Extending Answers using Discourse Structure

Wauter Bosma
Human Media Interaction
University of Twente
Postbus 217
7500AE Enschede, the Netherlands
bosmaw@cs.utwente.nl

Abstract

Research on Question Answering is focused mainly on classifying the question type and finding the answer, while presenting the answer in a way that suits the user's needs has received little attention. This paper shows how existing question answering systems can be improved by exploiting Rhetorical Structure Theory-based summarization techniques in order to extract more than just the exact answer from the document in which the answer resides. The output is an extensive answer, which also provides additional information related to the question, and which may give the user an opportunity to assess the accuracy of the answer (is this what I am looking for?). A first experiment confirms that the proposed summarization method performs better than a baseline summarization method.

1 Introduction

This paper presents a novel approach to applying summarization techniques to extend answers provided by a question answering engine. As recent studies show, much can be gained by integrating existing techniques of question answering (QA) and text summarization (Burger et al. 00; Mori et al. 04).

A question answering system pinpoints an *answer* to a given question in a set of documents. A *response* is then generated for this answer, and presented to the user (c.f. Hirschman & Gaizauskas 01). Discussion of the task of pinpointing the answer is beyond the scope of this paper. I will assume that the sentence which best matches the question, the *answer sentence*, is located by a QA system in a corpus of text documents. What remains is the task of generating an appropriate response and presenting it to the user.

Question answering systems traditionally try to find an 'exact answer'. This is also the focus of large-scale question answering evaluation programs such as TREC (Voorhees & Tice 00). An exact answer is a "text string consisting of a complete answer and nothing else" (Voorhees 03). Strings that contain a correct answer with additional text are considered 'inexact'.

Studies have shown, however, that users appreciate receiving more information than *only* the exact answer (Burger et al. 00). Consulting a question answering system is only part of a user's attempt to fulfill an information need: it's not the end point, but a step in what has been called a 'berry picking' process, where each answer/result returned by the system may motivate a follow-up step (Bates 90). The user may not only be interested in the answer to the question, but also in related information. The 'exact answer approach' fails to show leads to related information that might also be of interest to the user. Lin et al. (03) show that when searching for information, increasing the amount of text returned to users significantly reduces the number of queries that they pose to the system, suggesting that users utilize related information from supporting text.

Both commercial and academic QA systems tend to present more to the user than only the exact answer, but the sophistication of their responses varies from system to system. There are three degrees of sophistication in response generation: giving the exact answer, giving the answer plus context, and giving an extensive answer. The first is the most basic form of answer presentation. The second includes text surrounding the exact answer as well, which may allow the user to assess the

accuracy of the answer extraction, and thus to verify whether the answer is correct (Lin et al. 03). The extensive answer approach aims at not just including the immediate context, but generating a response in a more intelligent way, aiming at optimizing the amount of useful information while maintaining verifiability.

This paper presents a summarization technique based on Rhetorical Structure Theory (RST, Mann & Thompson 88) which can be used to create extensive answer presentations. This is done by transforming the rhetorical structure of the document to be summarized into a weighted graph, in which each vertex represents a sentence. This graph is used to select the sentences which are most salient with respect to the answer. Furthermore, a small study has been conducted in which users evaluated the verifiability, usefulness and relevance of the information that was presented in response to a question.

The paper is structured as follows. First, research concerning query-based summarization is discussed in section 2, and a brief description of RST is given in section 3. Section 4 discusses the proposal to answer extension and section 5 discusses an evaluation experiment where subjects judged the verifiability, usefulness and relevance of the information that was presented in response to a question. This paper concludes with a discussion and possible follow-ups on this research in section 6.

2 Query-based Summarization

The work presented in this paper focuses on the generation of query-based extracts. *Query-based* (as opposed to generic), because the summarization is tailored to suit the user's declared information needs, while a generic summarization is intended to reflect only the writer's communicative intent as conveyed by the source document. And the results are *extracts* (as opposed to abstracts), because the summarization only takes care of extracting portions of the source document, while abstracting also involves rewriting or rephrasing text.

While creating an extract for a particular answer, a candidate sentence can only be included if something is known about the relation between the candidate sentence and the answer. Indications of a relation between two sentences may be based on statistical measures of text similarity, such as the number of denotations of mutual concepts (Erkan & Radev 04). This paper focuses on the use of rhetorical relations.

