Experiments in Multidocument Summarization

Barry Schiffman
Columbia University
New York, NY 10027, U.S.A.
bschiff@cs.columbia.edu

Ani Nenkova
Columbia University
New York, NY 10027, U.S.A.
ani@cs.columbia.edu

Kathleen McKeown
Columbia University
New York, NY 10027
kathy@cs.columbia.edu

ABSTRACT

This paper describes a multidocument summarizer built upon research into the detection of new information. The summarizer uses several new strategies to select interesting and informative sentences, including an innovative measure of importance derived from the analysis of a large corpus. The system also computes concept frequencies rather than word frequencies as an additional measure of importance. It merges these strategies with a number of familiar summarization heuristics to rank sentences. The initial version of the summarizer performed successfully in the evaluation reported at the Document Understanding Conference last year, although the system addressed only the content of the summary and not the presentation. We also discuss here the procedures we are developing to improve the presentation and readability of the summaries.

General Terms

Multidocument summarization

Keywords

Importance metrics, text generation, text analysis

1. INTRODUCTION

The extreme variety in multidocument summarization makes it hard to anticipate all of the demands that would be made on a general purpose summarization system. The input set of articles may be very loosely tied together; for example, it may have the same general topic, such as earthquakes in recent years or third-world debt, or it may be focused on one person or entity but over a long time span. On the other hand, the input set may be narrowly focused on a particular event, such as a mayoral election. The training and test sets for multidocument summarization developed by NIST for the Document Understanding Conference demonstrate the large number of possible variations.

Summarization of document clusters that are tightly focused on a single event is the research focus of Columbia's MultiGen [11], a summarization system that uses information fusion and similarities across input articles. However, since many of the documents in the DUC training corpus in 2001 were only loosely connected, we needed to develop an alternative summarization strategy. Our approach for handling other types of document clusters builds on experiments we were carrying out in detecting new information. One problem for new information detection is that most sentences in one article will be to a great extent different from the sentences in another article [1]. In moving to summarization, we emphasize the problem of determining which of the many different sentences are actually important enough to be included in a general-purpose summary. In turn, our experiments in sentence extraction for summarization will help refine our techniques in new information detection. These experiments were implemented in the DEMS summarizer, the Dissimilarity Engine for Multidocument Summarization.

DEMS produces summaries by extracting the top-ranked sentences until the desired length is met. To do the ranking, DEMS scans all the sentences in the input set of articles and assigns values to features, some of which try to measure the inherent importance and interest of the thought. Some of the features were derived from a large background corpus and will be described in later sections. The combination of innovative features with established techniques, such as increased weight to sentences near the beginning of articles, resulted in a reasonably successful experimental summarizer.

The Columbia Summarizer, used in the DUC evaluation, used different strategies for different types of document clusters. It first examined the document clusters and routed sets to MultiGen if the articles were determined to be on the same event; document clusters on topically related, but different events, were routed to DEMS. In the evaluation in 2001, DEMS was assigned 29 of the 30 sets. In the 2002 evaluation, which is not yet complete, DEMS was assigned 33 of the 59 clusters of articles.

DEMS is now used in tandem with MultiGen in the Newsblaster on-line news browsing system [12]. Document sets with high similarity scores, as measured by the clustering module, are routed to MultiGen, and DEMS gets all others.

In the remainder of this paper, we focus entirely on our work in DEMS, presenting first the overall system and then the evaluation results, both in DUC and in a follow-up evaluation that we carried out of an enhanced DEMS.

2. IMPORTANT AND INTERESTING

Our approach relies on three main techniques to compute significance – identifying importance-signaling words through an analysis of lead sentences in a large corpus of news, identifying high-content verbs through a separate analysis of subject-verb pairs in a

HLT '02, San Diego, Calif., USA
.

Figure 1: DEMS architecture. The darker boxes represent the system as used at the Document Understanding Conference in 2001. The lighter boxes are the modules added since then.

