The Impact of Frequency on Summarization

Ani Nenkova
Department of Computer Science
Columbia University
New York, NY 10027
ani@cs.columbia.edu

Lucy Vanderwende
Microsoft Research
One Microsoft Way
Redmond, Washington 98052
lucyv@microsoft.com

Abstract

Most multi-document summarizers utilize term frequency related features to determine sentence importance. No empirical studies, however, have been carried out to isolate the contribution made by frequency information from that of other features. Here, we examine the impact of frequency on various aspects of summarization and the role of frequency in the design of a summarization system. We describe SumBasic, a summarization system that exploits frequency exclusively to create summaries. SumBasic outperforms many of the summarization systems in DUC 2004, and performs very well in the 2005 MSE evaluation, confirming that frequency alone is a powerful feature in summary creation. We also demonstrate how a frequency-based summarizer can incorporate context adjustment in a natural way, and show that this adjustment contributes to the good performance of the summarizer and is sufficient means for duplication removal in multi-document summarization.

1 Introduction

There has been a growing interest in summarization in the past years, and in particular, the large amount of on-line news and information has led to the development of numerous multi-document summarizers for newswire[1], as well as online systems such as NewsInEssense[2] and the Columbia Newsblaster[3] that run on a daily basis. The main problem an extractive summarization system needs to solve is content selection, i.e., deciding which sentences from the input documents are important enough to be included in a summary. Even systems that go beyond sentence extraction and use generation techniques to reformulate or simplify the text of the original articles need to decide which simplified sentences should be chosen, or which sentences should be fused together or rewritten (Barzilay et al., 1999; Jing, 2000; Vanderwende et al., 2004). The usual approach for identifying sentences for inclusion in the summary is to train a binary classifier (Kupiec et al., 1995), a Markov model (Conroy et al., 2004), or directly assign weight to sentences based on a number of features and heuristically chosen feature weights and pick the most highly weighted sentences (Schiffman et al., 2002; Lin and Hovy, 2002). Invariably, term frequency is among the features used to determine sentence importance, but the relative contribution that term frequency makes is generally not reported. In this paper, we study the association of frequency of words and content units in the input and the likelihood with which they will appear in a human summary (see section 2). The study shows that frequency is indeed a powerful predictor for human choices in content selection, but it does not completely explain these choices. In section 3, we discuss SumBasic, a summarization system that we created to exploit frequency exclusively, in order to isolate the contribution of frequency. In section 4 we test the performance of SumBasic against that of the other systems that entered the 2004 Document Understanding Conference (DUC), and the 2005 Multilingual Summarization Evaluation (MSE) that was carried out as a part of the Machine Translation and Summarization Evaluation

[1] See for example http://duc.nist.gov

[2] http://lada.si.umich.edu:8080/clair/nie1/nie.cgi

[3] http://newsblaster.columbia.edu

workshop at ACL 2005[4]. We also discuss how directly modeling frequency in summary creation further eliminates the need for a separate duplication removal component, which we discuss in section 5. We conclude with a discussion of findings and future work.

2 Frequency and human summarization behavior

One of the issues studied ever since the inception of automatic summarization in the 60s was that of human agreement (Rath et al., 1961): different people can choose different content for their summaries (Halteren and Teufel, 2003; Radev et al., 2003; Nenkova and Passonneau, 2004). More recently, others have studied the degree of overlap between input documents and human summaries (Copeck and Szpakowicz, 2004; Banko and Vanderwende, 2004). In this section, we focus on frequency, investigating the association between content that appears frequently in the input, and the likelihood that it will be selected by a human summarizer for inclusion in a summary. This question is especially important for the multi-document summarization task, where the input consists of several articles on the same topic and usually contains a considerable amount of repetition of the same facts across documents. We first discuss the impact of frequency in the input at the word level, and then look at frequency at a semantic level, using semantic content units.

