A Novel Rhetorical Structure Approach for Classifying Arabic Security Documents

Hassan I. Mathkour, *Member, IAENG*

***Abstract*—Security Documents classification is aimed at securing documents from being illegally disclosed. Classifying a portion of a document as a 'secret' depends on the type of effect its disclosure will have in an organization. In this respect, Information is classified according to their critical semantic (i.e. its context or value and intended uses or audience at particular time or situation). Understanding the semantic of a document is not an easy task. The rhetorical structure theory (RST) is one of the leading theories that have been applied successfully in text processing and understanding. In this paper, we will describe a novel approach to automatically classify Arabic Security documents using RST.**

***Index Terms*—RST, Information classification, Arabic security documents, automatic document classification.**

I. INTRODUCTION

Profitability of organizations is ultimately dependent on the effectiveness with which they exchange, process, control, manage and, more importantly, protect their data and information so that only authorized persons access them. All these processes require that the right information be made available to the authorized persons at the right place and at the right time [1]. To this end, one needs to provide a well-defined methodology for specifying the criteria that govern the security classification. In addition, it necessary to re-evaluate the value and importance of the information being classified to reassign the security as this changes with time depending on the organization. This will require defining appropriate procedures and protection requirements for the security reassignment. Not all documents are of the same importance and so not all information in those documents requires the same degree of protection. This implies that different portions of information are assigned different security classifications. The assignment follows different schemes depending on the nature of the organization.

The first step in information classification is to identify a senior member of management as the authorized person of the particular information to be classified. The second step is to develop a classification policy. The policy should describe the different classification labels, define the criteria for assignina particular label to information, and then list the required security controls for each classification [1].

Volubility of the information in the document and their timeliness are among the factors that influence the assignment of document security classification. Laws, national security issues and other regulatory requirements are important considerations when classifying documents. Furthermore, in classifying documents, the security requirements of specific categories of documents, the various processing stages of documents, such as draft and final, and the contents and structure of documents must be clearly articulated and specified. A review of the classification of particular information should be done periodically to ensure that its classification is still appropriate, and also to ensure the security controls required by the classification are in place [2][3][4].

Documents are classified according to their importance [5], or their potential effect on a specific organization. Classified information is sensitive information to which access by particular classes of people is restricted by either law or regulation. A formal security clearance is required to handle classified documents or access classified data [3][4].

The common information security classification labels that are used by the business sector are public, sensitive, private, and confidential. Government and its agencies adopt classification labels such as Classified, Unclassified, Sensitive but Unclassified, Restricted, Confidential, Secret, Top Secret and their non-English equivalents [6][4].

Top secret- The compromise of this information or material would likely cause a tremendous effect on the organization.

Secret- The compromise of this information or material would likely cause a big effect (but less than the above class) on the organization and should not be disclosed by anyone.

Confidential- Indicates that the information is private information which should not be disclosed by a person other than to the intended one.

The term "unclassified" was not formally a classification but could be used to indicate positively that information or material did not carry a classification.

Companies have documents that carry private information. They need to label them with one of the labels described above to inform the recipient about their importance. Classifying the information in a certain document depends mainly on its semantic (i.e. its context or value and intended uses or audience at particular time or situation). An example is that part of the document that contains the salaries of the executive board members. In some cases, this information

Manuscript received March 30, 2009. This work was supported in part by the the research center of the college of computer and information sciences in King Saud University.

H. I. Mathkour is with the Department of Computer Science, King Saud University, Riyadh, 11653 Saudi Arabia, phone: 966-467-0797; fax: 966-467-5423。

may be considered as secret in a document of the financial analysis of the company. This same information may be considered as unclassified if mentioned in the context of the richest people in the city [2][3][4].

It is most likely that the classification is done manually by some people who follow certain guidelines. This process may take long time when the document is huge –hundreds of pages. However, performing the classification automatically would make the process faster and easier. The idea of automating the process has the challenge of understanding the semantic of the information. In this paper, we propose a novel approach that is intended to address this challenge.

