

# Summarizing Text Documents: Sentence Selection and Evaluation Metrics

Jade Goldstein<sup>†</sup> Mark Kantrowitz\* Vibhu Mittal\* Jaime Carbonell<sup>†</sup>  
*jade@cs.cmu.edu mkant@jprc.com mittal@jprc.com jgc@cs.cmu.edu*

<sup>†</sup>Language Technologies Institute  
Carnegie Mellon University  
Pittsburgh, PA 15213  
U.S.A.

\*Just Research  
4616 Henry Street  
Pittsburgh, PA 15213  
U.S.A.

**Abstract** Human-quality text summarization systems are difficult to design, and even more difficult to evaluate, in part because documents can differ along several dimensions, such as length, writing style and lexical usage. Nevertheless, certain cues can often help suggest the selection of sentences for inclusion in a summary. This paper presents our analysis of news-article summaries generated by sentence selection. Sentences are ranked for potential inclusion in the summary using a weighted combination of statistical and linguistic features. The statistical features were adapted from standard IR methods. The potential linguistic ones were derived from an analysis of news-wire summaries. To evaluate these features we use a modified version of precision-recall curves, with a baseline derived from a theoretical analysis of text-span overlap based on random selection. We illustrate our discussions with empirical results showing the importance of corpus-dependent baseline summarization standards, compression ratios and carefully crafted long queries.

This paper is eligible for the *Best Student Paper* award.

## 1 Introduction

With the continuing growth of the world-wide web and online text collections, it has become increasingly important to provide improved mechanisms for finding information quickly. Conventional IR systems rank and present documents based on measuring relevance to the user query (e.g., [5, 19]). Unfortunately, not all documents retrieved by the system are likely to be of interest to the user. Presenting the user with summaries of the matching documents can help the user identify which documents are most relevant to the user's needs. This can either be a *generic* summary, which gives an overall sense of the document's content, or a *query-relevant* summary, which presents the content that is most closely related to the initial search query.

Automated document summarization dates back at least to Luhn's work at IBM in the fifties [13]. Several researchers continued investigating various approaches to this problem through the seventies and eighties (e.g., [17, 25]). The resources devoted to addressing this problem grew by several orders of magnitude with the advent of the world-wide web and large scale search engines. Several innovative approaches began to be explored: linguistic approaches (e.g., [1, 2, 4, 12, 14, 15, 18]), statistical and information-centric approaches (e.g., [6, 9, 16, 24]), and

combinations of the two (e.g., [3, 24, 26]). The TIPSTER Phase III Program, an information retrieval initiative of the US Defense Department funded several of these projects on summarization [27].

Almost all of this work (with the exception of [12, 15, 18, 23]) focused on "summarization by text-span extraction", with sentences as the most common type of text-span. This technique creates document summaries by concatenating selected text-span excerpts from the original document. This paradigm transforms the problem of *summarization*, which in the most general case requires the ability to understand, interpret, abstract and generate a new document, into a different and possibly simpler problem: *ranking sentences* from the original document according to their salience or their likelihood of being part of a summary. This kind of summarization is closely related to the more general problem of information retrieval, where documents from a document set (rather than sentences from a document) are ranked, in order to retrieve the best matches.

Human-quality summarization, in general, is difficult to achieve without natural language understanding. There is too much variation in writing styles, document genres, lexical items, syntactic constructions, etc., to build a summarizer that will work well in all cases. An ideal text summary includes the relevant information for which the user is looking and excludes extraneous and redundant information, while providing background to suit the user's profile. It must also be coherent and comprehensible, qualities that are difficult to achieve without using natural language processing to handle such issues as coreference, anaphora, etc. Fortunately, it is possible to exploit regularities and patterns – such as lexical repetition and document structure – to generate reasonable summaries in most document genres without having to do any natural language understanding.

This paper focuses on text-span extraction and ranking using a methodology that assigns weighted scores for both statistical and linguistic features in the text span. Our analysis illustrates that the weights assigned to a feature may differ according to the type of summary and corpus/document genre. These weights can then be optimized for specific applications and genres. To determine possible linguistic features to use in our scoring methodology, we evaluated several syntactical and lexical characteristics of news-wire summaries. We used statistical features that have proven effective in standard mono-lingual information retrieval techniques. Next, we outline an approach to evaluating summarizers that in-

cludes: (1) a theoretical analysis for base-line performance of a summarizer that can be used to measure relative improvements in summary qualities by either modifying the weights on specific features, or by incorporating additional features, and (2) a modified version of Salton’s 11-pt precision/recall method [22]. One of the important parameters for evaluating summarizer effectiveness is the desired compression ratio; we also analyzed the effects of different compression ratios. Finally, we describe empirical experiments that support these hypotheses.

