The Columbia Multi-Document Summarizer for DUC 2002

Kathleen McKeown **Regina Barzilay** **Sasha Blair-Goldensohn**
David Evans **Vasileios Hatzivassiloglou** **Judith Klavans**
Ani Nenkova **Barry Schiffman** **Sergey Sigelman**

Department of Computer Science, Columbia University, New York, NY 10027

1 Introduction

The Columbia Summarizer for DUC 2002 is based on the multi-document summarization system that we developed for DUC 2001 (McKeown et al., 2001). It is a composite system that uses different summarization strategies depending on the type of documents in the input set. This year, we have four different strategies, one for single events, one for multiple related events, one for biographies and one for discussion of an issue with related events. The changes that we implemented this year fell into four categories: routing of document sets to strategies, statistical techniques for determining similarity and important content, editing of text for production of abstracts, and changes affecting formatting errors. In addition to producing abstracts as summaries, we also created a version which produces extracts.

2 Routing

In the Columbia Summarizer, MultiGen (Barzilay et al., 1999; Hatzivassiloglou et al., 1999; McKeown et al., 1999; Barzilay et al., 2001; Hatzivassiloglou et al., 2001) is used for generating summaries of a set of documents on a single event while DEMS (Dissimilarity Engine for Multidocument Summarization) (Schiffman et al., 2002) is used in three different configurations for generating summaries of a set of documents on multiple events, on issues and for biographical documents.

The router determines whether to send articles to MultiGen or DEMS based solely on date. Following the DUC guidelines, we are assuming that all document sets containing documents that were published within seven days are on a single event. These document sets are sent to MultiGen. All other document sets are sent to DEMS. DEMS does an additional test to determine whether the majority of articles within a set center around references to a single person. To do this, it tests for a high percentage of the same named entity references (using a shorthand for named entities that checks for capitalized words).

We considered doing an additional test for similarity between the documents in a document set. We would have done use this using similarity metrics that we currently use for document clustering in the tracking and clustering stage of Newsblaster (McKeown et al., 2002). However, the guidelines indicate that the DUC training data is significantly different from testing data this year and we also suspect that the data we used for training for Newsblaster is significantly different from DUC 2002 testing data. Therefore, we felt thresholds for similarity could not be reliably determined.

3 Statistical Techniques

We made changes only to the statistical techniques used in DEMS to determine content. Rather than fine tune the DEMS parameters to the categorization of articles provided by NIST, we examined the training articles to see if a more useful breakdown might give improved summaries.

The examination of the 60 document sets (or clusters) used as training and testing corpora in 2001 revealed that there were three general types: (1) single event tracked over a long period of time, usually about a particular person; (2) multiple events of a similar nature; and (3) discussion of an issue with some related events. Examples of these three, respectively, are Elizabeth Taylor's bout with pneumonia, various marathon runners and races, and gun control.

We tuned statistical parameters for extracting sentences to the three types. Intuitively, type 1 requires some extra weight to both the main character (or entity) in the set and also needs to pay attention to the publication date so that the outcome is included in the summary. For example, did Elizabeth Taylor recover? Type 2 requires a broad brush approach, achieved by putting more emphasis on first sentences, and no emphasis on the target or publication date. Type 3 summaries improved when the parameters emphasized the concepts most frequently found in the set. In all types of sets, summaries were more coherent if outlier articles, which didn't fit in the categorization scheme, were left out.

We computed similarity between each pair of articles in the input set, calculated over Concept Sets (Schiffman et al., 2002). When the span of similarity values is too wide, the set is usually type 2, multi-event. When the span is very narrow, it is either type 1 or type 3, and these could be distinguished easily by examining whether the most frequent concept was a named entity or not. Categorization into these three types is made automatically, after the sentences are ranked but before the summaries are actually assembled.