2.1 Word frequency

In order to study how frequency influences human summarization choices, we used the 30 input sets for the multi-document summarization task from DUC 2003. For each set, NIST provided four human-written summaries and the submissions of all participating automatic summarizers. Table 1 shows the percentage of top frequency content words (a stop list was used to exclude pronouns, function words etc from consideration) from the input documents that also appear in the human models. In order to compare how many of these matches are achieved by a good automatic summarizer, we picked one of the top performing summarizers and computed how many of the top frequency words from the input documents appeared in its automatic summaries, which is also shown in table 1.

	Top 5	Top 8	Top 12
human	94.66%	91.25%	85.25%
machine	84.00%	77.87%	66.08%

Table 1: Percentage of the top n frequency words from the input documents that appear in the four human models and in a state-of-the-art automatic summarizer (average across 30 input sets)

These figures suggest two things—1) the high frequency words from the input are very likely to appear in the human models, confirming that frequency is one of the factors that impacts a human's decision to include specific content in a summary, and 2) the automatic summarizer we examined includes fewer of these high frequency words and the overlap with the human models can be improved if the inclusion of these words is targeted. Trying to maximize the number of matches with the human model is reasonable, since on average across the 30 sets, the machine summary contained 30 content words that did not match any word in a human model.[5]

2.2 Word frequency and human agreement

In the previous section we observed that the high frequency words in the input will tend to appear in *some* human model. But will they be words that the humans will agree on, and that will appear in *many* human summaries? Here, in order to qualify the frequency of words, we used the probability of a word in an input set—$\frac{n}{N}$, where n is the number of times the word occurred in the input (its frequency) and N is the number of all words in the input. It is obvious that the high-frequency words will be those with high-probability in the input.

We found that in fact the words that human summarizers agreed to use in their summaries include the high frequency ones and the words that appear in only one human summary tend to be low frequency

[4] `http://www.isi.edu/~cyl/MTSE2005/MLSummEval.html`

[5] Even though no rigorous study of the issue has been done, it can be considered that the content words that do not match any of the models describe "off-topic" events. This is consistent with the results from the quality evaluation of machine summaries in which human judges perceived more than half of the summary content to be "unnecessary, distracting or confusing".

used in	number words	average probability
4	200	0.0116
3	307	0.0049
2	703	0.0029
1	3107	0.0011
–	——	———
0	29816	0.0007

Table 2: Number of summaries that used the word, number of words in the class, the average probability of the words in the class. The numbers are computed across the 30 DUC'03 sets.

words (see table 2). At the same time, a large percentage of the words in the models were used by only one or two humans and had lower probabilities, which suggests that a successful automatic summarizer should also incorporate some mechanism that in certain cases will allow low frequency words to have bigger impact on sentence weights.

In the 30 sets of DUC 2003 data, the state-of-the-art machine summary contained 69% of the words appearing in all 4 human models and 46% of the words that appeared in 3 models. This indicates that high-probability (high-frequency) words, which human summarizers will tend to select and thus will be rewarded during evaluation, are missing from the summary.

2.3 Frequency of semantic content units

In the previous two sections we established that high-frequency content words in the input will be very likely to be used in human summaries, and that there will be a consensus about their inclusion in a summary between different human summarizers. But the co-occurrence of *words* in the inputs and the human summaries does not necessarily entail that the same *facts* have been covered. A better granularity for investigating frequency impact is the semantic content unit—an atomic fact expressed in a text. We now briefly describe content units as introduced by (Nenkova and Passonneau, 2004) and then turn to studying content unit frequency.

Content units are atomic facts expressed in the text, such as "*an airplane crash occurred*", "*the crash happened of the coast of Nova Scotia*", "*the reason for the crash is still unkown"*.