## 2 Generating Summaries by Text Extraction

Human summarization of documents, sometimes called abstraction, produces a fixed-length *generic* summary that reflects the key points which the abstractor deems important. In many situations, users will be interested in facts other than those contained in the generic summary, motivating the need for *query-relevant* summaries. For example, consider a physician who wants to know about the adverse effects of a particular chemotherapy regimen on elderly female patients. The retrieval engine produces several lengthy reports (e.g., a 300-page clinical study), whose abstracts do not mention whether there is any information about effects on elderly patients. A more useful summary for this physician would contain query-relevant passages (e.g., differential adverse effects on elderly males and females, buried in page 211 of the clinical study) assembled into a summary. A user with different information needs would require a different summary of the same document.

Our approach to text summarization allows both generic and query-relevant summaries by scoring sentences with respect to both statistical and linguistic features. For generic summarization, a centroid query vector is calculated using high frequency document words and the title of the document. Each sentence is scored according to the following formula and then ordered in a summary according to rank order.

$$Score(S_i) = \lambda \sum_{s \in S} w_s * (Q_s \cdot S_i) + (1 - \lambda) * \sum_{l \in L} w_l * (L_l \cdot S_i)$$

where

S is the set of statistical features

L is the set of linguistic features

Q is the query

w is the weights for the features in that set

These weights can be tuned according to the type of data set used and the type of summary desired. For example, if the user wants a summary that attempts to answer questions such as who and where, linguistic features such as name and place could be boosted in the weighting. (CMU and GE used these features for the *Question and Answer* section of the TIPSTER formal evaluation with some success [27].) Other linguistic features include quotations, honorifics, and thematic phrases, as discussed in Section 4.

Furthermore, different document genres can be assigned weights to reflect their individual linguistic features, a method used by GE [24]. For example, it is a well known fact that summaries of newswire stories usually include the first sentence of the article (see Table 1). Accordingly, this feature can be given a reasonably high weight for the news-wire genre.

Statistical features include several of the standard ones from information retrieval: cosine similarity; TF-IDF weights; pseudo-relevance feedback [21]; query-expansion using techniques such as local context analysis [28] or thesaurus expansion methods (e.g., WordNet [7]); the inclusion of other query vectors such as user interest profiles; and methods that eliminate text-span redundancy such as Maximal Marginal Relevance [6].

## 3 Data Sets: Properties and Features

An ideal query-relevant text summary must contain the relevant information for which the user is looking as well as eliminate irrelevant and redundant information. A first step in building such summaries is to identify how well a summarizer can extract the pieces of articles that are relevant to a user query and the methodologies that improve summarizer performance. To this end we created a database of assessor-marked relevant sentences that may be used to examine how well systems could extract these pieces. This *Relevant Sentence Database* consists of 17 sets of 50 documents from the TIPSTER evaluation sets of articles spanning 1988-1991. Each set of 50 documents contains relevant and non-relevant documents for the topic.<sup>1</sup> (See Table 2)

Three evaluators ranked each of the sentences in the documents as relevant, somewhat relevant and not relevant. For the purpose of this experiment, somewhat relevant was treated as relevant and the final score for the sentence was determined by a majority vote. Sentences were marked as relevant to the topic if they received a majority vote. The document itself was also ranked as relevant or not relevant to the topic. The evaluators also marked the three most relevant sentences for each article as well as the one sentence that most embodied the subject matter of the article.

The second data set *Model Summaries*, was created from the training set for the Question and Answer portion of the TIPSTER evaluation as well as the three sets used in the formal evaluation (See Table 1). It consisted of four sets of model summaries, one set of 48 documents (the training set) and three of 30 documents (the formal evaluation set). Each “model” summary consists of sentences directly extracted from the document that answer a list of questions for the given topic. The four sets of articles are subsumed by the *Relevant Sentence* data.

