéressent.

Le résumé final sera généré selon le genre: résumé indicatif (Qui, Quoi, etc.) (Maâloul, 2007). Le résumé peut être aussi généré selon un profil utilisateur. En effet, un utilisateur peut préférer un résumé se focalisant sur les unités minimales importantes (*noyaux*) décrivent les relations définitoires alors qu'un autre peut s'intéresser à un résumé focalisant sur les passages conclusifs.

De cette manière, nous abordons la production d'un résumé dynamique en fonction des intérêts de l'utilisateur.

Afin de limiter le nombre de phrases tout en augmentant leur pertinence, nous proposons de réduire l'arbre RST en éliminant tous les descendants qui forment des relations non retenues pour le résumé final. Nous nous sommes inspirés de la technique de simplification utilisée par Udo Hahn pour déterminer le rôle d'une expression propositionnelle dans un document en vue de tirer la structure de discours d'un texte (Udo Hahn et al., 2000). Ainsi, l'extrait final conserve seulement les unités minimales noyaux restantes dans l'arbre RST après simplification.

En résumé, outre l'utilisation classique de la technique RST pour présenter un texte sous une structure hiérarchique, notre proposition tient en considération lors de la sélection des phrases du résumé, des besoins potentiels d'un utilisateur et ce en exploitant le type et la sémantique des relations rhétoriques (définition, évidence, condition, conclusion, etc.).

4 Le système ARSTRESUME

La méthode que nous avons proposée pour le résumé automatique de textes arabe a été implémentée à travers le système ARSTRESUME. L'architecture de ce système est représentée dans la figure 1.

Les textes traités par ARSTRESUME sont d'abord prétraités pour les préparer à une segmentation en titres, sections, paragraphes et phrases. Le texte segmenté fera l'appel à une base de frames rhétoriques afin de détecter les différentes unités *noyaux* et *satellites* du texte, ainsi que les types des

relations qui les relient. Nous obtenons ainsi un seul arbre rhétorique. Ce dernier sera élaboré en se basant sur un certain nombre de *règles et de schémas rhétoriques*.

Les phrases sélectionnées pour le résumé final dépendent des relations rhétoriques choisies par l'utilisateur (ou automatiquement par le système en cas ou l'utilisateur ne précise aucun choix).

La figure 1. Présente les principales phases du système ARSTRESUME.

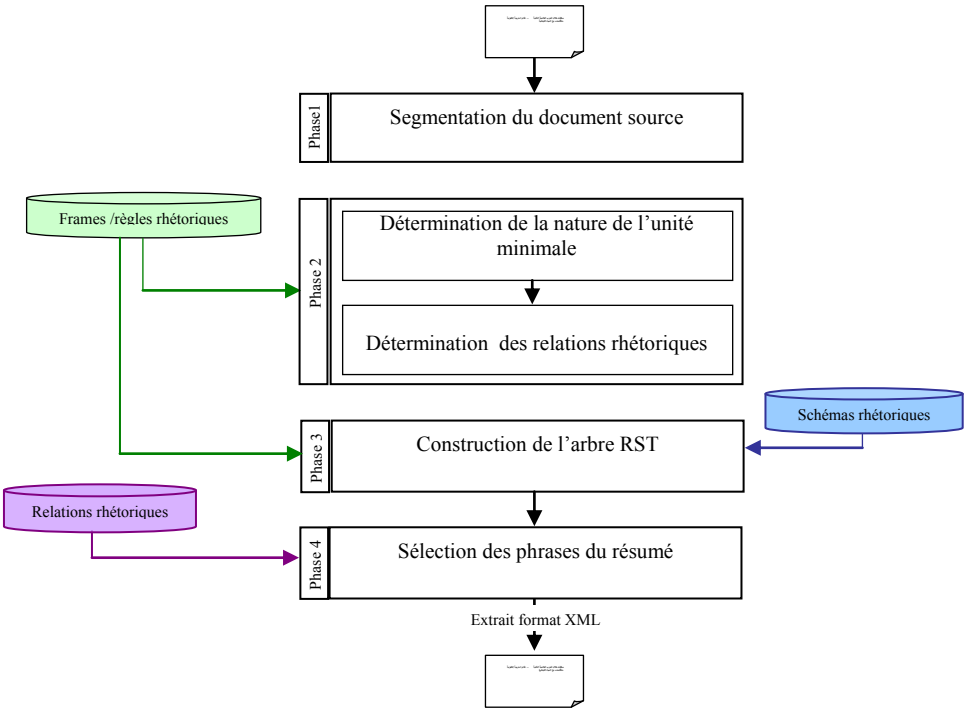

Figure 1 : Principales étapes de la méthode

4.1 Segmentation du document source

Cette phase consiste à hiérarchiser et à structurer le texte source en différentes unités plus petites : titres, sections, paragraphes et phrases.

