Journal of Biomedical Informatics: X 1 (2019) 100003

Contents lists available at ScienceDirect

Journal of Biomedical Informatics: X

journal homepage: www.journals.elsevier.com/journal-of-biomedical-informatics-x

Ultrasound reports standardisation using rhetorical structure theory and domain ontology

Nur Zareen Zulkarnain[a,b,*], Farid Meziane[a]

[a] School of Computing, Science and Engineering, University of Salford, Greater Manchester M5 4WT, UK
[b] Centre for Advanced Computing Technology, Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

ARTICLE INFO

Keywords:
Rhetorical structure theory
Rhetorical relation
Discourse parsing
Structured reporting
Ultrasound reporting
Ontology
Discourse markers

ABSTRACT

Ultrasound reporting plays an important role in diagnosis as images produced during an ultrasound examination do not give the whole view of the medical conditions. However, in practice there are many issues that are inherent to ultrasound reporting and the most important was identified to be the lack of standardisation when producing these reports. There is a resistance to change from some radiologists preferring the free writing style, making any attempt to computerise the processing of these reports difficult. This paper explores the possibility of using Rhetorical Structure Theory (RST) together with a domain ontology to transform free-form ultrasound reports into a structured form. It discusses a new approach in segmenting and identifying rhetorical relations that are more applicable to ultrasound reports from classical RST relations. The approach was evaluated on a sample ultrasound reports where the system's parsing was compared to the manual parsing performed by experts. The results show that discourse parsing using RST in ultrasound reports can be performed effectively using the support of a domain ontology. The results also demonstrate that the transformation of free-form ultrasound reports into a structured form can be performed with the support of RST relations identified and the domain ontology.

1. Introduction

Ultrasound is often used as the medical diagnostic imaging technique to evaluate a variety of conditions in the human body particularly the abdomen. Since images generated from an ultrasound examination do not give the whole view of the examination, the report produced by a radiologist is vital in diagnosing a patient's disease and the referring clinicians rely heavily on it [1]. Thus, the reports should be written clearly to ensure that diagnosis can be carried out correctly. Most ultrasound reports are written in free-form whereby findings are described in an essay-like document. According to Cramer et al. [2], there is a lot of interest lately in standardising ultrasound reports by using structured reporting instead of the standard free-form reports. Both radiologists and non-radiologists have also shown their preferences in using structured reporting as compared to free-form reports [3–7].

However, it was acknowledged that one major challenge to the adoption of structured reporting is the radiologists' resistance to change. This is supported by the study conducted by Tran et al. [8] where they saw higher usage of structured reporting among trainees as compared to staff radiologists with feedbacks expressing their difficulties to change and adopt the new reporting style. To give a better chance for the adoption of structured reporting, we argue that those who prefer to use free-form reports should be allowed to continue to do so and then develop a system that will allow the transformation of the free-text report into a structured one. An architecture to support the standardisation of ultrasound reports which also allows for the flexibility to write in both forms was proposed by Zulkarnain et al. [9].

This transformation is made possible by using the Rhetorical Structure Theory (RST) where rhetorical relations between text spans in the reports are identified and relevant information extracted and adapted to produce the structured report. RST is chosen as the mechanism to transform free-form reports into the structured form because of its strong reliance on the relationships between sentences and their components [10]. In our approach, RST is implemented with the support of a domain ontology, the Abdominal Ultrasound Ontology (AUO), as its knowledge base.

AUO was developed by reusing three existing biomedical ontologies namely, the National Cancer Institute Thesaurus (NCIT), the

* Corresponding author at: Centre for Advanced Computing Technology, Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia.
E-mail addresses: zareen@utem.edu.my (N.Z. Zulkarnain), f.meziane@salford.ac.uk (F. Meziane).

https://doi.org/10.1016/j.yjbinx.2019.100003

Available online 12 February 2019

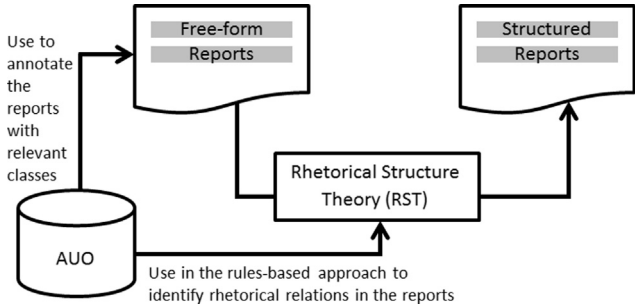

Fig. 1. Summary of the transformation process.

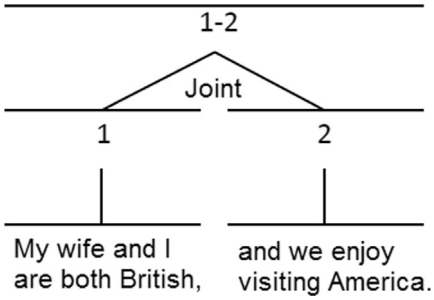

Fig. 2. JOINT relation signalled by "and" [16].

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) and the Radiology Lexicon (RadLex). The development of AUO is described in [11] together with the development methodology. AUO mainly consists of classes describing the abdominal ultrasound scanning and includes both the anatomy and pathologies of the abdominal area. It also includes technical terminologies that will assist in the reporting of the abdominal ultrasound scanning such as the word ultrasound itself as well as units of measurements.

AUO serve two purposes in achieving the standardisation of ultrasound reports: (i) to enforce the use of a standard terminology and (ii) to analyse the reports written in Natural Language (English free-text) with the aim of transforming them into a structured format. To achieve the second purpose, AUO is used to annotate the free-form reports with relevant classes that are then compared to the rules that are defined to identify existing rhetorical relations between the text spans in the report. This process is summarised in Fig. 1.

The rest of the paper is organised as follows: Section 2 briefly introduces RST and how discourse parsing was performed using several approaches. Then, Section 3 explains how RST could be adapted for use in ultrasound reports and presents the rules of some rhetorical relations that have been identified. In Section 4, we explain the development of a medical discourse parser that implements these rules using RST and AUO. In Section 5 we explain how the parser can be used in transforming free-form reports to structured form in the medical ultrasound reporting system as proposed in [9]. Finally, Section 6 will discuss the result of implementing RST on a sample of ultrasound reports by comparing the parsing of the system to the manual parsing completed by Natural Language Processing (NLP) experts. It will also discuss the evaluation and feedback gathered from a pair of specialists regarding the transformation.

2. Background

RST is a descriptive theory of a major aspect of organisation of natural text [10] pioneered by Mann, Matthiessen and Thompson in the 1980s [12]. It allows for the classification of texts' spans and the description of the relations between them that have independent functional integrity units also known as Elementary Discourse Unit (EDU) [13]. RST has the ability to present texts as coherent and illustrates the hierarchical structure of each part of the texts as having a role to play [12]. These roles are defined by rhetorical relations between texts' spans and can be identified as either a nucleus or a satellite. A nucleus is part of a text that can stand alone without its satellite; a satellite on the other hand loses its meaning without the nucleus [10]. A well written text is one that is coherent and has relations between each of its text spans and no text span is isolated [14].

In segmenting and identifying rhetorical relations, classic RST uses discourse markers and sentence structure as its indicators. However, in this paper, we describe a novel approach that utilises a domain ontology to perform the segmentation and identification of rhetorical relations in ultrasound reports. The next subsections will further discuss the different approaches of RST in discourse parsing.

2.1. Classic RST discourse parsing

Among the existing linguistic theories, RST has been shown to be effective in many computational linguistic applications [12]. RST allows for large texts to be broken down into smaller text spans where each individual segment has its own role to play in ensuring the coherence of a text. This coherence is ensured by the rhetorical relations that exist between each of these non-overlapping texts identified during the text analysis process. There are two tasks that are performed when analysing texts using RST namely, text segmentation and rhetorical relations identification.

The first task requires texts to be segmented into text spans or discourse units where each unit has an independent functional integrity [10]. Classic RST segments texts based on the sentence structure, cue words and punctuation marks. Marcu [15] for example segments texts into EDUs using discourse markers as an indicator to recognise the possible rhetorical relations and sets boundaries to perform segmentation. Mann and Taboada [16] also used the same method in their work where they segment a sentence using the discourse marker "and" to signal a JOINT relation as seen in Fig. 2. A similar approach was also taken by Carlson, Marcu and Okurowski [17] where texts are segmented into EDUs using lexical and syntactic clues as well as parts of speech to help determine the boundaries.

Using discourse markers, syntactic and parts of speech (POS) clues to segment text in an ultrasound report is a challenge since most ultrasound reports were not written in complete sentences. Sentences written in the reports are often short and straightforward, thus making it hard for traditional NLP tools such as POS tags to be applied [18]. In order to apply RST in ultrasound reports, the approach will need some modification. Instead of using syntactic clues or POS tags to segment text in ultrasound report, this paper discusses the possibility of using an ontology in addition to discourse markers such as "but", "otherwise" and "or" to segment a text. Correct segmentation of texts is important in developing a quality RST tree [19] because a correct segmentation will ease the process of identifying relevant relations between the segmented texts.

The second task involves the identification of rhetorical relations between the segmented text spans. In classic RST, 30 rhetorical relations were defined by Mann and Taboada [16]. In previous works [10,15,17], these relations were recognised using discourse markers as well as POS tags where each word in the text is assigned with its sentence parts such as nouns, verbs and adjectives. Even though discourse markers assisted in giving indications on the type of relations that exist between two text spans, not all relations are signalled with discourse markers [16].