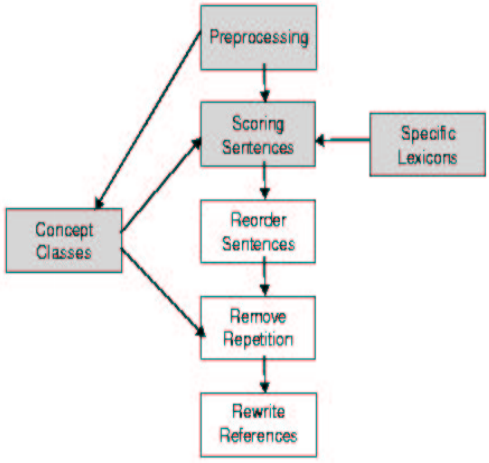
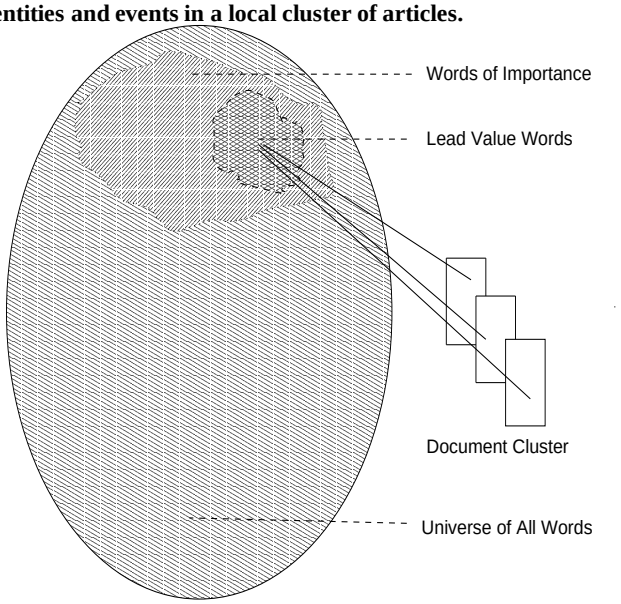


Figure 2: Lead words are those more likely to be found at the beginnings of news articles, and thus are likely to indicate importance in a general or global sense, without reference to the entities and events in a local cluster of articles.



news corpus, and finding the dominant concepts in the input clusters of articles – rather than frequent words. Figure 1 shows how the system is arranged.

2.1 Lead Values

It is well known that the lead sentences of news articles can often make excellent brief summaries [15], but it is difficult to decide which of them to choose in the multidocument setting. Further, the lead sentences do not always indicate the significance of the article. But, if we can identify topics that often appeared in the lead sentences, then by definition these topics should cover important and interesting information. We examined first a large corpus of New York Times articles from 1996, and later a corpus of Reuters articles from the same year to determine what features could distinguish lead sentences from the average sentence [18]. Using just the noninflected forms of the words as the features, we developed lists of 4,600 and 4,900 words from the two corpora that tended, with a reasonable measure of statistical significance, to be in the leads of articles more often than in the full text. We hypothesized that, on average, sentences with more high "lead words" would tend to reflect important events. Table 1 shows a sample of lead words from the Reuters corpus. The criteria for selecting the lead words is:

$$\frac{p(W_{inlead})}{p(W_{anywhere})} > 1$$

The ratios were checked for statistical significance with the binomial test and only those with ratios where $pvalue < 0.05$ were accepted for inclusion in the lexicon.

The lead words are used as binary values, and averaged over the entire article, so that the sentence richest in lead words gets the highest score. By using the lexicon of lead words we are often able to locate secondary topics of interest in the articles, and to make comparisons of importance across documents. Thus, the system goes beyond the simple technique of grabbing the lead sentences – a technique that forms one of the baselines in the DUC evaluation.

We regard the lead words as a beginning of our approach to assess importance in new-information detection, where a statement might have few entities in common with all the material previously seen. We make the distinction between *local importance*, that is importance in the context of the articles we have in a cluster, and *global importance*, or importance is some larger, universal context (Figure 2). The lead-words lexicon is then one way to identify information that is globally important. We are experimenting with others for use in both the summarizer and in new-information detection.

2.2 Verb Specificity

In an effort to put sentences with the maximum amount of content into the summaries, we borrowed an the idea of verb specificity comes earlier work on a a biographical summarizer [19]. In that work, we sought to retrieve from a large corpus a brief description and a short list of interesting events about a person or several people. We first extracted a short description of the person and used the head nouns in that description to select sentences with verbs closely associated with the kind of person the user was interested in, reasoning that these sentences would be more relevant.

