Journal of Computer Science 4 (9): 713-720, 2008
ISSN 1549-3636

Parsing Arabic Texts Using Rhetorical Structure Theory

Hassan I. Mathkour, Ameur A. Touir and Waleed A. Al-Sanea
Department of Computer Science, King Saud University, Riyadh, Saudi Arabia

Abstract: Problem Statement: Processing texts based on rhetorical structure theory has shown interesting results. Rhetorical Structure Theory (RST) improves the ability of extracting the semantic behind the processed text. Different applications such as information retrieval, text summarization, and text generation have proved to give better result using RST. The applicability of RST to process and understand texts has been studied in several languages, but little is devoted to the Arabic language. Given an Arabic text, the more accurate the Arabic rhetorical relations are extracted the more useful the subsequent text representation will be. This, in turn, leads to a better understanding of the text and, hence, better results. **Approach:** We show a framework of applying RST on Arabic language in order to rhetorically parse, understand, and summarize Arabic texts. We discuss a new approach that extracts the Arabic rhetorical relations that is based on studying the English relations, analyzing Arabic corpus and understanding and using the Arabic cue phrases. **Results:** We obtain rhetorical relations based on Arabic cues. We show how this approach contributes in improving the understanding of the Arabic text. The study addresses the relations that rise from cues that act as connectors among Arabic clauses as well as words. **Conclusion:** The introduced approach suggests that realizing text coherency in the process of obtaining Arabic rhetorical relations suits the characteristics of the Arabic language and avoids the disadvantages of previous approaches. The obtained Arabic rhetorical relations will make it possible to build rhetorical trees for Arabic texts to apply in text summarization and generation, information retrieval, and text segmentation while preserving the coherency of the text.

Key words: Rhetorical Structure Theory (RST), parsing Arabic texts, rhetorical relations, Arabic corpus analysis, Arabic cue phrases, Arabic text coherence.

INTRODUCTION

Rhetorical Structure Theory (RST)[11] was introduced to serve as a discourse structure in the computational linguistic field. It is intended to describe texts and offers an explanation of their coherence[14]. RST gives rise to rhetorical relations. Rhetorical relations can be described functionally in terms of the writer purposes and the writer assumptions about the reader. These rhetorical relations hold between adjacent and non-adjacent spans of texts.

The output of applying the rhetorical structure theory to a text is a tree structure that organizes the text based on the rhetorical relations[14]. This structure is called the rhetorical schema. Each relation connecting two spans of a text might be one of two cases: one of the two spans is more important to the reader than the other and it represents the semantic of the two spans, the other case is that the two spans have the same importance to the reader. In the first case, the important span is called the nucleus and the other span is called satellite. The other case is called multinuclear relation where both spans are considered nucleus. The process of parsing the text and building the rhetorical structure is called the rhetorical analysis. During the process of the rhetorical analysis, the elementary units that participate in building the rhetorical schema are determined and then the rhetorical relations that hold among these units are determined to connect related spans. Determining the potential relations that connects related spans could be done using several techniques[2,5,6]. One of such techniques is through the use of cue phrases[16]. Marcu[15] has given several cue phrases that can be used in the English language processing. Cue phrases have been used in various application including text segmentation[9,10,18] and text summarization[7].

MATERIALS AND METHODS

The process of identifying rhetorical relations and obtaining cue phrases presented in[13,15,16] was based on

Corresponding Author: Hassan I. Mathkour, Department of Computer Science, King Saud University, Riyadh, Saudi Arabia corpus analysis of English texts. Since the analysis was done on the English corpus, the rhetorical relations that were identified can serve in the processing and analysis of English text[7]. But there is no guarantee that the same set of relations work on other languages. Our aim is to apply such a theory on the Arabic language. We studied those relations in the Arabic corpus and Arabic literature[17]. Due to the differences between the Arabic and English languages, the English rhetorical relations cannot be used in their present forms for the Arabic text. For example, the relation concession is not known in the Arabic rhetoric and literature. This relation and the relation contrast might correspond to an Arabic relation that may be called (اســتدراك /Recalling). In the sequel we write the relation name in Arabic and its translation in English using the following format (Arabic relation name/English relation name). The following examples clarify the idea:

