

DISCOURSE THEORY AND INFORMATION ACCESS — THREE COMPUTATIONAL MODELS TO AID UNLOCKING INFORMATION IN TEXT

Fredrik Olsson
fredrik.olsson@sics.se

January 18, 2004

Abstract

This report scratches the surface of research on discourse processing in order to find out what computational models are available to aid unlocking important information in text. Three areas are identified; discourse segmentation, discourse parsing, and reference resolution, in each of which a particular algorithm with accompanying implementation is presented.

Introduction

The motivation for this, very brief investigation, is a desire to know whether theories about discourse can be of any practical help in information access. Subordinate questions of importance include: What are the active areas of research in textual discourse processing? Are there computational models available to exemplify these research areas?

The characteristics of text involve linguistic relations spanning sentence boundaries, connecting entities that may be several sentences apart. Although there is no uncontradicted definition of what a text is, it is fairly evident that a text must satisfy the criteria of *coherence* and *cohesion* (discussed by, e.g., Titscher et al. (2000)). Coherence is a relation that hold between textual elements that fit well together, according to some criteria; it can be thought of as textual semantics. Cohesion is the glue that holds propositions of a text together; it can be considered as the text-syntactic connectedness. It is clear that information expressed in natural language text rely on the discourse structure. Still, most attempts at pin-pointing what is important in a text are grounded on linguistic features at a sub-sentential level, omitting the fact that sentences are inter-connected to form a whole. Nonetheless, these kind of descriptions manage to capture a lot of what is important by, e.g., identifying keywords, gists, and summary (extracts). Despite the success of these sub-sentential methods, it seems that there is a need to gain a deeper understanding of textuality in order to form a more robust and linguistically informed way of identifying important entities and the relations in which they participate.

There exist a range of linguistic theories about discourse, e.g., Rhetorical Structure Theory (Mann and Thompson, 1988), Discourse Representation Theory (Kamp and Reyle, 1993), anaphora resolution (Mitkov, 2002), centering al-

gorithm (Sidner, 1983), coherence (Hobbs, 1985), cohesion (Halliday and Hasan, 1976), intentional structure (Grosz and Sidner, 1986), to name but a few. The list of efforts made to describe the workings of discourse is far too long to be exhaustively elaborated on in this report; we are only scratching the surface. The aim of this report is to decide whether there are computational models of discourse mature enough to be put to use in the quest for the aspects of importance of a text. Since the present report is concerned with written text, discourse theories regarding speech and dialogues are not taken into consideration.

Discourse theory and information access

Intuitively, there is an obvious role to play in information access for theories about discourse structure. This intuition is manifested in a small and well documented psycholinguistic study carried out by Marcu (1999). The study shows that the concepts of discourse structure and nuclearity *can* in fact be employed in text summarization to extract salient pieces of text. In particular, Marcu shows that there is a strong correlation between what readers find important in a text and the nuclei of the discourse structure. Marcu concludes that, even though discourse-based methods are good indicators of importance of text units, no single discourse theory is capable enough if the desire is to obtain perfect results.

Before pursuing an implementation of a theory, it is important to know what can be expected from it in terms of coverage and expressive power of the theory alone. If the theory does not live up to the expectations, the implementation of the theory obviously will not do so either. An experiment along these lines is described by Sparck Jones (1993), who conducted a study in which several approaches to text characterisation were manually applied. The purpose of the experiment was to determine *what* information the theories under investigation can make available for automatic summarisation, and *how* this information can be exploited for summarisation. Sparck Jones (1993) states that “*It seems evident that any kind of discourse organisation above the sentence must play a part in marking what is important and thus relevant to summarising*”. Sparck Jones lists a range of factors important in summarisation, such as the nature of the source text, the intended purpose of the summary, and the output factors related to material and format.

In her experiment, Sparck Jones employs three different, albeit not mutually exclusive types of information, capturing linguistic, domain, or communicative information. Due to the labor intensive nature of the experiment, ten rather short texts were used, three of which were considered simplistic, ranging from paragraph to page in length. In Table 1, the combination of information types and representation forms outlined and used by Sparck Jones is illustrated.

The conclusion of the paper by Sparck Jones (1993) is that there is no single representation of discourse structure that capture the information in a text needed to produce a good summary, and that a more comprehensive discourse model is required to do so.