In the biography work, we also experimented with the notion of verb specificity. If a verb was closely associated with only a few types of subjects, i.e. one that is highly specific, it would tend to convey information by itself in a sentence, and it would indicate a specific, well-defined event. For example, a verb like "arrest" suggests police activity. But less specific verbs (e.g., "be" or "do") occur with a wide range of subjects and objects.

We found in constructing the biographical summarizer that descriptions could be found with the aid of pattern matching. But representative events important to that person were more difficult to select. In a large corpus, a great number of sentences might mention the particular name the user was interested in, but not all of them were interesting or informative. The system ranked sentences on the basis of how closely tied the verbs were to the terms in the person's description. The association of subject nouns to verbs was computed on the basis of a large corpus study. Mutual information

Table 1: A sampling of the *Lead Value* Lexicon made from the 1996 Reuters news wire. The entries are also used only as binary values.

cynical	coaxing	eerie	renovator	cling	impressionism	cutter	tusker
worn-out	convalescent	vial	unplayable	waterlogged	syphilis	decathalon	dragonfly
gigantic	extricate	unbowed	cherry	waterborne	watershed	phenobarbital	reappearance
rivet	heady	beloved	placid	bloke	caravan	large-scale	windfall
petrol	dame	mend	truffle	chubby	enthral	enunciate	dank
stopgap	freak	pensive	meld	mortuary	well-kept	well-established	one-man
linguist	impresario	ostrich	possess	chump	crestfallen	menu	electronics
nationalize	restive	daub	vile	wizard	finalist	dishevelled	crossroad
autism	workable	reverberate	excitable	trawler	sizeable	insolvent	stewardess
rhyme	fluorescent	sharpen	infighting	setter	electrical	mesa	jeopardize
chunk	rude	rambunctious	polyglot	chivalry	statistical	impressionist	bloodbath
conscription	spectre	crowning	zealotry	intrusion	gutsy	westernmost	showpiece

statistics were collected from a year's worth of newswire.

The data suggested that many verbs were closely tied to a only a few classes of nouns. We derived a "verb specificity" measure that reflects how often the mutual information between a particular verb and one noun or another exceeded a threshold.

$$VerbSpecificity = \frac{Count(V_{mi>T})}{Count(V_{mi>0})}$$

In the DEMS summarizer, the highest verb specificity in a sentence is used as the feature, in order to give increased weight to sentences rich in content. The motivation is to identify sentences that convey a complete thought by themselves, without depending too much on the surrounding context.

2.3 Concepts

A key feature in DEMS was reliance on concepts instead of the individual nouns and verbs. Thus, while we also use two frequency measures, we count the occurrences of concepts rather than words. On a single word basis, many common nouns and verbs will occur only once in a document, but two or more of them might actually be referring to the same idea or entity. Thus, alone, the words might miss the point of the cluster, but when considered as a "Concept Set" would zero in on the main ideas.

To build the Sets, we collect equivalent nouns and verbs into classes of words that conceivably could refer to one another. At present, we use WordNet [13] synonyms, hypernyms and hyponyms, with some simple constraints to decide which words belong together. The constraints are imposed on words that have more than five senses, and therefore large numbers of synonyms – for example "matter" or "issue." The system has no means of disambiguating words and the sets tended to grow unacceptably large. By removing highly polysemous words, manageable sets are obtained. But we noticed that one sense often dominated some words with many senses. At the threshold of 5 senses, we lost words like "father." To restore some of the dropped words, we used the study of subjects and verbs mentioned above to find the words closest in usage to a number of polysemous words, and used that list in place of the WordNet entry.

This strategy, we felt, would provide a truer measure of "aboutness," and would produce higher weights for sentences that talk about the subject of the input set – or at least the topic that ties the articles together . In an ideal world, we would know which words referred to the key entities in the set, but at present we do not have a good way to resolve references either within documents or across documents. We find concept sets an acceptable alternative. Table 2 shows an example of the largest Concept Sets from one article.

Table 2: Sample concepts sets for one article.

war campaign warfare effort cause operation conflict	10
concern carrier worry fear scare	9
home base source support backing	7
arrive reach hand find receive	8
report announce	4
prepare mount launch plunge	4
anger fire fuel dismiss	4

2.4 Other Features

In addition, we used a number of other heuristics, many of which are found in other systems, such as additional value to sentences near the beginnings of articles, or the publication date to make sure the most recent information is included. We diminish the value of sentences that deviate too much from a normal size, which we set at 15 to 30 words, and of sentences with pronouns.

