Introduction to the Special Issue on Summarization

Dragomir R. Radev*
University of Michigan

Eduard Hovy†
USC/ISI

Kathleen McKeown‡
Columbia University

1. Introduction and Definitions

As the amount of on-line information increases, systems that can automatically summarize one or more documents become increasingly desirable. Recent research has investigated types of summaries, methods to create them, and methods to evaluate them. Several evaluation competitions (in the style of the National Institute of Standards and Technology's [NIST's] Text Retrieval Conference [TREC]) have helped determine baseline performance levels and provide a limited set of training material. Frequent workshops and symposia reflect the ongoing interest of researchers around the world. The volume of papers edited by Mani and Maybury (1999) and a book (Mani 2001) provide good introductions to the state of the art in this rapidly evolving subfield.

A summary can be loosely defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that. *Text* here is used rather loosely and can refer to speech, multimedia documents, hypertext, etc.

The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its informativeness. Luckily, information content in a document appears in bursts, and one can therefore distinguish between more and less informative segments. Identifying the informative segments at the expense of the rest is the main challenge in summarization.

Of the many types of summary that have been identified (Borko and Bernier 1975; Cremmins 1996; Sparck Jones 1999; Hovy and Lin 1999), indicative summaries provide an idea of what the text is about without conveying specific content, and informative ones provide some shortened version of the content. Topic-oriented summaries concentrate on the reader's desired topic(s) of interest, whereas generic summaries reflect the author's point of view. Extracts are summaries created by reusing portions (words, sentences, etc.) of the input text verbatim, while abstracts are created by regenerating

* Assistant Professor, School of Information, Department of Electrical Engineering and Computer Science and Department of Linguistics, University of Michigan, Ann Arbor. E-mail: radev@umich.edu.
† ISI Fellow and Senior Project Leader, Information Sciences Institute of the University of Southern California, Marina del Rey, CA. E-mail: hovy@isi.edu.
‡ Professor, Department of Computer Science, New York University, New York, NY. E-mail: kathy@cs.columbia.edu.

the extracted content. *Extraction* is the process of identifying important material in the text, *abstraction* the process of reformulating it in novel terms, *fusion* the process of combining extracted portions, and *compression* the process of squeezing out unimportant material. The need to maintain some degree of grammaticality and coherence plays a role in all four processes.

The obvious overlap of text summarization with information extraction, and connections from summarization to both automated question answering and natural language generation, suggest that summarization is actually a part of a larger picture. In fact, whereas early approaches drew more from information retrieval, more recent approaches draw from the natural language field. Natural language generation techniques have been adapted to work with typed textual phrases, in place of semantics, as input, and this allows researchers to experiment with approaches to abstraction. Techniques that have been developed for topic-oriented summaries are now being pushed further so that they can be applied to the production of long answers for the question-answering task. However, as the articles in this special issue show, domain-independent summarization has several specific, difficult aspects that make it a research topic in its own right.

2. Major Approaches

We provide a sketch of the current state of the art of summarization by describing the general areas of research, including single-document summarization through extraction, the beginnings of abstractive approaches to single-document summarization, and a variety of approaches to multidocument summarization.

2.1 Single-Document Summarization through Extraction

Despite the beginnings of research on alternatives to extraction, most work today still relies on extraction of sentences from the original document to form a summary. The majority of early extraction research focused on the development of relatively simple surface-level techniques that tend to signal important passages in the source text. Although most systems use sentences as units, some work with larger passages, typically paragraphs. Typically, a set of features is computed for each passage, and ultimately these features are normalized and summed. The passages with the highest resulting scores are sorted and returned as the extract.

Early techniques for sentence extraction computed a score for each sentence based on features such as position in the text (Baxendale 1958; Edmundson 1969), word and phrase frequency (Luhn 1958), key phrases (e.g., "it is important to note") (Edmundson 1969). Recent extraction approaches use more sophisticated techniques for deciding which sentences to extract; these techniques often rely on machine learning to identify important features, on natural language analysis to identify key passages, or on relations between words rather than bags of words.

The application of machine learning to summarization was pioneered by Kupiec, Pedersen, and Chen (1995), who developed a summarizer using a Bayesian classifier to combine features from a corpus of scientific articles and their abstracts. Aone et al. (1999) and Lin (1999) experimented with other forms of machine learning and its effectiveness. Machine learning has also been applied to learning individual features; for example, Lin and Hovy (1997) applied machine learning to the problem of determining how sentence position affects the selection of sentences, and Witbrock and Mittal (1999) used statistical approaches to choose important words and phrases and their syntactic context.