Keeping the findings of Sparck Jones (1993) and Marcu (1999) in mind, searching the literature gives at hand that there are (at least) three areas in which there exists robust computational models of discourse that can aid in finding important information in text: discourse segmentation, discourse pars-

	Bottom-up representation	Top-down representation
Linguistic information	RST Mann and Thompson (1988) Sidner’s focus algorithm Sidner (1983)	McKeown (1985) Maybury (1991) Rumelhart (1977)
Domain information	Lehnert (1992)	DeJong (1979) Tait (1983)
Communicative information	Grosz and Sidner (1986)	

Table 1: Grid of type of information and form of representation suggested by Sparck Jones (1993).

ing, and reference resolution. The remainder of this report will introduce a couple specific implementations on these matters, as well as explain how they can contribute to unlocking information in text.

Discourse segmentation

Efforts in segmentation of textual discourse include work by, e.g., Morris and Hirst (1991), Kozima (1993), Hearst (1994), Litman and Passonneau (1995), and Beeferman *et al* (1999). This section introduces the TextTiling algorithm developed by Hearst (1994; 1997), since it is readily available¹ and has proven useful in several applications and languages.

TextTiling

TextTiling use changes in lexical repetition patterns as indications of boundaries when segmenting expository texts into multi-paragraph subtopic structure. The algorithm impose a linear segmentation on the input text, as opposed to the hierarchical segmentation assumed by discourse theories such as Rhetorical Structure Theory (Mann and Thompson, 1988) and the intentional/attentional structure advocated by Grosz and Sidner (1986). The TextTiling algorithm consists of three major parts; tokenisation, lexical score determination, and boundary identification.

In the tokenisation step, the input text is divided into lexical units, and the units that are pre-defined as stop words (e.g. closed-class words, or high-frequency words) are filtered out. The remaining units are reduced to root form. The text is then divided into pseudosentences (also referred to as token-sequences) of a predefined length; this is a measure taken so as to avoid having to perform a, potentially problematic, normalisation w.r.t. sentence length in subsequent steps. Tokens are then stored in a lookup table in which the stop words are removed, yet contribute to the size of the token sequence.

For lexical score determination, three methods for score assignment are used; blocks, vocabulary introduction and chains. Block comparison is used to compare adjacent blocks of text to see how similar they are in terms of the number of words they share. The vocabulary introduction method assigns a score to

¹An implementation of TextTiling is available at <http://elib.cs.berkeley.edu/src/texttiles/>

each token-sequence based on the number of new words seen in the interval in which the sequence is the midpoint. The method using lexical chains is an extension of Morris and Hirst’s (1991); it keeps track of active chains of repeated terms and determines subtopic boundaries looking for places in the text where one set of chain ends and another starts.

Finally, Identification of boundaries is done in the same manner, regardless of which lexical score method was used. Basically, it involves assigning a score related to the depth of the valley (if any) constituted by the token-sequence under consideration. Such a depth score corresponds to how big the difference is between the cues for a subtopic on both sides of a given token-sequence. Deeper valleys get higher scores. The scoring function, along with all the details of the workings of the TextTiling algorithm are elaborated on by Hearst (1994; 1997).

The evaluation of the algorithm indicate that it exhibits *“acceptable performance when compared against human judgments of segmentation, although there is room for improvement”* (Hearst, 1997). In numbers, the acceptable performance translates to a highest recall of 78% (the corresponding precision was 52%), and a highest precision of 71% (in which case the recall was 59%).

One of the appealing properties of the TextTiling algorithm is that it relies on term repetition alone, not requiring other knowledge sources that the text itself and a general purpose stemmer or morphological analyser. This property is perhaps the one enabling other researchers to put the TextTiling algorithm on trial in a number of tasks in, for instance, automatic summarisation, e.g., (Barzilay and Elhadad, 1999), and to improve on information retrieval methods, e.g., stylistic variation (Karlgrén, 1996), and query expansion (Mandala et al., 1999).

Segmentation is also a pre-requisite for the parsing of discourse. In fact, the SPADE parser, introduced in the next section, comes with the option of running in segmentation mode in order to obtain only the segmented version of the raw input text.

Appendix B contains the output produced by the TextTiling algorithm when run on the example text presented in Appendix A.

Discourse parsing

If the body of research available in discourse segmentation is considered small, then the amount of research conducted in the area of discourse parsing is tiny. In essence, during the past few years there has been only a handful of researchers active in the field, the most productive of which is Daniel Marcu, who has been involved in a range of interesting publications about learning and non-learning approaches to discourse parsing, e.g., (Marcu, 2000a; Marcu, 2000b) and its application to summarization, e.g., (Carlson et al., 2001; Daume III and Marcu, 2002). As an example of a discourse parser, this section introduces the SPADE parser described by Soricut and Marcu (2003), the underlying theory of which is Rhetorical Structure Theory (Mann and Thompson, 1988).

