Automatic Text Summarization from Multiple Sources for Time Evolving Events

Stergos D. Afantenos*

Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens
stergos.afantenos@lif.univ-mrs.fr

Abstract. In this PhD work we present a fresh look at the problem of summarizing evolving events from multiple sources. After a discussion concerning the nature of evolving events we introduce a distinction between *linearly* and *non-linearly* evolving events. We present then a general methodology for the automatic creation of summaries from evolving events. At its heart lie the notions of *Synchronic* and *Diachronic* cross-document Relations (SDRs), whose aim is the identification of similarities and differences between sources, from a synchronical and diachronical perspective. SDRs do not connect documents or textual elements found therein, but structures one might call *messages*. Applying this methodology will yield a set of messages and relations, SDRs, connecting them, that is a graph which we call *grid*. We will show how such a grid can be considered as the starting point of a Natural Language Generation System. The methodology is evaluated in two case-studies, one for linearly evolving events (descriptions of football matches) and another one for non-linearly evolving events (terrorist incidents involving hostages). In both cases we evaluate the results produced by our computational systems.

1 Introduction

With the advent of the Internet, access to many sources of information has now become much more easier. One problem that arises though from this fact is that of the information overflow. Imagine, for example, that someone wants to keep track of an event that is being described on various news sources, over the Internet, as it evolves through time. The problem is that there exist a plethora of news sources that it becomes very difficult for someone to compare the different versions of the story in each source. Furthermore, the Internet has made it possible now to have a rapid report of the news, almost immediately after they become available. Thus, in many situations it is extremely difficult to follow the rate with which

*Dissertation Advisor: Panagiotis Stamatopoulos.

the news are being reported. In such cases, a text summarizing the reports from various sources on the same event, would be handy. In this paper we are concerned with the automatic creation of summaries from multiple documents which describe an event that evolves through time. Such a collection of documents usually contains news reports from various sources, each of which provides novel information on the event as it evolves through time. In many cases the sources will agree on the events that they report and in some others they will adopt a different viewpoint presenting a slightly different version of the events or possibly disagreeing with each other. Such a collection of documents can, for example, be the result of a Topic Detection and Tracking system (Allan et al. 1998).

The identification of similarities and differences between the documents is a major aspect in Multi-document Summarization (Mani 2001; Afantenos, Karkaletsis, and Stamatopoulos 2005a; Afantenos et al. 2005; Afantenos, Karkaletsis, and Stamatopoulos 2005b; Afantenos et al. 2007; Afantenos 2007). (Mani and Bloedorn 1999), for example, identify similarities and differences among *pairs* of isolated documents by comparing the graphs that they derive from each document, which are based heavily on various lexical criteria. Our approach, in contrast, does not take into consideration isolated pairs of documents, but instead tries to identify the similarities and differences that exist between the documents, taking into account the time that the incidents occurred and the document source. This enables us to distinguish the document relations into *synchronic* and *diachronic* ones. In the synchronic level we try to identify the similarities and differences that exist between the various sources. In the diachronic level, on the other hand, we try to identify similarities and differences across time focusing on each source separately.

Another twofold distinction that we made through our study (Afantenos et al. 2005) concerns the type of evolution of an event, distinguishing between *linear* and *non-linear* evolution, and the rate of emission of the various news sources, distinguishing between *synchronous* and *asynchronous* emission of reports. Figure 1 depicts the major incidents for two different events: a linearly evolving event with synchronous emission and a non-linearly evolving one with asynchronous emission of reports. Whereas in the linearly evolving events the main incidents happen in constant and possibly predictable quanta of time,[1] in the non-linear events we can make no predictions as to when the next incident will occur. As

[1]This means that if the first news story q_0 comes at moment t_0, then we can assume that for each source the story q_n will come at time $t_n = t_0 + n * t$, where t is the constant amount of time that it takes for the news to appear.

you can see in Figure 1 we can have within a small amount of time an explosion of incidents followed by a long time of sparse incidents, etc.