Query-based summarization has been applied in information retrieval (c.f. Chali 02; Saggion et al. 03), but also in multi-document summarization (Mani & Bloedorn 97). In multi-document summarization—like in question answering—the source documents of the summarization are not written to satisfy the information need expressed by the query at hand.

Mani & Bloedorn (97) used graphs to formalize relations between sentences inside a document for multi-document summarization. A spreading activation algorithm used this graph to perform a query-based summarization, given a starting node that is selected for the query. Although Mani & Bloedorn (97) aim at summarizing by formalizing relations between concepts, a graph-based algorithm can also be used for generating RST-based summaries as answers to questions, as demonstrated in this paper.

3 Rhetorical Structure Theory

The proposed method exploits discourse structure in order to determine whether a sentence is included in the answer presentation. One of the most influential theories of discourse structure is the Rhetorical Structure Theory, developed by Mann & Thompson (88). For the purpose of text summarization, RST has theoretical and pragmatic advantages over other theories (e.g. Grosz & Sidner 86; Wolf & Gibson 05). Good levels of agreement have been measured between human annotators of

RST, which indicates that RST is well defined (Mann & Thompson 88; den Ouden 04). Furthermore, a corpus of RST-annotated documents is available, which can be used for training and evaluating RST-based summarization algorithms (Carlson et al. 02). Another advantage of RST is that RST defines coherence relations very formally and elaborately, which makes computational applications easier to develop.

According to RST, a rhetorical relation typically holds between two contiguous text spans, of which one span (the *nucleus*) is more central to the writer's intention than the other (the *satellite*), whose sole purpose is to increase the reader's understanding or belief of what is said in the nucleus. In some cases, two related spans are of equal importance, in which case there is a *multinuclear* relation between them. The related spans form a new span, which can in turn participate in a relation with another span. A full RST analysis is a hierarchical tree. The smallest units of discourse are *elementary discourse units* or *edus*, which are, in this paper, sentences.

RST has already been used to facilitate summarization (Marcu 97). In his summarization effort, Marcu used the nuclearity of relations in the rhetorical structure to determine which sentence is more salient, but he also explored other features as additional indicators of importance, such as sentence length (Marcu 97, 98). The following section shows how an algorithm capable of generating query-based summarizations can be used with the same RST features as the algorithms of Marcu, which are intended for generic summarization.

Because the proposed approach relies on the availability of RST analyses, it can only be used in real applications if there is a way to automatically create such an analysis from text. Automatically analyzing discourse structure in general is a hard problem. Although there is still much room for improvement, Marcu & Echihabi (02) show that this can be done using a small set of RST relations.

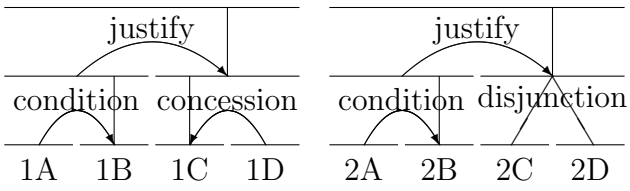

Figure 1: Rhetorical structure examples.

4 An Approach to Query-Based Summarization Using RST

This section describes a two-step approach to query-based summarization. First, the relations between sentences are defined in a discourse graph. Then, this graph is used to perform the summarization. During the first step, the rhetorical structure is transformed into a graph representation. The second step exploits a graph search algorithm in order to extract the most salient sentences from the graph. The starting node of the search is the node representing the answer sentence.

The summary should consist of the most salient sentences, given the answer sentence. This can be realized by determining the *distance* between the answer sentence and each of the other sentences. The sentences which are most closely related to the answer sentence are included in the summarization. The distance between sentences is measured by their distance in the RST graph. RST defines relations between two spans of text, which can be used to derive the distance from one sentence to another.

The most nuclear sentence of an RST analysis is the sentence which is most central to the writer's purpose. The graph ensures that, similarly to Marcu's approach, a nucleus is preferred over a satellite: in both summarization approaches, a satellite cannot be included in a generic summarization without its nucleus. The consequence is that in the specific case that the entry point of the summarization—the answer sentence—is the most nuclear sentence in the RST analysis, the result resembles the result of the summarization approach

by Marcu (97). However, the graph-based approach is more general in the sense that the summarization can start from any specific sentence rather than only the most nuclear sentence of the analysis.