SPADE

SPADE, which is short for sentence-level parsing of discourse, is freely available², and requires the syntactic parser by Charniak (2000)³. SPADE is trained using a publicly available corpora (Linguistic Data Consortium, 2002) made up of texts from the Wall Street Journal, and in which each sentence is analysed according to syntax as well as discourse.

The approach taken by Soricut and Marcu (2003) to building sentence-level discourse trees involve two sub-tasks: discourse segmentation and discourse parsing. Discourse segmentation is sub-divided into sentence segmentation, which is a well-researched problem not addressed further by Soricut and Marcu, and sentence-level discourse segmentation. The latter takes as input a sentence and outputs its elementary discourse unit boundaries. The sentence-level discourse segmentation module consists of two components; a statistical model which assigns a probability of the insertion of a discourse boundary after each word, and a segmenter, which employs the assigned probabilities to insert boundaries.

The input to the discourse parser is a discourse segmented lexicalised syntactic tree, and the output is a discourse parse tree. Here too, two components are used; a parsing model and a discourse parser. The parsing model assigns probability to all potential parse trees. And the discourse parser sets out to find the best parse tree for each input sentence. The parser uses dynamic programming to implement a bottom-up approach to search through the space of possible parse trees.

The performance results indicate that the weak point in the approach taken by Soricut and Marcu is that of discourse segmentation (achieving at best 83% F-score where the upper-limit, constituted by the inter-annotator agreement between humans on the same task, is 98.3%). On the other hand, the result imply that the syntactic parser is not a bottle-neck. The performance figures for the discourse parser as a whole peaks at 49% F-score when used with a set of 18 rhetorical labels to assign to the input. Human inter-annotator agreement is reported to be 77% on the same task. Overall, Soricut and Marcu (2003) claim that their “*discourse model is sophisticated enough to match near-human levels of performance.*”

Appendix C contains an excerpt of the output produced by the SPADE parser when invoked on the example text presented in Appendix A.

Reference resolution

The term reference resolution refers to the two related concepts of coreference and anaphora. Coreference is defined as a relation holding between noun phrases if they refer to the same entity (see e.g. (Hirschman and Chinchor, 1997)), and anaphora is understood as the presupposition of something that has gone before and that points back to some previous item (Halliday and Hasan, 1976). The difference between the two is, as pointed out by van Deemter and Kibble (2000), that coreference is an equivalence relation, while anaphora is an irreflexive, non-symmetrical and non-transitive relation. This imply that the phenomenon of

²SPADE can be downloaded from <http://www.isi.edu/~marcu/>

³Charniak’s parser is available at <ftp://ftp.cs.brown.edu/pub/nlparser/>

anaphora is sensitive to context, while coreference is not.

Anaphora resolution has been used in summarization, both to help deciding on important sentences to extract, and to tie up dangling references in automatically constructed abstracts. The most notable usage of reference resolution is that in the information extraction systems reported on in the MUC series, where coreference play a part in merging of extracted facts from text.

This section introduces the MARS anaphora resolution system⁴, described in detail by Mitkov *et al* (2002), as well as the coreference resolution capabilities of the LingPipe software package⁵, which is based on work by Baldwin (1995).

MARS

MARS is a re-implementation of an earlier, not entirely automatic approach to anaphora resolution developed by Mitkov (1998). The MARS algorithm, in its current incarnation, consists of five steps.

The first processing step involves preprocessing the input text using the functional dependency grammar parser from Conexor, which yields part-of-speech tags, lemmas, syntactic functions, grammatical number, and dependency relations present in the input.

In the second step, anaphoric pronouns are marked, while non-anaphoric and non-nominal occurrences of the pronoun *it* are removed.

In step three, then, antecedent candidates to each anaphoric pronoun are extracted. The scope of extraction is limited to a distance of three sentences prior to the position the pronoun. The candidates are put to further tests, and the ones passing are propagated to the next level of competition.

During step four, the candidates in the competing set of potential antecedents to a given anaphoric pronoun, undergo in total 14 further tests (preferential and impeding factors). Each of the tests assign a numerical score to each antecedent candidate, reflecting the confidence the system has the candidate is the antecedent to the pronoun under consideration.