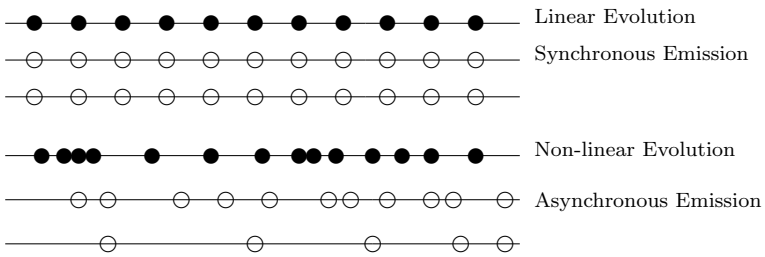

Fig. 1. Linear and Non-linear evolution

In order to represent the various incidents that are described in each document, we introduce the notion of *messages*. Messages are composed from a name, which reflects the type of the incidents, and a list of arguments, which take their values from the domain ontology. Additionally, they have associated with them the *time* that the message refers to, as well as the document *source*.

The distinction between linear and non-linear evolution affects mainly the *synchronic relations*, which are used in order to identify the similarities and differences between two messages from different sources, at about the same time. In the case of linear evolution all the sources report in the same time. Thus, in most of the cases, the incidents described in each document refer to the time that the document was published. Yet, in some cases we might have temporal expressions in the text that modify the time that a message refers to. In such cases, before establishing a synchronic relation, we should associate this message with the appropriate time-tag. In the case of non-linear evolution, each source reports at irregular intervals, possibly mentioning incidents that happened long before the publication of the article, and which another source might have already mentioned in an article published earlier. In this case we shouldn't rely any more to the publication of an article, but instead rely on the *time* tag that the messages have. Once this has been performed, we should then establish a *time window* in which we should consider the messages, and thus the relations, as synchronic.

2 Some Definitions

In our approach (Afantenos et al. 2004; Afantenos and Karkaletsis 2004; Afantenos et al. 2005; Afantenos, Karkaletsis, and Stamatopoulos 2005b; Afantenos et al. 2007; Afantenos 2006; Afantenos 2007) the major building blocks for representing the knowledge on a specific event are: the *ontology* which encodes the basic entity types (concepts) and their instances; the *messages* for representing the various incidents inside the document; and the *relations* that connect those messages across the documents. More details are given below.

Ontology. For the purposes of our work, a domain ontology should be built. The ontology we use is a taxonomic one, incorporating *is-a* relations, which are later exploited by the messages and the relations.

Messages. In order to capture what is represented by several textual units, we introduce the notion of *messages*. A message is composed from four parts: its *type*, a list of *arguments* which take their values from the concepts of the domain *ontology*, the *time* that the message refers, and the *source* of the document that the message is contained. In other words, a message can be defined as follows:

$$\texttt{message\_type (arg}_1\texttt{ , ... , arg}_n\texttt{)}$$
$$\text{where } \texttt{arg}_i \in \text{Domain Ontology}$$

Each message m is accompanied by the time (m`.time`) that it refers and its source (m`.source`). Concerning the source, this is inherited by the source of the document that contains the message. Concerning the time of the message, it is inherited by the publication time of the document, unless there exists a temporal expression in the text that modifies the time that a message refers. In this case, we should interpret the time-tag of the message, in relation to that temporal expression. A message definition may also be accompanied by a set of *constraints* on the values that the arguments can take. We would like also to note that messages are similar structures (although simpler ones) with the templates used in the MUC.[2] An example of a message definition will be given in the case study we present in section 3.

[2] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html`

Relations. In order to define a relation in a domain we have to provide a *name* for it, and describe the conditions under which it will hold. The name of the relation is in fact *pragmatic* information, which we will be able to exploit later during the generation of the summary. The conditions that a relation holds are simply some rules which describe the *temporal distance* that two messages should have (0 for synchronic and more than 1 for diachronic) and the characteristics that the arguments of the messages should exhibit in order for the relation to hold.

Furthermore, it is crucial to note here the importance that time and source position have on the relations, apart from the values of the messages' arguments. Suppose, for example, that we have two identical messages. If they have the same temporal tag, but belong to different sources, then we have an *agreement* relation. If, on the other hand, they come from the same source but they have chronological distance one, then we speak of a *stability* relation. Finally, if they come from different sources and they have chronological distance more than two, then we have no relation at all. We also do not have a relation if the messages have different sources and different chronological distances. Thus we see that, apart from the characteristics that the arguments of a message pair should exhibit, the source and temporal distance also play a role for that pair to be characterized as a relation. In section 3 we will give concrete examples of messages and relations for a particular case study.