Example 1: The text in Table 1 was taken from[12] as an example of the relation concession. The text is translated to Arabic to illustrate the use of the relation in the Arabic texts. The numbers in the subscripts indicate the unit number

Example 2: The text in Table 2 was taken from [6] as an example of the relation contrast between units 1 and 2. The text is translated to the Arabic languages to illustrate the relation in Arabic texts. The numbers in the subscripts indicate the unit numbers.

Table 1: An example the relation concession

(Concern that this material is harmful to health or the environment may be misplaced.)$_1$ (Although it is toxic to certain animals, evidence is lacking that it has any serious long-term effect on human beings.)
(اعتبر هذه المادة مضرة بالصحة او ان البيئة قد تتأثر بها)$_1$(مع انها سامة لبعض الحيوانات، لا يوجد دليل على ان لها أي تأثيرات طويلة المدى على الانسان)$_2$

Table 2: An example of the relation contrast

(Animals heal,)$_1$ (but trees compartmentalize.)$_2$ (They endure a lifetime of injury and infection by setting boundaries that resist the spread of the invading microorganisms.)$_3$
(الحيوانات تندمل)$_1$(لكن الاشجار تنقسم)$_2$(فهم يتحملون الجروح و الإصابات طوال حياتهم بوضع حدود تقاوم إنتشار غزو الميكروبات)$_3$

Table 3: A relation that connects clause with a word (preposition link)

I found my friend in the car	وجدت صديقي في السيارة

Table 4: A relation connecting 2 clauses

I found my friend in his brother’s home	وجدت صديقي في بيت أخيه

Table 5: An example of the relation recall

I will go to work, but I will not attend the meeting	سأذهب اليوم إلــى العمــل، و الاجتماع لن أحضرلكني

Because of the differences in the rhetoric, literature and relation concepts between the two languages, we started by studying the Arabic corpus to extract some Arabic rhetorical relations that reflect the essence of the Arabic texts.

Our approach to extract the Arabic rhetorical relations consists of three phases. These are:

- Studying the English relations
- Analyzing the Arabic corpus
- Understanding and using the Arabic cues

First, we extracted some of the Arabic relations from the English relations. The process consists of three steps as shown in (Fig. 1). We pick an English relation, then we scan the Arabic rhetoric and literature references[1,3,4,8] for this relation, we also scan the Arabic corpus collected from[1,3,4,8] to see if this relation is explicitly signaled. If so, the relation is added to the Arabic relations list; otherwise, the relation is ignored.

In the second phase, we looked into the Arabic rhetoric and literature references that have been written by Arabic language scholar for the relations that connect the Arabic clauses. Those relations fall into two categories:

- Connectors that connect clauses as well as words.
- Connectors that connect clauses only.

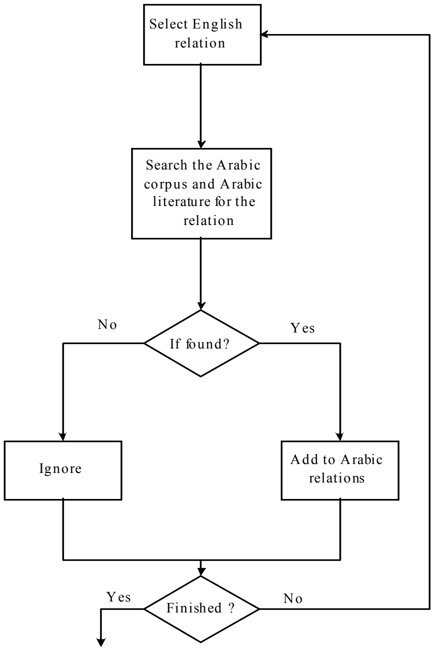


Fig. 1: The process of extracting Arabic relations from English relations.

