IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 5, 2020, accepted May 17, 2020, date of publication May 26, 2020, date of current version June 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997881

CAST: A Cross-Article Structure Theory for Multi-Article Summarization

NOUF IBRAHIM ALTMAMI AND MOHAMED EL BACHIR MENAI
Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
Corresponding author: Nouf Ibrahim Altmami (naltmami@su.edu.sa)

This work was supported by the Research Center of the College of Computer and Information Sciences, King Saud University.

ABSTRACT Over the last decade, discourse relations, also referred to as rhetorical or coherence relations, have been used to improve a range of natural language processing applications. Researchers have devised several theories, including rhetorical structure theory and cross-document structure theory, to examine relations between generic text units in single and multiple documents, respectively. In this paper, we propose a cross-article structure theory (CAST), that extends the benefit of discourse relations to multi-scientific article applications. It is based on the rhetorical structure theory (RST) and the cross-document structure theory (CST). The insight that underpins CAST is to consider both intra-section and cross-section relations. At the outset, these relations are classified based on the structural features of the article (that is, their appearance within each section type) and then the relations between text portions across multiple articles are classified. The practicality of the theory is showcased by solving a problem that consists to identify the types of relations which exist between each pair of sentences in related sections of different articles. A CAST bank was created and the k-nearest neighbors algorithm was used to develop two classifiers based on CAST and CST, respectively. The performance results obtained markedly demonstrate the role of the specific relations to scientific articles in CAST. Other applications of CAST could address the redundancy and readability problems, which represent main issues for different tasks, such as the summarization of multiple articles.

INDEX TERMS Cross-document structure theory, discourse relations, multi-article summarization, rhetorical structure theory.

I. INTRODUCTION

As rich and reliable sources of information, scientific articles play an essential role in various fields. Currently, numerous scientific articles are published online daily, and therefore, it is sometimes a complex matter for researchers to identify specific articles of interest. Even when query-based searches, field restrictions, and other advanced search techniques are used, the number of matching articles that are retrieved may exceed human processing capabilities. Therefore, to improve the information retrieval process and to aid in promoting high-quality, efficient, and effective research, it is worthwhile to understand how text portions within or between articles relate to one another. These relations could be used in weighting sentences or even articles by classifying text into important and less-important text. Additionally, they could help in ranking articles by navigating through the text's structure and then retrieving the most related articles. Moreover, we can use this information to avoid retrieving redundant information and thus facilitate the processing of multi-document phenomena [1], [2]. Rhetorical structure theory (RST), initially developed by Mann and Thompson [3], can be used to examine the connections that hold between portions of text within the same document, whereas Radev's cross-document structure theory (CST) [4] permits the same for multiple documents. Both RST and CST can be employed with generic text, and each one highlights contrasting types of relations that exist between text spans. Both are also associated with varying methods for the identification of relations. Additionally, RST and CST offer diverse systems the capability to identify significant content in text spans, and by examining relations between text fragments, the theories can detect areas that contain similar content. Examples of relations from RST and CST are presented as following.

A so-called '*Condition*' relation, as classified by RST, occurs between two sentences when one contains

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

a conditional statement that is essential to the occurrence or appearance of another sentence. In the following example, a condition relation exists between S_1 and S_2, where the latter specifies a condition that is essential to the appearance of the former.

S_1: *I will pass that exam.*

S_2: *If I study hard.*

Another type of relation, namely, an '*Equivalence*' relation proposed by CST, occurs when two sentences contain exactly the same information but are composed using a different syntax, grammar, or group of lexical elements. For example, an equivalence relation holds between S_3 and S_4, in which the two sentences are from different sources.

S_3: *The witness testified that he did not see the accused man commit the crime.*

S_4: *The witness testified that the accused man was not seen by him near the place where the crime was committed.*

The purpose of the present paper is to combine CST and RST for use with scientific articles, as opposed to generic texts, and to devise cross-article structure theory (CAST). An important point to note, therefore, is that various structural similarities exist across almost all scientific articles. Specifically, most scientific articles begin with an abstract, followed by an introduction, a literature review, a discussion of the methodological aspects of the study, a description of the experiments, the results, and the findings. Scientific articles typically close with a discussion of the findings and concluding remarks [5]. Every structural element within a scientific article, ranging from the abstract to the conclusion, is associated with specific characteristics and types of information.

The nature of language, and in fact, its power derives in large part from the intelligible relations that exist within and between sentences and their lexical elements, whether these sentences or elements are adjacent or not [6]. When one's task is to summarize a text or a portion of it, discovering the relations that exist between any given sentences that it contains is a valuable starting point. This approach allows the investigator to eliminate redundant information and to focus only on those sentences within the text that elucidate its underlying substance or general meaning. This is similar to the act of applying RST between two sentences but inside the sections of a single article (i.e., identifying intra-section relations). On the other hand, it is important to recognize that various types of relations exist between portions of text taken from multiple related articles. A case in point is when one author publishes an updated version of his/her prior work that contains new information or an additional component, and as a result of which, a clear relationship exists between the old and the new one. In such a case, it is likely to be true that the reader is only interested in the updated areas within the new article. Additionally, when one author references the work of another, this indicates that a relation exists between the articles. For example, the author could be highlighting that the work being cited provides supporting evidence for a claim or that it is the source of the ideas, insights, or concepts he or she is drawing on.

In light of discourse analysis theories, this work addresses three main points: first, we combined some relations from RST [4], CST [5], and Trigg's links [7] for use with scientific articles. We then classified the final relation set based on its existence within each section type and then across multiple sections of the same type from different articles. This combination and classification led to the production of CAST. The proposed theory was then tested by applying it to detecting the relations that exist between pairs of sentences from topically related sections of scientific articles. This effort differs from other works such as [4], [5], [7] in the way we combine three different kinds of relations and in the organization of the resultant set of relations within a single article and across multiple articles. Moreover, this work used a scientific text, which differs from a generic text.

The rest of the paper is organized as follows. Section II presents essential background information. Section III examines related prior studies. Section IV describes the proposed theory. Section V presents a case study and some experimental results. Finally, conclusions are drawn up in Section VI with an outlook on potential future research.

II. BACKGROUND

The purpose of a research article is to report on original work, whether theoretical or empirical. Such articles are regularly produced in academic fields, such as the natural or social sciences. Various types of research article exist, including review articles, meta-analyses, commentaries, and original research. This section provides a brief overview of the general structure of scientific articles, examines the multi-article summarization task, and explains RST and CST.

A. THE STRUCTURE OF SCIENTIFIC ARTICLES

Writing a scientific article is an essential step that researchers must undertake to ensure that their results can be accessed by other researchers in the academic community. A key difference exists between articles published in scientific fields and those published with generic text, most notably with respect to their structural features. Broadly speaking, a sizeable majority of all scientific articles adopt a uniform structure; the rationale for this being to promote standardization and accessibility in terms of the reception of the information each article contains. At the same time, the use of a uniform structure in scientific articles allows a wide readership to engage with the paper at a particular 'level' [8]. As a case in point, certain readers may only want an overview of the study; in which case, they will focus on the abstract. Other readers may be seeking specific information about the methodological aspects of the study; in which case, they can turn immediately to the corresponding section of the scientific article.

In terms of the structure of scientific articles, most are separated into the following distinct sections: abstract, introduction, literature review, methodology, results, discussion,

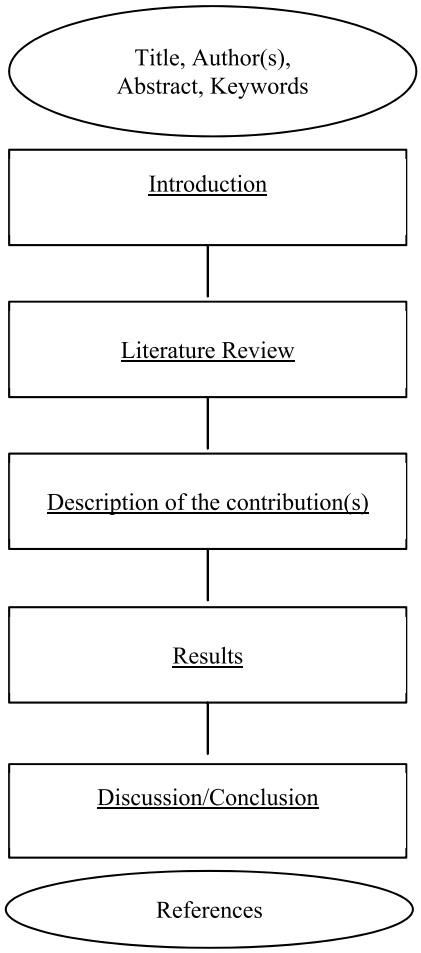

FIGURE 1. Structure of a scientific article.