4 Regeneration and Editing of Summary Sentences

MultiGen generates summary sentences by cutting and pasting phrases from themes, groups of sentences that are determined to be similar. These phrases are the "intersection" of the theme sentences. To determine the intersection, it does an alignment of the parse trees of theme sentences to determine similar phrases. For this version of the summarizer, we adjusted the thresholds on the similarity during the tree comparisons, giving different types of constituents different thresholds. Circumstantials, for example, were given low thresholds since they can easily be removed without affecting grammaticality or correctness of the output. Subjects and objects, however, were given high thresholds in order to avoid having the wrong reference being selected for the subject of a summary sentence.

For DEMS, we added the ability to rewrite noun phrases in the extracted sentences. IBM's NOMINATOR system (Ravin et al., 1997) was used to extract named entities in the input cluster. Titles and premodifier descriptions are also extracted, like "Actress Elizabeth Taylor." After ordering the sentences, references are substituted so that the longest variant of the name, possibly including a title, appears at the first mention of the name and subsequent repeated variants are substituted with the shortest most common name variant.

5 Postprocessing

We added a postprocessing phase to fix capitalization and punctuation errors, such as putting comma tokens back next to the preceding word. In addition, in this stage, we also implemented the ability to generate extractive as well as abstractive summaries. For DEMS, this simply meant repressing the ability to rewrite references. For MultiGEN, this meant extracting a representative sentence from each theme instead of generating a sentence from the intersection of similar phrases. This process was complicated, however, by the fact that the DUC data set did its own sentence identification and numbering. Since we had already implemented our own strategies for part-of-speech tagging and sentence splitting, which affected many stages in the summarizer, we felt that it was safer to compare our summary sentences to identified sentences in the DUC data set, using word overlap to determine the most likely match and select the identifying number. We ran the Columbia summarizer in both modes for the evaluation.

6 Evaluation Results

Extracts For the extracts, we measured precision and recall, both micro-averaging across all sentences for the produced summaries or model summaries, respectively, in the entire evaluation collection, and macro-averaging by computing precision and recall for each summary and averaging those across the collection. The results, shown in Table 1 for extracted summaries of all sizes, indicate that our system, 24, came in second if precision is the more im-

System code	Recall		Precision	
	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged
19	20.70% (1)	21.30% (1)	20.66% (3)	21.14% (3)
20	15.21% (5)	15.78% (5)	14.82% (5)	15.26% (5)
21	20.63% (2)	20.48% (2)	24.90% (1)	25.84% (1)
24	**18.23% (3)**	**17.91% (3)**	**22.11% (2)**	**22.26% (2)**
28	15.83% (4)	16.05% (4)	18.12% (4)	19.23% (4)

Table 1: Evaluation scores on extracts for the top five systems, across all summary sizes. Systems listed in order of system code, with Columbia's scores in bold. Ranks shown in parentheses among all 10 systems submitting extracts.

System code	Recall		Precision	
	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged
19	18.62% (1)	18.38% (1)	18.67% (3)	18.51% (3)
20	12.43% (5)	12.22% (5)	12.86% (5)	12.78% (5)
21	17.24% (2)	16.26% (2)	21.18% (1)	20.91% (1)
24	**15.67% (3)**	**14.65% (3)**	**19.49% (2)**	**19.05% (2)**
28	12.93% (4)	12.32% (4)	15.01% (4)	15.23% (4)

Table 2: Evaluation scores on extracts for the top five systems, on 200 word summaries. Systems listed in order of system code, with Columbia's scores in bold. Ranks shown in parentheses among all 10 systems submitting extracts.

System code	Recall		Precision	
	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged
19	22.78% (2)	22.81% (1)	22.66% (3)	22.46% (3)
20	18.00% (5)	17.61% (5)	16.78% (5)	16.39% (7)
21	24.02% (1)	22.65% (2)	28.61% (1)	28.31% (1)
24	**20.80% (3)**	**19.58% (3)**	**24.73% (2)**	**23.80% (2)**
28	18.73% (4)	17.97% (4)	21.23% (4)	21.19% (4)

Table 3: Evaluation scores on extracts for the top five systems, on 400 word summaries. Systems listed in order of system code, with Columbia's scores in bold. Ranks shown in parentheses among all 10 systems submitting extracts.