We settled on a set of weights experimentally, although we recognize these may not be ideal. Using machine learning techniques to derive weights or a set of rules for combining the features would incur a tremendous cost of preparing a training corpus.

Here is a list of the additional features:

Location A negative value that penalizes sentences that appear late in the document.

Publication Date Additional value to the most recent documents, on the assumption that users will want the most up-to-date information.

Target Indicates the presence of the central personage in the document cluster, if one exists.

Length A penalty for sentences that are below a minimum (15 words) and above a maximum (30 words). Short sentences are often require some introduction or reference resolution, or else are a kind of interjection. Long sentences can cover multiple thoughts that are often found elsewhere in the document cluster in single sentences.

Others Indicates the presence of any named entity, weighted to the frequency of that entity across all documents.

Pronoun A negative value on sentences that have pronouns in the beginning of the sentence.

Role A positive value in cases where a pronoun follows a named entity.

Lead Value, Set Frequency, Local Frequency and Verb Value are combined with each of the above features in a weighted sum of 11 features in total. The weights are determined experimentally, and have been revised periodically. Different sets of weight form different configurations, with Target and Publication Date emphasized for the biographical sets.

2.5 Ordering

In the initial version of the summarizer used at DUC 2001, we had no time to try to make the output read better or clearer.

Since the conference last year, we implemented a simple ordering scheme to group the sentences from different documents together and then order those from a particular document in order of appearance.

In addition we check for repeated sentences in the output. Repetition was not a serious problem in in the DUC evaluation because the documents were not very similar and the sentences did match across articles. We used only string matching and thus allowed nearly identical sentences in the summary, an occurrence which did happen. In Newsblaster, this repetition is much more of a problem since most of the articles in a cluster are very close matches.

DEMS now uses the Concept Sets instead of individual words to measure the unordered overlap of content words between each new sentence proposed for the summary and all previously selected sentences. By using the Concept Sets, two sentences could be perfectly matched without having any specific words in common. The overlap threshold in DUC 2002 was set at 60%, but it is reduced to 40% in Newsblaster, where the document clusters are usually more cohesive and contain many more passages that have large amounts of overlap.

2.6 Named Entities

References to named entities pose yet another problem to multidocument summarization. Summaries often contain cryptic mentions of people's last names or acronyms and as a result the text becomes unclear. Both intuition and corpus studies (e.g., [14]), suggest that it would be most natural to include people's full name and description at their first mention, and use shorter ways, e.g. last name only, to refer to them at subsequent mentions.

We experimented with an initial algorithm for modification of references to people within the summary. IBM's NOMINATOR system [16] was used to extract named entities from the input cluster.

Titles and premodifier descriptions are also extracted, for example "French President Francois Mitterand", "Actress Elizabeth Taylor", "Los Angeles Police Chief Daryl Gates" (Table 3). In this way, a list of all possible ways to refer to a named entity occurring in the summary is constructed. After ordering the sentences in the summary as described in the previous section, references are substituted so that the longest variant of the name, possibly including a title, appears as first mention and subsequent repeated variants are substituted with the shortest most common name variant.

"Shortest most common" is defined as the shortest way to refer to a person that occurs at least N times in the input documents, and N = number of times the person was referred to by name in the input/number of name variants for that person. This restriction is imposed so that atypical names such as nicknames can be avoided.

Even this simple algorithm seems to improve the clarity and readability of the summary. Immediate future work will focus on coming up with more sophisticated ways of picking variants for subsequent mentions and also ways for identifying and including postmodifying descriptions at first mention.

Table 3: Count of references to the actress Elizabeth Taylor in a cluster in the DUC 2001 test corpus.

Actress Elizabeth Taylor	1
Ms. Taylor	1
Elizabeth	1
Liz Taylor	3
Elizabeth Taylor	8
Liz	11
Taylor	31
Miss Taylor	33

3. RELATED WORK

The closest system in spirit to DEMS is NEATS [7] which used topic signatures. Topic signatures are derived from co-occurence statistics on preselected documents, while DEMS relies on a WordNet-based dictionary. The goal of topic signatures and Concept Sets is similar: to give increased weight to salient statements in the document cluster. Topic signatures focus on associations, like the association between words like restaurant and waiter. Concept Sets focus on equivalent meanings like restaurant and eatery and strives to make finer-grained distinctions. The two systems also use similar ordering schemes. However, NEATS has no equivalent mechanism to the DEMS features that try to assess global importance.