Finally, in step five, the candidate with the highest score is selected as the antecedent of the pronoun.

MARS was evaluated on a set of documents consisting of computer hardware and software technical manuals in English, reaching an average success in the high 50's. It should be noted that the definition of success rate used by Mitkov *et al* (2002), is defined as the ratio between the number of anaphoric pronouns that was resolved correctly and the number of anaphoric pronouns in the text (that is, not all pronouns present in the text).

Appendix D contains the output produced by the web version of the MARS system when invoked on the example text available in Appendix A.

LingPipe

The LingPipe version 1.0 software package from Alias i, consists of several language processing modules; a statistical named entity recogniser (for English news and genomics), a heuristic sentence splitter, and a heuristic within-document coreference resolution system. The latter is what is interesting to this report. As indicated in the documentation of the LingPipe package, the

⁴MARS is available at <http://clg.wlv.ac.uk/MARS/>

⁵LingPipe is available at <http://www.alias-i.com/lingpipe/>

coreference resolution system is based on the CogNIAC system described in the PhD thesis by Breck Baldwin (1995). His thesis is mainly concerned with the resolution of anaphoric expressions, and the underlying theoretical assumptions of CogNIAC is that of Centering Theory (Brennan et al., 1987; Grosz et al., 1995). The core algorithm used in CogNIAC can be outlined in four steps. The first of which involves picking out a noun phrase for anaphora resolution. In the second step, the discourse is structured into sets and ranked according to salience.

The third step results in a set of antecedent candidates; the potential candidates that do not match, according to some criteria, the anaphoric noun phrase under investigation are eliminated.

In the final step, the most salient set of antecedent candidates is picked out. If it contains only one member, that member is chosen as the antecedent. However, if it contains more than one element, processing is stopped since it is assumed that each anaphoric expression can have at most one antecedent.

There are no performance figures available for the LingPipe package on the coreference resolution task. However, the results reported on named entity recognition for a variety of languages and domains indicate that LingPipe is indeed a serious effort.

Appendix E contains the output produced by the LingPipe software package when run in coreference resolution mode on the example text presented in Appendix A. However, since LingPipe did not manage to solve any coreference relations in that text, the output of another test run is included as well.

Conclusions

The working questions for this report has been:

1. Can theories about discourse be of any practical help in accessing textual information?
2. What are the active areas of research in textual discourse processing?
3. Are there computational models available in these areas?

The answer to the first question is *yes*. This report has pointed out, and exemplified manners in which theories about textual discourse can play a role in information access.

In response to question two and three, three areas of research on textual discourse has been identified and presented, in each of which there exist robust and available computational models of discourse processing. The areas under investigation in this report are: *discourse segmentation*, represented by an implementation of the TextTiling algorithm; *discourse parsing*, represented by SPADE; and *anaphora and coreference resolution*, represented by MARS and LingPipe, respectively.

Of these four programs, two could be considered linguistically informed in the sense that they are grounded in a prevalent discourse theory, while the other two are not. SPADE (based on Rhetorical Structure Theory) and the coreference resolution part of LingPipe (based on a modified variant of Centering Theory) constitute the informed implementations, while TextTiling and MARS constitute the un-informed ones. In order to illustrate what kind of information

the implementations under investigation are able to find, the report also include example output from each of them when run on the same, randomly chosen input text.

References

- Baldwin, Frederick Breckenridge. 1995. *CogNIAC: A discourse processing engine*. Ph.D. thesis, University of Pennsylvania.
- Barzilay, Regina and Elhadad, Michael. 1999. Using lexical chains for text summarization. In Mani, Inderjeet and Maybury, Mark T. editors, *Advances in Automatic Text Summarization*, chapter 10, pages 111–121. MIT Press.
- Beeferman, Doug; Berger, Adam and Lafferty, John. 1999. Statistical models for text segmentation. *Machine Learning*, 34:177–210. Special Issue on Natural Language Learning (Claire Cardie and Raymond Mooney, eds.).
- Brennan, Susan E.; Friedman, Marilyn Walker and Pollard, Carl J. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162, Stanford University, Stanford, CA, USA. ACL.
- Carlson, Lynn; Conroy, John M.; Marcu, Daniel; O’Leary, Dianne P.; Okurowski, Mary Ellen; Taylor, Anthony and Wong, William. 2001. An empirical study of the relation between abstracts, extracts, and the discourse structure of texts. In *Proceedings of the Document Understanding Conference*, New Orleans, LA, USA, September 13-14.
- Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Annual Meeting*, pages 132–139, Seattle, Washington, USA, April 29 - May 3. ACL.
- Daume III, Hal and Marcu, Daniel. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA, July 7-12. ACL.
- DeJong, G. F. 1979. *Skimming Stories in Real Time: An Experiment in Integrated Understanding*. Ph.D. thesis, Yale University.
- Grosz, Barbara J. and Sidner, Candace L. 1986. Attention, intention and the structure of discourse. *Computational Linguistics*, 12:175–204.
- Grosz, Barbara J.; Joshi, Aravind K. and Weinstein, Scott. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.
- Halliday, Michael AK. and Hasan, Ruqaiya. 1976. *Cohesion in English*. Longman, London.
- Hearst, Marti A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, Los Cruces, NM, USA, June.