This is even more pertinent in ultrasound reports where words used are limited and contain medical terms. Indeed, RST requires texts to be segmented into text spans or EDUs that have independent functional integrity. This means that an EDU should be in the form of a clause, rather than a clause fragment. However, this is not the case in ultrasound reports. For example, in a sample report, there is this sentence "Normal appearances of the liver, gallbladder, kidneys, pancreas and

spleen." which if according to the clause rule, does not make a clause since it does not have a verb. Then again, what constitute an EDU differs from one researcher to another. The classic RST by Mann and Thompson [10] put emphasize in defining EDU as having independent functional integrity which essentially are clauses. However, as presented by Carlson, Marcu and Okurowski [17] in their paper, different researchers have different views on the size and what constitute an EDU which some of them view as a sentence, prosodic units, a clause or an intentionally defined discourse segments. Yet, they all agreed that EDU are non-overlapping spans of text.

Since ultrasound reports are often not written in proper syntactic structure as doctors themselves sometimes write in fragment clauses, rhetorical relations identification is much more difficult using the classic RST technique. Thus, this paper explores the possibility of identifying these relations using a rule-based approach that applies discourse markers together with classes from a domain ontology.

2.2. Automatic RST discourse analysis

Discourse analysis using RST is performed by first segmenting the text into text spans based on discourse boundaries that have been set. These segmented text spans are then be parsed to identify rhetorical relations that exist between them. Discourse analysis was performed manually by Mann and Thompson [10]. Even though their work was successful, Marcu [13] argued that their work was too informal and not suitable to be automated. Thus, there were several works that have attempted to automate this process.

One of the earliest to have developed an automatic discourse parser was Daniel Marcu [13,15]. He proposed a surface-based algorithm that is able to perform the following three tasks (i) identifying cue phrases and breaking sentences using these cue phrases, (ii) hypothesising rhetorical relations between texts as well as (iii) producing an RST Tree. He developed what he called a shallow analyser that is able to determine elementary discourse units of a text, the rhetorical relations that exist between these units as well as the identification of the unit as a nucleus or a satellite. The term shallow analyser was introduced because it does not use the traditional parsing and tagging techniques. Instead, Marcu argued that discourse parsing using RST can be performed automatically by relying only on the coherency and connectivity of the text. Assuming that texts are well formed, he stated that it is sufficient to segment text and hypothesising relations using discourse markers.

However, Marcu found that discourse markers are sometimes ambiguous depending on the relation they are signalling as well as the size of the whole text. The more complex the text is, the more ambiguous it becomes to automatically detect rhetorical relations. To solve the ambiguity problem, Marcu took advantage of the text coherency where he looks for similarity measure of co-occurrences of words. If the measure exceeds a certain threshold, then a relation holds. When two text spans mention the same topics and share the same words, they will be assigned with the ELABORATION or BACKGROUND relation. However, if two text spans mention two different subjects, they will be assigned with a JOINT relation. To evaluate the accuracy of his parser, Marcu used the result from a manual discourse analysis completed by three judges with two third majority. Overall, Marcu's shallow analyser achieved 80.8% recall and 89.5% precision which is quite high compared to other discourse parsers at that time.

A more current automatic discourse parsers were developed by Feng and Hirst [14,20] and Joty et al. [21,22] and applied on the text level. The first version of Feng and Hirst's parser [14] enhances the HILDA [23] discourse parser by including their own rich linguistic features such as cue phrases, production rules and contextual features combined with features from the work of Lin et al. [24] such as the dependency parse feature. Since HILDA has already achieved a high F-score of 93.8%, Feng and Hirst have focused on improving the RST tree building task instead. Adopting the same methodology as HILDA, Feng and Hirst used a greedy bottom-up approach with two Support Vector Machine (SVM) classifiers in their parser. The first classifier is the binary structure classifier which evaluates whether there is a relation that holds between two text spans and merges them into a new subtree. The other classifier, the multiclass relation classifier on the other hand, evaluates which relation should then be assigned.

In 2013, Joty et al. [21] developed a text level discourse parser similar to Feng and Hirst's [14] first version of a text level discourse parser but instead of using SVM, they have used two CRF models to build RST trees using an optimal parsing algorithm. One of the two models was used for intra-sentential parsing which produces subtrees for each sentence, and the other for multi-sentential parsing which combines these subtrees into a text level RST Tree. Joty et al. claimed that separating the parsing of both intra- and multi-sentential is much more effective then when they are combined. In parsing the text, Joty et al. have taken a non-greedy approach and optimal [25] compared to HILDA and Feng and Hirst's parser which used a suboptimal and greedy approach. Their approach uses a probabilistic Cocke-Kasami-Younger (CKY)-like bottom-up algorithm that resulted in a globally optimal RST Tree and received an overall 55.73% relation assignment accuracy when tested on two types of corpus which were news articles (RST-DT corpus) and instructional how-to-do manual.

Feng and Hirst [20] however argued that their approach is inefficient as it takes too much time since the CKY parser they used searches all possible paths. Therefore, in 2014, Feng and Hirst [20] aimed to improve the discourse parser developed by Joty et al. [21] using Conditional Random Fields (CRF) as local classifiers. To reduce time, Feng and Hirst used a greedy bottom-up approach adopting four local models with two dimensions, (i) scope of the model, either intra-sentential or multi-sentential and (ii) the purpose of the model, either to determine structures or relations. Feng and Hirst [20] also introduced a novel feature in their recent parser which is a post-editing process that allows the RST tree to be modified based on top-down information such as the depth of the tree which can only be obtained after the tree has been fully built. This feature doubles the time required to parse the text. However, it is outweighed by the fact that it enhances the performance of the parser to close to 90% of human performance.

In 2015, Joty et al. [22] have rebranded their discourse parser together with their discourse segmenter [25] as CODRA[1] which stands for "a complete probabilistic discriminative framework for performing rhetorical analysis in accordance to RST" and modifies their parsing algorithm to search for the k most probable RST trees for the text. This modification allows the parser to store and track k-best candidates simultaneously instead of storing just a single best parse. When tested on the RST Discourse Treebank (RST-DT)[2] corpus, their k-best intra-sentential reranking parser improved the accuracy significantly by 13.45% (base accuracy is 79.77%) for 30-best. However, when tested on document level, their k-best parser did not give much improvement to the base accuracy (55.83%) with improvement as little as 1.91% for 30-best.

2.3. Ontology usage in RST discourse analysis

Whilst existing works on automatic discourse analysis have achieved high accuracy, none that we know have been tested on biomedical corpus especially on ultrasound reports. We argue that because of the nature of most ultrasound reports where sentences are often not grammatically complete, the conventional method of parsing text would not achieve high accuracy. Thus, it has been proposed to use an ontology to support the discourse parsing by using a rule-based approach. The usage of ontology in RST discourse analysis has not been explored much in previous works. One notable work would be the

[1] http://alt.qcri.org/demos/Discourse_Parser_Demo/.
[2] http://catalog.ldc.upenn.edu/LDC2002T07.

discourse parser developed by Bärenfänger et al. [26] which used two ontologies, a taxonomy of rhetorical relations (RRSet Ontology) and an ontology version of GermaNet[3] as a knowledge base for the discourse parser.

The RRSet ontology consists of 70 rhetorical relation types where 44 of them were basic types and the rest were subtypes. In their work, the ontologies were used to help determine the type of relation that exists between text spans. For example, in determining an ELABORATION subtype, ELABORATION-CONTINUATION where it is not indicated by any discourse markers, the ontology will be consulted to find synonymy or pertainymy between the two text spans. Their work is different from the one reported and carried out in this research. Instead of using an ontology to consult about the rhetorical relation types, our work uses the ontology to annotate the corpus with relevant classes. The annotated corpus will then be compared with a set of rules defined using classes from the same ontology to determine the type of rhetorical relations that exist. The next section will elaborate further on these rules.

3. Rules for identifying rhetorical relations

The total number of reports used in this study was 100.60 reports were randomly selected as training data to identify text span boundaries as well as the rhetorical relations that exist between them. The remaining 40 reports were used as testing data to evaluate the performance of both the segmentation and the rhetorical relation identification process. A text analysis was conducted on the three datasets in order to understand their characteristics by examining the average word count, average sentence count, average maximum sentence length, average minimum sentence length and average token size. The summary of these statistics is given in Table 1. We note that the training data mostly have fewer words per report as compared to the testing data. This means that the training data is generally shorter as it contains less sentences per report as well as less words per sentence. This was pure coincidence as the splitting of the data set into training and testing data sets was random.

The 60 training data reports were manually parsed to identify their structure and how the texts in these reports can be segmented into text spans. Discourse parsing, which is the process of identifying discourse relations between discourse units [20] was manually performed on these reports. They were first annotated with relevant classes from AUO before being segmented and drawn into RST trees and rhetorical relations were identified between their text spans. From the discourse parsing, seven rhetorical relations were identified which are PREPARATION, RESTATEMENT, JUSTIFY, ELABORATION, LIST, JOINT and CONTRAST relations based on the definitions by Mann [10].