RST analyses as weighted graphs

It is relatively straightforward to derive a graph from a rhetorical structure. Although RST is not designed as a computational framework, graph theory is very suitable for this purpose. A rhetorical structure tree can be converted to a discourse graph by means of the following steps.

1. For each sentence in the rhetorical structure, create a vertex associated with it.

2. For each directed relation, create an edge from each of the sentences of the nucleus to each of the sentences of the satellite of the relation.

A sentence is a nuclear sentence of a text span if it is not part of any sub span (of the text span) which participates as a satellite in a directed relation with any other sub span. A text span can have multiple nuclear sentences if multinuclear relations are involved. For instance, in the RST diagram on the left in Figure 1, the set of nuclear sentences of the entire document (denoted as 1A:1D) contains only sentence 1C. The right diagram shows a rhetorical structure in which the set of nuclear sentences of 2A:2D consists of sentences 2C and 2D.

The result of the transformation is an acyclic directed graph of which the vertices correspond to sentences, and the edges define relations between them. The left diagram in Figure 2 shows an example of a rhetorical structure diagram and a discourse graph that was created for this rhetorical structure as described above. During the transformation from RST to graph, part of the structural information is lost because sentences of the graph are directly connected to other sentences, while in RST, one end of a relation

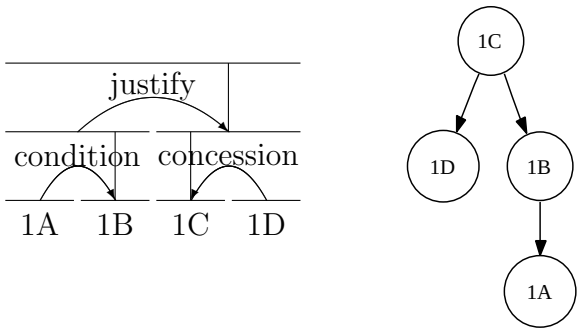

Figure 2: Rhetorical structure example and a discourse graph created for this rhetorical structure.

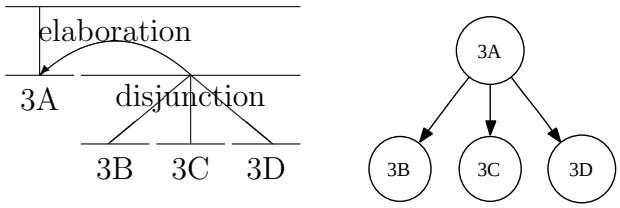

Figure 3: Rhetorical structure containing a multinuclear relation and the corresponding discourse graph.

can also span more than one sentence. If in RST one sentence is related to a text span of two sentences, the graph construction algorithm connects it to the nucleus of the two sentences in the discourse graph. In practice, this means that if the inclusion of a sentence in a summarization is justified by a rhetorical relation, the nucleus of that relation must be included in the summarization as well. This is in line with Mann and Thompson's definition of directedness of relations, which states that a nucleus of a directed relation has meaning without the satellite, but not the other way round.

In the case of a multinuclear relation, as in Figure 3, each of the sentences participating in the multinuclear relation (in the example: sentences 3B, 3C and 3D) is connected with the nucleus of the multinuclear span. That is, in the example, sentence 3A is connected to each of the sentences 3B, 3C and 3D, but sentences 3B–3D are not directly mutually connected. The reason for this is that in terms

of RST, there is a (multinuclear) relation between the sentences 3B, 3C and 3D, but they are mutually independent: if we know that 3B contains relevant information in a particular context, there is no way to be sure that, to any extent, 3C or 3D is relevant as well, based on the relevance of 3B and the multinuclear relation between the three sentences.

Now that we have a discourse graph T, we assume that given two sentences $a, b \in T$ for which there is a path from a to b, we can say that they are related, and therefore if a is relevant to the answer, b is also relevant to the answer. If a path contains more than one edge, the sentences are related only indirectly and an indirect relation is weaker than a direct relation between two sentences.

The strength of a relation between two sentences could be calculated by just counting the number of edges in the path between the vertices of the sentences. However, it may be the case that there is more than one sentence with an equally long path to the starting point of the summarization. In that case, the two sentences would be equally likely to be included in the summarization, although there might be other indications of one sentence being better suited for inclusion in the summarization than the other.