portant factor and third if recall is the more important measure. System 21 beat us on both recall and precision, while system 19 beat us on recall but not on precision. System 28 ranks consistently fourth on all measures, and system 20 fifth; these two systems are clearly separated from the top three by at least two points. This relative ranking also holds if we look at the subsets of 200 word extracts and 400 word extracts separately (Tables 2 and 3). Micro- or macro-averaging makes very little difference in the relative performance of the top five systems in the vast majority of cases; the one exception is recall for system 21 which moves from second to first on the 400 word summaries when micro-averaging is used.

Looking at all summaries independent of size, humans did better than systems in most cases on recall (7 out of 9), but by only a small margin (7.13 percentage points in the best case). On precision, only 4 out of 9 humans beat the top system when micro-averaging is used and two when macro-averaging is used. The difference in the best case, 3.28 percentage points, is even smaller. The numbers of hu-

System code	Coverage		Precision		Topic-related unmarked units
	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged	
19	21.20% (1)	18.72% (1)	74.52% (2)	71.11% (2)	38.56% (6)
20	16.75% (4)	14.12% (5)	57.19% (6)	56.75% (6)	39.58% (5)
24	**17.90% (2)**	**17.68% (2)**	**69.84% (3)**	**69.73% (3)**	**39.77% (4)**
26	17.01% (3)	15.53% (3)	65.96% (4)	64.94% (5)	46.69% (1)
28	15.61% (5)	15.42% (4)	79.72% (1)	78.90% (1)	31.19% (7)

Table 4: Evaluation scores on abstracts for the top five systems, across all summary sizes using length-adjusted mean coverage. Systems listed in order of system code, with Columbia's scores in bold. Ranks shown in parentheses among all 8 systems submitting abstracts.

System code	Coverage		Precision		Topic-related unmarked units
	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged	
19	27.83% (1)	25.22% (1)	74.52% (2)	71.11% (2)	38.56% (6)
20	15.40% (5)	17.53% (5)	57.19% (6)	56.75% (6)	39.58% (5)
24	**17.87% (4)**	**19.37% (4)**	**69.84% (3)**	**69.73% (3)**	**39.77% (4)**
26	22.28% (2)	22.24% (2)	65.96% (4)	64.94% (5)	46.69% (1)
28	22.09% (3)	22.09% (3)	79.72% (1)	78.90% (1)	31.19% (7)

Table 5: Evaluation scores on abstracts for the top five systems, across all summary sizes using unmodified mean coverage. Systems listed in order of system code, with Columbia's scores in bold. Ranks shown in parentheses among all 8 systems submitting abstracts.

mans exceeding system performance on recall and precision remains relatively constant when we focus on either the 200 word or 400 word summaries, although the best humans increase their difference from the top system in the former case (to 9.74 points for macro-averaged recall, for example), and reduce it in the latter case.

Abstracts For the abstracts, we computed both unadjusted and length-adjusted coverage using the definitions provided by NIST, using both micro- and macro-averaging as defined earlier. We also calculated precision (micro- and macro-averaged), and we've included our score on related but unmarked units, which indicates how many of the system summary sentences were related to the topic of the summary. For macro-averaging we employed the mean coverage within each summary rather than the median. Since a large percentage of model units are not covered at all in any peer summary, the median is often very low and obscures differences in coverage between systems for the model units that they do cover.

Table 4 shows the length-adjusted scores across all articles, where we rank second in coverage and third in precision, regardless of whether micro- or macro-averaging is used. We are fourth in the score of related but unmarked units. System 19 has done the best on abstracts, ranking first on coverage and second on precision, using these calculations. However, system 19 does worse on unmarked, related units ranking sixth. System 28 (which ranked fourth on extracts) is a distant third here, ranking first on precision but fourth (macro-averaging) or fifth (micro-averaging) on coverage and seventh on the related but unmarked units. System 21, which performed best on extracts, did not participate in the abstracts evaluation.