Summarization content units and the Pyramid evaluation method

Recent work on summarization evaluation (Halteren and Teufel, 2003; Radev et al., 2003) has emphasized that it is necessary to determine not only what information can be included in the summary, but also how important this information is. The pyramid method (Nenkova and Passonneau, 2004) has shown how different summary content units (SCUs) can be assigned weights and how highly weighted units can be considered as more essential for a summary than not so highly weighted ones. The approach using content unit annotation shows how the notion of importance can emerge when multiple human summaries are used to construct the gold standard. The more people agree that a certain SCU should be included in a summary, the more important this SCU is.

Content unit frequency

Evans and McKeown (2005) annotated 11 sets of input documents and human written summaries for content units following the approach of (Nenkova and Passonneau, 2004). They kindly provided us with their annotation and we were able to measure how predictive the frequency of content units in the documents is for the selection of the content unit in a human summary. As in our study for words, we looked at the top n most frequent content units in the inputs and calculated the percentage of these that appeared in any of the human summaries. Similarly, of the top 5 content units, 96% appeared in a human summary across the 11 sets. The respective percentages for the top 8 and top 12 content units were 92% and 85%. Thus content unit frequency is highly predictive for inclusion in a human summary.

This fact leads to the question: can then frequency in the input be directly used to assign weights to content units, making the creation of human summaries unnecessary for evaluation? We indeed evaluated an automatic system using the pyramid method on the 11 sets, using both a pyramid directly derived from the input documents and a pyramid built from human summaries as the original method prescribes. We looked at the correlation between the scores assigned by the two methods for the 11 summaries. Pearson's correlation coefficient was 0.83 (p-value$=0.0103$) and Spearman's correlation was 0.81 (p-

value$= 0.0348$). Both correlations are significant at 95% level of significance, but still the correlation is not perfect. These figures suggest the same conclusion as studying frequency at the word level—frequent content units are often included in human summaries, but frequency alone cannot fully explain human summarization behavior.

3 SumBasic: a frequency-based summarization system

All of the findings from the previous section suggest that frequency in the inputs is strongly indicative of whether a word or a fact will be used in a human summary. We thus set out to construct a summarization system that exploits frequency information exclusively, in order to isolate the contribution of frequency information. One formal method to capture this phenomenon would model the appearance of words in the summary under a multinomial distribution. That is, for each word w in the input vocabulary, we associate a probability $p(w)$ for it to be emitted into a summary. The overall probability of the summary then is

$$\frac{N!}{n_1!...n_r!}p(w_1)^{n_1} \cdot ....\cdot (w_r)^{n_r} \quad (1)$$

where N is the number of words in the summary, $n_1 + ... + n_r = N$ and for each i, n_i is the number of times word w_i appears in the summary and $p(w_i)$ is the probability that w_i appears in the summary. In order to confirm the hypothesis that human summaries maximize the probability of their vocabulary under a multinomial model, we computed the log-probability of all human and machine summaries from DUC'03. The log-probabilities of summaries produced by human summarizers were overall higher than for those produced by systems and the fact that the top five highest log-probability scores belong to humans indicate that some humans indeed employ a summarization strategy informed by frequency.[6]

In general, computing the optimal choice of sentences that would lead to the highest summary probability is infeasible. The input often contains over 300 sentences, occasionally thousands, and forming all the possible 100 word summaries (3 or 4 sentences) and choosing the one with highest probability cannot be done reasonably quickly. The following algorithm with a greedy search approximation describes SumBasic, a summarizer that uses a frequency-based sentence selection component, with a component to re-weight the word probabilities in order to minimize redundancy:

human1:	-198.65	system6:	-213.65
human2:	-205.90	human9:	-215.65
human3:	-205.91	system7:	-215.92
human4:	-206.21	system8:	-216.04
human5:	-206.37	system9:	-216.20
system1:	-208.21	system10:	-216.24
human6:	-208.23	system11:	-218.53
human7:	-208.90	system12:	-219.21
system2:	-210.14	system13:	-220.31
system3:	-211.06	system14:	-220.93
human8:	-211.95	system15:	-223.03
system4:	-212.57	system16:	-225.20
system5:	-213.08	human10:	-227.17

Table 3: Average log probabilities for different summarizers in DUC'03. All summaries were truncated to 80 words to neutralize the effect of deviations from the required length of 100 words

Step 1 Compute the probability distribution over the words w_i appearing in the input, $p(w_i)$ for every i; $p(w_i) = \frac{n}{N}$, where n is the number of times the word appeared in the input, and N is the total number of content word tokens in the input.