An automatic security document classifier system should involve pre-processing, text understanding, comprehending multi-level security grades followed by the application stages. The system that we propose is through an approach that realizes the degree of importance of given text via a rhetorical tree. It then tags the respective sections with the appropriate security level according to the security standards (policies set by organization) that is defined priori. For the purpose of realizing the importance of portions of documents, we make use of the rhetorical structure theory (RST) [7][8]. In [9] we introduce the idea of using RST in document classification. In here, we elaborate the concept, discuss the system, and report on test cases.

The remaining of the paper is organized as follow: Section 2 highlights related work. Section 3 introduces the concept of the proposed technique and gives an overall description of the system. Section 4 discusses an experimental study that was conducted on the system and the results obtained. The experiment was conducted on Arabic texts. Section 5 concludes the paper and highlight future work.

II. Related Work

Little work and limited research has been made in the area of security classification of documents. The most updated research [10] incorporates section level security classification of documents only, a top-bottom approach. Damiani et al. [11] proposed an access control model for XML documents. The model is confined to DTD (Data type definition) level only. In this model, each DTD is associated with particular information that is contained in the document, and this is used to decide which part can and cannot be accessed by user.

Information Security Management Tool, developed by the University of Auckland (New Zeeland) for the security protection of the local University official documents, fragments documents into classes with certain implementation of security levels ranging from top to bottom). The NIST, National Institute of standards and Technology (Under Secretary of Commerce for Technology, United States), Tool was developed by NIST for the information protection in documents trafficking between their mother institute and its child institutes. Microsoft Office Document Classifier [12] classifies and labels Office documents, using document markings that identify the existence of confidential and private information. This tool helps in stopping information leakages of hidden content in Microsoft Word documents, and encourages proper handling of sensitive information. Titus Labs Document Classification (Titus Labs Document Classification for Microsoft Office) [13] is a classification tool for Microsoft Office that allows government and military customers to manage the classification, distribution and retention of valuable corporate documents. Users are constrained to select from a dropdown menu the appropriate classification labels for their documents, spreadsheets, or presentations. ArticSof (Cryptographic tool) [14] is used for email classification, encryption and digital signature. SECLORE (Document’s right management tool) [15] uses dynamic rights for distributed document usage control.

For the purpose of realizing the importance of portions of documents, we make use of the rhetorical structure theory (RST). RST was originally developed as part of studies of computer-based text generation [7]. It has been developed to serve as a discourse structure in the computational linguistic field. RST gives the coherence in text [16]. It is intended to describe texts, rather than the process of creating or reading and understanding them. It uses a set of rhetorical relations that associate spans of text in an attempt to identify the importance of the portions. The rhetorical relations can be described functionally in terms of the writer purposes and the writer assumptions about the intended reader. These rhetorical relations hold between two adjacent spans of texts (although there are some exceptions).

The output of applying the rhetorical structure theory to a text is a tree structure that organizes the text based on the rhetorical relations [16]. Each relation connecting two spans of a text may be of two cases. In the first case, the relation connects two spans where the first is identified as a nucleus, representing the semantic of the two spans (i.e., this is more important to the reader than the other span), and the second span is called satellite. In the second case, the two spans have the same importance to the reader. In this latter case, the relation is called multinuclear relation where both participating 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 the rhetorical relations that hold among these units are also determined to connect the two spans. Determining the potential relations that connects the two spans could be done using several techniques; one of which is determining the rhetorical relations through the cue phrases [17]. Marcu [18] has given cue phrases that can be used in the English language processing. The process of building the rhetorical structure may lead to more than one structure, consequent upon the nature of the natural language text that stipulates that more than one relation could be assumed to connect two spans. At the end, the emergent structures are most likely closed, but in some cases; they may lead to ambiguity.

We selected RST as the basis for our technique as it allows classifying texts on the basis of the importance of its constituents which serves the purpose of assigning security labels to parts of a give document that is available electronically. There are other text classification techniques that have explored in other applications such as in [19][20][21].