3 Methodology

At the heart of Multi-document Summarization (MDS) lies the process of identifying the similarities and differences that exist between the input documents. Although this holds true for the general case of MDS, for the case of summarizing *evolving events* the identification of the similarities and differences should be distinguished, as we have previously argued (Afantenos 2006; Afantenos et al. 2004; Afantenos, Karkaletsis, and Stamatopoulos 2005b; Afantenos et al. 2007; Afantenos et al. 2005) between two axes: the *synchronic* and the *diachronic* axes. In the synchronic axis we are mostly concerned with the degree of agreement or disagreement that the various sources exhibit, for the same time frame, whilst in the diachronic axis we are concerned with the actual evolution of an event, as this evolution is being described by one source.

The initial inspiration for the SDRs was provided by the *Rhetorical Structure Theory* (RST) of Mann & Thompson (Mann and Thompson 1987; Mann and Thompson 1988). Rhetorical Structure Theory—which

was initially developed in the context of "computational text generation"[3] (Mann and Thompson 1987; Mann and Thompson 1988; Taboada and Mann 2006)—is trying to connect several *units of analysis* with relations that are semantic in nature and are supposed to capture the intentions of the author. As "units of analysis" today are used, almost ubiquitously, the clauses of the text. In our case, as units of analysis for the Synchronic and Diachronic Relations we are using some structures which we call *messages*, inspired from the research in the Natural Language Generation (NLG) field. Each message is composed of two parts: its *type* and a list of *arguments* which take their values from an *ontology* for the specific domain.

Concerning the SDRs, in order to formally define a relation the following four fields ought to be defined (see also (Afantenos et al. 2007)):

1. The relation's type (*i.e.* Synchronic or Diachronic).
2. The relation's name.
3. The set of pairs of message types that are involved in the relation.
4. The constraints that the corresponding arguments of each of the pairs of message types should have. Those constraints are expressed using the notation of first order logic.

The name of the relation carries *semantic* information which, along with the messages that are connected with the relation, are later being exploited by the NLG component (see (Afantenos et al. 2007)) in order to produce the final summary.

The methodology we propose consists of two main phases, the *topic analysis phase* and the *implementation phase*. The topic analysis phase is composed of four steps, which include the creation of the ontology for the topic and the providing of the specifications for the messages and the SDRs. The final step of this phase, which in fact serves as a bridge step with the implementation phase, includes the annotation of the corpora belonging to the topic under examination that have to be collected as a preliminary step during this phase. The annotated corpora will serve a dual role: the first is the training of the various Machine Learning algorithms used during the next phase and the second is for evaluation purposes. The implementation phase involves the computational extraction of the messages and the SDRs that connect them in order to create a directed acyclic graph (DAG) which we call *grid*. The architecture of the summarization system is shown in Figure 2.

[3] Also referred to as Natural Language Generation (NLG).

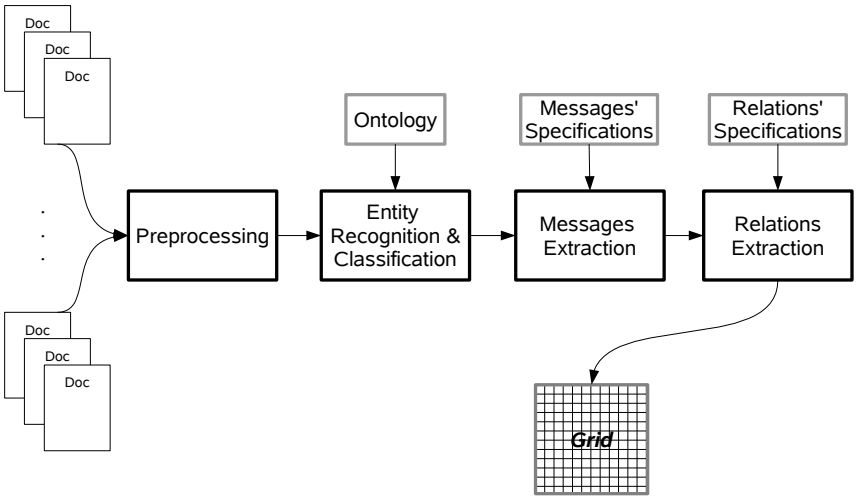

Fig. 2. The summarization system.