PREPARATION, RESTATEMENT, JUSTIFY and ELABORATION relations are all mononuclear while LIST, JOINT and CONTRAST relations are multinuclear. These relations have been validated by an ultrasound expert to ensure their relevance in the production of ultrasound reports. These reports were then analysed again together with the relations that have been identified to design a set of rules that are able to identify all seven relations using discourse markers and relevant classes from AUO. In the following subsections, we will describe how these relations were identified and what they represent. In the later sections, we will assess the accuracy of these rules during the testing phase.

3.1. Preparation relation

PREPARATION relation is one of the seven relations identified in the sample ultrasound reports and serves as a precedence in order to prepare the readers to what they are about the read. In a normal text or paragraph, one example of a PREPARATION relation is between the title and the rest of the text where the rest of the text is the nucleus and the title is the satellite. This is the same in the case of ultrasound reports where 89 PREPARATION relations were identified in the 60 testing data and each title prepares the audience for the content. From the analysis, 87 out of 89 PREPARATION relations identified started with a title and followed by colons (:). Consider the example below:

S1. US Abdomen:
S2. Normal liver echo pattern with no focal lesion demonstrated.
S3. No evidence of gall stones or dilatation of the bile ducts.
S4. Both kidneys are normal in size and echo pattern with no mass lesion or evidence of obstruction.
S5. Conclusion:
S6. Normal examination.

In this example, it is clear that S1 is the title of the report which enables the readers to know that the report is about an abdominal ultrasound examination. S1 has a PREPARATION relation with a list of findings which are S2 until S4. S5 on the other hand has a PREPARATION relation with S6 where it prepares the readers for the conclusion. Therefore, a rule can be stated that if a certain text is followed by a semicolon then it has a PREPARATION relation with all the texts following it until the next PREPARATION relation or end of the paragraph.

3.2. Restatement relation

Another type of relation that was identified in the ultrasound reports is the RESTATEMENT relation where the writer re-expresses a sentence using another sentence. The RESTATEMENT relation was found in only 11 out of the 100 acquired sample ultrasound reports where only 2 were found in the 60 training data while the other 9 were found in the testing data. This is often signalled by the appearance of a “main or principal diagnosis” title together with a “conclusion” title in an ultrasound report. Consider the following example:

S1. US Abdomen:
S2. Normal appearances of the liver, gallbladder, kidneys, pancreas and spleen.
S3. The aorta was normal in caliber.
S4. The CBD was within normal limits (3 mm).
S5. Main or principal diagnosis:
S6. Normal abdominal ultrasound scan.
S7. Conclusion:
S8. No abnormality found.

This report details the list of findings of an abdominal ultrasound with S2, S3 and S4. It then reports the main finding in S6 using S5 as the title that prepares the readers. The report then gives a conclusion of the report in S8. It can be perceived that in this report, the “main or principal diagnosis” is actually already the conclusion of all the findings listed above which is a “normal abdominal ultrasound scan”. This statement is then repeated with a “conclusion” of “no abnormality found”. Thus, it can be inferred that the “conclusion” restates the “main or principal diagnosis” in the report. Most reports that have been acquired have the same pattern when both the “main or principal diagnosis’ and “conclusion” were recorded in the report. If one of these titles is absent, then the RESTATEMENT relation will not hold.

3.3. Justify relation

The JUSTIFY relation allows a reader to accept the nucleus based on the justification given by the satellite. In the context of an ultrasound report, the JUSTIFY relation gives reason as to why there is such finding. In identifying the JUSTIFY relation, Rule 1 denotes normal findings while Rule 2 denotes abnormal findings.

Rule 1: Normal Findings

[3] http://www.sfs.uni-tuebingen.de/GermaNet/.

Table 1
Summary of the text analysis for the three data sets.

	Average word count	Average sentence count	Average max sentence lenght	Average min sentence lenght	Average token size
Training data	62.88	9.15	14.28	1.53	6.48
Testing data	82.73	14.15	15.12	1.07	6.41
Both data	70.82	11.15	14.63	1.34	6.45

negative + biospecimen/disease/disorder/finding $\overset{justify}{\rightarrow}$ organ/ body part + positive

Rule 2: Abnormal Findings.

biospecimen/finding + organ/body part $\overset{justify}{\rightarrow}$ disease/disorder.

In Rule 1, if a text span contains a negative word (ex: no, without) and a word annotated with the class "biospecimen", "disorder", "disease" or "finding" followed by or preceded by another text span that contains a word annotated with "organ" or "body part" and a positive word (ex: normal, unremarkable) then it denotes a JUSTIFY relation in a normal finding. Consider the following example:

TS1. Liver is normal in echo pattern,
TS2. with no focal lesion.

In this example, the word "no" shows the absence of "lesion"; which is annotated with the class "finding" in TS2. This justifies the "normal" (positive word) condition of the "liver"; which is annotated with the class "organ" in TS1. Fig. 3 illustrates the RST tree for this relation.

Rule 2 denotes an abnormal finding. It states that if a text span contains a word annotated with the class "biospecimen" or "finding" and a word annotated with the class "organ" or "body part" followed by or preceded by a text span with a word annotated with "disorder" or "disease", then there is a JUSTIFY relation. Consider the following example:

TS1. Subtle dilatation of the collecting system of the left kidney,
TS2. hydronephrosis grade 1.

In this example, the "dilatation" of the "left kidney" in TS1 which was annotated with "finding" and "organ" respectively justifies the "disorder" found in TS2 which is "hydronephrosis" as seen in Fig. 4. Out of the 60 sample ultrasound reports, there were a total of 72 JUSTIFY relations found with 48 following these two rules. This number includes both direct JUSTIFY relations as well as JUSTIFY relations that were nested into either themselves or other relations. Another 24 JUSTIFY relations found in the reports were the relation between a list of findings and the conclusion of the report where the list of findings justifies the conclusion. This is for instance in the example given in Section 3.1 where the findings in S2 until S4 justified the conclusion in S6.

3.4. Elaboration relation

The ELABORATION relation gives further information on the text span. Unlike other RST relations, ELABORATION relation has many subtypes as it can be found in many text spans without being signalled by any discourse markers [26]. There are several subtypes of ELABORATION relation that could exist in an ultrasound report where each has different signals or cue words and gives elaborations on different aspects of the text span.

However, there are three prominent subtypes that have been identified in the 60 sample ultrasound reports used as training data. The first subtype is an ELABORATION relation that gives further information on the location of a finding. This is often signalled by a spatial qualifier such as the word "within". This type of ELABORATION relation is represented by Rule 3.

Rule 3: Spatial Qualifier.

spatial qualifier + organ/body part $\overset{elaboration}{\rightarrow}$ biospecimen/finding.

In Rule 3 it is stated that a combination of a spatial qualifier and an organ or body part elaborates a biospecimen or a finding by letting us know the location of the biospecimen or finding. To understand this better, consider the following example:

TS1. No gall stones are seen,
TS2. within the gallbladder.

In this example, we can see from the RST tree in Fig. 5 that TS1 can be understood by the reader without TS2. However, TS2 elaborates further on the finding in TS1 by letting the reader know the exact location of where the "gall stones" were found. The second subtype of ELABORATION relation is defined in Rule 4.

Rule 4: Unit of Measure.

unit of measure $\overset{elaboration}{\rightarrow}$ biospecimen/finding/organ/body part.

This subtype of ELABORATION relation is the easiest one to recognise and gives further information on the measurement of the "finding" or the "organ". It is often signalled by words such as "measures" and "measuring" or parenthesis that contains a measurement or simply a unit of measure such as "mm" and "cm". Consider the following example:

TS1. The gallbladder contains a single stone,
TS2. measuring 7 mm.

Just like the example before this, TS1 can be understood without TS2 where readers would know that there is a "single stone" in the "gallbladder" (see Fig. 6). TS2 however extends the information on TS1 by giving the measurement of the stone. The last subtype of

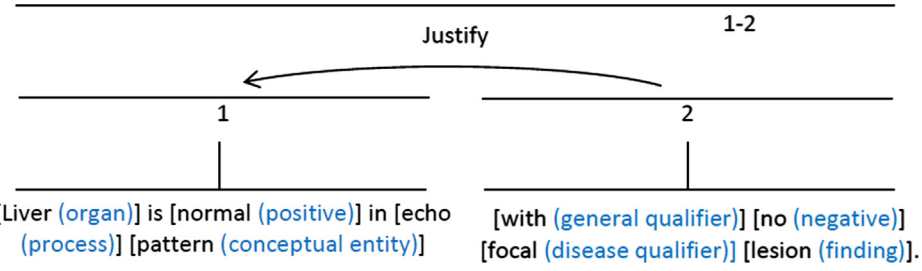

Fig. 3. Example of a JUSTIFY relation in a normal finding.

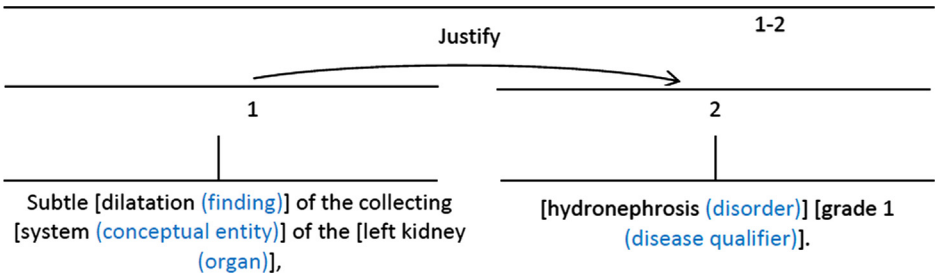

Fig. 4. Example of a JUSTIFY relation in an abormal finding.