If we use instead the unmodified coverage metric, we score somewhat lower on coverage, ranking fourth under both micro- and macro-averaging (Table 5). System 19 retains the top position, and system's 28 position improves slightly to third.

7 Issues with the Evaluation

DUC has provided a more focused collection of document sets this year, identifying separate types such as single events and multiple events and restricting the time span of the articles in each set. However, our analysis of the test documents and the scores obtained by various systems indicates that we still have a considerable way to go in order to focus on document sets that fully utilize today's systems capability of absorbing large amounts of information and taking advantage of the repetition of closely related text.

The single event sets were often broader than the kinds of single event sets that we routinely process in Newsblaster, our online version of DUC input. Given the number of articles we get on one day alone, we are able to produce much more focused sets than are available in DUC; events in news stories do not typically span more than a couple of days, much less the seven in DUC's single events or the even broader document sets provided for multi-event categories. Furthermore, Newsblaster input sets often have far more documents (e.g., often reaching 60 or 70) than in DUC, where the largest input set for single events was 16. Given the lack of training data on single events (there were only two sets from last year that could be considered single event), it was impossible to predict any kind of system parameters a priori.

For the multi-event input sets, the breadth of the document sets made it extremely difficult to determine what a good summary should be. Most systems did poorly on less coherent document sets. For example, if we divide the 2002 data according to the time span criterion (7 days or less versus more than 7 days), the average macro-averaged mean coverage for all systems across all summary lengths is 21.89% on the smaller time span set, and the macro-averaged precision is 72.55%. In contrast, the same measures for the document sets spanning more than 7 days are 15.25% and 57.86Even the humans seemed to have a hard time determining how to produce a good summary. Their summaries were often wildly different from each other and the abstracts did more generalization than could be expected to be produced by systems that rely on extracted sentences and even phrases.

In the remainder of this section, we describe problems with the single-event document sets and then show by example, the difficulties in summarizing the multi-event document sets in which documents were most diverse.

7.1 Single-event document sets

Multigen, as well as other summarization systems (Mani and Bloedorn, 1997; Carbonell and Goldstein, 1998), assumes that repetition of information is an indicator of importance. One would ideally want such systems to produce summaries similar to those produced by human summarizers. Thus, a pertinent question is whether or not abstracts manually produced by the DUC assessors themselves contain information repeated in more than one article. To investigate this question, we examined five sets of articles of the type ”single event” (D061, D092, D095, D096, D101) and their summaries produced by humans. For each sentence in a summary, we examined how many articles contain information conveyed in the sentence. Note that repetition does not require that the exact same sentence or phrase be used, but that the same information be conveyed regardless of the wording that is used. This analysis was performed manually to ensure the accuracy of the matching process. Our experiment revealed that from 29 summary sentences only 11 (37.9%) sentences contained repeated information. This result seems to refute our basic assumption about repetition-based content selection.

However, a closer look at this data gives an interesting insight on the connection between input type and a summarization strategy. While all the sets we examined are of single event type, the events vary greatly in granularity. For example, the articles in the set D095 describe a school shooting event and are similar in style and in the type of information included (i.e., the event description, sketches of victims and the murderer). It is not surprising that the same information often appears more than once across these articles. From seven summary sentences, five contain information which appear in several articles. However, the ratio of repeated information is lower in summaries of more heterogeneous sets of articles. The set of articles D096 contains articles on several topics related to the Super Bowl, and there is little overlap in their content. Interestingly, only two out of seven summary sentences appeared in more than one article. While our analysis is limited in scope, it suggests that repetition-based content selection is a good strategy for ho-

Source	Date	Subject	Appointment Type
AP	17-05-89	Senate endorses Secchia for Italy	Political
AP	15-07-89	Gildenhorn named for Switzerland	Contributor
AP	01-08-89	Spiers to the U.N.	Career
AP	02-01-90	Hinton to Panama	Career
AP	20-04-90	Wells to be named for Nicaragua	Woman, Career
AP	10-05-90	Shlaudeman to Nicaragua – Wells withdrawn	Career
FBIS	22-03-94	Egan to Jordan	Career
LA	02-02-89	Lilley to China	Former CIA
LA	03-02-89	Negroponte to Mexico	Career
LA	26-07-89	Zappala, Spain; Sember, Australia; Newman, N.Z.	Political
WSJ	03-02-92	Pickering to India	Career