Table 6: Examples of the cue phrases features

Arabic rhetorical relation	Relation name in English	Cue phrases that signal the relation
شرط	Condition	لو -إذا -إن
العطف	Joint	Determined by word co-occurrence If (co-occurrence < threshold) This relation exists
التفســـير	Interpretation	بمعـنى أن - أي - حيث
الإسـتدراك	Antithesis	لكن - غير أن - إلا أن
التعليــل	Justification	بسـبب - لأن
التوكيـد	Confirmation	بل
الترتيــب	Sequence	من ثم - ثم
النتيجـة	Result	نسـتنتج -نتيجـة لهـذا - نتيجـة لـذلك - لذا - إذن
التمثيــل	Example	علـى سـبيل المثـال - مثلا
القاعدة	Base	بما أن - بناءا علـى
التفصـــيل	Explanation	خاصةً -لاسيما

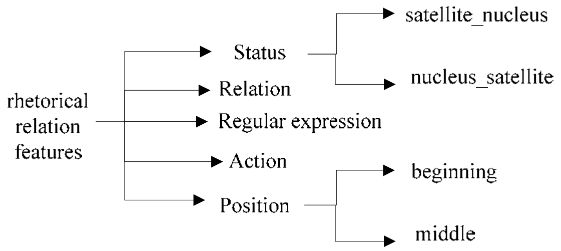


Fig 2: Features that define the rhetorical relations.

The following examples explain the two categories:

Example 3: The sentence in Table 3 includes the Arabic relation (جار و مجـرور /preposition link) that connects a clause with a word.

Whereas in the sentence in Table 4 includes the same relation, but in this case it connects two clauses.

Example 4: The following sentence (Table 5) includes a relation ا(اسـتدراك /Recalling) that connects two clauses only.

We select the relations that belong to the second category since we are targeting the relations between the clauses or sentences.

In the third phase, we scan the corpus to obtain the words that are considered connector in the Arabic language. These are known as cue phrases in the literature[3]. Next, we examine the relations that they signal. If a relation belongs to the second category, then we add it to the Arabic relations; otherwise it is ignored. Examples 3 and 4 illustrate this phase. The connectors (underlined) are extracted, then the relations they signal are studied. Consequently, we make the decision of including these relations or not.

A relation is signaled by a set of cue phrases[13,15,16]. We studied the Arabic rhetorical relations shown in Table 6 and observed the cue phrases that signal each one. The set of cues that we generated are used in the rhetorical parser of the Arabic text. Each cue phrase signals a rhetorical relation between two units based on some features. These features are extracted from the corpus analysis of the cue phrases. In the sequel, we present some of these features Fig. 2.

- Status: Specifies the rhetorical status of the units that are linked by the cue phrase. Its value must be either satellite_nucleus, or nucleus_satellite indicating that either the designated cue phrase connects two units where the first one is satellite and the second one is nucleus or the first one is nucleus and the second one is satellite
- Position: Specifies the position in the text where the cue phrase must be located. Its value is either beginning indicating that the designated cue phrase is located in the beginning of the statement, or middle indicating that the cue phrase is located in the middle of the statement
- Action: Specifies the action that this cue phrase has in determining the elementary units.
- Relation: Specifies the relation that this cue phrase signals
- Regular expression: Specifies the regular expression of the cue phrase.

RESULTS

Table 6 lists eleven new relations that rise in Arabic texts and that were obtained using our approach. Table 7 shows an example of the described features for the cue phrase) إن /if). As shown in the table, when the cue phrase comes in the Beginning (B) of the statement, it connects the unit that the cue phrase إن is located in as Satellite (S) with the one that comes next to it which is considered as Nucleus (N). The other case is when the cue phrase إن is located in the Middle (M) of the statement; it connects the unit that comes before the one that this cue phrase is located in as nucleus, with the unit where the cue is located as satellite.