In order to remedy this situation, we can assign weights to edges in the discourse graph. A greater rhetorical distance is reflected by a greater weight. A low weight of the path from a to b indicates a high probability that b is relevant, given that a is relevant. The total weight of the path from a to b is denoted as $weight(a, b)$. The weight of a path between two sentences is defined iff there is a path that connects them. The weight of a path is the sum of the weights of the edges in the path.

Given the entry point of the summarization (the answer sentence), the shortest path from this sentence to any other sentence defines the relevance of the other sentence to the final answer.

The distance between two sentences is affected by the weights of the edges that connect the nodes corresponding to the sentences in the discourse graph. These weights are determined by using features of the rhetorical structure from which the graph was created, such as features of the text spans on either side of the relation for which the edge was created. The weight of an edge also depends on features of the sentence corresponding to the vertex which is targeted by the edge. The only constraint is that all weights are non-negative. Due to space constraints, the exact weight computation cannot be given in detail.

An important factor of the weights of edges is the size of the span which corresponds to the target node of the edge. This is inspired by the observation that if the author spends many words on a specific issue, the author apparently considers it relatively important.

Also, the length of a sentence is taken into account. In extractive summarization, it is usually better to include long sentences than short sentences, because short sentences typically contain more anaphora, which makes it more difficult to produce coherent results.

The features of the rhetorical structure that are considered for determining the weights are limited to the features for which there is evidence that they contribute to the quality of a summarization. Further research may motivate the use of other features as well.

An Example

This example shows how three sentences can be extracted from a text, based on its RST analysis, and given the entry point of the summarization. In a QA context, the entry point would be the answer sentence. Two of the extracted sentences are direct or indirect satellites of the answer sentence, the third is the answer sentence itself. The RST analysis of the following (segmented) text is shown in Figure 4. The entry point for the extraction is sentence 4E. This sentence could

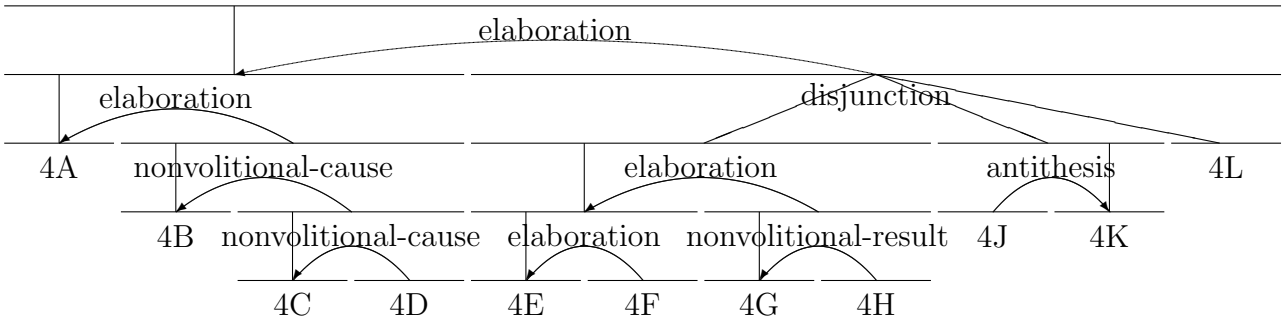

Figure 4: Rhetorical structure tree of the text fragment.

for instance be the QA output for the question: "what can be the cause of RSI?"

[A high pressure of workload, stress and repeatedly carrying out the same operation for a long period of time are the most important factors causing RSI to develop.]^{4A} [In the Netherlands the work pressure has increased with approximately 1.5% per year.]^{4B} [This is the result of shorter working hours in the eighties and nineties of the twentieth century.]^{4C} [Despite fewer working hours, the same quantity of work had to be finished.]^{4D} [A possible explanation of the development of RSI as a result of frequently repeated movements which are performed with low exertion is that the movement always involves contraction of the same muscles.]^{4E} [This happens for instance when working with a display device.]^{4F} [The motorial entities can be damaged because of oxygen lack and the impossibility of removing waste products.]^{4G} [Eventually they can cease to function and the muscle will lose strength.]^{4H} [There are however also indications that the complaints do not arise from damaged muscles.]^{4J} [Instead, they supposedly arise from abnormalities in the response of the brain to signals from the muscles.]^{4K} [Another possibility is that psychological factors can lead to symptoms of RSI.]^{4L}