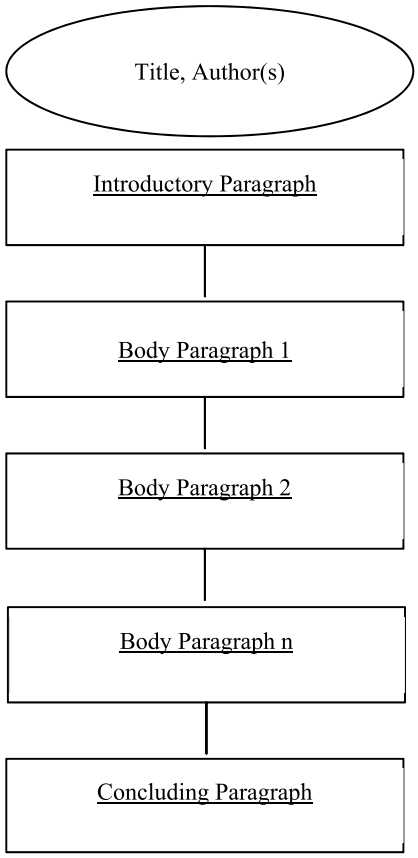

FIGURE 2. Structure of a general article.

and conclusion. Fig. 1 shows the general structure of a scientific article and Fig. 2 presents the main structure of a general article. Regarding the abstract, this segment summarizes the core features of the article in sequence, beginning with a problem statement (or specification of the research question and objectives) and following this with the methods, findings, results, contributions, and conclusions. For the introduction, this portion provides an overview of the research context (whether empirical or theoretical); explains the significance of the research; presents the research aim, question, and objective; and where relevant, specifies null and alternative hypotheses. In the literature review, previous studies relevant to the topic of the scientific article are examined, and in most cases, critically assessed. The methodology section details the research design, as well as its techniques and procedures, and the information is presented in a logical way to ensure the reader's understanding of key points.

In the results section, attention is drawn to the noteworthy outcomes obtained by implementing the study's methods, and the information needed to address the research questions, and where relevant, the hypotheses are given. Typically, the results section of a scientific article starts with text content, after which tables and figures are given. The penultimate section, the discussion, interprets the results against those reported elsewhere in the literature, provides an interpretation of the main research problem that has been addressed, and examines the potential contributions of the findings. Depending on the article in question, the last paragraph of the discussion section may be used to conclude the article, or a standalone section is dedicated to presenting concluding remarks. The purpose of this closing section is to review the principal points of the article, to highlight implications, findings, or limitations, and to offer recommendations for further research.

It is worth noting that, due to the abovementioned uniform structure that is adopted in most scientific articles, certain articles, especially those published in the same field, have a high likelihood of containing similar – if not the same – information (e.g., in the literature review section). The information reported in the sections that follow (e.g., methodological information) tends to differ because, in order to be published, most articles must have a unique focus. In certain sections, relations can be identified between different articles, and in two different articles, it may be the case that the same design is being reproduced, which necessitates that an equivalence relation holds between much of the content each one contains. Additionally, wherever one paper cites an article that has also been cited by another article, and when the citation is made for the same reason, this also creates a relation. These possibilities demonstrate that contrasting relations can exist

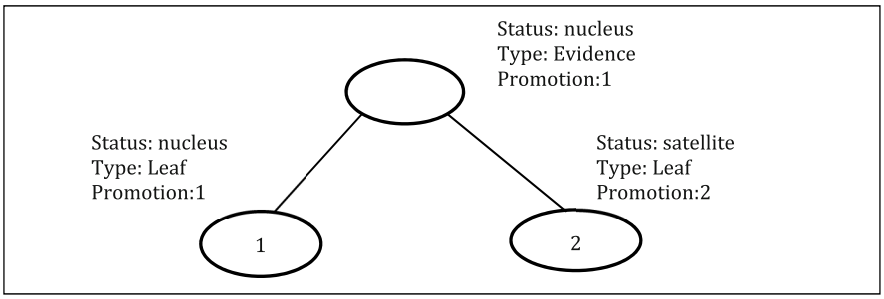

FIGURE 3. Example for the feature status, type, and promotion for an evidence relation that connects two leaf spans.

between different articles, and at the same time, between any given portions of text within an article.

B. SUMMARIZATION OF MULTIPLE SCIENTIFIC ARTICLES

Summarizing a set of scientific articles is a distinct task when compared to summarizing a set of generic texts [5], [9]. Even if each scientific article within the set addresses the same topic, each paper presents different information [9]. Therefore, the task is to present the research questions and the different arguments that are used. Additionally, the structure of these articles is another problem that must be considered [5]. The abstract associated with each article serves as a short summary, and therefore, summarizing these abstracts will lead to incoherent text. Multiple introductions about the same or similar topics would contain overlapping information. Therefore, in any attempt to summarize multiple scientific articles, the task is to identify a reasonable way in which to synthesize information from each of the articles, and in turn, to generate a brief overview of the important points. One of the main challenges associated with this task arises when it is necessary to compare varying perspectives on the same literature [9]. Summarizing multiple related work sections requires a presentation of different viewpoints towards the same reference, all the while avoiding redundancy. In some sections, including the methodology, results, and discussion, unique information is presented. Hence, in every section type, the objective is to offer a summary of every article while retaining coherency and readability. In the conclusion section, two problems are usually addressed: first a summary of the study's main points, and second, an overview of future research directions. When summarizing multiple scientific articles, it is important to generate a single conclusion that synthesizes every possibility highlighted in each of the articles in question. For the references section of a scientific article, differentiating between the same reference that is presented using different referencing styles across several articles is another challenge that must be addressed.

C. RHETORICAL STRUCTURE THEORY

As a theory that describes texts and their structural features, rhetorical structure theory (RST) was initially formulated by Mann and Thompson [3]. RST identifies relations between two adjacent text spans (with some exceptions) in a hierarchical fashion. A binary tree is generated after applying RST to a text, where the leaves of the tree denote elementary textual units that organize the text based on rhetorical relations [10]. The key elements of RST are text spans and relations. The span can be conceptualized as a nucleus (N), which denotes the unit in the relation characterized by the greatest level of importance. Alternatively, it can be regarded as a satellite (S), which may be another nucleus (multiple nuclei) or a simple satellite, the purpose of which is to provide information to the nucleus. For a nucleus and satellite, when summarizing the text, it is possible to retain only the significant span. In the case of multiple nuclei, both must be retained. A range of approaches can be used to analyze text and identify the relations that exist within it, including cue phrases [11], [12]. When constructing the rhetorical structure, the nature of the relations between a pair of text spans informs the binary trees that are produced.

Fig.3 offers an example of the feature status, type, and promotion associated with an evidence relation that is linked to leaf spans. If two text spans present an evidence relation, this means that the writer utilizes S to heighten the credibility of N in the reader's perception. Every node has a *status* (i.e., nucleus or satellite), a *type* (i.e., the rhetorical relation holding between the pieces of text that the node spans), and a *salience or promotion set* (i.e., the group of units representing the key aspect of the text that the node spans) [12]. These salient units are specified in a bottom-up mode. The salient unit of a leaf node is itself, whereas the union of the salient units of an internal node's nuclear children form its salient units. As a case in point, the promotion set of the node that spans units (1, 2) in Fig. 3 has 1 as its salient unit because only the node that corresponds to span (1, 2) is a nucleus, whose salient unit is 1. The text's fundamental units are those situated in the root node's promotion set. Therefore, the units at a specific level in a node's promotion set are characterized by a higher level of importance when compared to those situated at any level underneath. This provides the user with the ability to construct summaries of the text in question for a range of granularities [13]. In this way, units in the root node's promotion set generate a concise summary, whereas units in

the root node's promotion set and those situated in the initial level of the RS-tree generate a comparatively long summary. The length of the summary increases with the corresponding level of the RS-tree.

D. CROSS-DOCUMENT STRUCTURE THEORY

Trigg [7] published a pioneering study in the field of link type classification for scientific articles. The researcher proposed 80 link types and classified these as either normal or commentary links. Normal links connect nodes that lie in the same scientific work or in different works, whereas commentary links connect an external portion of text about a node to the node in a scientific article. Cross-document Structure Theory (CST) was first proposed by Radev [4], and it was based on Trigg's research [7] but with generic text. CST assigns a relation type to the link, where each relation type can exist between words, phrases, sentences, or entire documents. In Radev's study [4], 24 relation types were proposed, including contrast, judgment, and translation relations. Table 1 lists these relations. In CST, relations may have two distinct directionalities: symmetrical or asymmetrical.

TABLE 1. CST Relations by Radev [4].

Identity	Judgment
Equivalence	Fulfillment
Translation	Description
Subsumption	Reader profile
Contradiction	Contrast
Historical background	Parallel
Modality	Cross-reference
Attribution	Citation
Summary	Refinement
Follow-up	Agreement
Elaboration	Generalization
Indirect speech	Change of perspective

Directionality relies upon the semantic nature of the connection, which means that symmetrical directionality occurs when both fragments are influenced similarly by the relation, such as the equivalence relation (which states that two text segments have similar contents), whereas asymmetrical directionality occurs when one unit influences the other, such as with a historical background relation (which states that a text portion provides information necessary for the proper understanding of some other text). Fig. 4 shows examples of these two relations among sentences from different sources. CST is applicable to various tasks in natural language processing (NLP), including text summarization [4] and multi-document parsing [14].

'Equivalence' relation:
S_1: *The witness testified that he did not see the accused man commit the crime.*
S_2: *The witness testified that the accused man was not seen by him near the place where the crime was committed.*

'Historical background' relation: (directionality: from S_3 to S_4)
S_3: *Someone left a coffee cup in my office;*
S_4: *would the owner please come and get it?*

FIGURE 4. Examples of CST relations.

III. LITERATURE REVIEW

A. DISCOURSE RELATIONS

Trigg [7] and Trigg and Weiser [15] were among the first researchers to explore multi-document relations. Their aim was to express the basic structure of a scientific article by capturing semantic relations among textual spans and various levels of information. To achieve their goal, the researchers classified the following types of textual spans: chunks and tocs.