Table 7: Examples of the Arabic cue phrases features

Regular Expression	Position	Relation	Status	Action
(\sإن)	B	شرط	S_N	Normal then comma
(\sإن\s)	M	شرط	N_S	Normal

Table 8: An example with cue phrase whose action is nothing

(The meeting of enhancing the company’s software will be deferred until the end of the vacation,)$_1$
(this is due to the absence of most of the employees.)$_2$
(سيتم تأجيل الاجتماع المخصص لمشروع تطوير برامج الشركة الى ان تنتهي الإجازة.)$_1$ (أى أنه بسبب عدم توفر العدد الكافي من الحضور تم تأجيل الأجتماع)$_2$

Table 9: A text unit boundary based on cue phrase with action nothing followed by a cue phrase that has an action

(A team of engineers will visit the company headquarter to evaluate the existing equipments)$_1$ (that is, if the team was formed in the appropriate time)$_2$
(سيقوم فريق من المهندسين بزيارة لمقر الشركة للتعرف على المعدات الموجودة لديها)،$_1$ (إذا تم تشكيل فريق في الوقت المحدد)$_2$

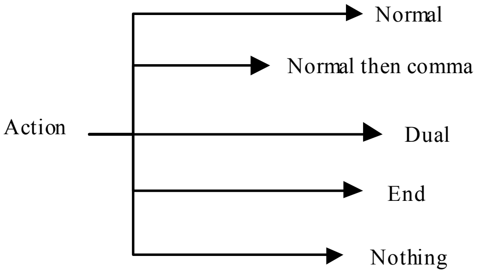


Fig 3: Actions used by the Arabic cue phrases

Before building the rhetorical relations, the text units that those relations are toiled for should be determined. Since the rhetorical relations are signaled by specific cue phrases, the determination of the text units is based on cue phrases too[13,15,16]. The cue phrases are considered as the units connectors. Each cue phrase has a specific action on determining the text units. The result is a set of actions that are associated with the cue phrases. We present the actions that are used by the Arabic cue phrases to determine the text units of a discourse Fig. 3

- Normal: Adds a unit boundary before the cue phrase
- Normal then comma: Adds a unit boundary before the cue phrase and another unit boundary after the first occurrence of a comma, a semicolon or an end of a statement. If a comma is met and it is followed by the cue phrases (و /and) or (أو /or) it adds a unit boundary after the next occurrence of the three markers
- Dual: Same as the action normal if the cue phrase is not preceded by another cue phrase, action normal then comma is applied otherwise
- End: Adds a unit boundary after the cue phrase
- Nothing: Bypass the cue phrase, this action is associated with a cue phrase that does not signal a rhetorical relation but helps other cues in determining the text units

In the case that a cue phrase is followed by another cue phrase, we consider the first cue phrase as the unit boundary determiner if its action is not nothing and the cue phrase that follows is by passed. The example of Table 8 illustrates this case:

In this example, two cue phrases (underlined) exist in unit$_2$, the first one (أي أنه /that is) signals the relation (تفســيـر /interpretation) and the second one (بسبب/ due to) signals the relation (تعليــل /justification). In such a case, the first cue phrase is considered as the unit boundary determiner. It will apply its action. Consequently, this cue phrase will signal the rhetorical relation for unit$_2$.

In the case that a cue phrase has the action nothing and it is followed by another cue phrase that has a different action, the second cue phrase will determine the text unit boundary based on its action. However, the unit determiner is put before the first cue phrase that has the action nothing. The example of Table 9 highlights this case.

In this example, two cue phrases (that is and if) exist in unit$_2$, the first one (هذا /that is) has the action nothing and doesn't signal a discourse relation. The second one(إذا /if) signals the relation condition. Consequently, the second cue phrase determines the rhetorical relation for unit$_2$. It also determines the text unit boundary according to its action. However, the unit determiner will deal with the first cue phrase so that unit boundary is added before it instead of the second one. The same idea is applied in case if a cue phrase is followed by more than one cue phrase. The Arabic relations mentioned in Table 6 differ in terms of the units that they connect. Some relations connect two units. These are called binary relation. In other cases, they connect one unit with one or more other units. These are called bulky relations.