Pour notre corpus constitué de textes en format HTML, nous utilisons un segmenteur pour la langue arabe basé sur les signes de ponctuation et sur un ensemble de balises HTML (
, <P> et </P>, <Div> et </Div>, etc.). Notons que la segmentation de textes arabes ne peut pas s'appuyer uniquement sur les signes de ponctuation mais elle se base aussi sur les conjonctions de coordination et un certain nombre de mots outils (Belguith et al, 2005). Cette étape de segmentation fournit en sortie un texte en format XML enrichi avec des balises encadrant les titre : <Titre>…</Titre>, les sections : <Section>…</Section>, les paragraphes : <Paragraphe>…</ Paragraphe> et les phrases : <Phrase>…< /Phrase>.

4.2 Détermination de la relation et de la nature de l'unité minimale

Cette étape a un double objectif; premièrement de lier deux unités minimales adjacentes entre elles, dont l'une possède le statut de *noyau* – segment de texte primordial pour la cohérence – et l'autre a le statut *noyau* ou *satellite* – segment optionnel, et deuxièmement la détermination des relations rhétoriques qui existent entre les différentes unités minimales juxtaposées d'un même paragraphe.

Les relations sont déduites à partir de la base des *frames rhétoriques*. Ainsi, les frames sont des règles rhétoriques formées par des critères linguistiques et heuristiques. Ces règles rhétoriques sont appliquées pour construire par la suite les différents arbres rhétoriques possibles.

4.3 Détermination de l'arbre RST

Afin de construire les différentes structures hiérarchiques (*arbre RST*) décrivant l'organisation structurelle du texte source, cette étape fait appel un certain nombre de *règles et de schémas rhétoriques*.

Les *règles rhétoriques* sont utilisées afin d'hiérarchiser et d'affiner l'arbre RST. Elles utilisent des heuristiques, adoptées après observation des résultats. Nous donnons ici à titre représentatif une règle rhétorique.

SI (un **indicateur déclencheur** se trouve au début de phrase) ALORS La phrase annotée est en relation avec le passage qui la précède

Tableau 2 : Exemple de règle rhétorique

Les *schémas rhétoriques* décrivant l'organisation structurelle d'un texte, quel que soit le niveau hiérarchique de ce dernier, permettent de lier un *noyau* et un *satellite*, deux ou plusieurs *noyaux* entre eux, et un *noyau* avec plusieurs *satellites* (Marcu, 1999).

Les différentes structures du texte sont donc définies en termes de compositions d'applications de schémas, et ce de manière itérative.

Les *schémas rhétoriques* se présentent sous la forme de cinq *modèles de schémas* (voir Figure 2.) qui peuvent être utilisés récursivement pour décrire des textes de taille arbitraire.

Généralement, le schéma le plus fréquent illustré est celui liant un *satellite* unique à un noyau *unique*.

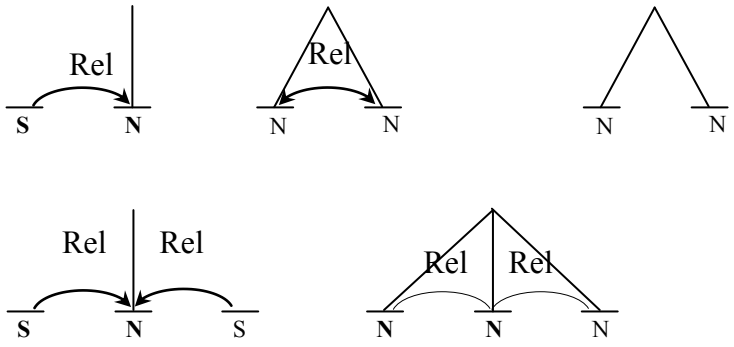

Figure 2 : Schémas rhétoriques de base RST (Mann et al., 1988)

L'exemple suivant présente une interprétation RST (Figure 3.) déduite à partir des modèles de schémas présentés précédemment relatifs au paragraphe suivant.

(1) تشتهر مدينة صفاقس بتقديم أطباق ثمار البحر على أنواعها. (2) **عندما** يرتاد زوار مدينة صفاقس، (3) **فإنهم** يطلبون باستمرار أطباق ثمار البحر و**خاصة** طبق المحار والإخطبوط المشوي على الفحم.

(1) La ville de Sfax est connue par la présentation des plats de fruits de mer de tout type. (2) **Lorsque** les visiteurs se rendent à la ville de Sfax, (3) **ils** demandent régulièrement les plats de fruits de mer et **surtout** le plat d'huître et de poulpe grillé sur le charbon.

Il est a signaler que le jugement d'appartenance à la relation rhétorique "*Evidence* / قاعدة" est attribué aux unités minimales (1) et (2). Cette attribution est faite en se basant sur *l'indicateur déclencheur* de recherche **Lorsque** / **عندما**. Alors que la relation rhétorique "Condition / شرط" est attribuée aux unités minimales (2) et (3). Cette attribution est faite en se basant sur *l'indicateur déclencheur* de recherche **surtout** / **خاصة** est *indices complémentaires* **Ils** / **فإنهم**.

