AUTOMATIC SUMMARIZATION OF ARABIC TEXTS BASED ON RST TECHNIQUE

Mohamed Hédi Mâaloul[1], Iskandar keskes[2]
[1]*Laboratoire LPL, 5 avenue Pasteur - BP 80975,13604 Aix-en-Provence, France*
mohamed.maaloul@lpl-aix.fr
blache@lpl-aix.fr

Lamia Hadrich Belguith[2], Philippe Blache[1]
[2]*LARIS- MIRACL Laboratory, FSEGS, BP 1088, 3018, Sfax - Tunisia*
iskandarkeskes@gmail.com
l.belguith@fsegs.rnu.tn

Keywords: Rhetorical Structure Theory, rhetorical relations, linguistic markers, automatic summarization of Arabic texts.

Abstract: We present in this paper an automatic summarization technique of Arabic texts, based on RST. We first present a corpus study which enabled us to specify, following empirical observations, a set of relations and rhetorical frames. Then, we present our method to automatically summarize Arabic texts. Finally, we present the architecture of the ARSTResume system. This method is based on the Rhetorical Structure Theory (Mann, 1988) and uses linguistic knowledge. The method relies on three pillars. The first consists in locating the rhetorical relations between the minimal units of the text by applying the rhetorical rules. One of these units is the nucleus (the segment necessary to maintain coherence) and the other can be either nucleus or satellite (an optional segment). The second pillar is the representation and the simplification of the RST-tree that represents the entries text in hierarchical form. The third pillar is the selection of sentences for the final summary, which takes into account the type of the rhetorical relations chosen for the extract.

1 INTRODUCTION

In the current context, we have to deal with a huge mass of electronic textual documents available through the net. We need tools offering fast visualization of the texts (so that the user can evaluate its relevance). Automatic summarization provides a solution which makes it possible to extract interesting information for an advantageous reuse. Indeed, the summary helps the reader to decide whether the original document contains the required information or not. Moreover, in some cases the reader does not need to read the totality of the original document, simply because the required information is in the summary (Mâaloul, 2007).

Automatic summarization approaches are inspired by various orientations. Some approaches relies on symbolic techniques (based on the analysis of the discourse and its discursive structure), some others are based on numerical treatments (based on a statistical, probabilistic calculation or even on training) (Amini, 2002).

In addition, the majority of automatic summarization systems mainly treat texts in English, French, German, etc. To our knowledge, there are very few systems that could handle Arabic. Thus, there is an increasing need to develop automatic summarization systems dedicated to Arabic to handle the increasing amount of electronic documents written in Arabic (Mâaloul, 2007).

Thus, the achievements in the field of automatic summary are generally set out again according to the approaches used. Mainly three approaches are distinguished: numerical, symbolic and hybrid. Our contribution is in the context of symbolic approach and we propose a system for the automatic summarization of Arabic texts which is based on a purely symbolic technique: RST technique (Mann, 1988). It is a question of detecting the semantic relations and the intentional relations which exist between the segments of a document. Indeed, the rhetorical analysis aims at establishing the relations as well as the relative importance of the sentences or propositions and the dependences on one another (Teufel, 1997).

Our method addresses the question of the *user's needs* since an information is not important in itself, but must correspond to the user's needs.

This paper is threefold. The first section presents the study corpus and the linguistic analysis which was based on this corpus to determine the rhetorical relations as well as the organization of the linguistic markers. The second section exposes the proposed method in order to present the text in the form of a hierarchical tree of rhetorical relations. The third section presents the process of summary generation. This process is carried out by the selection of the sentences containing the textual units corresponding to the indicative type of summary or the set of rhetorical relations chosen by the user.

2 LINGUISTIC ANALYSIS OF THE STUDY CORPUS

An automatic summary requires as a preliminary step a linguistic analysis of the corpus. The essential goal of this analysis is to determine the surface linguistic units which represent *linguistic markers* launching research (research the valid rhetorical relations) as well as their corresponding *validation markers*. These *linguistic markers* are independent from a particular field and are organized in *rhetorical relations.* In order to illustrate the characteristics of our study field, we present the various stages of our study carried out on a corpus of newspaper articles.