ELABORATION relation is when a "finding" is elaborated by another "finding" as stated by Rule 5.

Rule 5: Finding within a finding.

biospecimen/finding $\overset{elaboration}{\rightarrow}$ biospecimen/ finding + organ/body part.

Consider the following example as illustrated in the RST tree in Fig. 7:

TS1. Multiple small gall stones at GB neck,
TS2. with no signs of inflammation.

In this example, TS1 reports a "finding" of "multiple small gall stones at GB neck". Instead of elaborating on where the "gall stones" are or the size of it, this subtype of ELABORATION relation describes another finding in relation to the finding in TS1. Instead of giving the reason for the finding in TS1, TS2 further elaborates the finding with another finding which makes it different from the JUSTIFY relation explained in the previous section.

3.5. List relation

The LIST relation is a multinuclear relation which gives a list of items to the readers. In a multinuclear relation, each text span is a nucleus and plays an important role. There are no specific rules in identifying this relation. However, in ultrasound reports, it is assumed that all the text spans that follows a title, is a list of that title. For example, all text spans following the "US Abdomen:" title up until the text span before the title "Conclusion:" is assumed to be the list of findings for the abdominal ultrasound report.

3.6. Joint relation

The JOINT relation is another multinuclear relation that can be found in an ultrasound report. This relation is important in ensuring that when a text is segmented, it does not lose its meaning. The JOINT relation was found 157 times which makes it the most used relation in this research. A JOINT relation is often signalled by "AND" or "OR". However, not all text span segmented with "AND" or "OR" denotes a JOINT relation. Rules 6, 7 and 8 are some examples of identifying the JOINT relation.

Rule 6: Joint relation between two organs.

organ $\overset{joint}{\leftrightarrow}$ organ + finding

Rule 7: Joint relation between two biospecimens.

biospecimen $\overset{joint}{\leftrightarrow}$ biospecimen + organ

Rule 8: Joint relation between two anatomy qualifiers.

anatomy qualifier $\overset{joint}{\leftrightarrow}$ anatomy qualifier.

Consider the following example:

TS1. There were multiple calculi,
TS2. and sludge within the gallbladder.

In the example given in Fig. 8, TS1 contains a "biospecimen" but without an "organ" whereas TS2 contains a "biospecimen" together with an "organ". This denotes a JOINT relation where the "organ" in TS2 is jointly referred to by the "biospecimen" in both TS1 and TS2. With the JOINT relation, it is possible to assert that there are two findings in the sentence which are "There were multiple calculi within the gallbladder" and "There were sludge within the gallbladder" without losing any important information. Most sentences that contain the discourse markers "AND" and "OR" signals a JOINT relation unless it follows Rule 9.

Rule 9: "AND"/"OR" that is not a joint relation.

TS1: organ + finding
TS2: organ + finding

If each text span has a pair of "organ" and "finding" or "organ" and "biospecimen", then it does not have a JOINT relation since both text spans have enough information without needing to share any other information from the other text span. This is instead a LIST relation. The following example illustrates this case:

TS1. The bile ducts were not dilated,
TS2. and the liver texture appears satisfactory.

In TS1, the "bile duct" is an "organ" and "dilated" is a "finding"

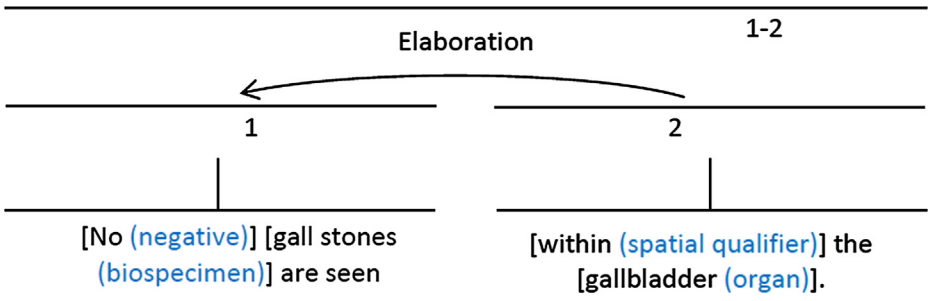

Fig. 5. Example of an ELABORATION relation with spatial qualifier.

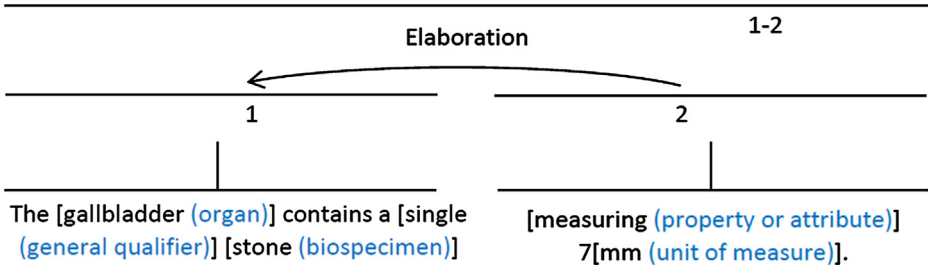

Fig. 6. Example of an ELABORATION relation with unit of measure.

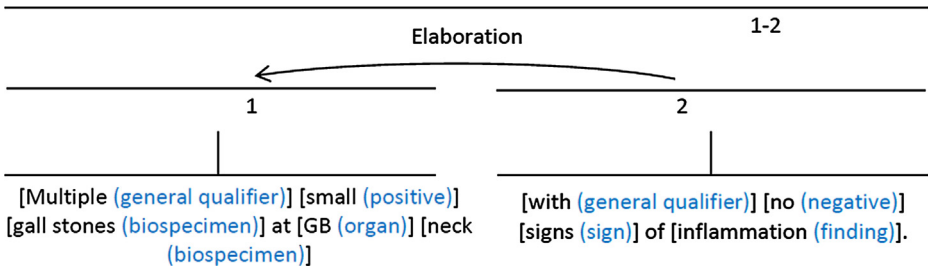

Fig. 7. Example of an ELABORATION relation where a finding elaborates another finding.

while in TS2, the "liver" is an "organ" and "satisfactory" is a "positive finding". Since each text span has its own pairs of organ and finding, thus this is not a JOINT relation but only a list of findings (see Fig. 9).

3.7. Contrast relation

The CONTRAST relation is another rhetorical relation that can be found in ultrasound reports. However, the frequency of use of this relation is small as only 7 relations were found. The CONTRAST relation is often signalled by the word "but" and "otherwise" and it gives contradicting findings.

All the examples provided in this section demonstrated that it is possible to identify rhetorical relations using a rule-based approach and an ontology. These rules were generated from the set of 60 sample ultrasound reports used as training data. The next section will explain further on how these rules can be used together with the ontology to perform discourse parsing on medical reports.

4. Applying RST and ontology in medical discourse parsing

Discourse parsing is the task of segmenting texts based on certain word boundaries in order to find rhetorical relations between them which in turn signals the coherence of the text. Bärenfänger et al. considered discourse parsing as an iterative task whereby annotated texts are submitted as an input that produces an output text with another annotation [26]. Traditionally, discourse parsing was carried out using texts that have been annotated with POS tags on corpus related to newspaper and magazine articles as well as academic journals. Bärenfänger et al. [26] however, tried to implement discourse parsing with a different approach. In their work, two OWL ontologies were used to perform certain tasks of discourse parsing. In this research, we aim to use a similar approach whereby the AUO is used instead of the POS tags or treebanks in annotating texts as well as in implementing the rhetorical rules defined in previous sections. The next subsections will elaborate further on the ontology that was used as well as the process taken in applying RST and ontology in parsing ultrasound reports.

4.1. Annotating relevant classes

The first step in executing discourse parsing using RST and AUO is to annotate the reports with relevant classes. Not all words will be annotated since we are only interested in classes that are involved in the rhetorical relation rules. Punctuations such as full stops (.), colons (:), and commas (,) were first separated from the text before splitting the paragraph into single spans.

The classes in the ontology have at most three words combination. Therefore, once the paragraph is split into single words, the next step is to combine the single words into three words combination and it will then be compared with the classes in AUO to see if there is a match. Once a match is found, the text will then be annotated with the class or its parent or ancestors that are relevant to the rules. The following ultrasound report will be used throughout this section as an example to explain the implementation of RST and ontology in medical discourse parsing:

> "US Abdomen: There were multiple tiny calculi in the neck of the gallbladder. The CBD appeared normal. The pancreas was obscured by gas. No abnormality was seen in relation to the spleen or kidneys.

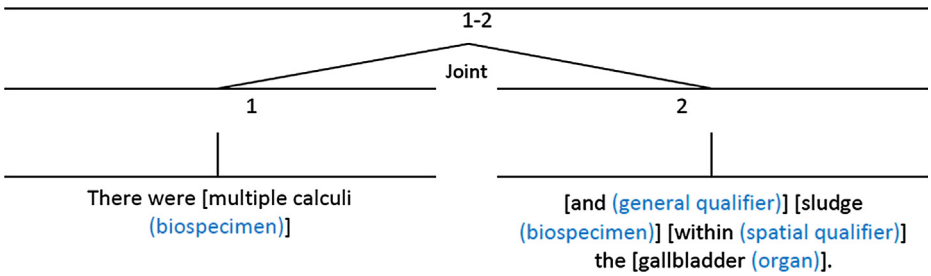

Fig. 8. Example of a JOINT relation.