The former textual span referred to a sentence, paragraph, or even an entire document, whereas the latter was used to refer to more than one chunk. Additionally, the researchers proposed two types of connection links: first a normal link, which connected nodes that laid within a single article or several articles, and second, a commentary link, which drew a connection between the external portion of text about a node and the node contained within the article. Finally, the researchers identified the following types of directionality for the abovementioned connection links: first physical directionality, which referred to the way in which the relation was drawn, and second, semantic directionality, which instead depended on the connection's meaning.

Allan [16] published an approach for identifying content-based relations among multiple documents. The researcher reported on a group of connections and categorized these as manual, automatic, or pattern-matching. The identification of manual connections may require human intervention, which stems from the impracticalities associated with the use of computational strategies in this context.

In contrast, automatic links might require more sophisticated procedures, but they can still be identified using computational procedures. Finally, pattern-matching connections may be distinguished by drawing on straightforward or expounded pattern-matching procedures (e.g., word matching). In contrast to Trigg [7] and Allan [16], McKeown and Radev [17], [18] formulated a multi-document summarizer that relied on a set of semantic relations. The relations in question were comparable to those reported on by the previous researchers, but they included features that were directly relevant for multiple documents. These features operated in a domain-dependent field in which a template was automatically populated with data pertaining to terrorist attacks. Following the completion of the templates, the semantic relations existing among them were manually specified.

As an extension of Trigg's [7], Allan's [16], and Salton *et al.*'s [19] works, Radev [4] proposed CST. The approach was based on the idea of RST [3] but included

multiple related documents. In contrast to RST, CST does not depend on the writing style; the latter assigned a relation type to the link that could exist between words, phrases, sentences, or entire documents based on the documents' structures. The method works on generic text with an arbitrary domain. We have briefly explained CST [4] in Section II(D). Similar to McKeown and Radev [18] and Afantenos *et al*. [20], Afantenos [21] proposed semantic relations between textual units based on templates but for the football domain. The authors insisted that these relations were specific for each domain or subject. For this goal, two types of relations were proposed: synchronic and diachronic. The former was very similar to CST in that it described an event, at a specific period, among numerous information sources. The later, diachronic relations depicted the development or progress of an event in one information source, through a timeframe. This model has never had a programmed application outside of the football domain. The works by Afantenos *et al.* [20] and Afantenos [21] need a corpus that should be based on a certain topic and several events — that should be summarized — and that is evolving and being described by more than one source. Additionally, the topic's ontology (i.e., the types of entities in the corpus that the summaries concentrate on) needs to be specified. Furthermore, the system needs to specify the message type, which should contain the entity type and event-specific role.

Zhang *et al*. [22] and Zhang and Radev [23] were among the first researchers to automate the task of identifying CST relations [4]. They proposed two classifiers for this task, the first of which determined whether a given pair of sentences was associated with a CST relation (irrespective of its type). As a result, the purpose of the classifier was simply to identify whether CST relations existed or not. Regarding the second classifier, this sought to document the existence of the CST relation, and in addition, identify its type. Lexical, semantic, and syntactical features were used to complete this task, and the training data were 41 news texts from CST Bank [24]. Only the following types of relations were employed in the experiment: 'Description', 'Follow-up', 'Equivalence', 'Elaboration', 'Overlap', and 'Subsumption'. A special type was also included to designate situations in which no relation was identified. In terms of the F-measure for the types of relations, the average value was 0.25.

Finally, in the study conducted by Miyabe *et al*. [25], the researchers formulated an approach that could be used to facilitate the automatic identification of 'Equivalence' and 'Transition' relations in the Japanese language. The process involved grouping the sentences based on their similarity and then checking for the existence of an equivalence relation. In turn, lexical and syntactical features were used to identify transition relations (i.e., relations in which a pair of sentences have identical content with different numerical values). The identification of equivalence relations was effective when compared to that of transition relations, as indicated by the F-measures of 0.76 and 0.46, respectively.

Radev's CST [4] has become extremely prominent in multi-document analysis [14], [26], [27]. This is particularly true for the area of multi-document summarization. Despite this, the approach has received substantial criticism due to its subjectivity and ambiguity. This reaction has prompted various researchers to propose refinements and extensions to CST. One such group of researchers, Maziero *et al*. [28], first formalized the original CST relation definitions based on two factors: relation directionality and restrictions. In turn, the researchers pruned and combined certain relations based on their meaning. Finally, they organized the proposed relations in a hierarchical fashion, whereby relations were classified based on their semantic nature. To achieve this, two relation types were used: first *content*, which included all the relations that referred to similarities and differences in the contents of the textual units and which can be further subdivided into *redundancy*, *complement*, and *contradiction*; and second, *form*, which included those relations that dealt with superficial aspects of the text (e.g., relations concerned with writing style and citations).

B. MULTI-DOCUMENT SUMMARIZATION

There are many works explored the use of CST in the field of multi-document summarization [4], [29]–[34]. A four-stages multi-document summarizer has been proposed by Radev [4] to investigate the usefulness of the proposed CST relations. In the first stage, the set of documents is clustered based on topic. The second stage is document analysis, in which the document tree representation is generated. Following this is the third stage, which includes the creation of CST relations. The final stage is summary generation. The generated summary can be for user preferences or generic summary, based on the CST relations of a portion of the text. Later, this methodology was followed by Zhang *et al.* P[32], who showed that incorporation of CST relations with a multi-document summarizer produces better results. Otterbacher *et al.* [33] also supported the usefulness of CST relations in multi-document summarization. Through the use of sentence ordering, the authors observed more coherent summaries. Castro and Pardo [34] also proposed a CST-based summarizer for a multi-document task. They proposed five content-selection operators: context, contradiction, authorship, evolving events, and redundancy. Experimental results on Brazilian Portuguese news texts showed that a CST-based summarizer produced more informative summaries.

Similar to the work presented in this paper, Cardoso [29] and Cardoso and Pardo [30], [31] combined the RST [3] and the CST [4] and proposed some methods based on these theories to address the problem of information relevance in automatic multi-document summarization. The RST model details major aspects of the organization of a text and indicates the relevant discourse units. The CST model, meanwhile, describes semantically related textual units from texts on related topics. The authors used the CSTNews corpus [35], composed of 2,088 sentences written in Brazilian Portuguese, and manually annotated it with RST and CST relation(s).

Some of the proposed methods have been based only on RST relations. The second group of methods has based on combining RST and CST, and the last group has been based on both in addition to subtopics. The related contribution of their works to CAST is the combination of RST with CST, in which redundancy across multiple texts is controlled by means of CST relationships, whereas RST is used to remove irrelevant information and make room for more information.

C. MULTI-ARTICLE SUMMARIZATION

Multi-article summarizations have become more important recently due to explosive growth in scientific publications and the frequent presence of pertinent information in multiple articles. A certain amount of redundancy is found in this type of summarization, because the contributions from a target article may be described in multiple texts. Thus, the demand for identifying important differences among documents is high. Agarwal *et al.* [36] summarized a collection of papers cited within the same target article and proposed an interactive multi-document summarizer called SciSumm. It creates a query-based summary that comprises four modules. First, text tiling generates tiles of text that are relevant to the citation context. The clustering module then groups these tiles into labeled clusters. A convenient and comprehensive description of each cluster is provided using these labels. Ranking is then applied to the clusters based on their relevance to the generated query. Finally, the clusters with the highest scores from the previous module are generated through summary presentation. Chen and Zhuge [37] made additional progress by taking advantage of multiple citations appearing in one paragraph or section. The main contribution of their work was to expand article citations using CFDSumm, a multi-document summarization system that exploits a set of terms co-occurring in a list of citations according to the common-fact phenomenon. In a recent study, Sun and Zhuge [38] proposed a summarization system based on the semantic network. This network is built to represent the semantic link (or type of relation) between the nodes of the scientific paper (i.e., sections, subsections, paragraphs, sentences, and words). The authors focused on three particular types: *is-part-of*, *similar-to*, and *co-occurrence*. The sentences were then ordered using a graph-ranking algorithm on the constructed semantic-link network, and the top-k ranked sentences were selected for the final summary. The experimental results demonstrated the effectiveness of this system. In addition, the *is-part-of* relation was shown to be more helpful for short summaries than for long ones, and more effective with longer papers containing more structural information.

IV. CROSS-ARTICLE STRUCTURE THEORY (CAST)

As reported in several studies, including Bosma [27], Verberne *et al.* [14], and Mittal [28], discourse relation analysis has recently been applied successfully in various areas. Cross-article structure theory (CAST), the theory presented in this paper, combines RST [3] and CST [4] with several of the links proposed by Trigg [7]. The purpose of CAST is to assist in natural language processing (NLP) tasks, including the summarization of scientific articles. This particular task remains the focal point of the present paper.

It is useful to consider both intra-document (i.e., intra-section) and cross-document (i.e., cross-section) relations when summarizing multiple scientific articles. In terms of the summarization process itself, it aims to present the key points and ideas reported in a text in an accurate and concise way, all the while retaining the overall meaning. In writing, combining sentences as naturally as possible is an essential part of fluent communication. To do this, writers usually combine sentences using linking words and connectives. When the purpose and type of specific linking words are identified, it is then possible to determine the type of relation that exists between a pair of sentences. In turn, a decision can be made about whether to retain each sentence or only one sentence. The hypothesis that guides the present study is the utilization of the relations that exist between the sentences in each section of one article. In this way, it is possible to generate an abbreviated version of each section, which contains only the important sentences. In turn, discoveries can be facilitated about the relations that exist across the sections of an arbitrary number of articles. In doing this, the important sentences in every article are retained, while redundant information is eliminated. Fig. 5 provides an overview of these relations.