We applied our methodology in two different case studies. The first case study concerned the description of football matches, a topic which evolved linearly and exhibited synchronous emission of reports, while the second case study concerned the description of terroristic incidents with hostages, a topic which evolved non-linearly and exhibited asynchronous emission of reports.[4] The preprocessing stage involved tokenization and sentence splitting in the first case study and tokenization, sentence splitting and part-of-speech tagging in the second case study. For the task of the *entities recognition and classification* in the first case the use of simple gazetteer lists proved to be sufficient. In the second case study this was not the case and thus we opted for using what we called a *cascade of classifiers* which contained three levels. At the first level we used a binary classifier which determines whether a textual element in the input text is an instance of an ontology concept or not. At the second level, the classifier takes the instances of the ontology concepts of the previous level and classifies them under the top-level ontology concepts (e.g. `Person`). Finally at the third level we had a specific classifier for each top-level

[4]On the distinction between linearly/non-linearly events and synchronous/asynchronous emission of reports the interested reader is encouraged to consult (Afantenos 2006; Afantenos, Karkaletsis, and Stamatopoulos 2005b; Afantenos et al. 2007; Afantenos et al. 2005).

ontology concept, which classifies the instances in their appropriate sub-concepts; for example, in the `Person` ontology concept the specialized classifier classifies the instances into `Offender, Hostage`, etc. For the third stage of the messages' extraction we use in both case studies lexical and semantic features. As lexical features in the first case we used the words of the sentences (excluding low frequency words and stop-words) while in the second case study we used only the verbs and nouns of the sentences as lexical features. As semantic features in the first case study we used the number of the top-level ontology concepts that appear in the sentence, while in the second case study we enriched that with the appearance of certain trigger words in the sentence. Finally, the extraction of the SDRs is the most straightforward task, since the only thing that is needed is the translation of the relations' specifications into an appropriate algorithm which, once applied to the extracted messages, will provide the relations that connect the messages, effectively thus creating the grid. In Table 1 we present the statistics of the final messages and SDRs extraction stages for both case studies.[5]

	Case Study I	Case Study II
Messages	Pr : 91.12% Rc : 67.79% FM : 77.74%	Pr : 42.96% Rc : 35.91% FM : 39.12%
SDRs	Pr : 89.06% Rc : 39.18% FM : 54.42%	Pr : 30.66% Rc : 49.12% FM : 37.76%

Table 1. Precision, Recall and F-Measure for the extraction of the Messages and SDRs for both case studies.

The creation of the grid can be considered as completing—as we have previously argued (Afantenos et al. 2007)—the *Document Planning* phase of a typical architecture of an NLG system (Reiter and Dale 2000).

4 Conclusions

In this PhD we have presented a novel approach concerning the summarization of multiple documents dealing with evolving events. One point we focused particularly on was the automatic detection of the Synchronic and Diachronic Relations. As far as we know, this problem has never been

[5]For more details, critique of those results and comparison with related work the interested reader is encouraged to consult (Afantenos 2006; Afantenos et al. 2007).

studied before. The closest attempt we are aware of is (Allan, Gupta, and Khandelwal 2001) work, who create what they call temporal summaries. Nevertheless, this work does not take into account the event's evolution. Additionally, they are in essence agnostic in relation to the source of the documents, since they concatenate all the documents, irrespective of source, into one big document in which they apply their statistical measures.

In order to tackle the problem of summarizing evolving events, we have introduced the notions of messages and Synchronic and Diachronic Relations (SDRs). Messages impose a structure over the instances of the ontology concepts found in the input texts. They are the units of analysis for which the SDRs hold. Synchronic relations hold between messages from different sources with identical reference time, whilst Diachronic relations hold between messages from the same source with different reference times. We have provided definitions for the notions of *topic, event* and *activities*, borrowing from the terminology of Topic Detection and Tracking research. We also drew a distinction concerning the evolution of the events, dividing them into linear and non-linear events. In addition, we made a distinction concerning the report emission rate of the various sources, dividing them into synchronous and asynchronous emissions. We also provided a formal framework to account for the notions of linearity and synchronicity. Finally, we have shown how these distinctions affect the identification of the Synchronic and Diachronic Relations.