2.1 Presentation of the study corpus

Our corpus of Arabic texts has been created from the web, by selecting newspaper articles[1]. These articles are of HTML type with a UTF-8 coding.

They were downloaded without restriction to their contents and their volume: we think that the more the corpus is varied, the more it is representative (it contains an important number of linguistic markers).

2.2 The rhetorical relation extraction

Thanks to the study corpus, we determine the *frames* of the rhetorical relations. These frames are rhetorical rules formed by the linguistic signals and the observed heuristic, which are mainly markers independent from a particular field but which have important values in a newspaper article (Alrahabi, 2006).

Such rhetorical rules are applied to build the rhetorical tree (the RST-tree). The markers forming the *frames* of a rhetorical relation have, a double role. First, to bind two adjacent minimal[2] units together, one of these units having the status of a *nucleus* (segment of paramount text for coherence) and the other one having the status of a *nucleus* or *satellite* (optional segment) (Christophe, 2001) and second the types of rhetorical relations which connect them.

We began our analytical study with the semantic analysis of the texts of the corpus. This study enabled us to locate a score of rhetorical relations formed by a set of *rhetorical frames*. *A rhetorical frame* is made up of *linguistic markers*.

These markers can be indexed in two types: *releasing indicators* and *complementary indexes* (Minel, 2002). The releasing indicators state important concepts which are relevant for the task of automatic summarization. The complementary indexes are required in a space defined starting from

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

[2] The authors of the RST define minimal units (span) as functionally independent units: they correspond generally to the proposals.

the indicator (in the vicinity of the indicator). They can thus act in the context in order to confirm or to cancel the rhetorical relation stated by the releasing indicator.

Table 1: Rhetorical frame specification

Name of relation:	{Specification/تخصيص}
Constraint on (1):	contains a complementary index (es) {but/بل, not/لم, no/لا, etc}
Constraint on (2):	contains the releasing index {such as / لاسيما}
Position of the releasing indicator:	In the middle
Minimal unit reserve:	(2)

From our corpus study, we enumerated the following rhetorical relations:

Table 2: List of rhetorical relations

List of rhetorical relations	Condition/شرط
	Concession/استدراك
	Enumeration/تفصيل
	Restriction/استثناء
	Confirmation/توكيد
	Reduction/تقليل
	Joint/ربط
	Obviousness/قاعدة
	Negation/نفي
	Exemplification/تمثيل
	Explanation/ تفسير
	Classification/ترتيب
	Conclusion/استنتاج
	Assertion/جزم
	Definition/تعريف
	Weighting/ترجيح
	Possibility/إمكان
	Restriction/ حصر
	Specification/تخصيص

Let us note that some of these rhetorical relations are common to those described by other automatic summarizations using RST (Mathkour, 2008).

The following example illustrates a sentence extracted from one of the newspaper articles of our study 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) But Albert Kasiry was not a resident in his room in that hotel solely, **but** he was one of the street people, and its famous cafes, (2) **such as** the cafe "Flor" in which he spends a few hours all alone or with passers-by.

This sentence contains a {specification/تخصيص} relation between the first minimal unit (1) and the second minimal unit (2).

A {specification/تخصيص} relation has generally a role of detailing what is indicated and confirming meaning and clarifying it.

The frame specification (table 1) is used to detect the rhetorical relation specification.

2.3 Organization of the rhetorical frames in rhetorical relations

In this section we explain how to build the rhetorical frames formed by markers (*releasing indicators* and *complementary indexes*) and classify them according to the rhetorical relations. Thus we have, in a rhetorical relation, a list of linguistic patterns made of a set of linguistic units of which the categories are sometimes heterogeneous (nouns, verbs, connectors, word tools, etc.) but which always fulfill the same discursive semantic functions.