- Hearst, Marti A. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March.
- Hirschman, Lynette and Chinchor, Nancy. 1997. Muc-7 coreference task definition (version 3.0). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*.
- Hobbs, Jerry R. 1985. On the coherence and structure of discourse. In Polyani, Livia editor, *The Structure of Discourse*. Ablex Publishing Corporation.
- Kamp, Hans and Reyle, Uwe. 1993. *From Discourse to Logic*. Kluwer Academic Publishers.
- Karlgren, Jussi. 1996. Stylistic variation in an information retrieval experiment. In *Proceedings of the 2nd International Conference on New Methods in Natural Language Processing*, Bilkent University, Ankara, Turkey, September.
- Kozima, Hideki. 1993. Text segmentation based on similarity between words. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL.
- Lehnert, W. G. 1992. Plot units: A narrative summarisation strategy. In *Strategies for Natural Language Processing*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Linguistic Data Consortium, . 2002. Rst discourse tree bank. LDC2002T07. FTP FILE. Philadelphia: Linguistic Data Consortium.
- Litman, Diane J. and Passonneau, Rebecca J. 1995. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL.
- Mandala, Rila; Teknobi, Tokunaga and Hozumi, Tanaka. 1999. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd Annual ACM Conference on Information Retrieval (SIGIR 99)*, Berkeley, CA, USA. ACM.
- Mann, William and Thompson, Sandra. 1988. Rhetorical structure theory: Towards a functional theory of text organisation. *Text*, 8(3).
- Marcu, Daniel. 1999. Discourse trees are good indicators of importance in text. In Mani, Inderjeet and Maybury, Mark T. editors, *Advances in Automatic Text Summarization*, chapter 11, pages 123–136. MIT Press.
- Marcu, Daniel. 2000a. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26(3):395–448, September.
- Marcu, Daniel. 2000b. *The theory and practice of discourse parsing and summarization*. MIT Press, Cambridge, Massachusetts, London, England.
- Maybury, Mark T. 1991. *Planning Multisentential English Text Using Communicative Acts*. Ph.D. thesis, University of Cambridge, Cambridge.
- McKeown, Kathleen R. 1985. *Text Generation*. Cambridge University Press.

- Mitkov, Ruslan; Evans, Richard and Orasan, Constantin. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution system. In Gelbukh, Alexander F. editor, *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, Mexico-City, Mexico, February.
- Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 869–875, Montreal, Canada. ACL.
- Mitkov, Ruslan. 2002. *Anaphora resolution*. Studies in Language and Linguistics. Longman/Pearson Education, London, New York, Toronto, Sydney, Tokyo, Singapore, Hong Kong, Cape Town, New Delhi, Madrid, Paris, Amsterdam, Munich, Milan, Stockholm.
- Morris, Jane and Hirst, Graeme. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, March.
- Rumelhart, D. E. 1977. Understanding and summarising brief stories. In D. Laberge and Samuels, S .J. editors, *Basic processes in reading: perception and comprehension*, pages 265–303. Lawrence Erlbaum Associates.
- Sidner, Candace L. 1983. Focusing in the comprehension of definite anaphora. In *Computational models of discourse*. MIT Press, Cambridge, MA.
- Soricut, Radu and Marcu, Daniel. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting*, Edmonton, Canada, May 27 - June 1. ACL.
- Sparck Jones, Karen. 1993. What might be in a summary? In Knorz, G; Krause, J and Womser-Hacker, C editors, *Proceedings from Information Retrieval 93: Von der modellierung zur Anwendung*, pages 9–26.
- Tait, J. I. 1983. *Automatic Summarising of English Texts*. Ph.D. thesis, Computer Laboratory, University of Cambridge. Also available as Technical Report 47.
- Titscher, Stefan; Meyer, Micheal; Wodak, Ruth and Vetter, Eva. 2000. *Methods of Text and Discourse Analysis*. SAGE Publications, London, Thousand Oaks, New Dehli.
- Deemter, van Kees and Kibble, Rodger. 2000. On coreferring: coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4):629–637, December.