Intra-section links refer to the connections between sentences in a single section of any given article (typically adjacent sentences). The definition of RST was adopted in this paper, and the relevant rhetorical relations were chosen for individual scientific articles existing between sentences in a particular section. Contrastingly, cross-section links exist between sentences from the same sections of different articles. Given that scientific articles are typically structured in the manner described in Section II(A), our process was initially to apply RST links [3] to provide a description of the relations among the sentences within a single section of an article. In turn, the relevant CST relations [4] with several of Trigg's links [7] were used to determine what relations exist between text portions from multiple articles.

In CAST, relations from different levels (that is, within a single article and across multiple articles) are used to remove redundant information and generate an optimal ordering of sentences in the resulting summary of multiple articles. CAST also enhances the process of sentence extraction. Sentences with the most relations among the multiple articles are classified as the most important. Determining the potential relations that may be exist between text spans from the same section can be done using different techniques [39]–[41]. Marcu [42] has specified different cue phrases that can be used with English language-processing applications such as text summarization [43] and text segmentation [44], [45]. A cue phrases-based approach is reasonable in the case of intra-section links determination because of the conventions of writing and the fact that authors usually tend to write using certain writing techniques. However, in the case of cross-section links, we cannot expect to observe

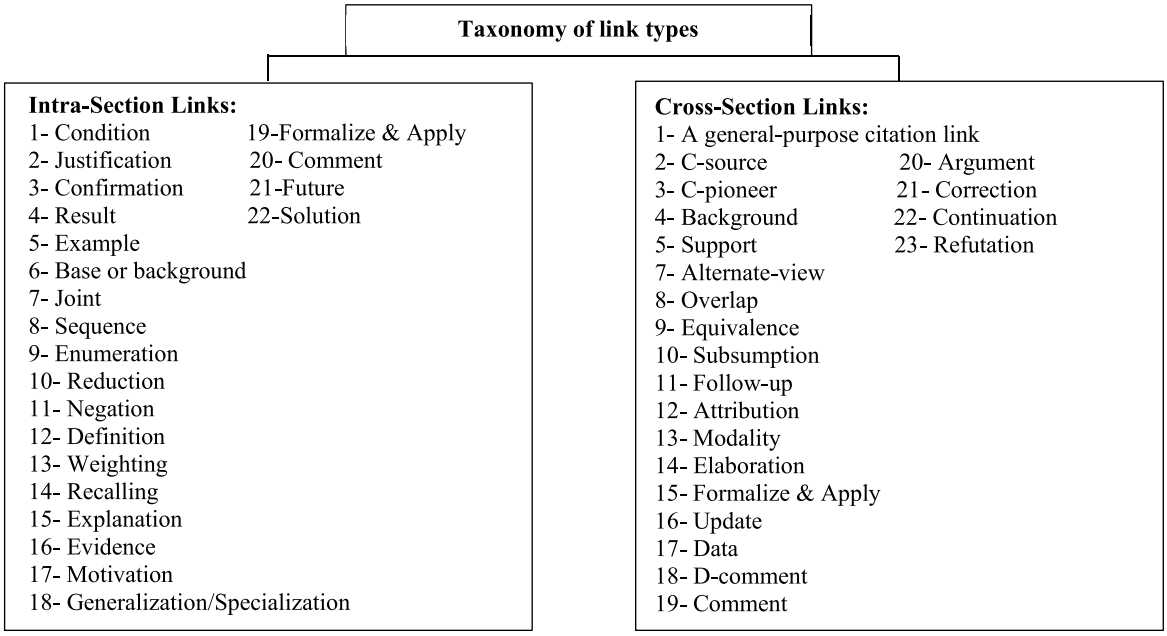

FIGURE 5. Taxonomy of link types.

TABLE 2. Sentences extracted from [46].

Paper	*Section*	*Sentence*	*Text*
Agirre and Soroa (2009)	A	1	Given an input context, the method first explores the whole LKB in order to find a subgraph which is particularly relevant for the words of the context.
		2	Then, they study different graph-based centrality algorithms for deciding the relevance of the nodes on the subgraph.
		3	As a result, every word of the context is attached to the highest ranking concept among its possible senses.
	B	1	The vector v can be non-uniform and assign stronger probabilities to certain kinds of nodes, effectively biasing the resulting PageRank vector to prefer these nodes.
		2	For example, if we concentrate all the probability mass on a unique node i, all random jumps on the walk will return to i and thus its rank will be high.

a static phrase in one text portion from one article that reliably indicates a specific relationship to some phrase in another text portion from a different but related article. Therefore, it may be worthwhile to look for deeper-level cues and pursue statistical approaches instead. Machine learning based approaches have been used widely to determine cross-document relations in works such as [22], [23], [25], [28] and this study (see Section V).

To illustrate the proposed theory, it is useful to consider a case in which there are multiple scientific articles that need to be summarized. The process begins with the introduction of the first article, after which RST is applied to the sentences this section contains to identify the type of relations existing between two adjacent spans of a text (e.g., by using cue phrases). As noted in section II(C), a tree structure is the output of RST, which arranges the text on the basis of rhetorical relations. Based on the predetermined level, only the sentences in the promotion set of a given level's node are considered since these are regarded as the most important units. As a result, unnecessary spans are excluded from the tree (that is, the satellites). In turn, the same process takes place for the other articles in the overall set. When the process is complete, what remains is a group of abbreviated introductions for each article, whereby unnecessary information has been eliminated. Following this, CST with Trigg's links [7] is applied to the remaining sentences, which permits the identification of cross-section relations. As such, a summary of the introductions can be generated while mitigating redundancy. Finally, the combined RST-CST process is applied to the remaining sections of the articles, leading to the identification of only those sentences that are considered the most important. At this point, it is worth considering a concrete example.

TABLE 3. Sentences extracted from [46], [47].

Paper	*Section*	*Sentence*	*Text*
Agirre and Soroa (2009)	Literature Review	1	Sinha and Mihalcea (2007) extend their previous work by using a collection of semantic similarity measures when assigning a weight to the links across synsets. They also compare different graph-based centrality algorithms to rank the vertices of the complete graph. They use different similarity metrics for different POS types and a voting scheme among the centrality algorithm ranks.
Agirre *et al.* (2009)	Literature Review	2	Graph-based methods using WordNet have recently been shown to outperform other knowledge-based systems. Sinha and Mihalcea (2007) use graph-centrality measures over custom built graphs, where the senses of the words in the context are linked with edges weighted according to several similarity scores.

Therefore, Table 2 presents three sentences from a single section in Agirre and Soroa's paper [46], as well as two sentences from a different section. Sentence A2 contains a 'Sequence' relation between A1 and A2. The role of a relation of this kind is to order the sentences and arrange them in a sequence (i.e., sentence A1 is followed by A2). By contrast, the relation between A3 and A2 is known as a 'Result' relation. This relation signals that A3 presents a consequence of the situation initially presented in A2. Thus, one can benefit from these relations when ordering the sentences, and in this way, enhance the readability of the generated summary. B1 has an 'Example' relation with B2. In some cases, one can omit B2, particularly if the aim is to generate a short summary.

In the light of this, and by identifying the relations that exist among sentences, it is possible to produce a shorter version of the text (i.e., an article with fewer words) without losing any of the main ideas and concepts.

Following the generation of an abbreviated version of each section within a particular article, CAST focuses on cross-section relations among the articles of interest in the set. The objective at this point is to summarize each article of interest and to combine each one's contents into a single piece of text. The critical point of this process is to eliminate redundant information while ensuring that the important ideas and pieces of information are included.

In Table 3, sentences from the literature reviews of Agirre and Soroa [46] and Agirre *et al*. [47] are presented alongside one another. The application of CST [4] with Trigg's links [7] reveals that an equivalence relation holds between the two sentences, thereby meaning that only one needs to be retained to produce an adequate summary. The method illustrated here can be implemented across the texts, ensuring that the specific types of relations identified are dealt with in an appropriate way during the summarization process (or for that matter, any NLP task).

In the sub-sections that follow, a classification of the relevant relation types is presented, and each type of relation is defined.

A. INTRA-SECTION LINKS

Given that scientific articles are generally structured in a uniform way, it is useful to offer a classification of the relations based on their appearance in each section type. For example, the so-called 'Future' relation is typically not found in the introduction section of a scientific article. As another example, 'Base' or 'Background' relations are not present in a conclusion section. A corpus of 50 subsets of scientific articles were gathered. Each subset contains a set of three related scientific articles. Our dataset covered a wide variety of topics from different fields, including text summarization, sentiment analysis, robot motion planning, facial recognition, e-learning, finance analysis, and Arabic dialect identification, among others. First, we examine each section type to search for the relation type that may exist between any pair of sentences within the same section. Then, we examine the different sections of the same type from three articles for the same purpose. Our observations from this examination leads to the classification presented in Table 4 and Table 5. It is worth noting that several of the categories overlap. The proposed taxonomy of intra-section relations is presented in Table 4, and its union generates the set illustrated in the left-hand portion of Fig. 5.

Table 6 offers definitions for the abovementioned relations. Several definitions were adapted from [3], and others, namely, those focusing on a single article, were adapted from [7].