We observed from the Arabic discourse that the relations (نتيجـــة /result), (تمثيــل /example) and (قاعدة /base) are bulky relations. The relation (نتيجـــة /result) normally connects one unit as a result of one or more units. Consider the example of Table 10. Note that unit$_1$ and unit$_2$ are parameters for the result mentioned in unit$_3$.

Table 10: An example of a bulky relation

(These days are the last days in the semester,)$_1$ (and we expect that many students will have good grades.)$_2$ (Consequently, the school is preparing a party to prize those students.)$_3$
(تعتبر هذه الايام الاخيرة للفصل الدراسي)$_1$ (و من المتوقع حصول الكثير من الطلاب
على نتائج جيدة هذا العام)$_2$(لذا تستعد هيئة التدريس هذه الايام الترتيب لحفل تكريم الطلاب
المتفوقين)$_3$

Table 11: A relation that connects one unit with fact units

(As known, fruits are essential source of vitamins that are necessary for human.)$_1$ (Doctors always recommend eating fruits.)$_2$ (The existence of this source over the year is one of the God’s merciful.)$_3$ (For example, you can find orange and apples in the shops in all over the year and they contain useful vitamins for the bodies.)$_4$
(من المعلوم ان الفواكه مصدر اساسي للفيتامينات الضرورية للجسم)$_1$(و دائما م
ينصح الاطباء لتناول الفواكه بشكل مستمر)$_2$(و من نعم الله على عباده ان هذا المصدر
متوفر طوال العام)$_3$(فعلى سبيل المثال نجد البرتقال و التفاح متوفران في الاسواق بشكل
مستمر وهي من الفواكه الغنية بالفيتامينات الضرورية للجسم)$_4$

Table 12: A relation that has several units built upon it

(Based on what HR manager recommended,)$_1$ (the company’s general manager decided to promote all the employees who have grade 5 and above,)$_2$ (and grant those employees certificates.)$_3$
(بناءا على ما رفعه مدير إدارة شئون الموظفين)$_1$(فإن الشركة قررت ترقية جميع
الموظفين الحاصلين على درجة تقييم 5 فما فوق)$_2$(و منح هؤلاء الموظفين شهادة تقدير
و تميز)$_3$

The relation (تمثيـ ـ ـل /example) normally connects one unit with one or more units that are facts. Consider the text fragment of Table 11.

In this example, unit$_4$ is related to all the units that came before it, namely, unit$_1$, unit$_2$ and unit$_3$. This is because unit$_4$ is an example of the facts in units 1, 2 and 3.

The relation) قاعـدة /base) normally connects one unit as a base that one or more units are built upon. The example of Table 12 illustrates the relation.

Note that unit$_1$ is related to both unit$_2$ and unit$_3$ as a base that the two units 2 and 3 are built upon.

The examples mentioned above showed that the three relations (نتيجــة /result), (تمثيـ ـ ـل /example) and (قاعدة /base) might connect one unit with several units and because of this we have to handle them depending on the position of the indicator. This takes place according to the following:

- If the cue phrase that signals the bulky relation comes in the first unit of the paragraph, this unit is related with all the subsequent units up to the end of the section
- If the cue phrase that signals the bulky relation comes in a unit in the middle or in the end of the paragraph, this unit is related to all the preceding units
- The other relations connect two units according to the position of the cue phrase

Text is principally coherent, which means that a unit in the middle of the text might be related to another unit in the beginning. In constructing the rhetorical relations based on cue phrases, the relation that will be built will relate adjacent units only, which will lead to a rhetorical representation that misses some important relations between units that are not adjacent. When dealing with building the rhetorical relations between the text units we recall what Marcu stated in[13] that An accurate determination of elementary units of a text and of the relations that hold among them is beyond the current state of the art in natural language processing.