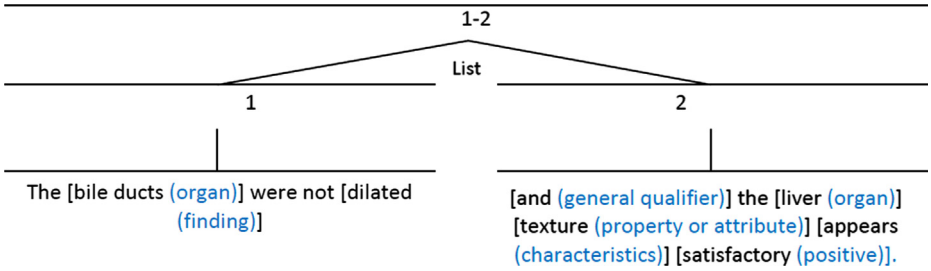

Fig. 9. Example of the occurence of AND/OR which does not signals a JOINT relation.

"[US Abdomen | #Imaging_Technique|]: There were multiple tiny [calculi | #Biospecimen|] in the [neck | #Body_Region|] of the [gallbladder | #Organ|]. The [CBD | #Organ|] [appeared | #Characteristic|] [normal | #Clinical_or_Research_Assessment_Answer|]. The [pancreas | #Organ|] was [obscured | #Visibility_Descriptor|] by gas. No [abnormality | #Finding|] was seen in relation to the [spleen | #Organ|] or [kidneys | #Organ|]. There was no [ascites | #Disease_or_Disorder|]. Conclusion: Tiny [calculi | #Biospecimen|] [within | #Spatial_Qualifier|] the [neck | #Body_Region|] of the [gallbladder | #Organ|]."

Fig. 10. Ultrasound report annotated with relevant AUO classes.

There was no ascites. Conclusion: Tiny calculi within the neck of the gallbladder."

The report is first annotated by the parser with relevant AUO classes as shown in Fig. 10. This can be exemplified using the word "CBD" where the direct parent of "CBD" is "Extrahepatic Bile Duct". However, because we are more interested in knowing the general class of "CBD", it is annotated with the class "Organ" instead, which is the ancestor of both "CBD" and "Extrahepatic Bile Duct'. This annotation makes it possible to compare the text spans in the report with the rhetorical relations rules to identify the existing relations. The same process will then be repeated for two words combination and one word to find the exact match of the words or their synonyms. Once the matching process is completed, the annotated single words will be combined to produce a complete annotated paragraph for further processing.

4.2. Segmenting ultrasound reports

The next step in parsing the ultrasound reports would be to segment the annotated paragraph syntactically. Soricut and Marcu [19] defined discourse segmentation as a process where texts are split into non-overlapping text spans with each having their own roles in rhetorical relations. The segmentation task presented in this research loosely follows Marcu's shallow analyser [13] which uses discourse markers instead of traditional POS tagging technique to segment texts into smaller text spans.

In our work, the same technique was used where the segmentation was conducted based on punctuations and signal words which act as the text span boundaries. However, in segmenting ultrasound reports, we found that most texts in the ultrasound reports are not well-formed. Most reports were written in short but straight to the point sentences at the cost of sentence structures being sometimes disregarded. Even so, this technique can still be applied in our work because the rhetorical relation rules presented in the previous sections serve as a constraint in determining the relations between the segmented text spans. Table 2 states the list of punctuations and signal words that have been recognised from the 60 sample ultrasound reports used for training. It also gives the corresponding relation that these punctuations and words often signal.

The process of segmenting paragraphs using this technique is quite straightforward. The paragraph will first be parsed to detect the occurrence of punctuation that follows another punctuation (e.g. ?.), a signal word that follows another signal word (e.g. compatible with, measure within), a punctuation that follows a signal word (e.g., and) or vice versa. If there are any, a dash (–) symbol will be inserted between the words as a flag so that oversegmentation does not occur. The parser will then try to locate a punctuation and set a boundary after the punctuation to be split. Next, a signal word is located and a boundary will be placed before the signal word for splitting. Fig. 11 shows an example of a segmented and annotated ultrasound report. The example illustrates that in TS1, the text span boundaries have been set after the colon (:) symbol whereas in TS6, the boundaries of the text span was before the signal word "or".

Table 2
Punctuation/signal words and the corresponding rhetorical relations.

Punctuation/signal words	Corresponding rhetorical relation
Fullstop(.)/Question Mark (?)	End of sentence
Colon (:)	PREPARATION, LIST
Comma (,)	JOINT, JUSTIFY, ELABORATION
Parenthesis	ELABORATION, JUSTIFY
With/compatible with/in keeping with/ associated with	JUSTIFY
Suggest/suggestive of	JUSTIFY
Could be	JUSTIFY
Measuring/measuring with/measures	ELABORATION
However/but/otherwise	CONTRAST
And/or	JOINT
Which/within	ELABORATION

4.3. Identifying rhetorical relations

After texts segmentation, the next task is to identify relations that exist between these text spans. Marcu [15] performed this by hypothesising the relations based on the appearances of discourse markers. If a discourse marker is absent, the co-occurrence of similarity will then be measured [15]. In this research, discourse markers were also used in identifying possible rhetorical relations that exist between text spans. The difference is that this research applied a rule-based approach. Discourse markers are sometimes ambiguous to which rhetorical relation they are signalling [15]. Hence, the rule-based approach was proposed to reduce this ambiguity. Even though discourse markers were highly used in this work, it does not depend on them entirely. Discourse markers signals rhetorical relation, however, the ontology classes and relation rules confirm it.

In identifying rhetorical relations between text spans, our discourse parser takes the annotated and segmented text spans as an input and outputs a list of all possible relations between these text spans. Fig. 12 shows the output of the discourse parser where the discourse relations have been identified for the ultrasound report example used in this section. The text spans will be parsed one by one and the parser will first look for the existence of any PREPARATION relation which is often signalled by colons (:). Almost all sample ultrasound reports that were used as training and testing have a title as its first text span. Most

Segmented Text

TS1. [US Abdomen | #Imaging_Technique|] :

TS2. There were multiple tiny [calculi | #Biospecimen|] in the [neck | #Body_Region|] of the [gallbladder | #Organ|] .

TS3. The [CBD | #Organ|] [appeared | #Characteristic|] [normal | #Clinical_or_Research_Assessment_Answer|] .

TS4. The [pancreas | #Organ|] was [obscured | #Visibility_Descriptor|] by gas .

TS5. No [abnormality | #Finding|] was seen in relation to the [spleen | #Organ|]

TS6. or [kidneys | #Organ|] .

TS7. There was no [ascites | #Disease_or_Disorder|] .

TS8. Conclusion :

TS9. Tiny [calculi | #Biospecimen|]

TS10. [within | #Spatial_Qualifier|] the [neck | #Body_Region|] of the [gallbladder | #Organ|] .

Fig. 11. Sample annotated and segmented ultrasound report.

Discourse Relations

TS1 PREPARE LIST (TS2-TS7)
TS5 JOINT TS6
(TS1-TS7) JUSTIFY (TS8 PREPARE TS9-TS10)
TS10 ELABORATE TS9

Fig. 12. List of relations identified using ontology and the rhetorical relation rules.

reports have between one to two titles that prepare the readers to understand its content and whenever there is a conclusion, it is often the last title in a report. The discourse parser takes particular attention on conclusions because they do not only have a PREPARATION relation with the text spans that comes after it but it also has a JUSTIFY relation with the text spans before it. This is because all the findings stated before the conclusion justifies it. This is clearly demonstrated in the complete RST tree for the example ultrasound report built using O'Donnell's RST Tool [27] (see Fig. 13) where TS1 - TS7 JUSTIFY TS8 - TS10.

Once the PREPARATION relation has been recognised, the parser will then identify the LIST relation which is all the text spans that comes after the title up until before the next title or until the last text span, whichever comes first. Then, the parser will look for the JOINT relation which is often signalled by "and", "or" and comma (,). Marcu [13] in his work has considered "and" as a highly ambiguous discourse marker. Even in our work, all three "and", "or" and comma (,) discourse markers were the most common signals whereby there were at least 165 occurrences. However, not all occurrences signal a JOINT relation. Hence, before a JOINT relation can be identified, the parser will first need to recognise the occurrences of "and" and "or" which are not a JOINT relation before attempting to identify the JOINT relations. Finally, the discourse parser will search for the remaining three relations which are JUSTIFY, ELABORATION and CONTRAST.

In most cases, a text span has a rhetorical relation with a text span immediately before or after it. A text span with a mononuclear relation often only has a relation with either a text span before or after it. This is unless it has a nested relationship whereby a text span has a relation with an immediate text before it and both of them have a relationship with another text span. The following sentence, that has been segmented into three text spans, can be used as an example to illustrates this: "[The gallbladder is well distended]$_{TS1}$ [and contains at least 4 stones]$_{TS2}$ [measuring just over 1 cm in diameter.]$_{TS3}$". In this case, TS2 has an ELABORATION relation with a text span right after it i.e TS3 elaborates TS2. Both TS2 and TS3 then have a nested relation with TS1 where TS1 JOINT (TS3 ELABORATE TS2). As for a text span with a multinuclear relation, it is common for the text span to have the same relation with both text spans before and after it. From the training and testing work completed on all the sample ultrasound reports, we did not come across a text span, say TS1, which skipped another text span, say TS2 and has a relation with TS3. The evaluation of both the segmentation and rhetorical relation processes will be presented in the next section.