Approaches involving more sophisticated natural language analysis to identify key passages rely on analysis either of word relatedness or of discourse structure. Some research uses the degree of lexical connectedness between potential passages and the remainder of the text; connectedness may be measured by the number of shared words, synonyms, or anaphora (e.g., Salton et al. 1997; Mani and Bloedorn 1997; Barzilay and Elhadad 1999). Other research rewards passages that include topic words, that is, words that have been determined to correlate well with the topic of interest to the user (for topic-oriented summaries) or with the general theme of the source text (Buckley and Cardie 1997; Strzalkowski et al. 1999; Radev, Jing, and Budzikowska 2000).

Alternatively, a summarizer may reward passages that occupy important positions in the discourse structure of the text (Ono, Sumita, and Miike 1994; Marcu 1997b). This method requires a system to compute discourse structure reliably, which is not possible in all genres. This technique is the focus of one of the articles in this special issue (Teufel and Moens 2002), which shows how particular types of rhetorical relations in the genre of scientific journal articles can be reliably identified through the use of classification. An open-source summarization environment, MEAD, was recently developed at the Johns Hopkins summer workshop (Radev et al. 2002). MEAD allows researchers to experiment with different features and methods for combination.

Some recent work (Conroy and O'Leary 2001) has turned to the use of hidden Markov models (HMMs) and pivoted QR decomposition to reflect the fact that the probability of inclusion of a sentence in an extract depends on whether the previous sentence has been included as well.

2.2 Single-Document Summarization through Abstraction

At this early stage in research on summarization, we categorize any approach that does not use extraction as an abstractive approach. Abstractive approaches have used information extraction, ontological information, information fusion, and compression.

Information extraction approaches can be characterized as "top-down," since they look for a set of predefined information types to include in the summary (in contrast, extractive approaches are more data-driven). For each topic, the user predefines frames of expected information types, together with recognition criteria. For example, an earthquake frame may contain slots for location, earthquake magnitude, number of casualties, etc. The summarization engine must then locate the desired pieces of information, fill them in, and generate a summary with the results (DeJong 1978; Rau and Jacobs 1991). This method can produce high-quality and accurate summaries, albeit in restricted domains only.

Compressive summarization results from approaching the problem from the point of view of language generation. Using the smallest units from the original document, Witbrock and Mittal (1999) extract a set of words from the input document and then order the words into sentences using a bigram language model. Jing and McKeown (1999) point out that human summaries are often constructed from the source document by a process of cutting and pasting document fragments that are then combined and regenerated as summary sentences. Hence a summarizer can be developed to extract sentences, reduce them by dropping unimportant fragments, and then use information fusion and generation to combine the remaining fragments. In this special issue, Jing (2002) reports on automated techniques to build a corpus representing the cut-and-paste process used by humans; such a corpus can then be used to train an automated summarizer.

Other researchers focus on the reduction process. In an attempt to learn rules for reduction, Knight and Marcu (2000) use expectation maximization to train a system to compress the syntactic parse tree of a sentence in order to produce a shorter but

still maximally grammatical version. Ultimately, this approach can likely be used for shortening two sentences into one, three into two (or one), and so on.

Of course, true abstraction involves taking the process one step further. Abstraction involves recognizing that a set of extracted passages together constitute something new, something that is not explicitly mentioned in the source, and then replacing them in the summary with the (ideally more concise) new concept(s). The requirement that the new material not be in the text explicitly means that the system must have access to external information of some kind, such as an ontology or a knowledge base, and be able to perform combinatory inference (Hahn and Reimer 1997). Since no large-scale resources of this kind yet exist, abstractive summarization has not progressed beyond the proof-of-concept stage (although top-down information extraction can be seen as one variant).

2.3 Multidocument Summarization

Multidocument summarization, the process of producing a single summary of a set of related source documents, is relatively new. The three major problems introduced by having to handle multiple input documents are (1) recognizing and coping with redundancy, (2) identifying important differences among documents, and (3) ensuring summary coherence, even when material stems from different source documents.

In an early approach to multidocument summarization, information extraction was used to facilitate the identification of similarities and differences (McKeown and Radev 1995). As for single-document summarization, this approach produces more of a briefing than a summary, as it contains only preidentified information types. Identity of slot values are used to determine when information is reliable enough to include in the summary. Later work merged information extraction approaches with regeneration of extracted text to improve summary generation (Radev and McKeown 1998). Important differences (e.g., updates, trends, direct contradictions) are identified through a set of discourse rules. Recent work also follows this approach, using enhanced information extraction and additional forms of contrasts (White and Cardie 2002).