To solve this problem, Marcu in[13,16] performed a corpus analysis of the English cue phrases and assigned a value called Maximal distance to each cue. This value holds the number of units found between the textual units that are involved in the rhetorical relation signaled by the designated cue phrase. The number is manually assigned to each cue by studying the maximum number of units that come between the participant units of the rhetorical relation in the corpus analysis of the cue phrases. For example, the cue phrase Although has been assigned the value 5 for its participant units in the relation elaboration. This means that the relation elaboration is hypothesized between the unit that contains the cue phrase and all the four units that come before (Maximal distance-1 as specified by Marcu[13,16]). This technique has some disadvantages including:

- Complexity in determining the cue phrases: To determine the cue phrases that signal the rhetorical relations, we have to investigate the occurrence of each cue in several texts to determine the value of the maximal distance
- Complexity associated with adding more cue phrases: The relations and cue phrases are open lists. It means that they are subject to future expansion. Each time a relation is added, it is corresponding cue phrases need to be investigated in terms of the maximal distance they have
- The semantic of the cue phrases depends on the context: The semantic of a certain cue phrase varies. A cue phrase in a certain context may connect certain number of units, whereas it may

connect different number of units in another context. Making each cue connects a fixed number of units for all the texts they appear in, does not reflect the fact that they depend on the context

To avoid such disadvantages, we followed a different approach. Our approach depends on the fact that the nuclearity relation among the text units is a transitive relation. This fact is based on (dilemma 1). In the sequel, we use the following notation: rhet_rel$_i$ (rel$_i$, u$_j$, u$_k$), where rhet_rel$_i$ defines the relation i that exists between the unit j and the unit k.

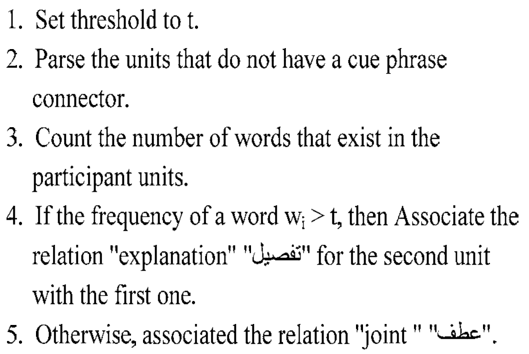
Dilemma 1: If there are three units u$_1$, u$_2$ and u$_3$ such that:
rhet_rel$_1$(rel$_1$, u$_2$, u$_1$) where u$_2$ is satellite and u$_1$ is nucleus,
and
rhet_rel$_2$(rel$_2$, u$_3$, u$_2$) where u$_3$ is satellite and u$_2$ is nucleus,
Then:
u$_1$ is more salient than u$_3$.

DISCUSSIONS

In our investigation and tedious analysis of Arabic corpus, we observed that in most cases, there is an implicit transitivity relation over the hypotactic Arabic rhetorical relations. Consider the following example (Table 13).

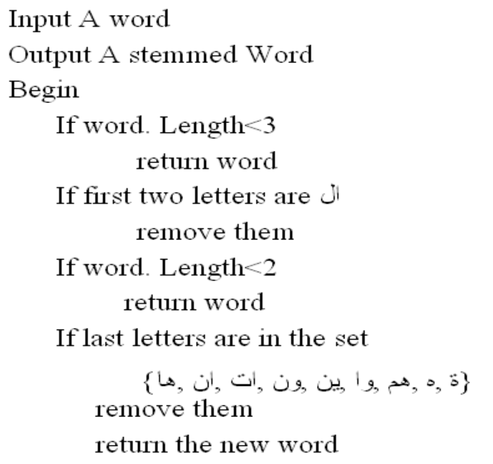
Table 13: An example of hypotactic Arabic rhetorical relation transitivity

(Khalid didn’t go shopping today ;)$_1$ (indeed, he did not get out of his home)$_2$ (because of the rainy weather.)$_3$ (لم يذهب خالد الى السوق هذا اليوم)$_1$(بل انه لم يخرج من البيت)$_2$(بسبب الامطار الغزيرة)$_3$



1. Set threshold to t.
2. Parse the units that do not have a cue phrase connector.
3. Count the number of words that exist in the participant units.
4. If the frequency of a word $w_i > t$, then Associate the relation "explanation" "تفصيل" for the second unit with the first one.
5. Otherwise, associated the relation "joint " "عطف".