III. The Proposed Approach

The rhetorical structure that is built from the rhetorical analysis process is represented as a binary tree that connects the two spans of a text [16]. Each node of the tree has a status (refered to as status information) which has the value of nucleus, satellite, a promotion which represents the most important text unit in this sub-tree, or the type, which represents the relation that connects the two spans [16]. The rhetorical structure tree could be built using the algorithm proposed in [18]. Marcu [22][17] argues that the promotion of the root of the whole tree gives the reader the most important unit(s) in the text. With this in mind, the classifier uses the status information to tag a certain portion of the document with proper security classification such as secret, confidential, etc. That is, the classifier uses the promotion to determine if the document is about some information that belongs to the specific classification. This information could be easily stated as a secret, and so on. It is possible that the promotion could be only one unit; therefore, the classifier may go some levels down if it wants to cover more units. The level the classifier should go depends on the importance of the information.

The proposed technique that could be used to classify the different parts of a certain document parses each paragraph in the document and builds the rhetorical tree that represents its structure. Then it determines what each paragraph is about by examining the promotion of the root of the tree. It uses the promotion to determine if the importance of the paragraph conforms to the user instructions. In such a case, the classifier labels the paragraph with the required classification. The technique could be adopted to work on pages or sections rather than paragraphs. One merely needs to know which text unit the technique could be applied on. The technique can also be adopted to go deeper to lower levels in the rhetorical tree to determine to which class this part of the text belongs. However, the main idea is to extract the promotion of the text unit. In cases such as in very sensitive documents, the user may desire to intervene in tagging portions of the document with the proper security class. In such cases, the classifier is able to determine the important parts of the text so that the user supplies the proper tags.

IV. An Overview of the Classifier System

Figures 1 and 2 depict details of the proposed system that is based on RST. The working of the system follows the following steps:

- RST Processor takes as input the text and builds its corresponding RS Tree.

- Tree Level Selector traverses the RST level by level starting from the root. Whenever a sentence (a text unit) is requested by the inference engine, the tree level selector picks the sentence from the current level down to the leave. As the tree is structured according to the most important sentence (nucleus-based), the probability is greater that a best fit of the classification of the text is at the higher level of the tree.

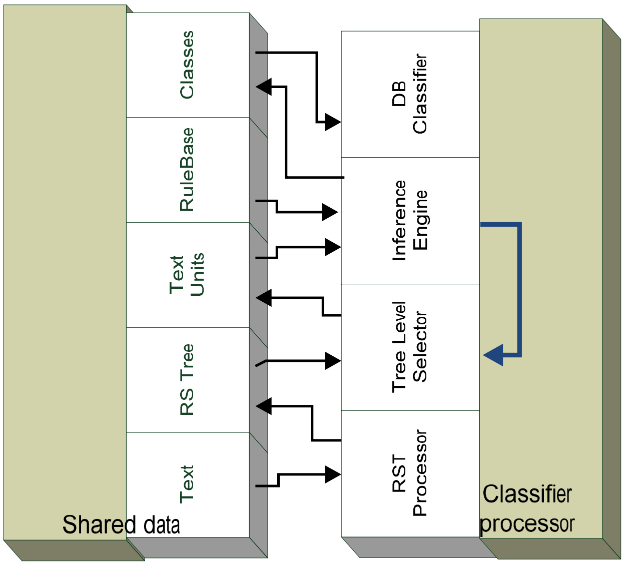
- Inference Engine processes the text unit based on the associate rules that are specific to the document type and decides whether the text can be classified at this stage or it requires the process of more text units. A looping communication with the Tree level Selector is established for this purpose.

- DB Classifier takes the result of the classification from the Inference Engine and update the database accordingly.

- RST Processor gets the relations, builds all the valid RS-trees, and then notifies the RS-tree Selector that the RS-trees are ready.

- RuleBase: This is a domain specific knowledge base that reflects the rules of the organization for the security classifications of their sensitive documents. It guided the classifier in tagging the proper security class in the absence of the intervention of the users.

- RS-trees Selector selects the most suitable RS-tree for Arabic text summarization.