We have presented our methodology behind the implementation of a system that extracts Synchronic and Diachronic Relations from descriptions of evolving events. This methodology is composed of two phases: the topic analysis phase and the implementation phase. We worked on two case-studies for a linearly and non-linearly evolving topic, which implement the proposed methodology (Afantenos et al. 2004; Afantenos et al. 2005; Afantenos, Karkaletsis, and Stamatopoulos 2005b; Afantenos et al. 2007; Afantenos 2007). While the results are promising in both cases, there is certainly room for improvement for certain components. The tools incorporated for the implementation include the WEKA platform for the training of the Machine Learning algorithms, as well as the ELLOGON platform used for the annotation stage of the topic analysis phase and the development of the module used in the extraction of the messages.

We have shown how the creation of the grid, *i.e.* the extraction of the messages and their connection via Synchronic and Diachronic Relations, forms essentially the Document Planning stage, *i.e.* the first out of the three stages of a typical Natural Language Generation (NLG) system (Re-

iter and Dale 2000). Finally, we have made comparisons with the related works (Afantenos et al. 2004; Afantenos and Karkaletsis 2004; Afantenos et al. 2005; Afantenos, Karkaletsis, and Stamatopoulos 2005b; Afantenos et al. 2007; Afantenos 2006; Afantenos 2007), emphasizing the relationship between *Rhetorical Structure Theory* and our approach. We have shown the respective similarities and differences between the two, highlighting the innovative aspects of our approach. These innovations are in line with what one of the creators of *Rhetorical Structure Theory* presents as one the points that ought to be considered for the future of RST in a recent paper entitled "Rhetorical Structure Theory: Looking Back and Moving Ahead" (Taboada and Mann 2006). Again, we would like though to emphasize that, while certain parts of our approach have been inspired by RST, the approach as a whole should not be considered as an attempt of improvement of RST. In a similar vein, our innovations, should not be considered as an extension of RST. Instead it should merely be viewed as a new kind of methodology to tackle the problem of summarization of evolving events, via Synchronic and Diachronic Relations.

As mentioned throughout this article and the related publications resulted from this PhD work, we have presented a general architecture of a system which implements the proposed approach. The implementation of the NLG subsystem has not been completed yet. The Micro-Planning and Surface Generation stages are still under development. The completion of the NLG component is an essential aspect of our current work. Even if the results of the entities-, message-, and relation-extraction components — which are part of the summarization core — yield quite satisfactory results, we need to *qualitatively* evaluate our summaries. Yet, this will only be possible once the final textual summaries are created, and this requires the completion of the NLG component.

As shown in the evaluation of the system's components, the results concerning the summarization core are quite promising. Obviously, there is still room for improvement. The component that seems to need most urgent consideration is the arguments filling component. Up to now we are using heuristics which take into account the sentences' message types, returned by the dedicated classifier, as well as the extracted entities, resulting from the various classifiers used. This method does seem to be brittle, hence additional methods might be needed to tackle this problem. One idea would be to study various Machine Learning methods taking into account previously annotated messages, *i.e.* message types and their arguments. Another module needing improvement is the entity-extraction

component, especially the first classifier (the binary classifier) of the cascade of classifiers presented.

Concerning the summarization core, as we have shown in the evaluation of the several components included in this system, the results are promising. Yet, there is still room for improvement. The component that seems to need an immediate consideration is the arguments filling one. Up till now we are using heuristics which take into consideration the message type of the sentence, as returned by the dedicated classifier, as well as the extracted entities, which are in turn the result of the various classifiers used. This method does not seem to perform perfectly, which means that additional methods should be considered in order to tackle that problem. An idea would be the investigation of various Machine Learning methods which would take into account previously annotated messages, *i.e.* message types with their arguments. An additional module that needs improvement is the entities extraction component, especially the first classifier (the binary classifier) in the cascade of classifiers that we have presented.