An analysis of the properties of human-written summaries can be used to improve the quality of machine-generated summaries. We analyzed articles and summaries from Reuters and the Los Angeles Times. Our analysis covered approximately 1,000 articles from Reuters, and 1,250 from the Los Angeles Times. The Reuters articles covered the period from 11/10/1997 through 11/25/1997. The Los Angeles Times articles covered the period from 1/1/1998 through 7/4/1998 (See Table 1). These summaries were *not* generated by sentence extraction, but were manually written. In order to analyze the properties of extraction based summaries, we converted these hand-written summaries into their corresponding extracted summary. This was done by matching every sentence in the hand-written summary to the smallest sub-set of sentences in the full-length story that contained all of the key concepts mentioned in that sen-

<sup>1</sup>Our analysis of this database is in progress; at the time of this submission, we were only able to analyze 4 of these datasets.

Data Set Comparison			
Property	Model Summaries	Reuters Summaries	Los Angeles Times Summaries
task	QandA	generic summaries	generic summaries
source	Tipster	human $\Rightarrow$ extracted	human $\Rightarrow$ extracted
number of docs	48	1000	1250
average no. of sent. per doc	22.6	23.10	27.9
median sentences per doc	19	22	26
maximum sentences per doc	51	89	87
mininum sentences per doc	11	5	3
query formation	questions	topic	-
summary as % of doc length	19.4%	20.1%	20.0%
summary includes 1st sentence	72%	70.5%	68.3%
average summary size (sent)	4.3	4.3	3.7
median summary size (sent)	4	4	4
typical summary length (75% of docs)	-	3-6	3-5

Table 1: Data Set Comparison

Relevant Sentence Data Set Comparison			
Property	Relevant Docs	Non-Relevant with Rel. Sent.	All Docs with Rel. Sent
number of docs	109	78	187
average sentences per doc	28	32	30
median sentences per doc	22	30	24
maximum sentences in doc	142	83	142
minimum sentences in doc	8	9	8
relevant sentences as % of doc length	72%	31%	55%
rel. sent. includes 1st sentence	84%	38%	65%
average rel. sent. per doc.	19	9	15
median rel. sent. per doc.	16	7	13
typical relevant sent. per doc (75% of docs)	8-33	2-19	4-27

Table 2: Relevant Sentence Data Set Properties

tence. Initially, this was done manually, but we were able to automate the matching process by defining a threshold value (typically 0.85) for the minimum number of concepts (keywords and noun phrases, especially named entities) that were required to match between the two. Detailed inspections of the two sets of sentences indicate that the transformations are highly accurate, especially in this document genre of news-wire articles.<sup>2</sup> Since hand-written summaries often employ complex syntactic sentential patterns with multiple clauses, we found that this transformation resulted in a 20% increase in summary length on an average.

#### 4 Empirical Properties of Summaries

Using the extracted summaries from the Reuters and the Los Angeles Times news articles, as well as some of the model summaries and *Relevant Sentence* data, we examined several properties of the summaries. Some of these properties are presented in Table 1. Others include the average word length for the articles and their summaries, lexical properties of the sentences that were included in the summaries (positive evidence), as well as lexical properties of the sentences that were not included

<sup>2</sup>The success of this technique depends on consistent vocabulary usage between the articles and the summaries, which, fortunately for us, is true for news-wire articles. Application of this technique to other document genres would require knowledge of synonyms and hypernyms, such as those provided the WordNet lexical resource [7].

in the summaries (negative evidence), and the density of named entities in the summary and non-summary sentences.

We found that summary length was independent of document length, and that compression ratios became smaller with the longer documents. This suggests that the common practice of using a fixed compression ratio is flawed, and that using a constant summary length is more appropriate. As can be seen in Figures 1 and 2, document compression ration decreases as document word length increases. The graphs are approximately hyperbolic, suggesting that the product of the compression and the document length (i.e., summary length) is roughly constant.

Tables 1 and 2 contain information about characteristics of sentence distributions in the articles and the summaries. Figures 3 and 4 show that the summary length in words is narrowly distributed around 85-90 words per summary, or approximately five sentences. The figures also show that there is somewhat more variation in summary length for the articles from the Los Angeles Times than the ones from Reuters.

We found that the summaries included indefinite articles more frequently than the non-summary sentences. Summary sentences also tended to start with an article more frequently than non-summary sentences. In particular, Table 3 shows that the token "A" appeared 62% more frequently in the summaries.