Below are some examples of frames distributed according to the rhetorical relations:

Table 3: Rhetorical frame negation

Name of relation:	{negation/نفي}
Constraint on (1):	contains a complementary index (es) {أما, بل, ولكن, لكنه, لكن, لكننا, لكنني, لكنهم}
Constraint on (2):	contains the releasing index {لم, ولم, لن, ليس, ليسوا, ليست}
Position of the releasing indicator:	In the middle
Minimal unit reserve:	(1)

Table 4: Rhetorical frame confirmation

Name of relation:	{confirmation/توكيد}
Constraint on (1):	contains a complementary index (es) {fascinating/ وإذ}
Constraint on (2):	contains the index release {لقد, إن, إنها , فإن, أنه, رغم , رغم, لئن, على}
Position of the releasing indicator:	Beginning
Minimal unit reserve:	(2)

3 PROPOSED METHOD

Our proposed method for automatic summarization of Arabic texts is mainly based on techniques of extraction using *linguistic criteria*.

Our corpus study showed that certain types of important minimal units are generally retained to summarize a newspaper article and that these minimal units can be located by using *frames* or *rhetorical rules*. We indexed these rhetorical frames in classes of rhetorical relations.

Our method mobilizes these linguistic resources and automatically uses the linguistic markers for better focusing the retrieval and the location of relevant information in a text. This stage of location makes it possible to attribute rhetorical labels to the various units of the original text.

Let us note that our proposition targets the potential needs of a user. Indeed, an information is not important in itself, but must correspond to the user's needs. We then offer, the user, the possibility to build its own routes through the text and this by choosing the rhetorical relations.

We are concerned with a dynamic summarization which can be generated according to a particular kind. The final summary will be generated by type: indicative summary (Which, What, etc.) (Maâloul, 2007). The summary can also be generated according to a user profile. Indeed, a user can prefer a summary focusing on the important minimal units (*nuclei*) describing the defining relations whereas another one can be interested in a summary focusing on the conclusive passages.

In this manner, we approach the production of a dynamic summarization according to the interests of the user.

In order to limit the number of sentences while increasing their relevance, we propose to reduce the RST tree by eliminating all the descendants which form relations not retained for the final summary. We took as a starting point the technique of simplification (Udo, 2000) to determine the role of a propositional expression in a document in order to draw out the structure of discourse from a text. Thus, the final extract preserves only the *nuclei* minimal units remaining in the tree RST after simplification.

In short, in addition to the traditional use of RST technique to present a text under a hierarchical structure, our proposition takes into consideration during the selection of the sentences of the summary, the potential needs of a user and this by exploiting the type and the semantics of the rhetorical relations (definition, obviousness, condition, conclusion, etc).

4 THE ARSTRESUME SYSTEM

The method that we proposed for automatic summarization of Arabic texts has been implemented through the ARSTResume system. The architecture of this system is represented in Figure 1.

The texts handled by ARSTResume are initially pretreated to prepare their segmentation in titles, sections, paragraphs and sentences. The segmented text will call upon a base of rhetorical frames in order to detect the various *nuclei* units and *satellites* of the text, as well as the types of the relations which connect them. Thus we obtain the rhetorical tree. This tree represents the text in hierarchical form. This will be developed and based on a set of rules and rhetorical diagrams.

The sentences selected for the final summary depend on the rhetorical relations chosen by the user (or automatically the system uses the indicative summary in case the user makes no choice).

This figure presents the principal phases of the ARSTResume system.

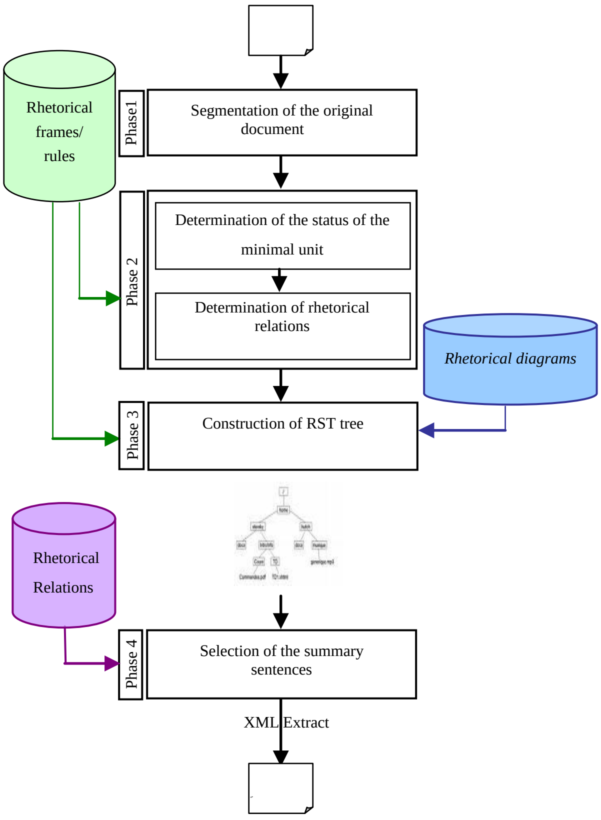
Rhetorical frames/ rules
Phase1
Segmentation of the original document
Phase 2
Determination of the status of the minimal unit
Determination of rhetorical relations
Rhetorical diagrams
Phase 3
Construction of RST tree
Rhetorical Relations
Phase 4
Selection of the summary sentences
XML Extract