Step 2 For each sentence S_j in the input, assign a weight equal to the average probability of the words in the sentence, i.e.
weight$(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{|\{w_i | w_i \in S_j\}|}$

Step 3 Pick the best scoring sentence that contains the highest probability word.

Step 4 For each word w_i in the sentence chosen at step 3, update their probability
$p_{new}(w_i) = p_{old}(w_i) \cdot p_{old}(w_i)$

Step 5 If the desired summary length has not been reached, go back to Step 2.

Steps 2 and 3 enforce the desired properties of the summarizer discussed in the previous section. Step 3

[6] Other humans might have other strategies, such as giving maximum coverage of topics mentioned in the input, even such mentioned only once. Human10 appears to have such a strategy for example (after examination of his summaries).

ensures that the highest probability word is included in the summary, thus each time a sentence is picked, the word with the highest probability at that point in the summary is also picked. Step 4 serves a threefold purpose:

1. it gives the summarizer sensitivity to context. The notion of "what is most important to include in the summary?" changes depending on what information has already been included in the summary. In fact, while $p_{old}(w_i)$ can be considered as the probability with which the word w_i will be included in the summary, $p_{new}(w_i)$ is an approximation of the probability that the word w_i will appear in the summary twice.

2. By updating the probabilities in this intuitive way, we also allow words with initially low probability to have higher impact on the choice of subsequent sentences.

3. The update of word probability gives a natural way to deal with the redundancy in the multi-document input. No further checks for duplication seem to be necessary.

Table 4 shows the number of the top frequency words from the DUC'03 documents that also appear in SumBasic summaries. As expected, these are much higher than the percentages for the non-frequency oriented machine summarizer, but even higher than in all four human models.

	Top 5	Top 8	Top 12
human	94.66%	91.25%	85.25%
machine	84.00%	77.87%	66.08%
SumBasic	96.00%	95.00%	90.83%

Table 4: Percentage of the top n frequency words from the input documents that appear in the four human models, a state-of-the-art machine summarizer and SumBasic, a new machine summarizer based on frequency.

A summary produced by SumBasic, alongside a human summary for the same set is shown in figure 1.

SumBasic SUMMARY
Former Chilean dictator Gen. Augusto Pinochet has been arrested by British police on a Spanish extradition warrant, despite protests from Chile that he is entitled to diplomatic immunity. Human rights and international law experts expressed enthusiastic support Saturday for the British arrest, and said it could have wide implications. Both President Eduardo Frei of Chile and President Eduardo Menem of Argentina have resisted the Spanish legal motions, arguing that they infringe on their nations' sovereignty. Baltasar Garzon, one of two Spanish magistrates handling probes into human rights violations in Chile and Argentina, filed a request to question Pinochet on Wednesday.

HUMAN SUMMARY
Augusto Pinochet, former Chilean dictator, was arrested in London on 14 October 1998. Pinochet, 82, was recovering from surgery. The arrest was in response to an extradition warrant served by Baltasar Garzon, a maverick Spanish judge. Pinochet was charged with murdering thousands, including many Spaniards. Pinochet is awaiting a hearing, his fate in the balance. Chile protested, insisting that Pinochet had diplomatic immunity. Britain disagreed. The reaction of the Chilean people was mixed. American scholars applauded the arrest, saying that it set a precedent for other terrorist dictators. Castro criticized the arrest, and called it unprecedented international meddling.

