AUTOMATIC SUMMARIZATION OF ARABIC TEXTS BASED ON RST TECHNIQUE

Mohamed Hédi Mâaloul[1], Iskandar Keskes[2], Lamia Hadrich Belguith[2] and Philippe Blache[1]
[1]*Laboratoire LPL, 5 avenue Pasteur - BP 80975, 13604 Aix-en-Provence, France*
[2]*ANLP Research Group- MIRACL Laboratory, FSEGS, BP 1088, 3018, Sfax, Tunisia*

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. Our method is based on the Rhetorical Structure Theory (Mann, 1988) and uses linguistic knowledge. It relies on three pillars. The first consists in locating the rhetorical relations between the minimal units of the text by applying 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 source 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 (Mâaloul, 2007).

In addition, the majority of automatic summarization systems mainly treat texts in English, French, German, etc. To our knowledge, there are 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 this language (Mâaloul, 2007).

The achievements in the field of automatic summarization are generally set out again according to the used approaches. Mainly three approaches are distinguished: numerical, symbolic and hybrid. Our contribution is in the context of the symbolic approach. We propose a system for automatic summarization of Arabic texts which is based on a purely symbolic technique: RST technique (Mann, 1988). The main idea of RST is to detect the semantic relations and the intentional relations 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 dependencies between them (Teufel, 1997).

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

2 LINGUISTIC ANALYSIS OF THE STUDY CORPUS

An automatic summary requires, as a preliminary step, a linguistic analysis of the corpus (newspaper articles). The main goal of this analysis is to determine the surface linguistic units which represent *linguistic markers* as well as their corresponding *validation markers*. These *linguistic markers* are independent from a particular field and are organized in *rhetorical relations.*

Hédi Mâaloul M., Keskes I., Hadrich Belguith L. and Blache P. (2010).
AUTOMATIC SUMMARIZATION OF ARABIC TEXTS BASED ON RST TECHNIQUE.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 434-437
DOI: 10.5220/0002976104340437

2.1 Presentation of the Study Corpus

Our corpus of Arabic texts has been created from the web, by selecting newspaper articles (http://www.daralhayat.com). These articles are of HTML type with an UTF-8 coding.

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 heuristics, which mainly represent markers independent from a particular field but have important values in newspaper articles (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 units together, one of these units having the status of a *nucleus* (segment of paramount text for coherence) and the other having the status of a *nucleus* or *satellite* (optional segment) (Christophe, 2001) and second to detect 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 set 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 segment defined starting from the indicator. Thus, they can act in the context in order to confirm or to cancel the rhetorical relation stated by the releasing indicator.

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

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)

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).

2.3 The Rhetorical Frame Organization

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. Indeed, in a rhetorical relation there is a list of linguistic patterns. These patterns are made of a set of linguistic units of which the categories are sometimes heterogeneous (nouns, verbs, connectors, word tools, etc.) but always fulfill the same discursive semantic functions.

Below an example of a frame distributed according to the rhetorical relations:

Table 3: Example of a 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)

3 PROPOSED METHOD

Our corpus study showed that some types of important *minimal units* are generally retained to summarize newspaper articles 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 aim is to propose a dynamic summarization which can be generate different kind of summaries (indicative, critical, informative, etc.) according to the user need. The final summary could be generated according to a user profile or according to the user choice. Thus, the user could choose the summary type or even the relation types to be considered when selecting the extract sentences (Maâloul, 2007). For example, 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 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 that will not be retained in 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 RST tree after simplification.

4 THE ARSTRESUME SYSTEM

The method that we proposed for automatic summarization of Arabic texts has been implemented through the ARSTResume system.

This figure presents the principal phases of the ARSTResume system.

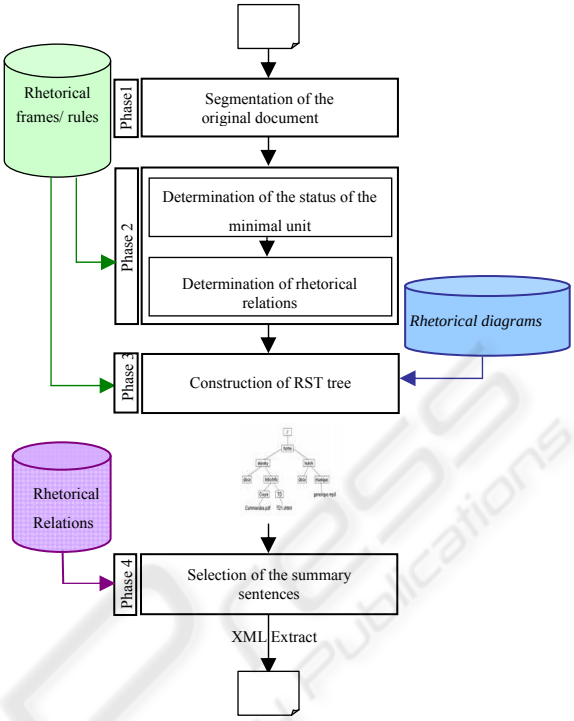

Figure 1: Principal stages of the ARSTResume system.

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.

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.

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 4: 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 preceds.
End if

The rule of table 4 can not linked between the minimal units but between sentences and paragraphs of text, because the *rhetorical diagrams* are insufficient to represent the full text.

The various structures of the text are thus defined in terms of compositions of applications of diagrams, etc.

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

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

4.4 Selection of the Summary Sentences

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

Basing on the analytical study that we have conducted on a hundred summaries performed by three experts, we determined a list of rhetorical relations for each summary type.

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 for a given summary type.

5 CONCLUSIONS AND PERSPECTIVES

In this paper, we proposed a method of automatic summarization of Arabic texts. Our method is implemented in the ARSTRemue system and is based on the RST technique (Mann, 1988), which uses purely linguistic knowledge. The goal of our proposal is to represent the text in the form of a tree in order to determine the nucleus sentences forming the final summary, then we generate the summary according to the types of rhetorical relations that correspond to the extract type (the extract type is either chosen by the user or determined from its profile; if not the indicative type is considered as a default type).

The ARSTResume evaluation has showed encouraging results based on 50 texts.

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.

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.

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.*