Fig 4: Steps of the word co-occurrences technique.



```
Input A word
Output A stemmed Word
Begin
    If word. Length<3
          return word
    If first two letters are ال
          remove them
    If word. Length<2
          return word
    If last letters are in the set
          {ة, ه, هم, وا, ين, ون, ات, ان, ها}
       remove them
       return the new word
```

Fig 5: An Arabic word stemming algorithm.

In the above example, unit$_2$ has the relation) توكيـــد /confirmation) with unit$_1$ and unit$_3$ has the relation) تعليـــل /justification) with unit$_2$. Further, unit$_3$ has the relation تعليـ ـل /justification) with unit$_1$.

Therefore, we hypothesized that the hypotactic Arabic rhetorical relations are transitive. Accordingly, we apply the following rules when rhetorically parse the Arabic text:

- If rhet_rel$_1$(rel$_1$, s$_1$, n$_1$) and rhet_rel$_2$(rel$_2$, s$_2$, n$_2$) such that n$_2$ = s$_1$ → new relations rhet_rel$_3$ (rel$_2$, s$_2$, n$_1$)
- If rhet_rel1(rel1, s1, n1) and rhet_rel2(rel2, s2, n2) such that n1 = s2→ new relations rhet_rel3 (rel1, s1, n2)

Since paratactic relations have both of their spans as nucleus, transitivity is not applied on them. Further, it should be noted that in some situations, there is no cue phrase connector between two units; in this case the relation is hypothesized between these units using a technique called word co-occurrences. It is summarized in Fig. 4. The parser parses the units that do not have a cue phrase connector and count the number of words that exist in the participant units; if the number of a certain word existence exceeds a certain threshold, the second unit is consider to have the relation (تفصـــيل /explanation) with the first one. If the number of the co-occurred words does not exceed the threshold, the relation (عطف /joint) is hypothesized between the units.

Note that when counting the co-occurred words, the words are not taken as they are since two similar words might have different morphemes. Thus the words are stemmed and then compared. The problem is that in Arabic language we lack an accurate stemming algorithm, but we developed a simple stemming algorithm that is suitable for word comparison but it is not fully accurate. The algorithm is explained in (Fig. 5). When the stemmed words are compared, the comparison algorithm checks the two words: If they have similar three or more letters, it considers them as similar words; otherwise they are considered not similar.

CONCLUSION

We showed a framework of applying RST on Arabic language in order to rhetorically parse and understand the Arabic texts. We discussed an approach to extract the Arabic rhetorical relations that is based on studying the English rhetorical relations, analyzing Arabic corpus and realizing the rhetorical impact of the Arabic cue phrases to facilitate and understand given Arabic texts. The Arabic corpus was collected from famous Arabic literature[1,3,4]. There are several applications to this work including: Text summarization and generation, information retrieval and text segmentation. We introduced an approach to deal with text coherency that suits the characteristics of the Arabic language and avoids the disadvantages of previous approaches. We are in the process of implementing the findings of this work to build a system to automate Arabic text summarization.

ACKNOWLEDGEMENT

This study is partially supported by the research center of the college of computer and information sciences in King Saud University, Riyadh, Saudi Arabia.