Appendix A: Example text

This section contains a text used to try out the discourse processing software introduced in this report. The text appeared as a news item at the CNN website (cnn.com) on January 15, 2004. It should be noted that none of the software programs have been trained on this text, and that the purpose of this section is to illustrate the type of output produced by the various approaches, not to evaluate their performance.

Mission managers are preparing to roll the Mars rover Spirit off its lander early Thursday, positioning it for departure down a rear ramp.

If all goes as planned, what is termed "egress" will begin at 1 a.m. ET Thursday, with mission controllers sending the rover commands to move three meters down a ramp and onto the Martian soil.

"We will be driving three meters forward on the surface of Mars and leaving our lander for good," said engineer Kevin Burke at NASA's Jet Propulsion Laboratory in Pasadena, California.

"Not without a parting shot, though. We do plan on taking a couple of images of our lovely delivery system, ... give the engineers their due, and get to see their hardware for the last time."

Mission controllers at the JPL expect those images to be transmitted back to Earth around 4 a.m.

Mission managers planned to roll the rover off the front ramp of the lander, but had to change that plan after an airbag failed to retract properly.

They trained for that contingency, however, and were prepared to turn the rover on the deck of the lander and roll it off via the rear ramp.

"What we are going to do as soon as we egress off the lander, the next thing we're planning on doing is deploying the robotic arm," said mission manager Jennifer Trosper.

"The first day we'll kind of hover over soil and take some microscopic images, and the second day we'll actually deploy the instruments on the soil, and then we'll swap instruments, and then we'll stow and get ready to drive."

Project scientists have likened Spirit to a robotic geologist. The rover's mission is to study rocks and soil in an effort to determine whether the cold, desert world once was a warm, wet planet.

Spirit will first analyze rocks and soil near the lander, eventually making its way toward a large crater about 300 feet away.

After exploring that area, the rover will literally "head for the hills," making its way toward an area called the "East Hill Complex."

"I cannot tell you that we are going to reach those hills," principal investigator Steve Squyres said Tuesday.

"Our requirement for how far we should be able to traverse over the course of the mission, was 600 meters. These hills are five times that far away. OK, so don't sit here and think, 'Oh, we're going to go to the hills.' We're going to go 'toward' the hills."

Squyres will also manage the scientific payload on Spirit's identical twin, Opportunity.

Opportunity is scheduled to complete the 300 million-mile trip to Mars next weekend. It will land on the opposite side of the planet from Spirit's landing site inside the Gusev Crater, a nearly 100-mile-wide pockmark just south of the Martian equator.

Spirit and Opportunity have considerably more mobility and capability than the most recent successful visitor to Mars.

The 1997 NASA mission included the Pathfinder lander, which beamed back thousands of images, and Sojourner, a toy-sized test rover that scurried around the rocks and boulders littering the landing site.

Each of the new rovers is built to explore nearly as much distance in several days as Sojourner covered in three months, about 100 yards.

Each comes equipped with eight cameras that should provide stunning panoramas of the Martian surface, with resolutions so sharp they retain crisp detail when blown up to the size of a movie screen, according to NASA.

Their microscopes, spectrometers and drills could unlock geologic secrets from billions of years ago, when scientists think the planet may have had conditions more suitable for life.

CNN.com writer/editor Richard Stenger contributed to this report.