DEMS differs from many summarization systems in that it uses no TF/IDF calculation. Lin and Hovy also found that TF/IDF was less effective than topic signatures in the summarization task [6]. In contrast, many summarization systems do use this measure of importance borrowed from Information Retrieval research, including those developed at the University of Texas [5] and the University of Michigan [3].

A number of systems use similarity to discern importance, but DEMS emphasizes statements that are different, and treats importance as a separate issue, giving no weight to similar passages. Systems that do measure similarity include MultiGen, the University of Texas system [5], focusing on information extraction techniques, and the ISI system [9], which used discourse structure. A group at CMU [4] uses cosine similarity of vectors in the MMR algorithm. A graph representation of several relationships between words is used to find similarities and differences between pairs of articles [8].

The use of the lead-words feature is related to a technique in the machine learning community, in which researchers have used existing corpora that are in some way preselected or partially annotated for one purpose or another. In an information extraction experiment Mark Craven at CMU [2] used what he called "weakly" labeled data to reduce the cost of annotating a training corpus. He was seeking a way to map medical texts into a structured data base. He used a database that contained links to related text articles. Another group working on information extraction at CMU, Seymour and others, sought to build a database of information about computer scientists from "distantly labeled data" composed of the header information on research papers [20]. Ellen Riloff learned textual-syntactic patterns for information extraction by comparing two corpora, a target containing the information she was interested in, and the other a general corpus [17]. The idea is that patterns of specialized words and syntactic structures will show up in greater numbers in the target corpus than in the general corpus.

Table 4: Overall preferences of human judges for reordered and named-entity substitution.

modified	48%
non-modified	13%
no pref	39%

Table 5: Breakdown of judges opinions.

	overall readability	*ordering*	*references*
modified	60%	78%	38%
non-modified	0%	0%	0%
no preference	40%	22%	62%

4. EVALUATION

4.1 At DUC

Our system was evaluated at DUC 2001. Figure 3 and Figure 4 show precision and recall respectively, compared with the averages of all automated systems. The systems were measured against a human-written summary for each cluster. The DEMS system scores are shown in both bar graphs in black and the average of all systems in white. Both charts list the sets in ascending order of the average score to try to reflect a rough scale of difficulty.

In an earlier analysis of the results, we found that either the sets or the human-summaries or the evaluators' judgments varied considerably [10]. In that analysis, the Columbia system was among the four top systems, which all performed in a relatively narrow range. The DUC organizers did not assign any kind of overall score to the participants.

4.2 After DUC

We carried out a post-DUC evaluation to determine whether the changes in presentation that we made to DEMS improved summary quality. The ten top-scoring summaries produced by DEMS for the DUC competition were chosen, the proposed ordering and name substitutions performed and the resulting pairs were given to human judges to assess. Results show that ordering considerably improves the summaries' quality. Named entity substitution also led to improvements, though they made smaller difference.

Three human judges were asked to compare the pairs of modified and non-modified text and to say which one they prefer in terms of 1) overall readability 2) ordering 3) references to people. The judges had the option to either express a preference to a text or state that the two variants are equal. The study was done on 200-word summaries. The overall distribution of preferences is shown in Table 4.

Since we did not define any criteria for good readability, good ordering or appropriate reference sequences, each human used his/her own understanding and the results suggest that those vary substantially– only in 27% of the cases did the opinions of the three judges coincide.

For each question, the majority answer was taken as final; that is, if two humans agreed in a judgment, that was taken as a final judgment for the text. Since there were three choices, the possibility existed that final judgment could not be reached because each human chose a different answer. This happened twice for judgments about references, once for ordering judgments and zero times for overall readability. These facts reveal which assessments are more difficult to make (slight errors in the substitution algorithm can cause disagreement). Table 5 shows the distribution of preferences according to majority opinion.