In the Reuters articles, the word "Reuters" appeared

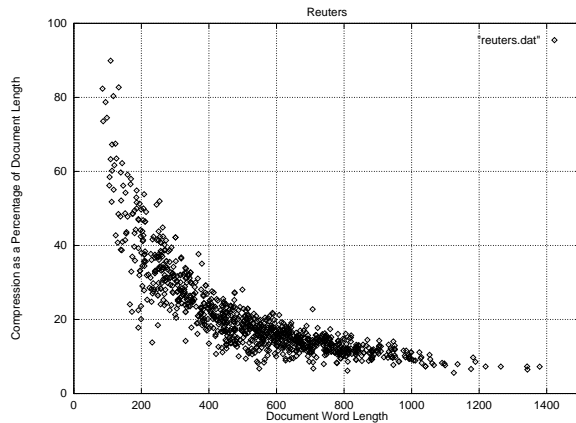


Figure 1: Compression Ratio versus Document Word Length (Reuters)

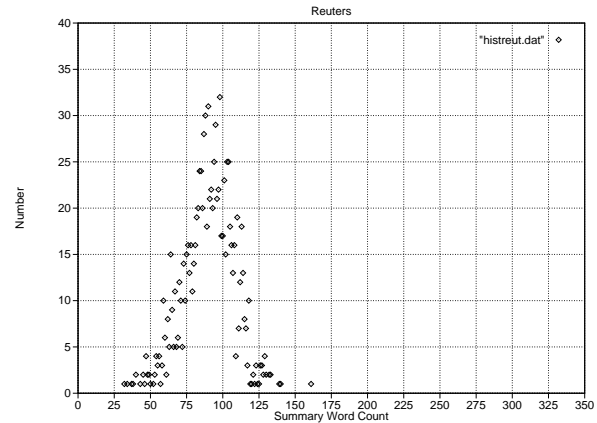


Figure 3: Distribution of Summary Word Length (Reuters)

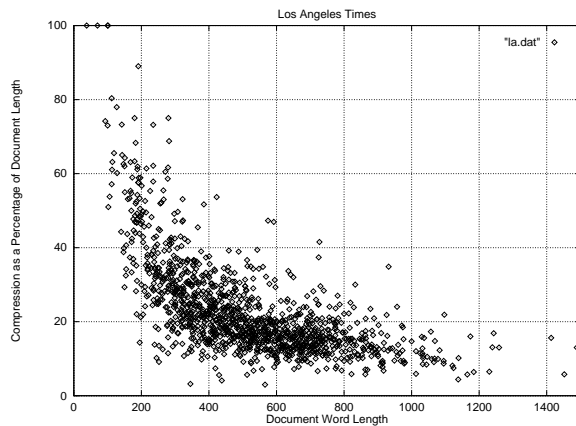


Figure 2: Compression Ratio versus Document Word Length (Los Angeles Times)

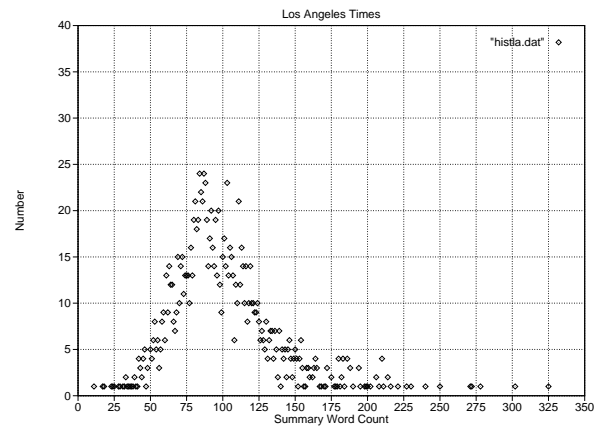


Figure 4: Distribution of Summary Word Length (Los Angeles Times)

Table 3: Table: Compares frequency of occurrence of word in summary sentences to frequency of occurrence in non-summary sentences. Calculated by taking the ratio of the two, subtracting 1, and representing as a percent.