To identify redundancy in text documents, various similarity measures are used. A common approach is to measure similarity between all pairs of sentences and then use clustering to identify themes of common information (McKeown et al. 1999; Radev, Jing, and Budzikowska 2000; Marcu and Gerber 2001). Alternatively, systems measure the similarity of a candidate passage to that of already-selected passages and retain it only if it contains enough new (dissimilar) information. A popular such measure is maximal marginal relevance (MMR) (Carbonell, Geng, and Goldstein 1997; Carbonell and Goldstein 1998).

Once similar passages in the input documents have been identified, the information they contain must be included in the summary. Rather than simply listing all similar sentences (a lengthy solution), some approaches will select a representative passage to convey information in each cluster (Radev, Jing, and Budzikowska 2000), whereas other approaches use information fusion techniques to identify repetitive phrases from the clusters and combine the phrases into the summary (Barzilay, McKeown, and Elhadad 1999). Mani, Gates, and Bloedorn (1999) describe the use of human-generated compression and reformulation rules.

Ensuring coherence is difficult, because this in principle requires some understanding of the content of each passage and knowledge about the structure of discourse. In practice, most systems simply follow time order and text order (passages from the oldest text appear first, sorted in the order in which they appear in the input). To avoid misleading the reader when juxtaposed passages from different dates all say "yesterday," some systems add explicit time stamps (Lin and Hovy 2002a). Other

systems use a combination of temporal and coherence constraints to order sentences (Barzilay, Elhadad, and McKeown 2001). Recently, Otterbacher, Radev, and Luo (2002) have focused on discourse-based revisions of multidocument clusters as a means for improving summary coherence.

Although multidocument summarization is new and the approaches described here are only the beginning, current research also branches out in other directions. Research is beginning on the generation of updates on new information (Allan, Gupta, and Khandelwal 2001). Researchers are currently studying the production of longer answers (i.e., multidocument summaries) from retrieved documents, focusing on such types as biographies of people, descriptions of multiple events of the same type (e.g., multiple hurricanes), opinion pieces (e.g., editorials and letters discussing a contentious topic), and causes of events. Another challenging ongoing topic is the generation of titles for either a single document or set of documents. This challenge will be explored in an evaluation planned by NIST in 2003.

2.4 Evaluation

Evaluating the quality of a summary has proven to be a difficult problem, principally because there is no obvious "ideal" summary. Even for relatively straightforward news articles, human summarizers tend to agree only approximately 60% of the time, measuring sentence content overlap. The use of multiple models for system evaluation could help alleviate this problem, but researchers also need to look at other methods that can yield more acceptable models, perhaps using a task as motivation.

Two broad classes of metrics have been developed: form metrics and content metrics. Form metrics focus on grammaticality, overall text coherence, and organization and are usually measured on a point scale (Brandow, Mitze, and Rau 1995). Content is more difficult to measure. Typically, system output is compared sentence by sentence or fragment by fragment to one or more human-made ideal abstracts, and as in information retrieval, the percentage of extraneous information present in the system's summary (precision) and the percentage of important information omitted from the summary (recall) are recorded. Other commonly used measures include kappa (Carletta 1996) and relative utility (Radev, Jing, and Budzikowska 2000), both of which take into account the performance of a summarizer that randomly picks passages from the original document to produce an extract. In the Document Understanding Conference (DUC)-01 and DUC-02 summarization competitions (Harman and Marcu 2001; Hahn and Harman 2002), NIST used the Summary Evaluation Environment (SEE) interface (Lin 2001) to record values for precision and recall. These two competitions, run along the lines of TREC, have served to establish overall baselines for single-document and multidocument summarization and have provided several hundred human abstracts as training material. (Another popular source of training material is the Ziff-Davis corpus of computer product announcements.) Despite low interjudge agreement, DUC has shown that humans are better summary producers than machines and that, for the news article genre, certain algorithms do in fact do better than the simple baseline of picking the lead material.