B. CROSS-SECTION LINKS

Following the identification of intra-section links, an abbreviated version of each section contained in every input article is produced. In turn, the next phase of CAST involves identifying the type of relation that exists between sentence pairs from the same sections in different articles. Table 6 provides an overview of the taxonomy of cross-section relations, while Table 7 offers a systematic definition of each relation. Trigg's [7] and Radev's [4] studies were consulted to specify these definitions, where the latter was used to

TABLE 4. Classification of intra-section relations.

	Introduction	*Related Work*	*Methodology*	*Experimental Results*	*Discussion*	*Conclusion*
Condition	√		√	√		
Justification	√	√	√	√	√	√
Confirmation	√	√		√	√	√
Example	√	√	√			
Base or background	√					
Joint	√	√	√	√	√	√
Sequence	√	√	√	√	√	
Enumeration	√	√	√	√	√	
Reduction	√			√	√	√
Negation	√	√	√	√	√	√
Definition	√		√	√		√
Weighting	√	√	√	√	√	
Recalling	√	√	√	√	√	
Explanation	√	√	√	√	√	
Evidence	√	√	√	√	√	
Motivation	√	√	√			
Generalization	√		√	√		
Specialization	√		√	√		
comment	√	√	√	√	√	√
Solution	√	√	√			
Result		√		√		√
Formalize & Apply			√			
Future						√

TABLE 5. Classification of cross-section relations.

	Introduction	*Related Work*	*Methodology*	*Experimental Results*	*Discussion*	*Conclusion*
Citation	√	√	√	√	√	√
Background	√					
C-source	√		√			
C-pioneer	√		√			
Support	√		√	√	√	√
Alternate-view	√		√		√	
Overlap		√				
Equivalence		√				
Subsumption		√				
Follow-up		√				
Attribution		√				
Modality		√				
Elaboration		√				
Formalize & Apply			√			
Update			√	√	√	
Data				√		
D-comment				√		
Comment				√	√	√
Refutation					√	√
Argument					√	√
Correction					√	
Continuation					√	√

focus on the relations existing between several scientific articles. To avoid the problems of subjectivity and ambiguity associated with CST [4], we differentiated between multiple types of relations, which were first proposed by Radev [4] and then revised by Maziero *et al.* [28]. According to [28], these problems are due to the similarity of some relation definitions and/or a lack of understanding about their semantic meanings. One of the main contributions of the present

TABLE 6. Intra-Section relations definitions.

Condition	Reader acknowledges that the situation expressed in satellite (S) is dependent on the situation expressed in nucleus (N)
Justification	Writer utilizes S to demonstrate that they were correct about N, thereby strengthening the reader's acceptance of N
Evidence	Writer utilizes S to heighten credibility of N in the reader's perception
Confirmation	S confirms a robust argument with premise N
Result	S proceeds logically from N
Example	S is a counterexample of N
Base or background	S contains every piece of information that a reader must know to heighten their understanding of N
Joint	A multinuclear schema in which no relation exists between S and N
Enumeration	Reader acknowledges a complete, ordered listing of each of the elements in a collection (multi-N)
Reduction	S is a concise version of N
Negation	S is a negation of N
Definition	S defines N
Weighting	S is a value assigned to N based on N's importance or significance
Recalling	Information retrieval procedure (i.e., S is a repetition of information referred to earlier in discussing N)
Explanation	S contextualizes some aspect of N
Sequence	An enumerated collection among the nuclei (multi-N)
Motivation	S increases reader's interest and commitment in N
Generalization/Specialization	Either (i) specific information is broadly applied, or (ii) broad information is specifically applied
Formalize & Apply	Refers to the act of formalizing and then applying a group of theoretical concepts to generate practical outcomes
Comment	S offers a general comment on N
Future	S discusses future work of N
Solution	Reader acknowledges that N solves the problem specified in S

work is the classification of these relations based on their existence across multiple articles. Additionally, we added certain Trigg's links [7] to describe more relations between scientific texts. This classification and avoidance of some types of relations address the aforementioned problems, as Zhang *et al.* [1] explained. The authors of [1] showed that, for example, the definition of 'Elaboration' relations was previously considered very similar to those of 'Refinement' and 'Description', as all of these relations cover text units that add more details about another text portion. Furthermore, the definition of the 'Fulfillment' relation was similar to that of 'Follow-up'. Thus, Maziero *et al.* [28] proposed pruning and combining some relations and produced a refined set of CST relations [4]. This refinement reduced the number of CST relations from 24 to 18. Based on these works [1], [28], we adopted the same procedure, thus avoiding the relations that may have similar meanings, and we further added some links proposed by Trigg [7] to capture other types of relations that may exist across scientific articles. As shown in Table 1, Radev [4] proposed two relations that almost give the same meaning: the 'Identity' relation, which connects two sentences with the same exact wording, and the 'Equivalence'

TABLE 7. Definitions of cross-section relations.

Citation	A general-purpose citation link
Background	Presents background information, referring to works by other authors (often entire works)
C-source	Points to the source of ideas and concepts, allowing the reader to verify whether certain claims are true
C-pioneer	Pays respects to the originators of certain ideas or concepts, or the discoverers of important findings or results
Support	Agrees with the ideas, results, or claims made in other articles
Alternate-view	Ideas are viewed in a novel way
Overlap	Spans of text overlap in terms of their content
Equivalence	A pair of text portions have identical information content that is presented differently
Subsumption	A sentence is included in another sentence, where the latter provides additional information
Follow-up	Additional information is offered regarding facts that have occurred at a later time
Attribution	A sentence offers attribution to the same information of another
Modality	A sentence is a qualified version of another
Elaboration	A sentence containing additional information
Formalize & Apply	Refers to the act of formalizing and then applying a group of theoretical concepts to generate practical outcomes
Update	One section highlights novel information from a different article
Data	A connection from an article to information given in another article
D-comment	A general comment on the data employed in another article
Comment	Generic comment link (e.g., focusing on the results reported in another article)
Refutation	Statements that view another article or set of ideas negatively
Argument	Link an argument to its conclusions
Correction	A link to a correction of erroneous information
Continuation	A link between a pair of sequential nodes

relation, which connects two sentences with the same content but different wording. Additionally, 'Agreement' and 'Support' are considered very similar. On the other hand, there are some relations that may be considered unnecessary for scientific articles: 'Reader profile', 'Translation' and 'Change of perspective'. 'Indirect speech' and 'Attribution' are very similar in that indirect speech needs a direct speech in one text portion of the pair. Moreover, 'Contradiction' with non-numerical information is difficult to detect automatically [28]. Thus, we make do with one relation that represent similar relations. As a result, we have those relations from CST with Trigg's links connecting text portions among scientific articles while avoiding ambiguous relations (i.e., relations with similar meanings).

There are important issues that must be addressed in multi-document summarization that are the same issues as those found with summarizing multiple articles. Redundancy, contradiction, information ordering, and complementarity are examples of these issues. Relations among text portions from multiple sources are used to avoid such problems [32]–[34]. Redundancy occurs when the same information is presented in different sources. 'Equivalence', 'Subsumption',

and 'Overlap' relations could be used to address this issue, as in [34]. Sometimes there is a situation in which inherent statements, actions, or ideas are inconsistent or contrary to one another. This situation can be overcome by determining 'Contradiction' relation among text portions. One of the main issues that is observed in the generated summaries of multiple documents is determining how to order set of sentences that comes from different sources. One technique that used to address this issue is the use of CST relations among text portions [33], [34]. As a case in point, let us consider a pair of sentences (S_1, S_2) that are from two different sources and have a background relation from S_1 to S_2. If this pair is selected to be in the final summary, S_1 will be placed immediately after S_2 which provides more clarity in the final summary. In some cases, one text provides complementary information about a fact presented in another text. This situation can be considered an 'Overlap' relation. The overlap relation means that spans of text overlap in terms of their content, but each text still provides unique information. Thus, in the final summary the overlapped content should be presented once, along with the union of the two texts' content [34]. One important property of scientific articles is the inclusion of citation sentences, which contain explicit reference(s) to other research articles. These sentences are used to summarize scientific articles, as in [36], [37]. Different viewpoints regarding the same literature need to be considered and combined to result in an accurate summary of the literature. However, these sentences may overlap or present the same information using different words (equivalent sentences). This kind of redundancy (i.e., redundancy of citation sentences) can be addressed by applying cross-article relations in which a 'Citation' relation detects the set of citation sentences and then, for each pair of sentences, 'Equivalence', 'Subsumption', or 'Overlap' relations will detect the redundancy between the two sentences.

V. CASE STUDY

A. PROBLEM DEFINITION

The problem tackled can be defined as follows: given a set of n pairs of sentences $S = \{(S_{i1}, S_{i2}), i = 1..n\}$ from topically related sections in different scientific articles, the objective is to identify the types of relation(s) that exist between each pair of sentences. This problem can be formulated as a multi-label classification problem, where the labels represent, in this case, the relations. The multi-label classification problem is decomposed into a set of binary classification problems (one for each relation).

B. FEATURES

Two sets of lexical and syntactical features were used to identify the discourse relations: one set S_A of 15 features for CST relations; another set S_B of 18 features for CAST relations. S_A consists of the set of features presented in subsections 1 and 2 whereas S_B includes S_A plus the set of features from subsection 3. In the following subsections, these features are discussed.

1) LEXICAL FEATURES

To determine the types of relations that exist between sentences, it is important to take into consideration the measure of covering information. A range of surface-level similarity features was used to evaluate the closeness of the lexical contents in the two sentences.