Article	Reuters	LA Times
the	-5.5%	0.9%
The	7.5%	10.7%
a	6.2%	7.1%
A	62.0%	62.2%
an	15.2%	11.7%
An	29.6%	38.3%

much more frequently in summary sentences than non-summary sentences. This is because the first sentence usually begins with the name of the city followed by “(Reuters)” and a dash. So this word is really picking out the first sentence. Similarly, the word “REUTERS” was a good source of negative evidence, because it always follows the last sentence in the article. Similarly, names of cities, states, and countries tended to appear more frequently in summary sentences in the Reuters articles, but not the Los Angeles Times articles.

Days of the week, such as “Monday”, “Tuesday”, “Wednesday”, and so on, were present more frequently in summary sentences than non-summary sentences.

Words and phrases common in direct or indirect quotations tended to appear much more frequently in the non-summary sentences. Examples of words occurring at least 75% more frequently in non-summary sentences include “according”, “adding”, “said”, and other verbs (and their variants) related to communication. The word “adding” has this sense primarily when followed by the words “that”, “he”, “she”, or “there”, or when followed by a comma or colon. When the word “adding” is followed by the preposition “to”, it doesn’t indicate a quotation. The word “according”, on the other hand, only indicates a quotation when followed by the word “to”. Other nouns that indicated quotations, such as “analyst”, “sources” and “studies”, were also good negative indicators for summary sentences. Personal pronouns such as “us”, “our” and “we” also tended to be a good source of negative evidence, probably because they frequently occur in quoted statements. Informal or imprecise words, such as “came”, “got”, “really” and “use” also appeared significantly more frequently in non-summary sentences.

Other classes of words that appeared more frequently

in non-summary sentences in our datasets included:

- Anaphoric references, such as “these”, “this”, “those”, etc. possibly because such sentences cannot introduce a topic.
- Honorifics such as “Dr.”, “Mr.”, and “Mrs.”, presumably because news articles often introduce people by name, (e.g., “John Smith”) and subsequently refer to them more formally (e.g., “Mr. Smith”) (if not by pronominal references).
- Negations, such as “no”, “don’t”, “never” and “nothing”.
- Auxiliary verbs, such as “was”, “could”, “did”, etc.
- Integers, whether written using digits (e.g., 1, 2) or words (e.g., “one”, “two”) or representing recent years (e.g., 1991, 1995, 1998).
- Evaluative and vague words that do not convey anything definite or specific, or that qualify a statement, such as “often”, “about”, “significant”, “lot”, “some” and “several”.
- Conjunctions, such as “and”, “or”, “but”, “so”, “although” and “however”.
- Prepositions, such as “at”, “by”, “for”, “of”, “in”, “to”, and “with”.

Named entities (proper nouns) represented 16.3% of the words in summaries, compared to 11.4% of the words in non-summary sentences, an increase of 43%. 71% of summaries had a greater named-entity density than the non-summary sentences.

For sentences with 5 to 35 words, the average number of proper nouns per sentence was 3.29 for summary sentences and 1.73 for document sentences, an increase of 90.2%. The average density of proper nouns (the number of proper nouns divided by the number of words in the sentence) was 16.60% for summary sentences, compared with 7.58% for document sentences, an increase of 119%. Summary sentences had an average of 20.13 words, compared with 20.64 words for document sentences. Thus the summary sentences had a much greater proportion of proper nouns than the document and non-summary sentences. As can be seen from Figure 5, summaries include relatively few sentences with 0 or 1 proper nouns and somewhat more sentences with 2 through 14 proper nouns.

## 5 Evaluation Metrics

Jones & Galliers define two types of summary evaluations: (i) intrinsic, measuring a system’s quality, and (ii) extrinsic, measuring a system’s performance in a given task [11]. Automatically produced summaries by text extraction can often result in a reasonable summary. However, this summary may fall short of an *optimal* summary in many ways: readable, useful, intelligible, appropriate length summaries from which the information that the user is seeking can be extracted.