Table 6: Documents in document set D119, a list of diplomatic appointments spanning two administrations.

mogeneous sets of articles, but it can be suboptimal for more heterogeneous sets. The DUC "single event" type sets are versatile in their structure and topic; thus, fine-grained classification among different types and multiple summarization strategies are required even within the single event type.

7.2 Summarization for diverse document sets

Document set D119 illustrates many of the difficulties with the evaluation. The 11 articles cover a time span of February 1989 to March 1994 (Table 6). The obvious tie is that they are all proposed or confirmed diplomatic appointments. All but one of the documents covers a single appointment discussing the particular issues of the nominee and the post for which he or she is proposed.

The summarizer, man or machine, has no clear guideposts on what point to draw from the collection. Neither repetition of information nor frequency of terms will help identify some meaningful topic. The human summarizer identified as *F* focused on politics. Here is his 100-word abstract:

> Some of President Bush's nominees for ambassadorships have been selected as a reward for large money contributions to the Republican Party, according to Democratic senators, and some retired diplomats consider a few not to be qualified. These nominees faced strong questioning from the Democrats during committee hearings but did manage to get a slim approval from the committee. Conservative senators objected to two of Bush's selections. Most of the president's picks won easy approval. Two of Bush's picks received highly favorable reaction: a Latin American expert for ambassador to Nicaragua, and the naming of Thomas Pickering, a well-known diplomat, to India.

However, this summary is misleading since of the 13 individuals discussed in the set, only five had political connections, like big donors, and the remaining eight were career diplomats or held other governmental positions, including one from the C.I.A. (see Table 6).

The human summarizer identified as *H* discerned the parallel themes and employed an economical writing style to identify three categories—veteran diplomats, contributors and others—and provide a list for each. Note that to generate a summary like this, which seems to be a good approach, a summarization system must be able to generalize from instances presented in the documents, a capability that is beyond any approach that relies primarily on extraction.

Instead, all the automatic summaries selected a small group of the individuals and presented them in no particular order. None were comprehensive like the *H* summary. The summary produced by DEMS, which at this point relies primarily on sentence extraction with rewriting of references, was typical:

> The President Bush administration is expected to name career diplomat Harry Shlaudeman as the first U.S. ambassador to Nicaragua in almost two years, a U.S. official says. Bush has decided to nominate John D. Negroponte, a veteran diplomat who helped direct U.S. aid to Nicaraguan rebels, to the key position of ambassador to Mexico, Administration officials said Thursday. Bush plans to name United Nations Ambassador Thomas Pickering as

U.S. envoy to India, and appoint the current head of the foreign service to succeed him.

The *F* summary was used as the model despite the fact that it does not seem to have been the best human summary produced. As a result, any system which focused more on political appointments would score higher. However, there seems to be no basis for this summary bias. There were three units in the DEMS summary, two of which were marked and one unmarked, meaning that two matched at least some part of the model summary. The unmarked sentence was given a score of 1.0 in relevance to the topic of the summary. It is unclear to us that this score is meaningful.

8 Suggestions for System Improvements

The results of the evaluation shed some light on improvements that we can make in the Columbia Summarizer for the future.

For multi-event document sets, we can improve coherence in the summaries produced by DEMS. Often sentences are extracted from different articles and placed side by side. In order to avoid the impression that these sentences related to the same event, we can prefix each sentence with the date and article tag to indicate its source. Second, we can develop a finer-grained classification of the type of input document set. We suspect that there are more than just three document set types in the input data and that if we can distinguish them, we can tune parameters to better adjust to input set. Finally, we feel we need to make more use of named entities in the unit selection. Currently, we use only capitalization to recognize named entities. If instead we did full named entity recognition at the beginning of preprocessing, then we would have more information to help in determining relatedness between sentences and important information. We also think that by adding coreference resolution (even a surface-based version that is not highly accurate), we can improve our techniques for sentence extraction.