- *Cosine Similarity*: This assesses the distance (or similarity) existing between S_1 and S_2. It can be found as a dot product of their vector representations that measure the cosine of the angle between them. The smaller the angle is, the higher the similarity. Cosine similarity is useful in measuring the similarity between any two sentences, even if they are far apart by the Euclidean distance, i.e., counting the common words (due to their sizes), since they may still be oriented closer together. It can be expressed as follows:

$$\cos(s_1, s_2) = \frac{\sum S_{1i} * S_{2i}}{\sqrt{\sum (s_{1i})^2} * \sqrt{\sum (s_{2i})^2}} \tag{1}$$

- *Word Overlap Ratio:* This measures the frequency of the words between S_1 and S_2 that match. It can be expressed as follows:

$$\text{word overlap ratio}(s_1, s_2) = \frac{\#\text{common words between } s_1 \text{ and } s_2}{\text{total number of words in } (s_1 + s_2)} \tag{2}$$

- *Sentence Length Difference:* This measures the difference between the lengths of S_1 and S_2 in terms of the number of lexical elements they contain. It can help in determining certain kinds of relations, such as 'Subsumption', since the value might show which sentence is more informative. This characteristic is computed using the following formula, where w denotes the word belongs to a sentence:

$$\text{Difference in length}(s_1, s_2) = \sum_{w \in s_1} w - \sum_{w \in s_2} w \tag{3}$$

- *Overlap Ratio:* This measures the number of words that S_1 and S_2 have in common. An estimate of the overlap ratio is employed to determine whether each word in one of the sentences is likely to be found in the other. This determination plays a valuable role in facilitating insight into the level of similarity between the two sentences. The overlap ratio shows the percentage of information coverage that each sentence has with respect to the other sentence. It could help to identify certain relations (e.g., 'Subsumption' and 'Overlap'). It can be expressed as follows, where w denotes the number of words in a given sentence:

$$\text{overlap ratio } (s_1, s_2) = \frac{\#\text{common words between } s_1 \text{ and } s_2}{\sum_{w \in s_2} w} \tag{4}$$

TABLE 8. Cue words and phrases used in our case study.

schema	according to	data	dataset	corpus	corpora	comparison	compare
describe	improve	outperform	similar to	pioneer	differ	baseline	in contrast to
the work of	their work	the study of	this corpus	a corpus of	formalizing	the result of	the author of
addresses	adds	argues	cites	claims	comments	compares	concludes
discusses	emphasizes	hypothesizes	illustrates	implies	indicates	observes	points out
figure	points	significant	return	result	system	highlight	technique
shows	affirms	supports	confirms	agrees	disagrees	refutes	rejects
criticizes	adapt	assign	conduct	process	method	model	procedure
judgment	mention	remark	review	report	comparison	contradiction	variation

$$\text{overlap ratio}\,(s_2, s_1) = \frac{\#\text{common words between } s_1 \text{ and } s_2}{\sum_{w \in s_1} w} \quad (5)$$

2) SYNTACTICAL FEATURES

At the level of syntax, the area of interest is the number of words with respect to the following parts of speech (POS): nouns, verbs, adverbs, proper nouns, and adjectives. For each type (p) from the aforementioned POSs, the following was computed: first the total word count with p POS type in S_1 was found, and second, the total word count with p POS type in S_2 was calculated. Hence, 5 features were used for every sentence, thereby resulting in a total of 10 features for any given sentence pair at this level. These features highlight the measuring of the word class coverage between the two sentences. They could help, for instance, to recognize the sentence that has more adjectives and perhaps expounds the other sentence. We used 'Stanford log-linear part-of-speech tagger', proposed by [48], which was implemented by the Stanford Natural Language Processing Group.[1] This software is a Java implementation of the log-linear part-of-speech tagger that worked with our system implementation, and its speed, performance, and usability had motivated us to use it.

3) FEATURES SPECIFIC TO SCIENTIFIC ARTICLES
4) AN ADDITIONAL SET OF FEATURES SPECIFIC TO SCIENTIFIC ARTICLES WAS USED

- *Cite:* It is a binary feature of 1 if a sentence contains a citation to another reference, and 0 value otherwise. Regular expressions were defined to capture different styles of citations.
- *Relative Section Position:* This feature shows the section headline that contains the sentence.
- *Cue Words and phrases:* Some words are usually used in scientific articles to describe, for example, data used ('Corpus', 'Data', and 'Dataset'). Also, some phrases are used to refers to figures to. Table 8 shows a list of 72 words and phrases. The words refer to figures and tables are extracted from Bhatia and Mitra [49] and the rest are observed by examining different scientific articles.

[1] https://nlp.stanford.edu/software/tagger.shtml

C. EXPERIMENTAL EVALUATION

Two experiments were conducted to detect the relations that exist between pairs of sentences from topically related sections of scientific articles using CST [24] and CAST, respectively. Two classifiers, KNN_A and KNN_B, were obtained by training the k-nearest neighbors algorithm (KNN) [50] on the same dataset using the two sets of features S_A and S_B, respectively (see Section V(B) for S_A and S_B). The dataset used is described as following.

1) DATASET

We have created a dataset, namely CAST Bank, that consists of 114 annotated pairs of sentences: 58 pairs of sentences from CST Bank [24] and 56 pairs of sentences from different related topical sections of scientific articles. CST Bank [24] is a collection of English-language documents with manual annotations with respect to CST relations. It is important for researchers to show that such data (i.e., data in which items are labeled with categories) are reliable, regardless of whether they are used to develop and test a computational model or to support a claim. To do so, a suitable measure is used to demonstrate the validity of the annotation task.

The annotation work for CST Bank was completed by 8 annotators, and the Kappa agreement measure [51] (which measures the agreement between two raters who each classify N items into C mutually exclusive categories) was identified as 0.53 (range: 0-1). Notably, this value is considered low since it is less than 0.6 [52]. However, from our point of view, 0.53 could be considered good for 8 annotators since it is difficult to reach full agreement among all eight judgments. For CAST, we adopted the same format used for CST Bank [24] (see Appendix). Due to a lack of human resources, the annotation of CAST Bank was performed by only two annotators, and the Kappa agreement measure obtained is 0.82. The CAST Bank covers 23 relations (see Fig.6). All the relations have 6 instances except the relations 'Citation', 'Judgment' and 'Description'. The relation 'Citation' has 14 instances as most of cross-section link instances has a citation. The relations 'Judgment' and 'Description' have only one instance as they were extracted from CST Bank [24].

2) EVALUATION METRICS

As discussed in Section V(A), the problem addressed in this study can be considered as a multi-label classification

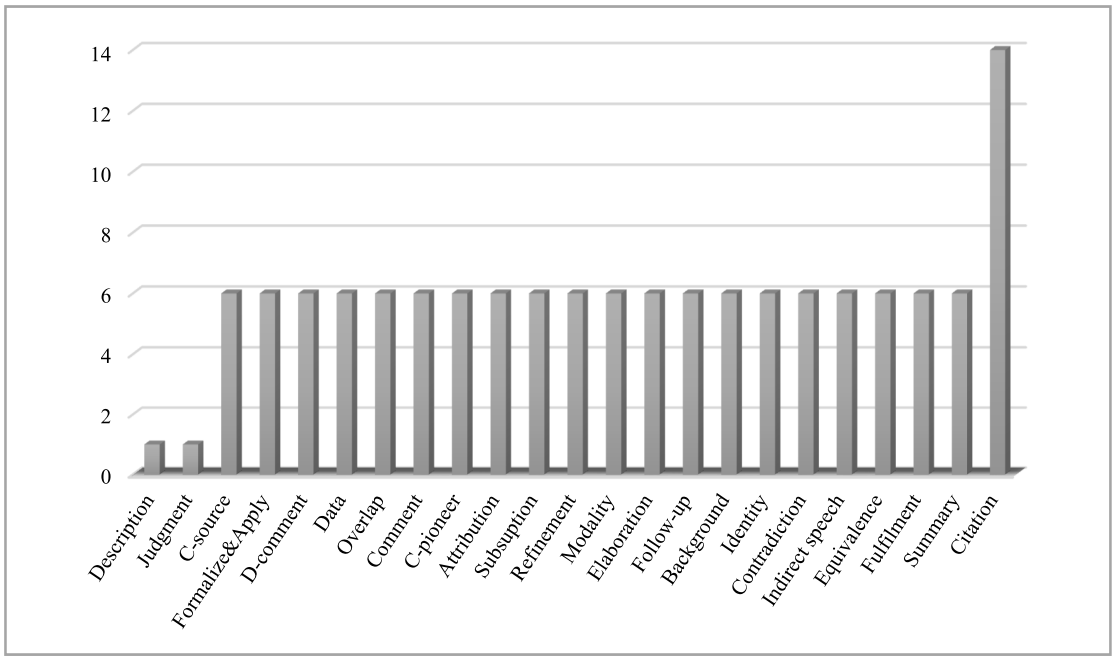

FIGURE 6. Distribution of number of instances per relation in CAST bank.

problem, where one relation is learned at a time and several relations can pertain to a single instance. The relation type represents the predicted label. The following evaluation metrics were used to evaluate the performance of the classifier: precision, recall, accuracy, and F-measure [53].