First, a discourse graph is created from an RST analysis (as shown in Figure 5). The graph contains weighted edges. For this graph, the total weight of the paths from sentence 4E to each sentence in the graph is calculated using Dijkstra's shortest paths algorithm (Dijkstra 59). A path in a graph is an alternating sequence of vertices and edges, beginning and ending with a vertex. For instance, in the graph of Figure 5, there is a path over three vertices and two edges from 4E to 4H. The weight of this path is the sum of the weights of all of its edges.

Only four sentences are reachable from 4E. Since the selection of sentences is based on the weight of their path from 4E, a sentence which is associated with an unreachable vertex cannot be included in the extract. In this case, the sentences with the cheapest path from the entry point 4E are selected. The selected sentences are filtered out, resulting in the discourse graph on the left in Figure 6. For the sentences in this graph, the rhetorical structure can be derived using the original RST analysis in Figure 4. The result is the rhetorical structure on the right in Figure 6. This rhetorical structure may be used for further processing, for example for the purpose of speech synthesis (den Ouden 04). The output of the extraction process would be the following text. The answer sentence is highlighted.

> **A possible explanation of the development of RSI as a result of frequently repeated movements which are performed with low exertion is that the movement always involves contraction of the same muscles.** This happens for instance when working with a display device. The motorial entities can be damaged because of oxygen lack and the impossibility of removing waste products.

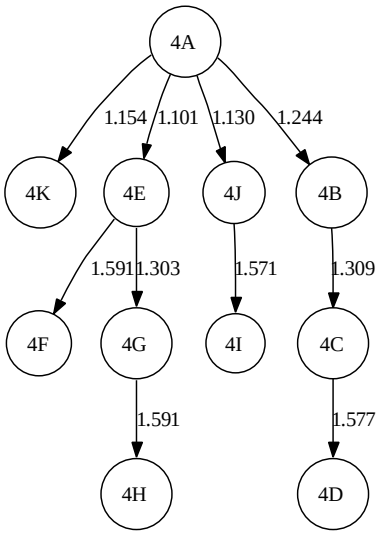

Figure 5: Weighted rhetorical structure graph of the text fragment. The vertex labels refer to their corresponding sentences. Edges are labeled by their weights.

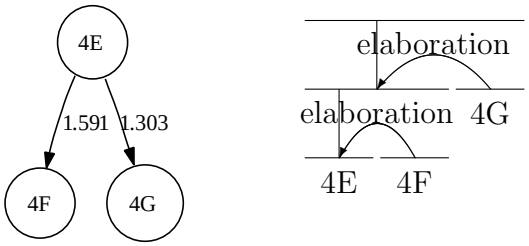

Figure 6: Extraction graph of the three sentences selected for inclusion in the summary, and the corresponding structure in RST notation, which is derived from the original RST analysis.

5 Evaluation

In order to find out whether the RST-based approach proposed here has any effects on the quality of the responses, a small user study has been carried out in which the RST-based method was compared with a baseline.

The quality of the summarizations was rated on three dimensions. First, the user was asked to indicate on a 5-point scale to what extent s/he was able to verify whether the answer was accurate. Secondly, the user was to judge how useful the provided information was with respect to the question. And finally, the user was asked how much irrelevant information was contained in the answer.

The study has been performed using the corpus of Carlson et al. (02), which is an RST-annotated corpus of news articles from the Wall Street Journal. The corpus also contains a collection of questions, each of which has an answer in one of the articles. A selection of questions was manually matched with an answer sentence in the corresponding article, and for each answer, an RST-based summarization was generated using the method described in this paper. Furthermore, a baseline answer presentation was generated, consisting of the answer sentence as well as the preceding and the successive sentence, following the linear order of the text (answer plus context).

We hypothesized that the RST-based summarization contains more useful and less irrelevant information than the baseline, while performing similarly in verifiability.

Fifteen people participated in the experiment. Each of them was presented with 12 different question-answer pairs, where the answer was generated according to either one of the strategies outlined above. The participants were prevented from having to evaluate two different summarizations that were generated for the same question.