An additional point that we would like to make concerns the nature of messages and the reduction of the human labor involved in the provision of their specifications. As it happens, the message types that we have provided for the two case studies, rely heavily on either verbs or verbalized nouns. This implies that message types could be defined automatically based mostly on statistics on verbs and verbalized nouns. Concerning their arguments, we could take into account the types of the entities that exist in their near vicinities. This is an issue that we are currently working on. Another promising path for future research might be the inclusion of the notion of messages, and possibly the notion of Synchronic and Diachronic Relations, into the topic ontology.

References

[Afantenos2006] Afantenos, Stergos D. 2006, December. "Automatic Text Summarization from Multiple Sources for Time Evolving Events." Ph.D. diss., Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece.

[Afantenos2007] ———. 2007, September. "Some Reflections on the Task of Content Determination in the Context of Multi-Document Summarization of Evolving Events." Edited by Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, *Recent Advances in Natural Language Processing (RANLP 2007)*. Borovets, Bulgaria: INCOMA, 12–16.

[Afantenos et al.2004] Afantenos, Stergos D., Irene Doura, Eleni Kapellou, and Vangelis Karkaletsis. 2004, May. "Exploiting Cross-Document Relations for

Multi-Document Evolving Summarization." Edited by G. A. Vouros and T. Panayiotopoulos, *Methods and Applications of Artificial Intelligence: Third Hellenic Conference on AI, SETN 2004*, Volume 3025 of *Lecture Notes in Computer Science*. Samos, Greece: Springer-Verlag Heidelberg, 410–419.

[Afantenos and Karkaletsis2004] Afantenos, Stergos D., and Vangelis Karkaletsis. 2004, December. "Linear Evolving Summarization: The first Results." Technical Report 2004/6, Institute of Informatics & Telecommunications, N.C.S.R. "Demokritos", Athens, Greece.

[Afantenos, Karkaletsis, and Stamatopoulos2005a] Afantenos, Stergos D., Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005a. "Summarization from Medical Documents: A Survey." *Journal of Artificial Intelligence in Medicine* 33 (2): 157–177 (February).

[Afantenos, Karkaletsis, and Stamatopoulos2005b] ———. 2005b, September. "Summarizing Reports on Evolving Events; Part I: Linear Evolution." Edited by Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, *Recent Advances in Natural Language Processing (RANLP 2005)*. Borovets, Bulgaria: INCOMA, 18–24.

[Afantenos et al.2007] Afantenos, Stergos D., Vangelis Karkaletsis, Panagiotis Stamatopoulos, and Constantin Halatsis. 2007. "Using Synchronic and Diachronic Relations for Summarizing Multiple Documents Describing Evolving Events." *Journal of Intelligent Information Systems*. Accepted for Publication.

[Afantenos et al.2005] Afantenos, Stergos D., Konstantina Liontou, Maria Salapata, and Vangelis Karkaletsis. 2005, May. "An Introduction to the Summarization of Evolving Events: Linear and Non-linear Evolution." *Natural Language Understanding and Cognitive Science NLUCS - 2005*. Maiami, USA, 91–99.

[Allan et al.1998] Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998, February. "Topic Detection and Tracking Pilot Study: Final Report." *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 194–218.

[Allan, Gupta, and Khandelwal2001] Allan, James, Rahuk Gupta, and Vikas Khandelwal. 2001. "Temporal Summaries of News Stories." *Proceedings of the ACM SIGIR 2001 Conference*. 10–18.

[Mani2001] Mani, Inderjeet. 2001. *Automatic Summarization*. Volume 3 of *Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

[Mani and Bloedorn1999] Mani, Inderjeet, and Eric Bloedorn. 1999. "Summarizing Similarities and Differences Among Related Documents." *Information Retrieval* 1 (1): 1–23.

[Mann and Thompson1987] Mann, William C., and Sandra A. Thompson. 1987. "Rhetorical Structure Theory: A Framework for the Analysis of Texts." Technical Report ISI/RS-87-185, Information Sciences Institute, Marina del Rey, California.

[Mann and Thompson1988] ———. 1988. "Rhetorical Structure Theory: Towards a Functional Theory of Text Organization." *Text* 8 (3): 243–281.

[Reiter and Dale2000] Reiter, Ehud, and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

[Taboada and Mann2006] Taboada, Maite, and William C. Mann. 2006. "Rhetorical Structure Theory: Looking Back and Moving Ahead." *Discourse Studies* 8 (3): 423–459 (June).