5. Transforming free-form reports to structured form

The medical discourse parser presented in the previous section serves as the basis in transforming the free-form reports to structured form. The first step that needs to be taken to perform this transformation is to manually pre-process the reports to remove the many obvious errors caused by the radiologists. Once this has been completed, the reports will be submitted to the system to be transformed into a structured form. The system will also perform a pre-processing phase whereby all the main titles of the reports such as "US Abdomen" and "Ultrasound: Abdomen" will be removed because the produced reports

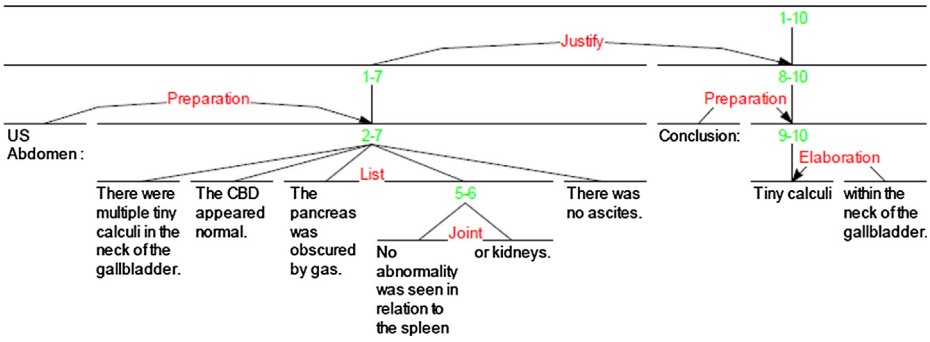

Fig. 13. RST tree of the sample ultrasound report.

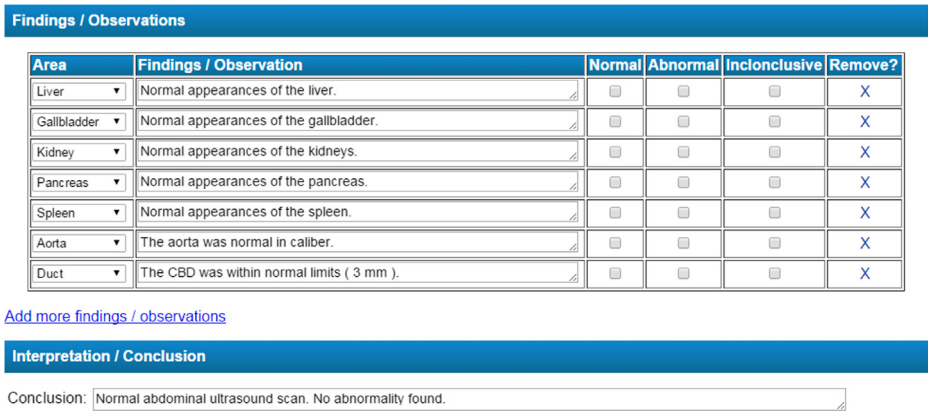

Fig. 14. A structured report generated from a free-form report.

will already have a standard title.

When a free-form report is submitted to the system, the relevant words in the report are annotated with classes from AUO as explained in Section 4.1. Next, the system will identify the type of sentences that have been split from the annotated paragraph whether it is a clinical history, finding/ observation, conclusion or further management based on signal word such as "history", "conclusion" and "suggest" as well as the PREPARATION and RESTATEMENT relations. All the sentences which were of these three types of information will be extracted from the free-form report and moved under suitable headings in the structured report.

Once this has been completed, the remaining sentences will be regarded as findings and observations. The aim of using RST in the transformation process is to group the findings in the report according to the area examined as shown in Fig. 14. For example, if the sonographer has examined the liver, pancreas and spleen of the patient, the findings and observations should be recorded according to the area examined instead of writing everything in one paragraph. In order to do this, another two out of the seven rhetorical relations presented in Section 3 were used which were the JOINT and ELABORATION relation.

The JOINT relation is used to separate several areas examined which were initially reported in one sentence into several separate sentences. The JOINT relation was the most used relation in transforming the free-form reports into structured reports. For example, a report stated that "The kidney, spleen and pancreas is normal". In this example, it was clear that there were three areas being examined which are "kidney", "spleen" and "pancreas". When RST is applied to this sentence, it will recognise that there exist a JOINT relation between the three organs because of the cue word "and" as well as the commas between the organs. This allows the system to separate the three organs into separate sentences without losing its observation. The system will start each sentences with the three different organs and because they have a JOINT relation, they will share the same observation which is "is normal". This transforms the initial sentence into three sentences which are "The kidney is normal", "Spleen is normal" and "Pancreas is normal".

The other relation being used in the transformation process is the ELABORATION relation. This relation is important in ensuring that any other extra information was not lost when it is being separated or joined with another sentence. For example, consider the sentence "Normal appearance of spleen (measuring 12.6 cm), head and body of pancreas and aorta (measuring 1.4 cm inner to inner)". This sentence consisted of three areas which were the spleen, head and body of pancreas and aorta. There was only one observation which was "normal appearance". A JOINT relation that exist between the three areas will separate the sentence into three which are "Normal appearance of spleen", "Normal appearance of head and body of pancreas" and "Normal appearance of aorta".

However, ELABORATION relation will allow the measurements of the organ to also be retained whereby the two sentences will become "Normal appearance of spleen (measuring 12.6 cm)" and "Normal appearance of aorta (measuring 1.4 cm inner to inner)". Once all the findings have been separated into the areas examined, they will be presented under the "Findings/ Observations" heading in the structured report. Finally, the complete structured report will be displayed to the radiologist.

6. Results and discussions

This section will first introduce a pre-processing phase and explain what was accomplished during this phase and how it impacts on the overall result of the parsing. It then presents the result of the report segmentation and identification of rhetorical relations where the system parsing was compared to the manual parsing performed by NLP experts. For the training data, the manual parsing was completed by a single expert. However, for the testing data, three experts were asked to manually parse all 40 sample ultrasound reports. The parsings which have a two third majority were then selected to be compared to those produced by the system.

6.1. The pre-processing phase

In the initial stage, RST and AUO were applied to the ultrasound reports without any cleaning or pre-processing of the data. After the annotation and segmentation stages were performed, the accuracy rate was found to be quite low. The reason for this was identified to be caused by human errors such as spelling mistakes, missing punctuations, missing spaces between words as well as abbreviations and symbols that are not recognised by the ontology. The number of occurrences of each error found in both training and testing data is summarised in Table 3.

Even though the total number of occurrences of these errors were not alarmingly high, it still had a huge impact on the evaluation result.

Table 3
Total of human errors found in 100 sample ultrasound reports.

Types of human errors	No of occurrences in training data	No of occurrences in testing data
Spelling mistakes	3	4
Missing or wrong punctuations	7	4
Missing spaces	1	5
Unrecognised abbreviations and symbols	1	5
Incorrect sentences	1	5
Total	13	23

Therefore, we have decided to include a pre-processing phase before submitting the ultrasound reports to the system to be parsed. During the pre-processing phase, we have manually corrected spelling mistakes as well as adding spaces between words and punctuations such as full stops and commas in places we believed appropriate. For example, the sentence "The spleen the kidney and pancreas is normal" should be segmented into three text spans which are "the spleen", "the kidney" and "and pancreas is normal". However, because there is no comma between "the spleen" and "the kidney", the system failed to segment it.

There were also several instances of incomplete sentences and two sentences that were combined without using any conjunction. An example of an incomplete sentence is "The spleen is of normal size and echotexture, measuring …". The radiologist writing the report has the intention to give the measurement of the organ but somehow did not. In this case, a full stop was included after the word "echotexture" to make the sentence more meaningful. Other than that, symbols and abbreviations were also automatically converted into words using regular expressions in the system. For example, symbols such as "/" and "&" were automatically changed to "or" and "and" respectively although this could be automated for future version of the system. The introduction of the pre-processing phase has improved the accuracy rate of the parser significantly. The results will be presented in the next section.

6.2. Report segmentation result

The accuracy of the system's report segmentation based on RST was evaluated in two stages. The first stage was without the introduction of the pre-processing phase while the second stage included the pre-processing phase. The evaluation was undertaken separately on the 60 training data reports and the 40 testing data reports where both were compared to a gold standard which is the experts' manual parsing. Both the training and testing data were evaluated by comparing the total number of text spans that were produced by the system for each report against the total number of text spans that were produced by the experts' manual parsing. If the total number of text spans segmented by the system matches the total number of text spans segmented by the experts, then the report is viewed as accurately segmented. Table 4 displays the result of the evaluation.

Without pre-processing, the training data achieved a 78.33% accuracy where 47 out of the 60 reports were segmented correctly. However, as for the testing data, the accuracy rate was very low where only half of the reports were segmented correctly. When the pre-processing phase was introduced, significant improvement of the accuracy was observed for both sets of data especially for the testing data. The accuracy rate increased from 78.33% to 88.33% for the training data and from 50% to 82.5% for the testing data.