Appendix B: Example TextTiling output

This is an excerpt of the output from the TextTiling software when run on the text in Appendix A, indicating the tiles identified.

```
<TILE 0 - FILE: ../cnn.original.txt START: 0 END: 2097>
Mission managers are preparing to roll the Mars rover Spirit off its lander early Thursday, positioning it for departure down a rear ramp.
:
"I cannot tell you that we are going to reach those hills," principal investigator Steve Squyres said Tuesday.
</TILE 0 - FILE: ../cnn.original.txt START: 0 END: 2097>
<TILE 1 - FILE: ../cnn.original.txt START: 2098 END: 3032>
"Our requirement for how far we should be able to traverse over the course of the mission, was 600 meters. These hills are five times that far away. OK, so don't sit here and think, 'Oh, we're going to go to the hills.' We're going to go 'toward' the hills."
:
The 1997 NASA mission included the Pathfinder lander, which beamed back thousands of images, and Sojourner, a toy-sized test rover that scurried around the rocks and boulders littering the landing site.
</TILE 1 - FILE: ../cnn.original.txt START: 2098 END: 3032>
<TILE 2 - FILE: ../cnn.original.txt START: 3033 END: 3645>
Each of the new rovers is built to explore nearly as much distance in several days as Sojourner covered in three months, about 100 yards.
:
CNN.com writer/editor Richard Stenger contributed to this report.
</TILE 2 - FILE: ../cnn.original.txt START: 3033 END: 3645>
```

Appendix C: Example SPADE output

This is an excerpt of the output from SPADE when run on the text in Appendix A. The excerpt corresponds to the five first paragraphs of the input text.

```
(Root (span 1 2)
  ( Nucleus (leaf 1) (rel2par span)
    (text _!Mission managers are preparing to roll the Mars rover Spirit off
      its lander early Thursday ,_!) )
```

```

    ( Satellite (leaf 2) (rel2par Manner-Means)
(text _!positioning it for departure down a rear ramp ._) )
)
(Root (span 1 6)
  ( Nucleus (span 1 4) (rel2par span)
    ( Satellite (span 1 3) (rel2par Condition)
      ( Satellite (span 1 2) (rel2par Attribution)
        ( Nucleus (leaf 1) (rel2par span)
(text _!If all goes_) )
          ( Satellite (leaf 2) (rel2par Background)
(text _!as planned ,_) )
        )
      ( Nucleus (leaf 3) (rel2par Background)
(text _!what is termed_) )
      )
    ( Nucleus (leaf 4) (rel2par Background)
(text _!' egress ' will begin at 1 a.m. ET Thursday ,_) )
    )
    ( Satellite (span 5 6) (rel2par Background)
      ( Nucleus (leaf 5) (rel2par span)
(text _!with mission controllers sending the rover commands_) )
        ( Satellite (leaf 6) (rel2par Enablement)
(text _!to move three meters down a ramp and onto the Martian soil ._) )
        )
    )
)
(Root (span 1 4)
  ( Nucleus (span 1 3) (rel2par span)
    ( Nucleus (leaf 1) (rel2par span)
(text _!'_) )
      ( Satellite (span 2 3) (rel2par Attribution)
        ( Nucleus (leaf 2) (rel2par Joint)
(text _!We will be driving three meters forward on the surface of Mars_) )
          ( Nucleus (leaf 3) (rel2par Joint)
(text _!and leaving our lander for good , ') )
          )
        )
      ( Satellite (leaf 4) (rel2par Joint)
(text _!said engineer Kevin Burke at NASA 's Jet Propulsion Laboratory in
Pasadena , California ._) )
      )
)
(Root (leaf 1)
(text _!' Not without a parting shot , though ._) )
)
(Root (span 1 3)
  ( Nucleus (leaf 1) (rel2par Joint)
(text _!We do plan on taking a couple of images of our lovely delivery
system , . . . give the engineers their due ,_) )
    ( Nucleus (span 2 3) (rel2par Joint)
      ( Nucleus (leaf 2) (rel2par Joint)
(text _!and get to see their hardware for the last time . ') )
        ( Nucleus (leaf 3) (rel2par Joint)
(text _! Mission controllers at the JPL expect those images
to be transmitted back to Earth around 4 a.m. ._) )
        )
    )
)

```

)

Appendix D: Example MARS output

This section contains an excerpt produced by the web version of the MARS system when run on the text presented in Appendix A. The excerpt corresponds to the first six paragraphs in the input text.

Mission managers are preparing to roll the Mars rover Spirit off its lander early Thursday , positioning it for departure down a rear ramp .

its appears in paragraph 1, sentence 1, from position 12 to position 12. It is singular. The antecedent is indicated to be **the Mars rover Spirit** in paragraph 1, sentence 1, from position 7 to position 10.

Mission managers are preparing to roll the Mars rover Spirit off its lander early Thursday , positioning it for departure down a rear ramp .

it appears in paragraph 1, sentence 1, from position 18 to position 18. It is singular. The antecedent is indicated to be **the Mars rover Spirit** in paragraph 1, sentence 1, from position 7 to position 10.