The largest task-oriented evaluation to date, the Summarization Evaluation Conference (SUMMAC) (Mani et al. 1998; Firmin and Chrzanowski 1999) included three tests: the categorization task (how well can humans categorize a summary compared to its full text?), the ad hoc task (how well can humans determine whether a full text is relevant to a query just from reading the summary?) and the question task (how well can humans answer questions about the main thrust of the source text from reading just the summary?). But the interpretation of the results is not simple; studies (Jing et al. 1998; Donaway, Drummey, and Mather 2000; Radev, Jing, and Budzikowska 2000)

show how the same summaries receive different scores under different measures or when compared to different (but presumably equivalent) ideal summaries created by humans. With regard to interhuman agreement, Jing et al. find fairly high consistency in the news genre only when the summary (extract) length is fixed relatively short. Marcu (1997a) provides some evidence that other genres will deliver less consistency. With regard to the lengths of the summaries produced by humans when not constrained by a particular compression rate, both Jing and Marcu find great variation. Nonetheless, it is now generally accepted that for single news articles, systems produce generic summaries indistinguishable from those of humans.

Automated summary evaluation is a gleam in everyone's eye. Clearly, when an ideal extract has been created by human(s), extractive summaries are easy to evaluate. Marcu (1999) and Goldstein et al. (1999) independently developed an automated method to create extracts corresponding to abstracts. But when the number of available extracts is not sufficient, it is not clear how to overcome the problems of low interhuman agreement. Simply using a variant of the Bilingual Evaluation Understudy (BLEU) scoring method (based on a linear combination of matching *n*-grams between the system output and the ideal summary) developed for machine translation (Papineni et al. 2001) is promising but not sufficient (Lin and Hovy 2002b).

3. The Articles in this Issue

The articles in this issue move beyond the current state of the art in various ways. Whereas most research to date has focused on the use of sentence extraction for summarization, we are beginning to see techniques that allow a system to extract, merge, and edit phrases, as opposed to full sentences, to generate a summary. Whereas many summarization systems are designed for summarization of news, new algorithms are summarizing much longer and more complex documents, such as scientific journal articles, medical journal articles, or patents. Whereas most research to date has focused on text summarization, we are beginning to see a move toward summarization of speech, a medium that places additional demands on the summarization process. Finally, in addition to providing full summarization systems, the articles in this issue also focus on tools that can aid in the process of developing summarization systems, on computational efficiency of algorithms, and on techniques needed for preprocessing speech.

The four articles that focus on summarization of text share a common theme: Each views the summarization process as consisting of two phases. In the first, material within the original document that is important is identified and extracted. In the second, this extracted material may be modified, merged, and edited using generation techniques. Two of the articles focus on the extraction stage (Teufel and Moens 2002; Silber and McCoy 2002), whereas Jing (2002) examines tools for automatically constructing resources that can be used for the second stage.

Teufel and Moens propose significantly different techniques for sentence extraction than have been used in the past. Noting the difference in both length and structure between scientific articles and news, they claim that both the context of sentences and a more focused search for sentences is needed in order to produce a good summary that is only 2.5% of the original document. Their approach is to provide a summary that focuses on the new contribution of the paper and its relation to previous work. They rely on rhetorical relations to provide information about context and to identify sentences relating to, for example, the aim of the paper, its basis in previous work, or contrasts with other work. Their approach features the use of corpora annotated both with rhetorical relations and with relevance; it uses text categorization to extract

sentences corresponding to any of seven rhetorical categories. The result is a set of sentences that situate the article in respect to its original claims and in relation to other research.

Silber and McCoy focus on computationally efficient algorithms for sentence extraction. They present a linear time algorithm to extract lexical chains from a source document (the lexical-chain approach was originally developed by Barzilay and Elhadad [1997] but used an exponential time algorithm). This approach facilitates the use of lexical chains as an intermediate representation for summarization. Barzilay and Elhadad present an evaluation of the approach for summarization with both scientific documents and university textbooks.

Jing advocates the use of a cut-and-paste approach to summarization in which phrases, rather than sentences, are extracted from the original document. She shows that such an approach is often used by human abstractors. She then presents an automated tool that is used to analyze a corpus of paired documents and abstracts written by humans, in order to identify the phrases within the documents that are used in the abstracts. She has developed an HMM solution to the matching problem. The decomposition program is a tool that can produce training and testing corpora for summarization, and its results have been used for her own summarization program.

Saggion and Lapalme (2002) describe a system, SumUM, that generates *indicative-informative* summaries from technical documents. To build their system, Saggion and Lapalme have studied a corpus of professionally written (short) abstracts. They have manually aligned the abstracts and the original documents. Given the structured form of technical papers, most of the information in the abstracts was also found in either the author abstract (20%) or in the first section of the paper (40%) or the headlines or captions (23%). Based on their observations, the authors have developed an approach to summarization, called *selective analysis*, which mimics the human abstractors' routine. The four components of selective analysis are *indicative selection*, *informative selection*, *indicative generation*, and *informative generation*.