TIPSTER has recently focused on both the intrinsic and extrinsic aspects of summarization evaluation [8]. The evaluation consisted of three tasks (1) determining

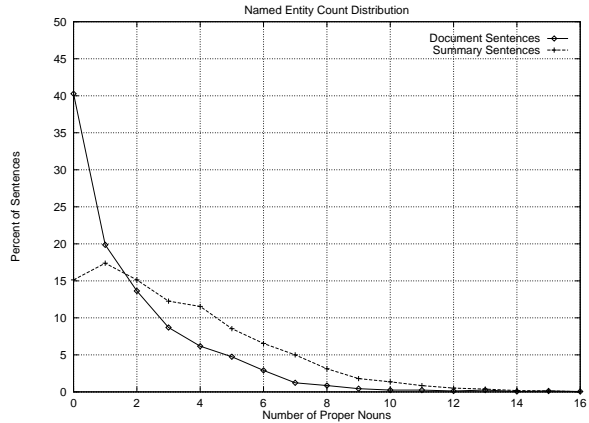


Figure 5: Number of Proper Nouns per Sentence

document relevance to a topic for query-relevant summaries (an indicative summary), (2) determining categorization for generic summaries (an indicative summary), (3) establishing whether summaries can answer a specified set of questions (an informative summary) by comparison to an ideal summary. In each task, the summaries were rated in terms of confidence in decision, intelligibility and length. Jing et al. [10] performed a pilot experiment (40 sentences) in which they examined the precision-recall performance of three summarization systems. They found that different systems achieved their best performance at different lengths (compression ratios). They also found the same results for determining document relevance to a topic (one of the TIPSTER tasks) for query-relevant summaries.

Any summarization system must first be able to recognize the relevant text-spans for a topic or query and use these to create a summary. Although a list of words, an index or table of contents, is an appropriate label summary and can indicate relevance, informative summaries need to indicate the relationships between NPs in the summary. We used sentences as our underlying unit and evaluated summarization systems for the first stage of summary creation – coverage of relevant sentences. Other systems [16,24] use the paragraph as a summary unit. Since the paragraph consists of more than one sentence and often more than one information unit, it is not as suitable for this type of evaluation, although it may be more suitable for a construction unit in summaries due to the additional context that it provides. For example, paragraphs will often solve co-reference issues, yet include additional non-relevant information. One of the issues in summarization evaluation is how to penalize extraneous non-useful information contained in a summary.

We used two data sets – described earlier, see Section 3 – to examine how performance varied for different features of summarization systems: (1) model summaries based on a person extracting sentences from the article that would provide the “best” summary for the article in regard to a series of questions *QandA summary*, and (2) relevant sentence data based on people marking all relevant sentences for an article on a provided topic.

In order to evaluate performance, we selected a baseline measure of random sentences. A theoretical analysis of the performance of random sentences reveals interesting properties about summaries (Section 6).

We used a modified version of the standard 11-point precision recall curves [22] to evaluate performance results. Since many documents have only a few relevant sentences, corresponding curves for summarization have a lot of intervals with missing data items. To remedy this situation, we implemented a step function for the precision values. This allowed the recall intervals that would not naturally be filled to be assigned an actual precision value. For example, in the case of two relevant sentences in the document, points 0-5 (the first five intervals) would all have the first precision value (naturally occurring at point 5) and points 6-10 would have the second value (naturally occurring at point 10). We interpolated the results of each query for the composite graph to form modified interpolated recall-precision curves.

In order to account for the fact that a compressed summary does not have the opportunity to return the full set of relevant sentences, we use a normalized version of recall and a normalized version of  $F_1$  as defined below.

Let

$M$  = Number of Relevant Sentences in Document

$J$  = Number of Relevant Sentences in Summary

$K$  = Number of Sentences in Summary

$P$  = Precision

$R$  = Recall

Then it follows that

$$P = \frac{J}{K} \quad (1)$$

$$R = \frac{J}{M} \quad (2)$$

$$F_1 = \frac{2 \cdot P \cdot R}{(P + R)} \quad (3)$$

$$R' = \frac{J}{\min(M, K)} \quad (4)$$

$$F_1' = \frac{2 \cdot P \cdot R'}{(P + R')} \quad (5)$$

## 6 Theoretical Analysis of Summary Properties

Current methods of evaluating summarizers often measure summary properties on absolute scales, such as precision, recall, and  $F_1$ . Although such measures can be used to compare summarization algorithms, they do not indicate whether the improvement of one summarizer over another is significant or not.