REFERENCES

1. Abdulmuttalib, H.M., 2003. Al-Nahr Al-Muiassar. Dar Al-Aafag Al-Arabiah, Cairo, Egypt.
2. Agichtein, E. and V. Ganti, 2004. Mining reference tables for automatic text segmentation. In the Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining Seattle, Aug. 2004, Washington, USA., pp: 20-29. http://doi.acm.org/10.1145/1014052.1014058.
3. Al-Ansari, I.H., 2003. Mugni Al-Labeeb an Kutub Al-Aareeb. 1st Edn., Al-Maktabah Al-Asriah for Publishing and Printing, Al-Mansourah, Egypt.
4. Aubadah, M.I., 1983. Al-Jumlah Al-Arabiah. 1st Edn., Munshat Al-Ma'aref, Alexandria, Egypt.
5. Chang, D.S. and K.S. Choi, 2005. Causal relation extraction using cue phrase and lexical pair probabilities. Lecture Notes in Comput. Sci., 3248: 61-70. DOI: 10.1007/b105612.
6. Chang, D.S. and K.S. Choi, 2006. Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities. Inform. Proces. Manage., 42 : 662-678. http://portal.acm.org/citation.cfm?id=1131995
7. Cristea, D., O. Postolache and L. Pistol, 2005. Summarisation through discourse structure. Lecture Notes in Comput. Sci., 3406: 632-644. http://springerlink.com/content/3ffaxv08lyl53w72/.
8. Gabawah, F., 1972. Iraab Al-Jumal wa Ashbah Al-Jumal. Dar Al-Qalam Al-Arabi, Damscuss, Syria (in Arabic). www.qalamarabi.com/About_dar.php
9. Golcher, F., 2006. Statistical text segmentation with partial structure analysis. Proceedings of 8th Conference on Natural Language Processing, Oct. 2006, Konstanz, Denmark, pp. 44-51. http://ling.uni-konstanz.de/pages/conferences/konvens06/konvens_files/abstracts/golcher.pdf .
10. Lamprier, S., T. Amghar, B. Levrat and F. Saubion, 2007. SegGen: A genetic algorithm for linear text segmentation. Proceeding of the 20th International Joint Conference on Artificial Intelligence, Jan. 2007, Hyderabad, India, pp: 1647-1653. http:// www.ijcai.org/papers07/Papers/IJCAI07-266.pdf
11. Mann, W.C. and S.A. Thompson, 1988. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8: 243-281. www.di.uniba.it/intint/people/fior_file/INTINT05/RST.pdf.
12. Mann, W.C., C.M. Matthiessen and S.A. Thompson, 1992. Rhetorical Structure Theory and Text Analysis. In: Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text, Mann W.C. and S.A. Thompson, (Eds.). John Benjamins Pub Co., Amsterdam/Philadelphia, pp: 406. ISBN-10: 1556192827.
13. Daniel Marcu, 1997. The Rhetorical Parsing Summarization and Generation of Natural Language Texts. Ph.D Thesis, Department of Computer Science, University of Toronto. http://www.int.gu.edu.au/kvo/reading/marcuphd.ps.gz.

14. Marcu, D., 1999. Discourse Trees are Good Indicator of Importance in Text. In: Advances in Automatic Text Summarization, Mani, I. and M. Maybury, (Eds.). The MIT Press, Cambridge, MA, pp: 434. ISBN: 0262133598.
15. Marcu, D., 1997. From discourse structure to text summaries. The Proceeding of the ACL'97/EACL'97 Workshop on Intelligence Scalable Text Summarization, July 11-11, Madrid, Spain, pp: 82-88. http://acl.ldc.upenn.edu/W/W97/W97-0713.pdf.
16. Marcu, D., 2000. The Theory and Practice of Discourse Parsing and Summarization. 1st Edn., The MIT press, Cambridge, MA., pp: 268. ISBN-10: 0262133725.
17. Mathkour, H., A. Touir and W. Al-Sanie, 2005. Automatic information classifier using rhetorical structure theory. Adv. Soft Comput., 31: 229-236. DOI: 10.1007/3-540-32392-9_24
18. Yang, C.C. and K.W. LI, 2005. A heuristic method based on a statistical approach for chinese text segmentation. J. Am. Soc. Inform. Sci. Tech., 13: 1438-1447.DOI:10.1002/asi.20237.