•

Figure 1: An overview of the classifier system

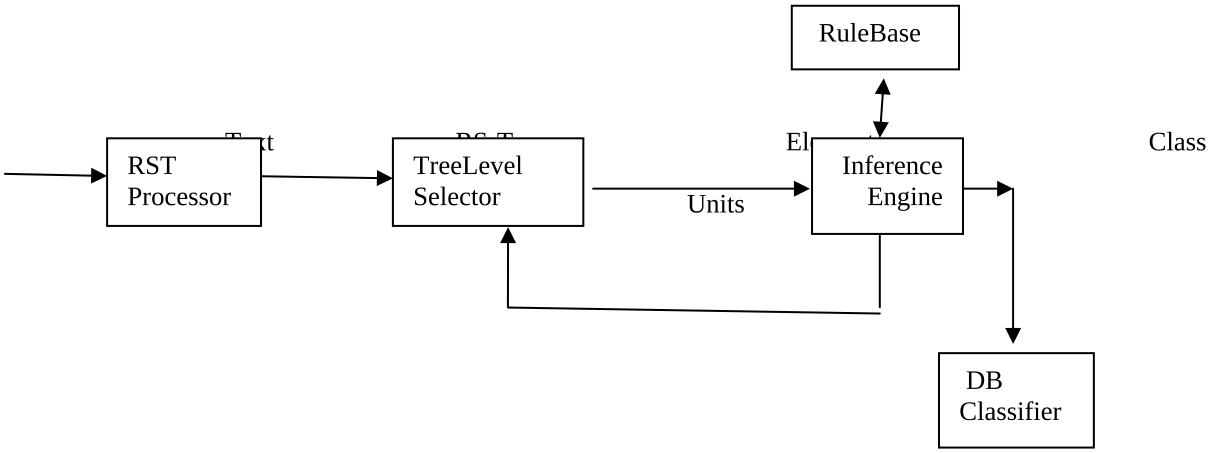


Figure 2: Interactions among the main components of the classifier system

V. Using the Cue Phrases in the Classifier System

We use the concept of cue phrases [16] to identify the rhetorical relations. The cue phrases were identified for the Arabic text based on the analysis of large Arabic corpus. Part of the cue phrases and the relations they determine are listed in Table 1. The listed cue phrases are only those that are related to texts used in the experiment for the security classifier of documents. Thus, they constitute only a partial list of the cues that were obtained from the corpus analysis.

Table 1: A Partial list of Cue Phrases.

Cue Phrase		Relation
1	و	Joint
2	حيث	Elaboration
3	لذا	Evidence
4	بالرغم	Concession
5	لكن	Contrast
6	على أن	Condition
7	بما أن	Background

[ستتم دراسة جميع المشاريع المطروحة]1 [بما أن العمل في المشروع السابق قد أوشك على الإنتهاء.]2

In the two examples above, the nucleus and satellite positions differ. In the first example, span (1) is the satellite and span (2) in the nucleus. In the second example, the nucleus and satellite exchange their positions –span (1) is the nucleus and span (2) is the satellite. Therefore, the relations that represent the two examples are:

rhet_rel (Background, 1, 2)
rhet_rel (Background, 2, 1)

The following two tables show the cue phrases in the two cases, and the satellite and nucleus positions in each case.

Cue Phrase	First Span	Second Span
و	Nucleus	Nucleus
حيث	Nucleus	Satellite
لذا	Satellite	Nucleus
بالرغم	Nucleus	Satellite
لكن	Nucleus	Nucleus
على أن	Nucleus	Satellite
بما أن	Nucleus	Satellite

Table 2: The two spans of the cue phrases appeared in the middle of the paragraph

Some cue phrases link two spans without respecting the positions these cue phrases appear in. Cue phrases 1 and 5 (Table 1) trigger a multinuclear relation in which both participating spans are nucleus. These two cue phrases always join the span that comes before them and the span that comes after them. The other cue phrases have two cases. They may join the span that comes before them and the span that comes after them, or they may join the span that comes after them (span 1) and the span that comes after span1 –the next two adjacent spans. This, however, depends on the token that comes before the cue phrase. If the token is a dot ("."), a new line ("\n"), or nothing (i.e. the first sentence in the document), the second case is applied; otherwise the first case is applied. The following example explains the two cases with the cue phrase (ĭŮ) (in English since):

[ĭŮ الشركة وقعت العقد،]1 [فإن العمل في المشروع سيبدأ بعد شهر من الآن]2.