The reason for the increment was due to a lot of reports which were under segmented in the initial stage where pre-processing was not conducted. This means that there were certain boundaries that the experts believed should be segmented but because there are no signal words or punctuation, the system failed to identify them. This is demostrated in an example of the sentence "Normal calibre aorta measuringe 1.5 cm in diameter." This sentence should be segmented into two text spans which are "'Normal calibre aorta" and "measuring 1.5 cm in diameter" where there is an ELABORATION relation between them. However, because the word "measuring" was misspelled, the system failed to recognise it and did not execute the segmentation. Accordingly, the percentage increased significantly when the pre-processing phase was introduced.

Table 4
The accuracy of report segmentation with and without pre-processing.

	Without pre-processing phase		With pre-processing phase	
	Accurately segmented reports	Percentage (%)	Accurately segmented reports	Percentage (%)
Training data	47	78.33	53	88.33
Testing data	20	50.00	33	82.50
Both data	67	67.00	86	86.00

Nevertheless, the accuracy rate was still less than 90%. This was substantially contributed by the oversegmentation of texts caused by the appearance of discourse markers that signals text boundaries. Oversegmentation of texts can be demonstrated in an example of the sentence "The CBD was within normal limits". In this sentence, the word "within" is a signal word for the ELABORATION relation. Therefore, it serves as a text boundary whereby the sentence will be split before the word "within". This produces two text spans which are "The CBD was" and "within normal limits". The segmentation of this sentence should not happen as both text spans do not have any meaning on their own and therefore, could not act as the nucleus in the relation.

The same is exhibited in the sentence "Spleen with normal size." where it was split into two text spans, "Spleen" and "with normal size". This is because of the cue word "with" which usually signals the JOINT relation. This implies that, for the problem of oversegmentation to be reduced, further rules should be defined to avoid sentences to be segmented in certain cases even though they contain a signal word.

6.3. Rhetorical relation identification result

The identification of the rhetorical relations between text spans in the sample ultrasound reports was also evaluated separately between the 60 training data and the 40 testing data. The evaluation was performed using the common information retrieval measurement methods which are precision, recall and F-score. In this research, precision will be the ratio of the relevant relations identified by the system to the total number of relevant and irrelevant relations identified by the system. Recall on the other hand will be the ratio of the relevant relations identified by the system to the total number of relevant relations identified by the experts. Finally, the F-score will be measured as the harmonic mean of both the precision and recall. The number of rhetorical relations identified by the experts and the system is summarized in Table 5 while the evaluation results are given in Table 6.

In the 60 training data, a total of 325 relations were identified by the experts. When these reports were submitted to the system, it managed to identify 310 relations between the text spans in the reports. Out of the 310 relations identified, 302 relations were similar to those identified by the experts in their manual parsing. This has resulted in a very high precision and recall of 97.4% and 92.9% respectively. As a result, the training data achieved an F-score of 95.1%.

The system has failed to correctly recognise 23 relations that were identified by the experts. This is caused by the fact that there were certain words which were not annotated with the relevant classes from the AUO. Two text spans which are "The liver is slightly hyperechoic" and "in keeping with fatty liver disease." can be taken as an example to illustrate this problem. There exist a JUSTIFY relation between the two text spans where the first text span justifies the second. For a JUSTIFY relation in an abnormal finding to be recognised, one of the words in the text span needs to be annotated with the class "finding" or a "biospecimen". However, because the word "hyperechoic" was not associated with any of the two classes, the system failed to recognise that it justified the existence of a "liver disease".

The precision and recall for the testing data were slightly lower compared to the training data which were 91.3% and 87.6% respectively. It also has a lower F-score of 89.4%. In the testing data, the experts have managed to identify 274 relations between the text spans in the 40 reports while the system managed to identify only 263 relations. Out of these relations identified by the system, 240 relations were similar to the ones identified by the experts. This means that the system failed to correctly recognise 34 relations between the text spans.

Table 5
The number of rhetorical relations identified by the experts and the system.

	No of relations identified manually	No of relations identified by the system	No of similarities
Training data	325	310	302
Testing data	274	263	240
Both data	599	573	542

Table 6
Evaluation of the rhetorical relation identification process.

	Precision (%)	Recall (%)	F-score (%)
Training data	97.4	92.9	95.1
Testing data	91.3	87.6	89.4
Both data	94.6	90.5	92.5

A reason for this was the same as in the training data whereby there were some words which were not annotated with the relevant classes in AUO. In addition to this, another reason for the system failing to correctly identify all 34 relations was because most of these relations were a first occurrence. Therefore, the system was not trained to recognise such patterns of the relations. Accordingly, in order to improve the score of the system, the ontology will need to be enhanced to accommodate the words which were not annotated. In addition to that, more sample ultrasound reports should also be added to the training data so that the system is able to recognise more patterns of the relations.

In both the training and testing data, there were several relations that have been identified by the system but not by the experts. This happened mainly because of oversegmentation. One example is the sentence “Spleen with normal size”. As mentioned in the previous section, this sentence was wrongly segmented into two text spans which were “Spleen” and “with normal size”. This has resulted in the system identifying an ELABORATION relation between the two text spans because there is a measurement and no mention of an “organ” in the second text spans and there is an “organ” in the first text span without any “findings”, “disease” or measurement. To avoid such problems in the future, the rules used to perform the segmentation need to be improved to reduce oversegmentation.

6.4. Discussions

An ultrasound report was submitted to CODRA in order to compare its performance with our system. The output from the discourse segmentation was 13 text spans. This, however, was not accurate as it only segments sentence boundaries signalled by full stops without taking into consideration other cues such as commas and colons as well as signal words like “and”. CODRA is also able to produce a complete RST tree that shows the rhetorical relations identified between the text spans. However, when the report was submitted, CODRA produces an RST tree that identified only two rhetorical relations between the text spans which were TOPIC-COMMENT and ELABORATION relation as depicted in Fig. 15. This confirmed our initial assumption that CODRA will fail to correctly segment and identify rhetorical relations in a medical ultrasound report as it was not developed for this specific purpose. Hence, it was not necessary to submit all the reports and conduct a full comparison between our system and CODRA.

In evaluating the transformation of the free-form reports to structured form, a pair of ultrasound specialists were approached. In general, both specialists agreed that the translation of the free-form reports to structured form was good. Most information in the free-form reports has been transferred correctly into the structured form. However, there were four issues that have been raised.

First, the specialists mentioned that the structured reports have faithfully replicated the errors contained in the free-form reports. For example, there was a finding “normal size is normal parenchyma of the liver” where the specialists believed that this was not making any sense. Another is “16 mm solid foci seen inferior to liver” where the specialists argued that it was grammatically wrong and that the better word to be used is “focus” which is a singular of “foci”. This issue was indeed interesting and should not be taken lightly so that the final report produced is of good quality.

In addition to this, another issue that was raised was in terms of the area allocation of the findings. The specialists argued that the finding “No evidence of gallstones or dilatation of bile ducts” should be grouped under the area “Gallbladder” instead of “Duct”. Following this, necessary changes have been made to ensure that the findings are appropriately grouped. The third issue that has been brought up by the specialists was regarding the related information that would have been better grouped into the same area. This is for instance in the finding “Liver echogenicity appears normal. No hepatic mass lesion is seen” which should not be separated into two sentences. This is indeed a limitation of our system for the moment because it failed to recognise that the word “hepatic” is related to the word “liver”. The limitation can later be improved by enhancing AUO to include this information. Finally, the specialists mentioned the difficulty of classifying the area examined of the finding “No intra- or extra-hepatic biliary dilatation” where it could fall into both the area “Liver” as well as “Duct”. Further discussions on this matter should be initiated for a decision to be made whether this finding would better be grouped under the area “Liver”, “Duct” or both.

As a conclusion, the specialists believed that the translation of the free-form report to structured form was executed appropriately. Although there were several errors that have been detected, these errors were mostly reproduced from the initial free-form reports which were not within the control of this study. Errors caused by the system however, could be reduced by making minor changes to the codes as well as by enhancing the information available in AUO.

The results obtained in this research in our view justify the use of discourse analysis and the task is not a simple information extraction (IE) exercise. In most situation, this involves explanation and justifications of some facts and findings that a simple IE will fail to deal with in an appropriate way. Furthermore the full information spans over one or more sentences. We agree, that this is not the traditional setting of a discourse where RST is normally used, but nevertheless it is a case of the use of free text and some texts are long enough and well structured to be considered as a discourse.

The implementation of RST alongside an ontology such as AUO in the medical domain is a new approach that has not been explored before. For this reason, there are a lot of issues that can be discussed in order to improve its accuracy. An example of such an issue is regarding the usage of words in the reports. This has been brought up by the ultrasound specialist where words such as “Fossa” can be understood depending on the context of the report. However, if it was separated in one sentence, the word will be meaningless.

Consequently, the implementation of RST using an ontology in the medical domain can be massively improved with further involvement of experts from both the NLP and the medical fields. Limitations of the current implementation can also be further reduced if more sample ultrasound reports can be gathered.This is because they will allow for the system to be trained using more data and at the same time learn new patterns of rhetorical relations.