In the area of single-event document summarization, we were severely hampered by the lack of training data and the fact that the test data is quite different from the input sets given to the summarizer in Newsblaster. We made extensive changes in our generation component in the final week of the system development. The resulting generated summaries from MultiGen in the test data had more grammatical errors than usual. Clearly, we should have taken more time to do these changes, but the lack of training data was also a factor. For the future, we will improve the linearization component by performing extensive training and debugging on a variety of new document sets.

Second, given the lack of repetition in the DUC input document sets and the fact that MultiGen relies on finding similarities to produce a good summary, we need to develop a better scoring system to complement the similarity-based content selection of Multigen. In the current version, the salience of the theme is computed according to its size and its lexical chains score (Barzilay and Elhadad, 1997). While this strategy yields sufficient results on Newsblaster-type input where we have many articles on the same event in one day, a more elaborate scoring strategy is required for the heterogeneous input sets used in DUC. Our analysis of summaries produced by humans for DUC revealed that the summaries sometimes contain information stated in only one article. In the future, we plan to further analyze these sentences in order to identify their characteristics and we will modify our scoring function accordingly.

9 Conclusions

Our main conclusion from this round of DUC is that the input document set plays a large role in whether a good summary can be generated or not. The broader the input set, the more difficult it is to determine what a good summary is. Human summaries degenerate and agreement between humans is quite low. Even in single-event input sets, the input document sets are broader than would be expected. This observation is consistent with the detailed statistical analysis we performed last year on DUC 2001 data (McKeown et al., 2001). In that analysis, we showed that the input set was the most important factor in predicting the summary score; much more important than the system that did the summarization, or even than whether the summarization was done by an automated system or a human. Although DUC has taken steps in organizing this year's data in subcategories that try to account for this variability, too much variance remains within these categories. Furthermore, the fact that documents are hand-picked means that the kind of repetition and information overload that one finds in real-world environments is missing.

Our strong feeling is that summarization systems can achieve better results when the input can be automatically sorted and categorized to yield meaningful input set types. Thus, one possibility for future evaluations is to create a system which can automatically filter and cluster large quantities of online data creating document sets according to a defined set of criteria. These criteria should yield sets that fit the capabilities and goals of current summarizers. Thus, very loosely connected sets might be filtered out and the system might produce larger numbers of documents on a narrowly defined event. In this way, the quantity of input can be scaled to become closer to real-world environments. Such an approach would better test the capabilities of existing summarizers.

Acknowledgments

The research reported on within this paper was supported by the Defense Advanced Research Projects Agency under TIDES grant NUU01-00-1-8919. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the funding agency.

References

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, August. Association for Computational Linguistics.

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

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2001. Sentence ordering in multidocument summarization. In *Proceedings of the 1st Human Language Technology Conference*, San Diego, California.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, August.

Vasileios Hatzivassiloglou, Judith L. Klavans, and Eleazar Eskin. 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, Maryland, June. Association for Computational Linguistics.

Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. 2001. SIMFINDER: A flexible clustering tool for summarization. In *NAACL Workshop on Automatic Summarization*, pages 41–49. Association for Computational Linguistics.

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

Kathleen R. McKeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-99)*, pages 453–460, Orlando, Florida, July. American Association for Artificial Intelligence.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Barry Schiffman, and Simone Teufel. 2001. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the Document Understanding Conference (DUC01)*.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the 2nd Human Language Technology Conference*, San Diego, California.

Yael Ravin, Nina Wacholder, and Misook Choi. 1997. Disambiguation of proper names in text. In *Proceedings of the 17th Annual ACM-SIGIR Conference*.

Barry Schiffman, Ani Nenkova, and Kathleen R. McKeown. 2002. Experiments in multi-document summarization. In *Proceedings of the 2nd Human Language Technology Conference*, San Diego, California.