Figure 1: Principal stages of the ARSTResume system

4.1 Segmentation of the original document

This phase consists in treating on a hierarchical basis and structuring the original text in minimal units: title, sections, paragraphs and sentences.

For our corpus made up of texts in HTML format, we use a tokenizer for Arabic language based on the punctuation marks and a set of HTML beacons (<Br>, <P> and </P>, <Div> and </Div>, etc). Let us note that the segmentation of Arabic texts cannot only be based on the punctuation marks but it is also based on the coordinating conjunctions and some word tools (Belguith, 2005). This stage of segmentation provides as output a text in XML format enriched with beacons framing the title: <Titre> … </Titre>, sections: <Section> … </Section>, paragraphs: <Paragraphe> … < Paragraphe> and sentences: <Phrase> … < /Phrase>.

4.2 Determination of the rhetorical relation and the status of the minimal unit

This stage has a double objective; firstly to bind two adjacent minimal units together, of which one has the status of *nucleus* (segment of paramount text for coherence) and the other has the status of *nucleus* or *satellite* (optional segment), and secondly the determination of the rhetorical relations which exist between the various juxtaposed minimal units of the same paragraph.

The relations are deduced starting from the base of the *rhetorical frames*. Thus, the frames are rhetorical rules formed by linguistic and heuristic criteria. These rhetorical rules are applied to build the rhetorical tree thereafter.

4.3 Construction of RST tree

In order to build the various hierarchical structures (*RST trees*) describing the structural organization of the original text, this stage calls for a certain number of *rules* and *rhetorical diagrams*.

The *rhetorical rules* are used to prioritize and refine the RST tree. They use heuristics adopted after observing the results. We give here as representative a rhetorical rule.

Table 5: Example of a rhetorical rule

```
IF (index release is at the beginning of the sentence)
THEN
    The sentence is annotated in connection with the passage that precedes.
End if
```

This rule can not be linked between the minimal units but between sentences and paragraphs of text, because the *rhetorical diagrams* are insufficient to represent the full text.

The *rhetorical diagrams* describing the structural organization of a text, whatever the hierarchical level of this latter, make it possible to bind a *nucleus* and a *satellite*, two or several *nuclei* together, and a *nucleus* with several *satellites* (Marcu, 1999).

The various structures of the text are thus defined in terms of compositions of applications of diagrams, and this in an iterative way.

The rhetorical diagrams are represented under five *models of diagrams* (Figure 2) which can be recursively used to describe texts of arbitrary size.

Generally, the most frequent diagram illustrated is the one linking a single *satellite* to a single *nucleus*.

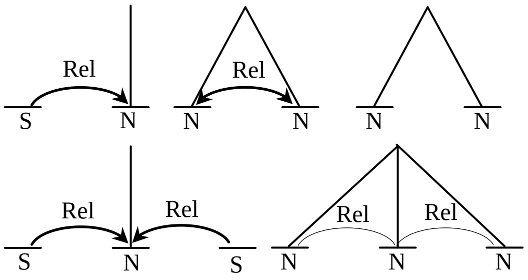

Figure 2: Basic rhetorical diagrams RST (Mann, 1988)

The following example presents an RST interpretation of the following paragraph.

(Figure 3) deduced from the diagram models presented previously.

(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) The town of Sfax is known for the presentation of the seafood dishes of any type. (2) **When** the visitors go to the town of Sfax, (3) **they** regularly ask for the seafood dishes and **especially** the dish of oyster and octopus roasted on charcoal.