The t-test was used to verify the hypotheses. It turned out that the RST-based summarizations were indeed significantly more verifiable, and contained less irrelevant information than the answer plus context ($p < 0.01$). Although the participants judged that the RST-based summarizations contained more useful information than the answers plus context, this difference was not significant.

The results suggest that RST-based summarization compares favorably to generating an answer presentation simply by including the answer sentence and surrounding sentences: using RST helps reducing the amount of irrelevant information, and increases the verifiability of the answer.

6 Conclusion

The presented approach to query-based summarization consists of two steps. First, the rhetorical structure tree is used to build graphs which determine the distances between individual sentences. Then, these graphs are used to decide which sentences are most relevant to the answer. These sentences are extracted to form an extensive answer presentation. An initial user study indicates that this method outperforms the baseline summarization method with respect to verifiability and the amount of relevant information, which are both crucial aspects of the quality of the answer.

Previous work on query-based summarization has mostly focused on extracting the set of sentences which best match the query. In contrast, we exploit the discourse structure in order to extract beyond the answer and include related (but relevant) information as well.

The advantage of the separation between formalization (graph construction) and extraction (graph search and sentence extraction) is that the latter is fairly generic: it can also be applied to discourse graphs that are not RST-based, for instance by creating graphs which are based on other models of discourse or onconceptual similarity relations between sentences. The conceptual graphs could be integrated with the RST-based graphs, in order to exploit all available indications of relevance.

Currently, the summarizations are created using a rule-based method which takes a small number of features of rhetorical structure of the answer document as input. Machine learning may be used to balance the features and to explore the use of other features.

Acknowledgements

This work is funded by the Netherlands Organization for Scientific Research (NWO).

References

(Bates 90) Marcia J. Bates. The berry-picking search: user interface design. In Harold Thimbleby, editor, *User Interface Design*. Addison-Wesley, 1990.

(Burger et al. 00) John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. Issues, tasks, and program structures to roadmap research in question & answering (q&a). NIST DUC Vision and Roadmap Documents, October 2000.

(Carlson et al. 02) Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers, 2002.

(Chali 02) Yllias Chali. Generic and query-based text summarization using lexical cohesion. In R. Cohen and B. Spencer, editors, *Advances in Artificial Intelligence: 15th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2002*, pages 293–302, Calgary, Canada, May 2002.

(den Ouden 04) Hanny den Ouden. *Prosodic realizations of text structure*. PhD thesis, University of Tilburg, December 2004.

(Dijkstra 59) E.W. Dijkstra. A note on two problems in connection with graphs. *Numerische Mathematik*, 1:269–271, 1959.

(Erkan & Radev 04) Güneş Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.

(Grosz & Sidner 86) Barbara J. Grosz and Candace L. Sidner. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July – September 1986.

(Hirschman & Gaizauskas 01) Lynette Hirschman and Rob Gaizauskas. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300, 2001.

(Lin et al. 03) Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, , and David R. Karger. What makes a good answer? the role of context in question answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction*, Zürich, Switzerland, 2003.

(Mani & Bloedorn 97) Inderjeet Mani and Eric Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97)*, pages 622–628, 1997.

(Mann & Thompson 88) William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

(Marcu 97) Daniel Marcu. From discourse structures to text summaries. In *The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 1997.

(Marcu 98) Daniel Marcu. To build text summaries of high quality, nuclearity is not sufficient. In *The Working Notes of the the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1–8, Stanford, CA, March 1998. American Association for Artificial Intelligence.

(Marcu & Echihabi 02) Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA, July 7-12 2002.

(Mori et al. 04) Tatsunori Mori, Masanori Nozawa, and Yoshiaki Asada. Multi-document summarization using a question-answering engine. In *Fourth NTCIR Workshop*, Tokyo, Japan, June 2004.

(Saggion et al. 03) Horacio Saggion, Kalina Bontcheva, and Hamish Cunningham. Robust generic and query-based summarization. In *10th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2003*, Hungary, April 2003.

(Voorhees 03) Ellen M. Voorhees. Overview of the trec 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, 2003.

(Voorhees & Tice 00) Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207, Athens, Greece, July 2000. ACM Press.

(Wolf & Gibson 05) Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based analysis. *Computational Linguistics*, 31(2):249–287, 2005.