Figure 1: Summaries produced for the same input by SumBasic and by a human.

4 Evaluation results

To evaluate the performance of SumBasic, we use the test data from two large common data set evaluation initiatives—the multi-document summarization task for DUC 2004 and the common test set provided in the 2005 MSE initiative.

DUC

While we used the data from the 2003 DUC conference for development, the data from the DUC 2004 was used as test data, which we report on here. We tested SumBasic on the 50 sets from the generic summary task in 2004 DUC. For the evaluation of this data we use the ROUGE-1 automatic metric, which has been shown to correlate well with human judgments (Lin and Hovy, 2003; Lin, 2004) and which was found to have one of the best correlations with human judgment on the DUC 2004 data (Over and Yen, 2004).

Using ROUGE-1, with stemming and stopwords removed, SumBasic performed significantly better than 12 of the other participating systems. Significant differences are those where the confidence intervals for the estimates of the means for the two

SYSTEM	MEAN ESTIMATE (95% CI)
peer 65	0.30822 (+- 0.01477)
SumBasic	0.30216 (+- 0.01687)
peer 34	0.29007 (+- 0.01742)
peer 124	0.28559 (+- 0.01766)
peer 102	0.28472 (+- 0.01750)
peer 81	0.27237 (+- 0.01724)
peer 44	0.26689 (+- 0.01654)
peer 55	0.26232 (+- 0.01654)
peer 93	0.25514 (+- 0.01979)
peer 140	0.24552 (+- 0.01980)
peer 117	0.24376 (+- 0.01866)
peer 120	0.24437 (+- 0.02068)
peer 138	0.23984 (+- 0.02175)
peer 11	0.23204 (+- 0.01987)
Baseline	0.20749 (+- 0.01923)
peer 27	0.18724 (+- 0.01912)
peer 123	0.18201 (+- 0.01639)
peer 111	0.07423 (+- 0.01067)

Table 5: DUC'04 ROUGE-1 (stemmed, stop-words removed) test set scores and their 95% confidence intervals for participating systems, the baseline, and SumBasic.

systems either do not overlap at all, or where the two intervals overlap but neither contains the best estimate for the mean of the other, which allows us to conclude that the means are different (Schenker and Gentleman, 2001). Those systems in table 5 that are not signficantly different from SumBasic are above the line, and those below the line are significantly different.

MSE

In April 2005, a multi-document summarization evaluation task was conducted as part of the Machine Translation and Summarization Workshop at ACL. The task is to produce a 100-word summary from multi-document inputs consisting of a mixture of English documents and machine translations of Arabic documents on the same topic. Some summarizers were specially modified to use redundancy to correct errors in the machine translations. We ran SumBasic without any modifications to account for the non-standard input.

The official evaluation metrics adopted for the workshop were the manual pyramid score, ROUGE-2 (the bigram overlap metric) and R-SU4 (skip bigram). The skip bigram metric measures the occurrence of a pair of words in their original sentence order, permitting up to four intervening words. The metric was originally proposed for machine translation evaluation and was shown to correlate well with human judgments both for machine translation and for summarization (Lin, 2004; Lin and Och, 2004).

system	pyramid	R-2	R-SU4	repetition
1	0.52859	0.13076	0.15670	1.4
28	0.48926	0.16036***	0.18627***	3.4***
19	0.45852	0.11849	0.14971***	1.3
SumBasic	*0.45274*	*0.12678*	*0.15938*	*0.6*
10	0.44254	0.13038	0.16568	1.2
16	0.45059	0.13355	0.16177	0.9
13	0.43429	0.08580***	0.11141***	0.4
25	0.39823	0.11678	0.15079	2.7***
4	0.37297	0.12010	0.15394	4.1***
7	0.37159	0.09654***	0.13593	0.4

Table 6: Results form the MSE evaluation. Pyramid scores and duplication is computed for 10 test sets, automatic scores for all 25 test sets