In the above example, the second case of linking the two spans is applied. In this case, the two consecutive spans that come after the cue phrase are linked. Now consider an example of the first case –in which the same cue phrase comes between the two spans it links.

ABLE 3: CUE PHRASES THAT CAN COME AT THE BEGINNING OF THE PARAGRAPH.

Cue Phrase	First Span	Second Span
حيث	Satellite	Nucleus
بالرغم	Satellite	Nucleus
على أن	Satellite	Nucleus
بما أن	Satellite	Nucleus

In a case that a sentence is not joined to any other sentence via a cue phrase, this sentence is joined to the one before it through the relation:

Rhet_rel (Joint, Second_sentence, First_sentence)

For example, in the following text, the parser will consider the second sentence as a case in which the author joins second sentence to the first one.

[التقـاريـر الماليـة الـصادرة مـن الإدارة الماليـة تفيـد بزيـادة المبيعات في هذه السنـة.]1 [هذه الزيادة ناتجة عن الإنتعاش الإقتصادي في السوق المحلية]2

VI. EXPERIMENTAL RESULTS

We have performed experiments on Arabic texts to demonstrate the working of classifier. The document units that we applied the classifiers on are the paragraphs. The technique examines each paragraph, determines what it is about, and then classifies it.

The following is an example of how the classifier works. As mentioned above, the classifier works on each paragraph and classifies it. Consider the paragraph below, which describes a company previous contracts and its intent to contract with an expert, and assume that the company considers such information to be classified as secret.

[وقعت الشركة في هذه السنة عدة عقود مع بعض شركات التقنية لتزويدها بالتجهيزات اللازمة.]1 [بالرغم من الجهود المبذولة من قبل إدارة التدريب،]2 [إلا أن موظفي الشركة لم يصلوا بعد للمستوى المأمول للتعامل مع هذه التقنية،]3 [لكن الموظفين لديهم الدافع للتعلم.]4 [لذا ستقوم الشركة بالتعاقد مع خبير تقني ليستفيد منه الموظفون]5 [بما أن تكلفة التعاقد أقل من تكلفة الدورات التدريبية،]6 [على أن لا تزيد مدة التعاقد على خمس سنوات]7 [و تشمل إقامة بعض الندوات العلمية،]8 [حيث تثري هذه الندوات معلومات الموظفين.]9

According to the description given about the cue phrases used, the following relations hold in the above paragraph:

rhet_rel (Joint, 2, 1)
rhet_rel (Concession, 2, 3)
rhet_rel (Contrast, 4, 3)
rhet_rel (Evidence, 4, 5)

rhet_rel (Background, 6, 5)
rhet_rel (Evidence, 7, 6)
rhet_rel (Joint, 8, 7)
rhet_rel (Elaboration, 9, 8)

Using the algorithm mentioned in [18], the classifier will build the RST that represents this paragraph. Figure 1 shows a generated tree that the classifier will use to classify this paragraph. The classifier will examine the promotion of the root node to find that sentence (5) is the most important unit in this paragraph. It then uses this fact as a basis to determine its security class and depending on the information obtained from the knowledge base. Since in our example any information regarding the company contracts are to labeled as secret, this paragraph will be classified as secret since the promotion indicates that the company is going to make a contract with an expert to train its employees, which is:

لذا ستقوم الشركة بالتعاقد مع خبير تقني ليستفيد منه الموظفون

If the user configures the classifier to classify this information under another class it will do so; otherwise it will be left unclassified.