Even though the experiment was relatively small, it allows us to draw some useful conclusions. First, it shows that even simple methods can improve the quality of the summary and this suggests that it's reasonable to look for ways to assess summaries not only based on content, but also on the basis of readability and naturalness of the text characteristics.

The benefit from name substitution is not immediately evident. There are two major reasons for that – even though the strategy is very sensible in theory, in practice two problems might arise. a) Not all summaries contain (many) names and thus it's not possible to apply substitutions. b) Errors might be introduced in the summary by the substitution algorithms because of occasional errors in the named entity recognition output. In order to get a better sense of the utility of the approach, we applied it on the entire 2001 DUC corpus, both training and testing, totaling 60 sets. Then we manually examined the substitutions suggested by the algorithm for all 60 summaries. We defined the notion of "good substitution" as a substitution of one of the following kinds—1) first mention is a last name, and full name and title are available 2) first mention is a last name and full name is available 3) first mention is an acronym and full name is available 4) full name is used after the first mention and it can be substituted with last name only. Substitutions were possible in 35 of the sixty summaries; a total of 83 good substitutions were suggested, 49 of which involved first mentions and 18 problematic substitutions were proposed.

5. CONCLUSION AND FUTURE WORK

Overall we were pleased with the results, since DEMS was built in about a month. We began work only after reviewing the DUC training sets, when we found that most of the clusters were too diverse for MultiGen, which had been in development for five years.

We are planning improvements along a number of broad areas:

- Categorization of the document clusters. The clusters of articles both in the DUC evaluations and those collected in the Newsblaster system are quite varied, and we need a finer-grained analysis of the relationship between the member articles than we are currently using. The weights for all the features can be easily set automatically once we determine the kind of set we are dealing with. If the set is centered around a person, more weight can be given to sentences which mention that person. If the set is extremely diverse, like a group of articles about different events of the same kind, like volcanoes or political assassinations, more emphasis can be given to global importance while no attention need be paid to publication date.

- Improvement of Concept Sets. We are planning to examine additional ways to refine the data in WordNet and to discover a method for adding new links between less common words. Although we have not conducted an evaluation of the Sets, it is clear that we frequently fail to group words that are in fact linked in the text, and at the same time inappropriate words are added to various sets. Ultimately, the solution would be a comprehensive system to resolve all nominal references in order to get an accurate count of the frequency of a concept in a document. Since we are dealing with sets of documents, the co-reference system would have to be able to operate across the documents.

- We are expanding our rewriting component and will eventually deploy a full-scale text generation capability to the system. The final shaping of the summary is a difficult problem for the kind of diverse document clusters that DEMS

Figure 3: Precision: Dems in the black, compared to average

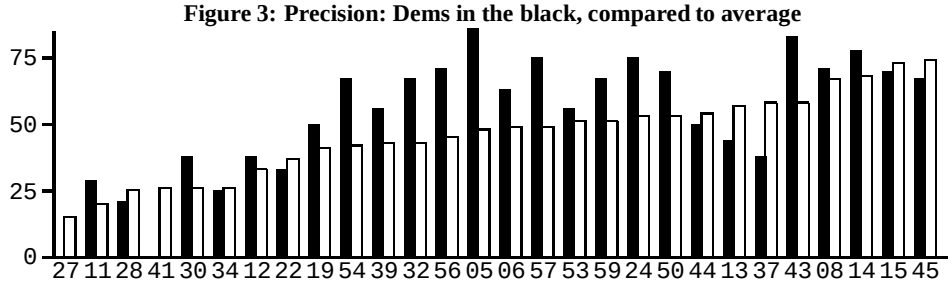
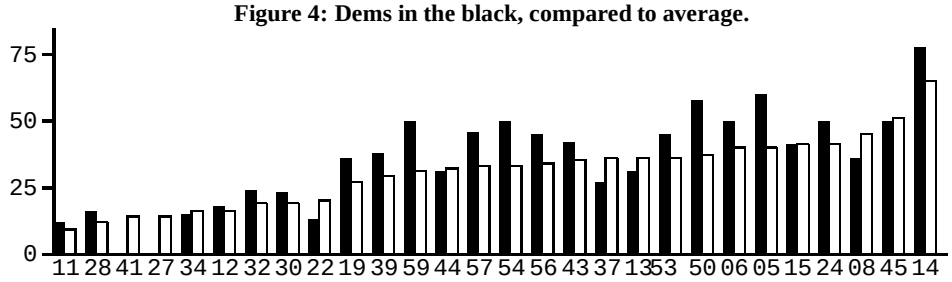


Figure 4: Dems in the black, compared to average.



deals with. Since the sentences across documents do not align well, DEMS cannot use an approach like that of Multi-Gen, which cuts and pastes similar phrases from sentences that cover the same ground. We are exploring alternative approaches to rewriting in this context.