The final article in the issue (Zechner 2002) is distinct from the other articles in that it addresses problems in summarization of speech. As in text summarization, Zechner also uses sentence extraction to determine the content of the summary. Given the informal nature of speech, however, a number of significant steps must be taken in order to identify useful segments for extraction. Zechner develops techniques for removing disfluencies from speech, for identifying units for extraction that are in some sense equivalent to sentences, and for identifying relations such as question-answer across turns in order to determine when units from two separate turns should be extracted as a whole. This preprocessing yields a transcript on which standard techniques for extraction in text (here the use of MMR [Carbonell and Goldstein 1998] to identify relevant units) can operate successfully.

Though true abstractive summarization remains a researcher's dream, the success of extractive summarizers and the rapid development of compressive and similar techniques testifies to the effectiveness with which the research community can address new problems and find workable solutions to them.

References

Allan, James, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of news topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–18.

Aone, Chinatsu, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. 1999. A trainable summarizer with knowledge acquired from robust NLP techniques. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 71–80.

Barzilay, Regina and Michael Elhadad. 1997.

Using lexical chains for text summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, July.

Barzilay, Regina and Michael Elhadad. 1999. Using lexical chains for text summarization. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 111–121.

Barzilay, Regina, Noémie Elhadad, and Kathy McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of the Human Language Technology Conference*.

Barzilay, Regina, Kathleen McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, 20–26 June, pages 550–557.

Baxendale, P. B. 1958. Man-made index for technical literature—An experiment. *IBM Journal of Research and Development*, 2(4):354–361.

Borko, H. and C. Bernier. 1975. *Abstracting Concepts and Methods*. Academic Press, New York.

Brandow, Ron, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

Buckley, Chris and Claire Cardie. 1997. Using empire and smart for high-precision IR and summarization. In *Proceedings of the TIPSTER Text Phase III 12-Month Workshop*, San Diego, CA, October.

Carbonell, Jaime, Y. Geng, and Jade Goldstein. 1997. Automated query-relevant summarization and diversity-based reranking. In *Proceedings of the IJCAI-97 Workshop on AI in Digital Libraries*, pages 12–19.

Carbonell, Jaime G. and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Alistair Moffat and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pages 335–336.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Conroy, John and Dianne O'Leary. 2001. Text summarization via hidden Markov models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 406–407.

Cremmins, Edward T. 1996. *The Art of Abstracting*. Information Resources Press, Arlington, VA, second edition.

DeJong, Gerald Francis. 1978. *Fast Skimming of News Stories: The FRUMP System*. Ph.D. thesis, Yale University, New Haven, CT.

Donaway, R. L., K. W. Drummey, and L. A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, Association for Computational Linguistics, 30 April, pages 69–78.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

Firmin, T. and M. J. Chrzanowski. 1999. An evaluation of automatic text summarization systems. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 325–336.

Goldstein, Jade, Mark Kantrowitz, Vibhu O. Mittal, and Jaime G. Carbonell. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *Research and Development in Information Retrieval*, pages 121–128, Berkeley, CA.

Hahn, Udo and Donna Harman, editors. 2002. *Proceedings of the Document Understanding Conference (DUC-02)*. Philadelphia, July.

Hahn, Udo and Ulrich Reimer. 1997. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In I. Mani and M. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 215–232.

Harman, Donna and Daniel Marcu, editors. 2001. *Proceedings of the Document Understanding Conference (DUC-01)*. New Orleans, September.

Hovy, E. and C.-Y. Lin. 1999. Automated text summarization in SUMMARIST. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 81–94.

Jing, Hongyan. 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics*, 28(4), 527–543.

Jing, Hongyan and Kathleen McKeown. 1999. The decomposition of human-written summary sentences. In

M. Hearst, F. Gey, and R. Tong, editors, *Proceedings of SIGIR'99: 22nd International Conference on Research and Development in Information Retrieval*, University of California, Berkeley, August, pages 129–136.

Jing, Hongyan, Kathleen McKeown, Regina Barzilay, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *Intelligent Text Summarization: Papers from the 1998 AAAI Spring Symposium*, Stanford, CA, 23–25 March. Technical Report SS-98-06. AAAI Press, pages 60–68.

Knight, Kevin and Daniel Marcu. 2000. Statistics-based summarization—Step one: Sentence compression. In *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence (AAAI-2000)*, pages 703–710.