VII. CONCLUSION

We have given an introduction to the information classification and its usage. We have shown its importance to ensure the privacy of an organization. The process of classifying the information might be slow and take too much effort when the document is huge. Developing an automated process of classification will definitely lead to faster classification of the documents. The classification process depends on the semantic of the text rather than the syntax; therefore, the technique that should be used to classify the information automatically should use its semantic. The rhetorical structure theory (RST) is one of the techniques that try to extract the semantic of the text. We have proposed a classification technique that uses this theory to extract the semantic of the text and then classifies the text. We have performed a experiments on certain Arabic texts, and has produced an optimistic result.

Currently, we are investigating the units of a document (paragraph, page, section) that could be used to build the rhetorical tree and then classify it. The depth level that the classifier goes for in looking for the promotions is a potential subject of future research. The research should also state the language the technique is being applied to, since the nature of one language may differ from another one. Techniques applied on certain language might not be applicable (or difficult to apply) in another language. We are also investigating the possibility of augmenting the RST techniques with text classification techniques such as statistical and lexical cohesion techniques to attain an improved outcome.

REFERENCES

[1] M. Gupta, Security Classification for Documents, Computers and Security, Volume 15, Number 1, 1996, pp. 55-71.

[2] J. Kajava, Information Security Standards and Global Business, Industrial Technology, 2006. ICIT 2006, IEEE International Conference on 15-17 Dec. 2006 pp. 2091 – 2095.

[3] W. H. Baker, Is Information Security Under Control?: Investigating Quality in Information Security Management, Security & Privacy, IEEE, Volume 5, Issue 1, Jan.-Feb. 2007, pp.36 – 44.

[4] Thomas R. Peltier, Information Security Policies, Procedures, and Standards: guidelines for effective information security management, Auerbach publications, Boca Raton, FL 2002.

[5] W. Nicolls, Implementing Company Classification Policy with the S/MIME Security Label, The Internet Society, 2002..

[6] R. W. Maule, Enterprise knowledge security architecture for military experimentation, IEEE International Conference on Systems, Man and Cybernetics, Oct. 2005, Volume 4, pp.3409 – 3414.

[7] William C. Mann, Sandra A.Thompson, Rhetorical structure theory: Toward a functional theory of text organization", Text, Vol. 8, No.3, pp.243-281, 1988.

[8] Bill Mann, An introduction to rhetorical structure theory (RST), http://www.sil.org/~mannb/rst/rintro99.htm, 1999.

[9] Mathkour, H., A. Touir and W. Al-Sanie, Automatic information classifier using rhetorical structure theory. Adv. Soft Comput., vol. 31, 2006, pp. 229-236.

[10] M. Alhammouri, and S. Muftic, A Design of an Access Control Model for Multilevel-Security Documents, 10th International Conference on Advanced Communication Technology (ICACT 2008), Feb. 2008, Volume 2, 17-20.

[11] Ernesto Damiani, Sabrina de Capitani di Vimercati, Stefano Paraboschi, and Pierangela Samarati, Design and implementation of an access processor for xml documents, The International Journal of Computer and Telecommunications Networking, Vol. 33, June 2000, pp. 59-75.

[12] http://www.re-soft.com/product/titus-classification-office.htm

[13] http://www.titus-labs.com/software/DocClass_default.html

[14] http://www.articsoft.com/

[15] http://www.seclore.com/

[16] Daniel Marcu, "Discourse trees are good indicator of importance in text", Advances in Automatic Text Summarization, The MIT Press, 1999, pp 123-136.

[17] Daniel Marcu, From discourse structure to text summaries, The Proceeding of the ACL'97/EACL'97 Workshop on Intelligence Scalable Text Summarization, Madrid, Spain, 11 July 1997, pp 82-88.

[18] Daniel Marcu, The theory and practice of discourse parsing and summarization, The MIT press, UK, 2000.

[19] Rıdvan Saraçoğlu, Kemal Tütüncü, Novruz Allahverdi, A new approach on search for similar documents with multiple categories using fuzzy clustering, Expert Systems with Applications, Volume 34, Issue 4, May 2008, pp. 2545-2554.

[20] Songbo Tan and Jin Zhang, An empirical study of sentiment analysis for Chinese documents, Expert Systems with Applications, Volume 34, Issue 4, May 2008, Pages 2622-2629.