- Ultimately we will add a module to highlight new information, as well as other contrasts, such as different perspectives, between documents.

6. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In *Proceedings of the ACM-SIGIR Conference*, 2001.

[2] M. Craven. Learning to extract relations from medline. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[3] H. J. Dragomir R. Radev and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extaction, utility-based evaluation and user studies. In *Proceedings of ANLP/NAACL-2000 Workshop on Automatic Summarization*, 2000.

[4] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of ANLP/NAACL-2000 Workshop on Automatic Summarization*, 2000.

[5] S. Harabagiu, D. Moldovan, P. Morarescu, F. Lacatusu, R. Mihalcea, V. Rus, and R. Girju. Gistexter: A system for summarizing text documents. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[6] C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Annual International Conference on Computational Linguistics*, 2000.

[7] C.-Y. Lin and E. Hovy. Neats: A multidocument summarizer. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[8] I. Mani and E. Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings, American Association for Artificial Intelligence 1997*, 1997.

[9] D. Marcu. Discourse-based summarization in duc-2001. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[10] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, B. Schiffman, and S. Teufel. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the Document Understanding Conference (DUC01)*, 2001.

[11] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. Towards multidocument summarization by reformation: Progress and prospects. In *Proceedings of American Association for Artificial Intelligence 1999*, 1999.

[12] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Human Language Technology Conference*, 2002.

[13] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312, 1990.

[14] M. Poesio and R. Vieira. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, 1998.

[15] L. Rau, R. Brandow, and K. Mitze. Domain-independent summarization of news. In *Summarizing Text for Intelligent Communication*, 1994.

[16] Y. Ravin, N. Wacholder, and M. Choi. Disambiguation of proper names in text. In *Proceedings of the 17th Annual ACM-SIGIR Conference*, 1997.

[17] E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the 13th National Conference on Artificial Intelligence*, 1996.

[18] B. Schiffman. Building a resource for evaluating the importance of sentences. In *Proceedings of the Third Annual Conference on Language Resources and Evaluation*, 2002.

[19] B. Schiffman, I. Mani, and K. J. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings European Association for Computational Linguistics 2001*, 2001.

[20] K. Seymore, A. McCallum, and R. Rosenfeld. Learning hidden markov model structure for information extraction. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

APPENDIX

A. EXAMPLES

Here is an example comparing a nonmodified summary, from a recent Newsblaster summery on weapons inspections in Iraq. Figure A shows an ordered summary for noun-phrase substitution is put in. Figure A shows the result after the references to people are reworked.

Figure 5: Summary without name substitution

Wolfowitz said Iraq is among countries trying to develop weapons of mass destruction that " would make September 11 pale by comparison."

In his State of the Union address, **Bush** named Iraq, Iran and North Korea part of an "axis of evil" bent on developing chemical, nuclear or biological weapons.

Wolfowitz said the **Bush** administration would prefer to find a diplomatic solution to the Iraq issue, but **Hussein** "has shown great resistance to accepting any reasonable outcomes."

Deputy Defense Secretary Paul Wolfowitz told CNN's "Novak, Hunt and Shields" that **President Bush** has not made a decision about military action against Iraq.

Figure 6: Summary with name substitution

Deputy Defense Secretary Paul Wolfowitz said Iraq is among countries trying to develop weapons of mass destruction that "would make September 11 pale by comparison."

In his State of the Union address, **President Bush** named Iraq, Iran and North Korea part of an "axis of evil" bent on developing chemical, nuclear or biological weapons.

Wolfowitz said the **Bush** administration would prefer to find a diplomatic solution to the Iraq issue, but **Iraqi President Saddam Hussein** "has shown great resistance to accepting any reasonable outcomes."

Wolfowitz told CNN's "Novak, Hunt and Shields" that Bush has not made a decision about military action against Iraq.