3) PERFORMANCE EVALUATION

Fig. 7 shows the performance results of the two classifiers KNN_A and KNN_B in terms of accuracy, precision, recall, and F-measure. The results clearly indicate that KNN_B outperformed KNN_A in terms of the four-performance measure used: accuracy of [KNN_B] is 0.91 vs. accuracy of [KNN_A] is 0.83; precision of [KNN_B] is 0.69 vs. precision of [KNN_A] is 0.62; recall of [KNN_B] is 0.71 vs. recall of [KNN_A] is 0.62; F-measure of [KNN_B] is 0.69 vs. F-measure of [KNN_A] is 0.61. The main differences between KNN_A and KNN_B is related to the three additional features used by KNN_B and related to scientific articles. Therefore, the impact of these features can justify the good performance of KNN_B.

Fig. 8 presents the F-measure results for the relation types identified by the two classifiers KNN_A and KNN_B. They distinctly highlight the superiority of KNN_B for identifying the different types of relations, particularly those associated to scientific articles, such as 'C-pioneer', 'C-source', 'Comment', and 'Data'.

4) DISCUSSION

Effective corpus analysis is the cornerstone of relation detection, where the frequency of each relation is sufficiently large

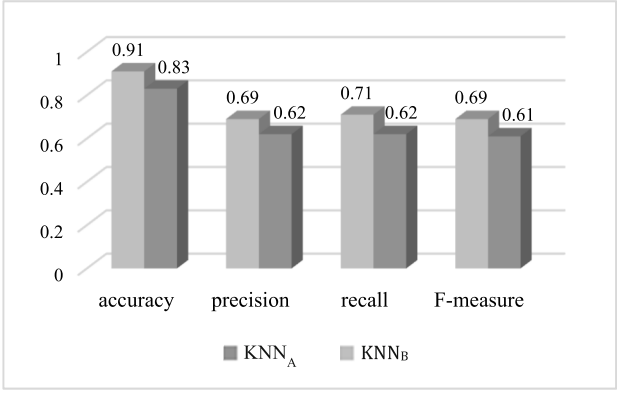

FIGURE 7. Performance results of KNN_A and KNN_B.

for any classifier to begin identifying the features of any given relation in an accurate way. In view of the disparities that exist between generic text and the text found in scientific articles, it is necessary to use scientific articles as an arena for further investigations of the relations existing within sections (in the same article) and between sections (in different articles).

The results indicated that certain relations must be explored in greater depth, particularly the 'Summary' and 'Fulfilment' relations. It is a complex task to determine whether one sentence is summarizing another sentence, given that a summary can be formulated in a range of ways.

Also, most of the cross-section relations connect two text portions (not sentences) or text segment with entire work

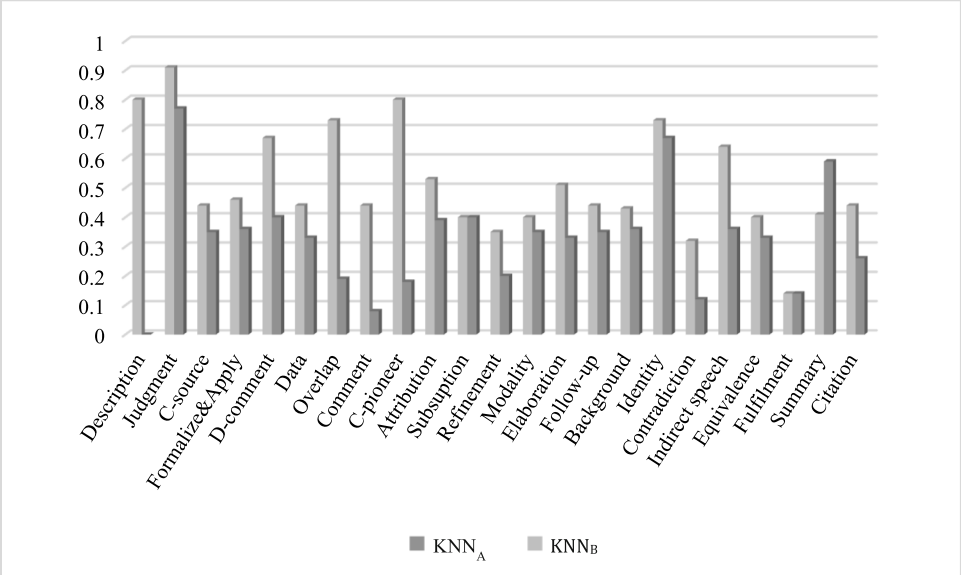

FIGURE 8. F-measure results for relation types identified by KNN_A and KNN_B.

(such as 'C-pioneer' and 'C-source'). Therefore, more features are needed to accurately capture the specific relations among the text portions.

VI. CONCLUSION

In this paper, CST and RST are exploited to expand the use of discourse relations with multiple scientific articles, as a result of which a cross-article structure theory (CAST) is proposed. Both intra-section and cross-section relations are considered in CAST. On the basis of the existence of these relations in each section of a scientific article, a classification of intra-section relations, as well as a taxonomy of cross-section relations among topically related articles, are presented. The results from the case study indicate that CAST generates promising results in terms of detection between relations among any given sentence pair in specific sections of scientific articles.

In future work, we intend to examine a larger body of scientific articles in order to possibly defining and classifying other relations, thereby leading to an extension of the CAST Bank. We also plan to compare the results with those of other studies, such as news text parsing, and to apply CAST to the summarization of multiple scientific articles.

APPENDIX

<TABLE>
<R>
SSENT = "In this work, we propose a refinement of the original CST."
TSENT = "We introduce CST (cross-document structure Theory) paradigm for multi-document analysis which takes into account the rhetorical structure of clusters of related textual documents."
SHEADLINE = "Introduction"
THEADLINE = "Abstract"
RELATION TYPE = "14"
</R>
<R>
SSENT = "Light stemming has shown competitive results in IR against root extraction-based stemmers [4]"
TSENT = "The stemmer introduced is a light stemmer, based on removing common prefixes and suffixes while keeping some of the word's distinctive features to minimize unnecessary conflation."
SHEADLINE = "Related Work"
THEADLINE = " Related Work "
RELATION TYPE = "10"
</R>
<R>
SSENT = "[3] proposed a framework for topic classification, which uses Linked Data for extracting semantic features."
TSENT = "Cano et al, [3] proposed a framework for topic classification, which uses Linked Data for extracting semantic features."
SHEADLINE = "Related Work"
THEADLINE = " Related Work "
RELATION TYPE = "4"
</R>
<R>
SSENT="Saggion and Lapalme organized this content into indicative or informative templates, to generate an article abstract."
TSENT = "This methodology allows the generation of indicative-informative abstracts integrating different types of information extracted from the source text."

SHEADLINE = ''Related Work''
THEADLINE = '' Methodology ''
RELATION TYPE = ''3''
</R>
<R>
SSENT = ''Slamet *et al.* (2018) proposed a simple system that automatically generates an article abstract for the Indonesian language.''
TSENT = ''The aim of this study is constructing automation for summarizing Indonesian articles as an alternative approach to an abstract.''
SHEADLINE = ''Related Work''
THEADLINE = '' Related Work ''
RELATION TYPE = ''3''
RELATION TYPE = ''2''
</R>
<R>
SSENT = ''According to [28], these problems are due to the similarity of some relation definitions and/or the absence of an understanding of their semantic meanings.''
TSENT = ''This subjectivity may be caused by various factors such as similarity among relation definitions or the lack of a proper understanding of the relations and their semantic nature.''
SHEADLINE = ''Introduction''
THEADLINE = '' Introduction ''
RELATION TYPE = ''9''
</R>
<R>
SSENT = ''Wolf and colleagues (Wolf *et al.*, 2005) have published a corpus of news articles annotated with coherence relations.''
TSENT = ''The Discourse GraphBank: A database of texts annotated with coherence relations. Philadelphia: Linguistic Data Consortium.''
SHEADLINE = ''Experiment''
THEADLINE = '' Title ''
RELATION TYPE = ''6''
RELATION TYPE = ''7''
</R>
</TABLE>

ACKNOWLEDGMENT

This work was supported by the Research Center of the College of Computer and Information Sciences, King Saud University. The authors are grateful for this support.