The pyramid method was used to evaluate only 10 of the test sets, while the automatic metrics were applied to all 25 test sets. None of the differences between systems were significant according to a paired t-test at the 95% level of significance. This is not surprising, given the small number of test points. Better results for significance will be achieved using permutation tests (Yeh, 2000), which will be done for the evaluation workshop. For the automatic metrics, significance was based again on the 95% confidence interval provided by ROUGE. One system was significantly better than SumBasic, and for each of the automatic metrics there were two systems that were significantly worse than SumBasic. The rest of the differences were not significant. Table 6 shows the averages for each system and the results that are significantly different from those for SumBasic are flagged by ***. During the annotation for the pyramid scoring, the content units that were repeated in an automatic summary were marked up, we include the average number of repeated SCUs per summary for all systems.

5 Duplication removal techniques and their impact on the summarizer performance

The need for duplication removal and/or re-ranking was first discussed in (Carbonell and Goldstein, 1998), but the benefits and efficiency of such modules have never been studied rigorously.

In order to examine the impact of the dupli-

cation removal method in the performance of a summarizer, we compared three systems: SumBasic, LexRank (Erkan and Radev, 2004) and DEMS (Schiffman et al., 2002). LexRank is one of the latest developed multi-document summarization systems and is based on a formalization in which each sentence in the input is represented by a node, the edges between two nodes are assigned a weight equal to the similarity between the sentences represented by the corresponding nodes. The PageRank (Page et al., 1998) algorithm is then used to assign weights to nodes (sentences). DEMS incorporates a large number of features, including frequency, lexicons for verb specificity and a lexicon with words that are more likely to occur in the first paragraphs of a news report, as well as preferences for choosing initial sentences. Both LexRank and DEMS use a postprocessing step to remove duplication among the top ranking sentences.

Table 7 shows the scores of summaries produced by SumBasic, LexRank, and DEMS, just based on initial sentence weighting (with no re-ranking for SumBasic or duplication removal for LexRank or DEMS), as well as the scores for the complete summarizers with duplication removal.

The results are interesting—without their duplication removal modules, the systems get ROUGE scores that are not statistically significantly different. On the other hand, after the duplication removal is added for LexRank and DEMS, and the re-ranking step for SumBasic, all systems get higher overall scores but SumBasic becomes significantly better than the other two. SumBasic's two components—a sentence weighting mechanism based on word frequency, and a mechanism to adjust weights after each selection—directly incorporate context by modeling the probability of a word appearing multiple times. This method appears more effective than having a separate component which removes duplication.

The 2005 MSE evaluation allows us to examine the issue of duplication removal in further detail. The Pyramid evaluation statistics include, for each system, the number of SCUs that are repeated in the summary; fewer repeated SCUs are desirable since fewer repetitions leave more room for a system to match additional SCUs.

The average duplication of content unit per summary for each system can be seen in table 6. The systems that dealt best with duplication were system 13, system 17 and SumBasic with average content unit repetition for 10 summaries of 0.4, 0.4, and 0.6 respectively. In fact, the duplication removal strategy of SumBasic was significantly better than those of three of the participating systems.[7].

SYSTEM	NO RERANK	WITH RERANK
LexRank	0.23924 (+- 0.02199)	0.24552 (+- 0.01980)
SumBasic	0.24227 (+- 0.01698)	0.30915 (+- 0.01604)
DEMS	0.25747 (+- 0.01734)	0.26689 (+- 0.01654)

Table 7: System estimated means for ROUGE-1 with and without their re-ranking on DUC 2004 test data/ duplication removal modules

6 Discussion

In this paper, we presented SumBasic, a frequency-based summarizer that seamlessly integrates content selection and re-ranking depending on context, i.e., the previous choices of summary content. In the process of evaluating and validating the new system, we discovered several factors, which although quite frequently mentioned in summarization literature rarely receive the full attention they seem to deserve.