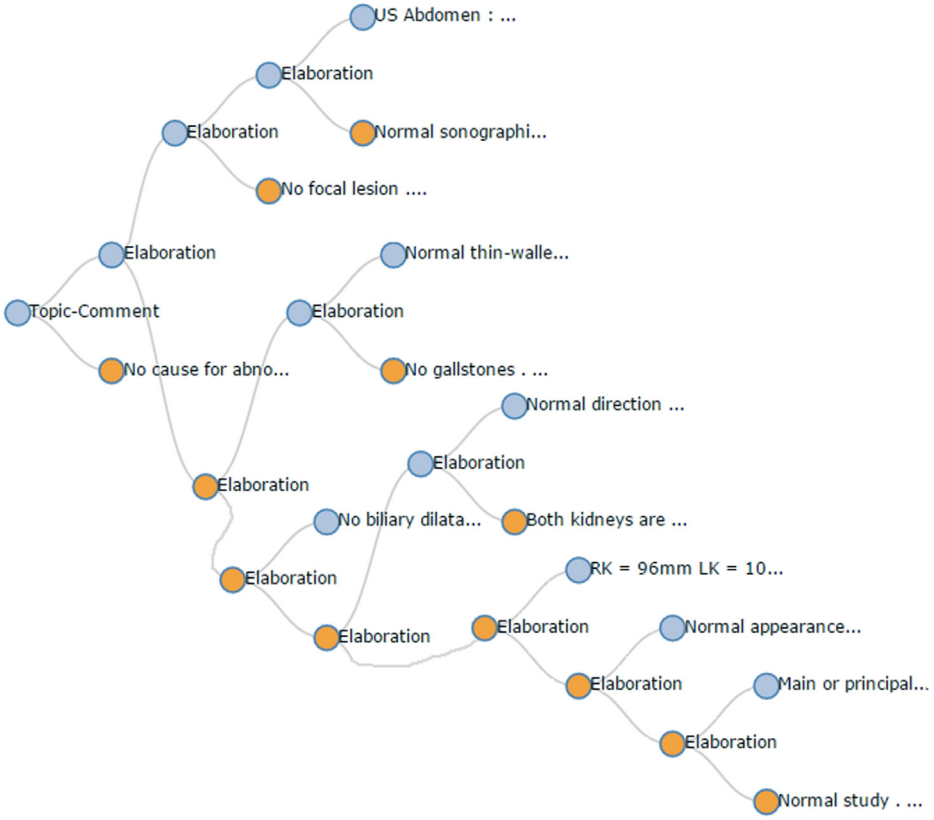

Fig. 15. RST tree produced by CODRA.

7. Conclusion

RST is a well-established theory in computational linguistics used to recognise relations between text spans in a coherent text where it uses discourse markers and sentence structure to recognise relations. This paper presented our approach in using an ontology and discourse markers to identify RST relations in ultrasound reports. Several rules have been designed as a guide to segment and recognise rhetorical relations in ultrasound reports. These rules have been designed based on an analysis performed on 60 out of the 100 sample reports collected. From the analysis, seven rhetorical relations have been identified which are PREPARATION, RESTATEMENT, JUSTIFY, ELABORATION, LIST, JOINT and CONTRAST. These seven rhetorical relation were then applied in the discourse parsing of the sample ultrasound reports. From the evaluation, the system achieved an accuracy of 88.33% and 82.50% for both the training and testing data for the segmentation process. As for the identification of rhetorical relations, the system achieved a precision of 97.42%, a recall of 92.92% and an F-score of 95.12% for the training data and a precision of 91.25%, a recall of 87.59% and an F-score of 89.38% for the testing data. This proves that implementing an ontology in the discourse parsing process of RST made it possible for RST to also be applied on medical reports. This then resulted in the possibility of automatically transforming free-form reports to structured form.

Although, the research presented in this paper is very promising and received positive reviews from radiologists, it has currently some limitations that will require improvements in future developments. We particularly highlights the following:

- The number of reports used in the current study is small and for future developments, a much larger sample is needed.
- Current trades favour the use of deep learning over rule based approaches. However, the use of deep learning requires large data sets and sometimes special hardware to support the complex computation behind it. However, the use of deep learning with the size of the dataset we had access to and used in this research together with the absence of a dedicated annotated corpus is not adequate and may not lead to improvements in the results.

This would be a completely different study that we are planning to continue as the obtained results are positively received by the health professionals.

Conflict of interest

We wish to confirm that there are no known conflicts of interest associated with this publication.

Acknowledgment

The authors would like to thank Gillian Crofts and Jan Dodgeon for providing the sample ultrasound reports and for evaluating the works produced in this research.

References

[1] H. Edwards, J. Smith, M. Weston, What makes a good ultrasound report? Ultrasound 22 (1) (2013) 57–60.
[2] J.A. Cramer, L.B. Eisenmenger, N.S. Pierson, H.S. Dhatt, M.E. Heilbrun, Structured and templated reporting: an overview, Appl. Radiol.
[3] L. Faggioni, F. Coppola, R. Ferrari, E. Neri, D. Regge, Usage of structured reporting in radiological practice: results from an Italian online survey, Euro. Radiol. (2016) 1–10.
[4] S.S. Naik, A. Hanbidge, S.R. Wilson, Radiology reports: examining radiologist and clinician preferences regarding style and content, Am. J. Roentgenol. 176 (3) (2001) 591–598.
[5] A.A.O. Plumb, F.M. Grieve, S.H. Khan, Survey of hospital clinicians' preferences regarding the format of radiology reports, Clin. Radiol. 64 (4) (2009) 386–394, https://doi.org/10.1016/j.crad.2008.11.009 395–396 <http://www.ncbi.nlm.nih.gov/pubmed/19264183> .
[6] L.H. Schwartz, D.M. Panicek, A.R. Berk, Y. Li, H. Hricak, Improving communication

of diagnostic radiology findings through structured reporting, Radiology 260 (1) (2011) 174–181, https://doi.org/10.1148/radiol.11101913 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3121011&tool=pmcentrez&rendertype=abstract> .
[7] C.L. Sistrom, J. Honeyman-Buck, Free text versus structured format: information transfer efficiency of radiology reports, Am. J. Roentgenol. 185 (3) (2005) 804–812.
[8] L. Tran, A. Wadhwa, E. Mann, Implementation of structured radiology reports, J. Am. College Radiol. 13 (3) (2016) 296–299.
[9] N.Z. Zulkarnain, G. Crofts, F. Meziane, An architecture to support ultrasound report generation and standardisation, Proceedings of the International Conference on Health Informatics, 2015, pp. 508–513, , https://doi.org/10.5220/0005252505080513.
[10] W.C. Mann, S.A. Thompson, Rhetorical structure theory: toward a functional theory of text organization, Text-Interdiscipl. J. Study Disc. 8 (3) (1988) 243–281.
[11] N.Z. Zulkarnain, F. Meziane, G. Crofts, A methodology for biomedical ontology reuse, in: E. Métais, F. Meziane, M. Saraee, V. Sugumaran, S. Vadera (Eds.), Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22–24, Springer International Publishing, 2016, pp. 3–14, , https://doi.org/10.1007/978-3-319-41754-7_1.
[12] M. Taboada, W.C. Mann, Rhetorical structure theory: looking back and moving ahead, Disc. Stud. 8 (3) (2006) 423–459.
[13] D. Marcu, The rhetorical parsing of natural language texts, Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, ACL, 1997, pp. 96–103.
[14] V.W. Feng, G. Hirst, Text-level discourse parsing with rich linguistic features, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, ACL, 2012, pp. 60–68.
[15] D. Marcu, The rhetorical parsing of unrestricted texts: a surface-based approach, Comput. Linguist. 26 (3) (2000) 395–448.
[16] W.C. Mann, M. Taboada, Rhetorical Structure Theory, 2005. < http://www.sfu.ca/rst/index.html > .
[17] L. Carlson, D. Marcu, M.E. Okurowski, Building a discourse-tagged corpus in the framework of rhetorical structure theory, Proceedings of the Second SIGdial Workshop on Discourse and Dialogue, vol. 16, 2001, pp. 1–10.
[18] W. Hua, Z. Wang, H. Wang, K. Zheng, X. Zhou, Understand short texts by harvesting and analyzing semantic knowledge, IEEE Trans. Knowl. Data Eng.
[19] R. Soricut, D. Marcu, Sentence level discourse parsing using syntactic and lexical information, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, ACL, 2003, pp. 149–156.
[20] V.W. Feng, G. Hirst, A linear-time bottom-up discourse parser with constraints and post-editing, in: ACL(1), 2014, pp. 511–521.
[21] S. Joty, G. Carenini, R.T. Ng, Y. Mehdad, Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis, in: ACL(1), 2013, pp. 486–496.
[22] S. Joty, G. Carenini, R.T. Ng, CODRA: A novel discriminative framework for rhetorical analysis, Comput. Linguist.
[23] H. Hernault, H. Prendinger, D.A. DuVerle, M. Ishizuka, T. Paek, HILDA: a discourse parser using support vector machine classification, Dial. Disc. 1 (3) (2010) 1–33.
[24] Z. Lin, M.-Y. Kan, H.T. Ng, Recognizing implicit discourse relations in the Penn Discourse Treebank, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1, ACL, 2009, pp. 343–351.
[25] S. Joty, G. Carenini, R.T. Ng, A novel discriminative framework for sentence-level discourse analysis, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, ACL, 2012, pp. 904–915.
[26] M. Bärenfänger, M. Hilbert, H. Lobin, H. Lüngen, OWL ontologies as a resource for discourse parsing, LDV Forum, vol. 23, 2008, pp. 17–26.
[27] M. O'Donnell, RSTTool 2.4: a markup tool for Rhetorical Structure Theory, Proceedings of the First International Conference on Natural Language Generation, vol. 14, ACL, 2000, pp. 253–256.