La RST va réagir à cet exemple comme suit et nous aurons comme résultat l'arbre suivant :

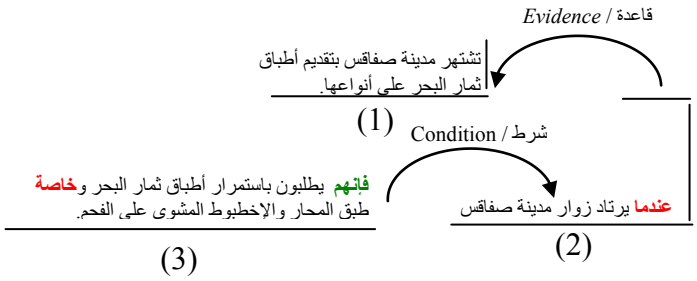

Figure 3 : Interprétation RST

4.4 Sélection des phrases du résumé

Pour le résumé, ce ne sont pas tous les noyaux qui sont considérés comme importants. En effet, l'étape de sélection des unités minimales importantes - noyaux, profite des relations entre les structures de discours pour en décider le 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. En cas où ce dernier ne précise aucun choix, le système détermine automatiquement les relations retenues pour le type de résumé indicatif.

Ainsi, suite à l'étude analytique menée sur une centaine de résumés réaliser par trois experts sur les documents du corpus, nous avons remarqué que généralement, un résumé indicatif est déterminé par cette liste relations rhétoriques.

Liste des relations rhétoriques	Condition / شرط
	Concession / استدراك
	Restriction / استثناء
	Confirmation / توكيد
	Evidence / قاعدة
	Négation / نفي
	Classement /ترتيب
	Affirmation / جزم
	Définition / تعريف

Tableau 3 : Liste des relations rhétoriques retenues pour le type de résumé indicatif.

La réduction de l'arbre RST se fait par la suppression de tous les descendants qui viennent d'une relation rhétorique non retenue (Udo Hahn et al., 2000).

5 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode de résumé automatique de textes arabes. Notre méthode se base sur la technique RST (Mann et al., 1988), qui utilise des connaissances purement linguistiques. Le but de notre proposition est de hiérarchiser le texte sous forme d'un arbre afin de déterminer les phrases *noyaux* formant le résumé final, qui tiennent compte des types de relations rhétoriques choisies pour l'extrait.

Notre travail s'est focalisé sur un type de texte particulier à savoir les articles de presse en format HTML. Le format XML a été abordé, mais pas assez suffisamment.

Comme perspective nous envisageons d'étendre notre évaluation sur un corpus plus large et d'étudier l'effet d'autres règles rhétoriques qui tiennent en consécration l'étiquetage morphosyntaxique des mots formant les unités minimales.

Nous envisageons aussi d'appliquer la méthode proposée à d'autres formats de documents comme le simple format texte.

Références

ALRAHABI M.(2006). "Annotation Sémantique des Énonciations en Arabe", *XXIV*^{ème} *Congrès en INFormatique des Organisations et Systèmes d'Information et de décision,* Hammamet-Tunisie.

AMINI M., ET GALLINARI P.(2002). "Apprentissage numérique pour le résumé de texte", Les Journées d'Étude de l'ATALA, Le résumé de texte automatique : solutions et perspectives, Paris France.

BELGUITH HADRICH L., BACCOUR L., MOURAD G.(2005), " Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules", Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles TALN'2005, Dourdan France, 6–10 Juin 2005, Vol. 1, p. 451–456.

CHRISTOPHE LUC.(2001), "Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte." TALN – Tours, France.

DOUZIDIA S.(2004). Résumé automatique de texte arabe, Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique.

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

WILLIAM C. MANN AND SANDRA A.(1988). Thompson. "Rhetorical structure theory: Toward a functional theory of text organization." Text, 8(3) : p. 243 – 281.

MARCU, D.(1997). The Rhetorical Parsing Summarization, and Generation of Natural Language Texts, PhD Thesis, Department of Computer Science, University of Toronto.

MARCU, D. "Discourse trees are good indicator of importance in text", *Advances in Automatic Text Summarization*, p. 123– 136.

HASSAN I. MATHKOUR, AMEUR A. TOUIR ET WALEED A. AL-SANEA. (2008), "Parsing Arabic Texts Using Rhetorical Structure Theory", *Journal of Computer Science* 4 (9): p.713–720.

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

UDO HAHN AND HOLGER SCHAUER (2000). "Phrases as carriers of coherence relations". *In Lila R. Gleitman and Aravind K. Joshi, editors, Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, p. 429-434. Philadelphia, PA, USA.