[21] Rey-Long Liu, Interactive high-quality text classification, Information Processing & Management, Volume 44, Issue 3, May 2008, pp. 1062-1075.

[22] Daniel Marcu, Building Up Rhetorical Structure Trees, The Proceeding of the Flexible Hypertext Workshop of the Eighth ACM International Hypertext Conference, 1997.

Hassan I. Mathkour obtained his M.Sc. and Ph.D. in computer science from the University of Iowa, USA. Currently, he is the head of the department of computer science in King Saud University. He has published several articles in journals and conferences

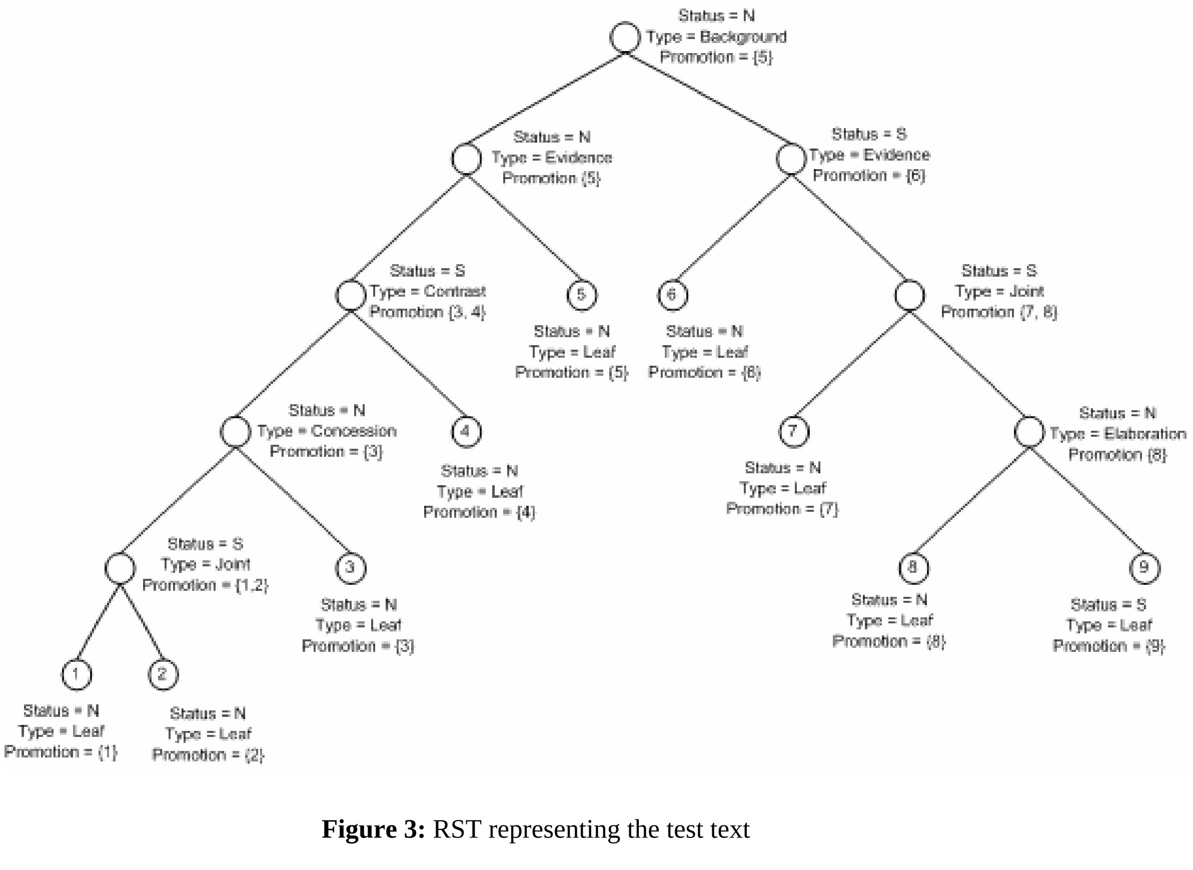


Figure 3: RST representing the test text