In our study of frequency and human summarization, we found that content units and words that are repeated often in the input will very likely be mentioned in a human summary. At the same time, frequency did not completely explain human choices and many words with relatively low frequency in the input did appear in the summaries. Also pyramid scores based on a pyramid built for the input documents rather than summaries and on a pyramid built from human summaries were shown to have significant, but still not perfect correlation. These findings indicate that the frequency feature is not adequate for capturing all the content that will appear in human summaries. It remains an open problem to investigate features that can help promote less frequent content into the machine summaries.

With the development of an automated evaluation metric, ROUGE and hopefully its successors, a cheap and fast way to compare systems and versions of systems has become available. The field

[7] For t-test at 95% level of significance.

will surely benefit from more thorough investigation of the usefulness of individual features. Our intention was to create a new baseline system, using frequency alone, with plans for including more features, such as sentence simplification and reference resolution. While we expected to obtain a reasonable summarizer, we did not expect SumBasic to do as well as it did. We hope that SumBasic will be adopted by other groups as a basic component in their summarization systems, and that others will join us in re-evaluating and isolating the contribution of the more interesting linguistic features that have been incorporated in the various existing systems.

Re-ranking and duplication removal can have a major impact on the final summarizer performance, so these modules should receive more attention in the future. From our experiments, it appears that directly modeling the probability of a word appearing in a summary, taking context in account, is a very effective method of re-ranking, without the need for a separate duplication removal component.

7 Acknowledgements

We would like to thank Barry Schiffman, Dragomir Radev and Gunes Erkan for giving us the intermediate output for their systems that helped us compare sentence rankings for DEMS, LexRank and SumBasic. We are also grateful to David Evans who provided us with his pyramid mark-up.

References

Michele Banko and Lucy Vanderwende. 2004. Using n-grams to understand the nature of summaries. In *Proceedings of HLT/NAACL'04*.

Regina Barzilay, 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*.

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 (SIGIR'98)*, pages 335–336.

John Conroy, Judith Schlesinger, Jade Goldstein, and Dianne O'Leary. 2004. Left-brain/right-brain multi-document summarization. In *Document Undersatnding Conference (DUC'04)*.

Terry Copeck and Stan Szpakowicz. 2004. Vocabulary agreement among model summaries and source documents. In *Proceedings of the Document Understanding Conference DUC'04*.

Gunes Erkan and Dragomir Radev. 2004. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

David Kirk Evans and Kathleen McKeown. 2005. Identifying similarities and differences across english and arabic news. In *Proceedings of the International Conference on Intelligence Analysis*.

Hans Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*.

Hongyan Jing. 2000. Sentence simplification in automatic text summarization. In *Proceedings of the 6th Applied NLP Conference, ANLP'2000*.

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

Chin-Yew Lin and Eduard Hovy. 2002. Automated multi-document summarization in neats. In *Proceedings of the Human Language Technology Conference (HLT2002)*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurance statistics. In *Proceedings of HLT-NAACL 2003*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop in Text Summarization, ACL'04*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*.

Paul Over and James Yen. 2004. An introduction to duc 2004 intrinsic evaluation of generic news text summarization systems. In *Proceedings of DUC 2004*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

Dragomir Radev, Simone Teufel, Horacio Saggion, and W. Lam. 2003. Evaluation challenges in large-scale multi-document summarization. In *ACL*.

G. J. Rath, A. Resnick, and R. Savage. 1961. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation*, 2(12):139–208.

Natalie Schenker and Jane Gentleman. 2001. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186.

Barry Schiffman, Ani Nenkova, and Kathleen McKeown. 2002. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*.

Lucy Vanderwende, Michele Banko, and Arul Menezes. 2004. Event-centric summary generation. In *Proceedings of the Document Understanding Conference (DUC'04)*.

Alexander S. Yeh. 2000. More accurate tests for the statistical significance of result differences. In *COLING*.