It has to be announced that the judgment of membership of the rhetorical relation "*Obviousness*/قاعدة" is attributed to the minimal units (1) and (2). This attribution is made while being based on the *releasing indicator* of research **When/عندما**. Whereas the rhetorical relation "Condition/شرط" is attributed to the minimal units

(2) and (3). This attribution is made while being based on the *releasing indicator* of research **especially/خاصة** and *complementary index* **them/فإنهم**.

The RST will react to this example as follows and we will obtain as a result the following tree:

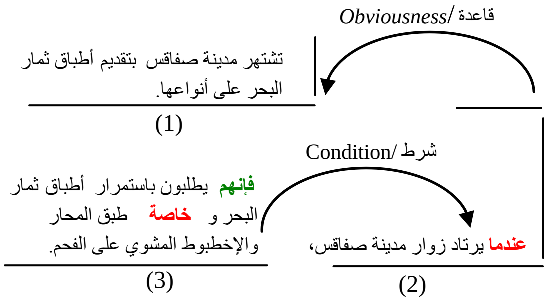
Obviousness/ قاعدة
تشتهر مدينة صفاقس بتقديم أطباق ثمار
البحر على أنواعها.
(1)
Condition/ شرط
فإنهم يطلبون باستمرار أطباق ثمار
البحر و خاصة طبق المحار
والإخطبوط المشوي على الفحم.
عندما يرتاد زوار مدينة صفاقس،
(3)
(2)

Figure 3: RST Interpretation

4.4 Selection of the summary sentences

For the summary, in fact not all the nuclei are considered to be important. Indeed, the stage of selection of the important minimal units - nucleus, benefits from the relations between the structures of discourse to decide the degree of their importance.

The final extract posts the nucleus units retained after the simplification of RST tree.

The simplification of the tree, will take into account the list of the relations retained by the user.

In case where this latter does not specify any choice, the system determines the relations automatically to retain according to the type of indicative summary to be generated.

Thus, following the analytical study conducted over a hundred summaries performed by three experts on the documents of the corpus, we noted that generally, an indicative summary is determined by this list of rhetorical relations, through an empirical study of the corpus.

Table 6: List of rhetorical relations selected for the indicative summary type.

List of rhetorical relations	Condition / شرط
	Concession / استدراك
	Restriction / استثناء
	Confirmation / توكيد
	Obviousness / قاعدة
	Negation / نفي
	Classification /ترتيب
	Assertion / جزم
	Definition / تعريف
	Restriction/ حصر

The reduction of RST tree is done by the removal of all the descendants which come from a rhetorical relation that has not been selected.

5 CONCLUSION AND PERSPECTIVES

In this paper, we proposed a method of automatic summarization of Arabic texts. Our method is implemented by ARSTRemue system and is based on the RST technique (Mann, 1988), which uses purely linguistic knowledge. The goal of our proposal is to treat on a hierarchical basis the text in the form of a tree in order to determine the nucleus sentences forming the final summary, which take into account the types of rhetorical relations chosen for the extract.

Our work focused on a particular type of texts (i.e., the newspaper articles in HTML format). The XML format was approached, but not rather sufficiently.

The ARSTResume has given encouraging results when we evaluated on a small corpus.

As a perspective, we plan to extend our evaluation on a larger corpus and to study the effect of other rhetorical rules which take into consideration the morpho-syntaxic features of the words forming the minimal units.

REFERENCES

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., 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, H., 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*, Vol. 1, p : 451–456, Dourdan-France.

Christophe, L., 2001. Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. *TALN – Tours*, France.

Mâaloul, M.H., 2007. Al Lakas El’eli / الآلي اللّخاص : Un système de résumé automatique de documents arabes, *IBIMA*.

Mann, W., C., 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*, p: 123– 136.

Mathkour, H., I., Touir A., Al-Sanie, W., 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.

Teufel, S., Marc, M., 1997. Sentence extraction as a classification task. *In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, p: 58-65, Madrid-Spain.

Udo, H., and Holger, S., 2000. Phrases as carriers of coherence relations, *In Lila R. Gleitman and Aravind K. Joshi, Proceedings of the 22nd Annual.*