Kupiec, Julian, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Research and Development in Information Retrieval*, pages 68–73.

Lin, C. and E. Hovy. 1997. Identifying topics by position. In *Fifth Conference on Applied Natural Language Processing*, Association for Computational Linguistics, 31 March–3 April, pages 283–290.

Lin, Chin-Yew. 1999. Training a selection function for extraction. In *Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM)*, Kansas City, 6 November. ACM, pages 55–62.

Lin, Chin-Yew. 2001. Summary evaluation environment. http://www.isi.edu/cyl/SEE.

Lin, Chin-Yew and Eduard Hovy. 2002a. From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of the 40th Conference of the Association of Computational Linguistics*, Philadelphia, July, pages 457–464.

Lin, Chin-Yew and Eduard Hovy. 2002b. Manual and automatic evaluation of summaries. In *Proceedings of the Document Understanding Conference (DUC-02) Workshop on Multi-Document Summarization Evaluation at the ACL Conference*, Philadelphia, July, pages 45–51.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165.

Mani, Inderjeet. 2001. *Automatic Summarization*. John Benjamins, Amsterdam/Philadelphia.

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

Mani, Inderjeet, Barbara Gates, and Eric Bloedorn. 1999. Improving summaries by revising them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 99)*, College Park, MD, June, pages 558–565.

Mani, Inderjeet, David House, G. Klein, Lynette Hirshman, Leo Obrst, Thérèse Firmin, Michael Chrzanowski, and Beth Sundheim. 1998. The TIPSTER SUMMAC text summarization evaluation. Technical Report MTR 98W0000138, The Mitre Corporation, McLean, VA.

Mani, Inderjeet and Mark Maybury, editors. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge.

Marcu, Daniel. 1997a. From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, July 11, pages 82–88.

Marcu, Daniel. 1997b. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, University of Toronto, Toronto.

Marcu, Daniel. 1999. The automatic construction of large-scale corpora for summarization research. In M. Hearst, F. Gey, and R. Tong, editors, *Proceedings of SIGIR'99: 22nd International Conference on Research and Development in Information Retrieval*, University of California, Berkeley, August, pages 137–144.

Marcu, Daniel and Laurie Gerber. 2001. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*, Pittsburgh, June. NAACL, pages 1–8.

McKeown, Kathleen, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the 16th National Conference of the American Association for Artificial Intelligence (AAAI-1999)*, 18–22 July, pages 453–460.

McKeown, Kathleen R. and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, July, pages 74–82.

Ono, K., K. Sumita, and S. Miike. 1994.

Abstract generation based on rhetorical structure extraction. In *Proceedings of the International Conference on Computational Linguistics*, Kyoto, Japan, pages 344–348.
Otterbacher, Jahna, Dragomir R. Radev, and Airong Luo. 2002. Revisions that improve cohesion in multi-document summaries: A preliminary study. In *ACL Workshop on Text Summarization*, Philadelphia.
Papineni, K., S. Roukos, T. Ward, and W-J. Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. Research Report RC22176, IBM.
Radev, Dragomir, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Çelebi, Hong Qi, Elliott Drabek, and Danyu Liu. 2002. Evaluation of text summarization in a cross-lingual information retrieval framework. Technical Report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, June.
Radev, Dragomir R., Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, April.
Radev, Dragomir R. and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.
Rau, Lisa and Paul Jacobs. 1991. Creating segmented databases from free text for text retrieval. In *Proceedings of the 14th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New York, pages 337–346.
Saggion, Horacio and Guy Lapalme. 2002. Generating indicative-informative summaries with SumUM. *Computational Linguistics*, 28(4), 497–526.
Salton, G., A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207.
Silber, H. Gregory and Kathleen McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4), 487–496.
Sparck Jones, Karen. 1999. Automatic summarizing: Factors and directions. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 1–13.
Strzalkowski, Tomek, Gees Stein, J. Wang, and Bowden Wise. 1999. A robust practical text summarizer. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, pages 137–154.
Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4), 409–445.
White, Michael and Claire Cardie. 2002. Selecting sentences for multidocument summaries using randomized local search. In *Proceedings of the Workshop on Automatic Summarization (including DUC 2002)*, Philadelphia, July. Association for Computational Linguistics, New Brunswick, NJ, pages 9–18.
Witbrock, Michael and Vibhu Mittal. 1999. Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, pages 315–316.
Zechner, Klaus. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4), 447–485.