" Not without a parting shot , though . We do plan on taking a couple of images of our lovely delivery system , ... give the engineers their due , and get to see their hardware for the last time. "

their appears in paragraph 4, sentence 2, from position 19 to position 19. It is plural. The antecedent is indicated to be **the engineers** in paragraph 4, sentence 2, from position 17 to position 18.

" Not without a parting shot , though . We do plan on taking a couple of images of our lovely delivery system , ... give the engineers their due , and get to see their hardware for the last time. "

their appears in paragraph 4, sentence 2, from position 26 to position 26. It is plural. The antecedent is indicated to be **the engineers** in paragraph 4, sentence 2, from position 17 to position 18.

Appendix E: Example LingPipe output

When run on the text in Appendix A, LingPipe did not recognise any coreferents. However, for sake of completeness, an excerpt of the output produced by the LingPipe package when run on the test text is given below. The excerpt corresponds to the first five paragraphs in the input text.

```
<?xml version="1.0" encoding="UTF-8"?><DOCUMENT>
<p>
<sent><ENAMEX id="0" type="ORGANIZATION">Mission</ENAMEX> managers are
preparing to roll the <ENAMEX id="1" type="LOCATION">Mars</ENAMEX>
```

rover Spirit off its lander early Thursday, positioning it for departure down a rear ramp.</sent>
</p>

<p>
<sent>If all goes as planned, what is termed "egress" will begin at 1 a.m. ET Thursday, with mission controllers sending the rover commands to move three meters down a ramp and onto the Martian soil.</sent>
</p>

<p>
<sent>"We will be driving three meters forward on the surface of <ENAMEX id="1" type="LOCATION">Mars</ENAMEX> and leaving our lander for good," said engineer <ENAMEX id="2" type="PERSON">Kevin Burke</ENAMEX> at <ENAMEX id="3" type="ORGANIZATION">NASA</ENAMEX>'s <ENAMEX id="4" type="ORGANIZATION">Jet Propulsion Laboratory</ENAMEX> in <ENAMEX id="5" type="LOCATION">Pasadena</ENAMEX>, <ENAMEX id="6" type="LOCATION">California</ENAMEX>.</sent>
</p>

<p>
<sent>"Not without a parting shot, though.</sent> <sent>We do plan on taking a couple of images of our lovely delivery system, ... give the engineers their due, and get to see their hardware for the last time."</sent>
</p>

<p>
<sent><ENAMEX id="0" type="ORGANIZATION">Mission</ENAMEX> controllers at the <ENAMEX id="7" type="ORGANIZATION">JPL</ENAMEX> expect those images to be transmitted back to <ENAMEX id="8" type="LOCATION">Earth</ENAMEX> around 4 a.m.</sent>
</p>

Example of coreference annotations

To show how coreference annotations may look like, the following text, supplied with the LingPipe software as an example, was run through the system:

George W. Bush believes he is the president. He is the commander in chief of the United States. His father used to work for the CIA. Barbara Bush believes she is his mother. She's married to his father, George H. Bush, who used to be the president.

The output produced looks like this:

```
<?xml version="1.0" encoding="UTF-8"?>
<DOCUMENT>
<P>
<sent><ENAMEX id="1" type="PERSON">George W. Bush</ENAMEX> believes
<ENAMEX id="1" type="MALE_PRONOUN">he</ENAMEX> is the
president.</sent>
<sent><ENAMEX id="1" type="MALE_PRONOUN">He</ENAMEX> is the commander
```

in chief of the <ENAMEX id="2" type="LOCATION">United States</ENAMEX>.</sent>
<sent><ENAMEX id="1" type="MALE_PRONOUN">His</ENAMEX> father used to work for the <ENAMEX id="3" type="ORGANIZATION">CIA.</ENAMEX></sent>
<sent><ENAMEX id="4" type="PERSON">Barbara Bush</ENAMEX> believes <ENAMEX id="4" type="FEMALE_PRONOUN">she</ENAMEX> is <ENAMEX id="1" type="MALE_PRONOUN">his</ENAMEX> mother.</sent> <sent><ENAMEX id="4" type="FEMALE_PRONOUN">She</ENAMEX>'s married to <ENAMEX id="1" type="MALE_PRONOUN">his</ENAMEX> father, <ENAMEX id="5" type="PERSON">George H. Bush</ENAMEX>, who used to be the president.</sent>
</P>
</DOCUMENT>