REFERENCES

[1] Z. Zhang, S. Blair-Goldensohn, and D. R. Radev, ''Towards CST-enhanced summarization,'' in *Proc. 18th Nat. Conf. Artif.*, 2002, pp. 439–446.
[2] Y. J. Kumar, N. Salim, A. Abuobieda, and A. T. Albaham, ''Multi document summarization based on news components using fuzzy cross-document relations,'' *Appl. Soft Comput.*, vol. 21, pp. 265–279, Aug. 2014.
[3] W. C. Mann and S. A. Thompson, ''Rhetorical structure theory: A theory of text organization,'' Inf. Sci. Inst., Univ. Southern California, Los Angeles, CA, USA, Tech. Rep. 8(3), 1987.
[4] D. Radev, ''A common theory of information fusion from multiple text sources step one: Cross-document structure,'' in *Proc. 1st SIGDIAL Workshop Discourse Dialogue*, 2000, pp. 74–83.
[5] S. Teufel and M. Moens, ''Summarizing scientific articles: Experiments with relevance and rhetorical status,'' *Comput. Linguistics*, vol. 28, no. 4, pp. 409–445, Dec. 2002.
[6] H. I. Mathkour, A. A. Touir, and W. A. Al-Sanea, ''Parsing Arabic texts using rhetorical structure theory,'' *J. Comput. Sci.*, vol. 4, no. 9, p. 713, 2008.
[7] R. H. Trigg, ''A network-based approach to text handling for the online scientific community,'' Ph.D. dissertation, Dept. Comput. Sci., Univ. Maryland, College Park, MD, USA, 1985.
[8] P. Pardede, ''Scientific articles structure,'' in *Proc. Sci. Writing Workshop*, 2012, pp. 1–15.
[9] N. Agarwal, K. Gvr, R. S. Reddy, and C. P. Rosé, ''Towards multi-document summarization of scientific articles: Making interesting comparisons with SciSumm,'' in *Proc. Workshop Autom. Summarization Different Genres, Media, Lang.*, 2011, pp. 8–15.
[10] D. Marcu, ''Discourse trees are good indicators of importance in text,'' in *Advances in Automatic Text Summarization*, vol. 293. Cambridge, MA, USA: MIT Press, 1999, pp. 123–136.
[11] D. Marcu, ''The rhetorical parsing summarization and generation of natural language texts,'' M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 1997.
[12] D. Marcu, *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press, 2000.
[13] D. Marcu, ''Improving summarization through rhetorical parsing tuning,'' in *Proc. 6th Workshop Very Large Corpora*, 1998, pp. 1–10.
[14] S. Verberne, L. W. J. Boves, N. H. J. Oostdijk, and P. Coppen, ''Discourse-based answering of why-questions,'' in *Proc. Traitement Automatique des Langues*, vol. 47, 2007, pp. 21–41.
[15] R. H. Trigg and M. Weiser, ''TEXTNET: A network-based approach to text handling,'' *ACM Trans. Inf. Syst.*, vol. 4, no. 1, pp. 1–23, Jan. 1986.
[16] J. Allan, ''Automatic hypertext link typing,'' in *Proc. the 7th ACM Conf. Hypertext*, 1996, pp. 42–52.
[17] K. McKeown and D. R. Radev, ''Generating summaries of multiple news articles,'' in *Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1999, pp. 381–389.
[18] D. R. Radev and K. R. McKeown, ''Generating natural language summaries from multiple on-line sources,'' *Comput. Linguistics*, vol. 24, no. 3, pp. 470–500, 1998.
[19] G. Salton, A. Singhal, M. Mitra, and C. Buckley, ''Automatic text structuring and summarization,'' *Inf. Process. Manage.*, vol. 33, no. 2, pp. 193–207, Mar. 1997.
[20] S. D. Afantenos, I. Doura, E. Kapellou, and V. Karkaletsis, ''Exploiting cross-document relations for multi-document evolving summarization,'' in *Proc. Hellenic Conf. Artif. Intell.*, 2004, pp. 410–419.
[21] S. D. Afantenos, ''Reflections on the task of content determination in the context of multi-document summarization of evolving events,'' in *Proc. Recent Adv. Natural Lang. Process.*, 2007, pp. 1–5.
[22] Z. Zhang, J. Otterbacher, and D. Radev, ''Learning cross-document structural relationships using boosting,'' in *Proc. 12th Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2003, pp. 124–130.
[23] Z. Zhang and D. Radev, ''Combining labeled and unlabeled data for learning cross-document structural relationships,'' in *Proc. Int. Conf. Natural Lang. Process.*, 2004, pp. 32–41.
[24] D. R. Radev, J. Otterbacher, and Z. Zhang, ''CST Bank: A corpus for the study of cross-document structural relationships,'' in *Proc. LREC*, 2004, pp. 1–4.
[25] Y. Miyabe, H. Takamura, and M. Okumura, ''Identifying cross-document relations between sentences,'' in *Proc. 3rd Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2008, pp. 1–8.
[26] W. Bosma, ''Query-based summarization using rhetorical structure theory,'' *LOT Occasional Ser.*, vol. 4, pp. 29–44, Dec. 2005.
[27] N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek, ''Sentiment analysis of hindi reviews based on negation and discourse relation,'' in *Proc. 11th Workshop Asian Lang. Resour.*, 2013, pp. 45–50.
[28] E. G. Maziero, M. L. D. R. C. Jorge, and T. A. S. Pardo, ''Revisiting cross-document structure theory for multi-document discourse parsing,'' *Inf. Process. Manage.*, vol. 50, no. 2, pp. 297–314, Mar. 2014.
[29] P. C. F. Cardoso, ''Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo,'' M.S. thesis, Universidade de São Paulo, São Paulo, Brazil, 2014.
[30] P. Cardoso and T. A. Pardo, ''Joint semantic discourse models for automatic multi-document summarization,'' in *Proc. 10th Brazilian Symp. Inf. Hum. Lang. Technol.*, 2015, pp. 81–90.

[31] P. C. F. Cardoso and T. A. S. Pardo, "Multi-document summarization using semantic discourse models," *Procesamiento del Lenguaje Natural*, vol. 56, pp. 57–64, Feb. 2016.
[32] Z. Zhang, S. Blair-Goldensohn, and D. R. Radev, "Towards CST-enhanced summarization," in *Proc. AAAI/IAAI*, 2002, pp. 439–446.
[33] J. C. Otterbacher, D. R. Radev, and A. Luo, "Revisions that improve cohesion in multi-document summaries: A preliminary study," in *Proc. ACL Workshop Autom. Summarization*, vol. 4, 2002, pp. 27–36.
[34] B. Schiffman, A. Nenkova, and K. McKeown, "Experiments in multidocument summarization," in *Proc. 2nd Int. Conf. Hum. Lang. Technol. Res.*, 2002, pp. 74–82.
[35] P. C. F. Cardoso, E. G. Maziero, M. L. Jorge, E. M. Seno, A. Di Felippo, L. H. M. Rino, M. das Graças V. Nunes, and T. A. S. Pardo, "CSTnews—A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese," in *Proc. 3rd RST Brazilian Meeting*, 2011, pp. 88–105.
[36] N. Agarwal, K. Gvr, R. S. Reddy, and C. P. Rosé, "Towards multi-document summarization of scientific articles: Making interesting comparisons with SciSumm," in *Proc. Workshop Autom. Summarization Different Genres, Media, Lang.*, 2011, pp. 8–15.
[37] J. Chen and H. Zhuge, "Summarization of scientific documents by detecting common facts in citations," *Future Gener. Comput. Syst.*, vol. 32, pp. 246–252, Mar. 2014.
[38] X. Sun and H. Zhuge, "Summarization of scientific paper through reinforcement ranking on semantic link network," *IEEE Access*, vol. 6, pp. 40611–40625, 2018.
[39] E. Agichtein and V. Ganti, "Mining reference tables for automatic text segmentation," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2004, pp. 20–29.
[40] D.-S. Chang and K.-S. Choi, "Causal relation extraction using cue phrase and lexical pair probabilities," in *Proc. Int. Conf. Natural Lang. Process.*, 2004, pp. 61–70.
[41] D.-S. Chang and K.-S. Choi, "Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities," *Inf. Process. Manage.*, vol. 42, no. 3, pp. 662–678, May 2006.
[42] D. Marcu, "From discourse structures to text summaries," in *Proc. Intell. Scalable Text Summarization*, Madrid, Spain, 1997, pp. 82–88.
[43] D. Cristea, O. Postolache, and I. Pistol, "Summarisation through discourse structure," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*, 2005, pp. 632–644.
[44] F. Golcher, "Statistical text segmentation with partial structure analysis," in *Proc. KONVENS*, 2006, pp. 44–51.
[45] S. Lamprier, T. Amghar, B. Levrat, and F. Saubion, "SegGen: A genetic algorithm for linear text segmentation," in *Proc. IJCAI*, 2007, pp. 1647–1652.
[46] E. Agirre and A. Soroa, "Personalizing pagerank for word sense disambiguation," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2009, pp. 4–33.
[47] E. Agirre, O. L. De Lacalle, and A. Soroa, "Knowledge-based WSD on specific domains: Performing better than generic supervised WSD," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, pp. 1–6.
[48] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol. (NAACL)*, vol. 1, 2003, pp. 173–180.
[49] S. Bhatia and P. Mitra, "Summarizing figures, tables, and algorithms in scientific publications to augment search results," *ACM Trans. Inf. Syst.*, vol. 30, no. 1, pp. 1–24, Feb. 2012.
[50] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis-nonparametric discrimination: Consistency properties," California Univ. Berkeley, Berkeley, CA, USA, Tech. Rep. 4, 1951.
[51] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Comput. Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.
[52] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Comput. Linguistics*, vol. 34, no. 4, pp. 555–596, Dec. 2008.
[53] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

NOUF IBRAHIM ALTMAMI was born in Riyadh, Saudi Arabia. She received the bachelor's degree (Hons.) and the M.Sc. degree in computer science from King Saud University, Riyadh, Saudi Arabia, in 2007 and 2012, respectively, where she is currently pursuing the Ph.D. degree. She has held a lecturing position at Shaqra University, Al-Muzhimiah, Saudi Arabia. Her research interests include natural language processing, machine learning, and AI applications.

MOHAMED EL BACHIR MENAI received a Ph.D. degree in computer science from Mentouri University of Constantine, Algeria, and University of Paris VIII, France, in 2005. He also received a Postdoctoral degree "Habilitation Universitaire" in computer science from Mentouri University of Constantine, in 2007 (this is the highest academic qualification in Algeria, France and Germany).

He is currently a professor in the department of computer science at King Saud University. His main interests include satisfiability problems, evolutionary computing, natural language processing, machine learning, and AI in medicine.