One possible solution to this problem is to derive a relative measure of summarization quality by comparing the absolute performance measures to a theoretical baseline of summarization performance. Adjusted performance values are obtained by normalizing the change in performance relative to the baseline against the best possible improvement relative to the baseline. Given a baseline value  $b$  and a performance value  $p$ , the adjusted performance value is calculated as

$$p' = \frac{(p - b)}{(1 - b)} \quad (6)$$

Given performance values  $g$  and  $s$  for good and superior algorithms, a relative measure of the improvement of the superior algorithm over the good algorithm is the

normalized measure of performance change

$$\frac{(s' - g')}{g'} = \frac{(s - g)}{(g - b)} \quad (7)$$

For the purpose of this analysis, the baseline is defined to be an ‘‘average’’ of all possible summaries. This is equivalent to the absolute performance of a summarization algorithm that randomly selected sentences for the summary. It measures the expected amount of overlap between a machine-generated and a ‘‘target’’ summary.

Let

$L$  be the number of sentences in a document,

$M$  be the number of summary-relevant sentences in the document, and

$K$  be the target number of sentences to be selected for inclusion in the summary.

Let  $P_i(L, M, K)$  be the probability of selecting  $K$  sentences such that  $i$  of them are from the set of  $M$  relevant sentences. Then  $P_i(L, M, K)$  is the product of the number of ways to select  $i$  sentences from the  $M$  relevant sentences, multiplied by the number of ways to select the remaining  $K - i$  sentences from the  $L - M$  non-relevant sentences, and divided by the number of ways to select  $K$  sentences from the  $L$  sentences in the document. Thus

$$P_i(L, M, K) = \frac{\binom{M}{i} \binom{L - M}{K - i}}{\binom{L}{K}} \quad (8)$$

Let  $E(L, M, K)$  be the expected number of relevant sentences. Then

$$\begin{aligned} E(L, M, K) &= \sum_{i=0}^M i \cdot P_i(L, M, K) \\ &= \sum_{i=0}^M i \cdot \frac{\binom{M}{i} \binom{L - M}{K - i}}{\binom{L}{K}} \end{aligned}$$

But  $\binom{M}{i} = \frac{M}{i} \cdot \binom{M - 1}{i - 1}$ , so

$$\begin{aligned} E(L, M, K) &= \sum_{i=0}^M M \cdot \frac{\binom{M - 1}{i - 1} \binom{L - M}{K - i}}{\binom{L}{K}} \\ &= \frac{M}{\binom{L}{K}} \cdot \sum_{i=0}^M \binom{M - 1}{i - 1} \binom{L - M}{K - i} \end{aligned}$$

From the identity<sup>3</sup>

$$\sum_{i=0}^M \binom{M - 1}{i - 1} \binom{L - M}{K - i} = \binom{L - 1}{K - 1}$$

it follows that

$$E(L, M, K) = \frac{M}{\binom{L}{K}} \binom{L - 1}{K - 1}$$

and hence

$$E(L, M, K) = \frac{M \cdot K}{L} \quad (9)$$

<sup>3</sup>To prove this combinatoric identity, equate the coefficients of  $x^{K-1}$  on both sides of  $(1 + x)^{M-1} \cdot (1 + x)^{L-M} = (1 + x)^{L-1}$ .

From this we may calculate the precision and recall values as

$$\textit{precision} = \frac{E(L, M, K)}{K} = \frac{M}{L} \quad (10)$$

$$\textit{recall} = \frac{E(L, M, K)}{M} = \frac{K}{L} \quad (11)$$

From this it follows that

$$F_1 = \frac{2 \cdot M \cdot K}{L \cdot (M + K)} \quad (12)$$

This formula relates  $F_1$ ,  $M$ ,  $K$ , and  $L$ . Given three of the values, the fourth can be easily calculated. In particular, the value of a baseline  $F_1$  can be calculated from  $M$ ,  $K$ , and  $L$ .

Incidentally, the value of recall derived above is the same as the document compression ratio. The precision value in some sense measures the degree to which the document is already a summary, namely the density of summary-relevant sentences in the document. The higher the baseline precision for a document, the more likely any summarization algorithm is to generate a good summary for the document. The baseline values measure the degree to which summarizer performance can be accounted for by the number of sentences selected and characteristics of the document.

It is important to note that much of the analysis presented in this section, especially equations 6 and 7, is independent of the evaluation method and can also apply to evaluation of document information retrieval algorithms.

## 7 Effect of Compression Ratios

Dataset	Document length words/chars	Summary compression words/chars	Extracted compression words/chars
Reuters	476/3054	0.20/0.20	0.25/0.24
LA Times	511/3158	0.16/0.18	0.20/0.20
CS Monitor	804/4993	0.09/0.09	0.09/0.08

Table 4: Compression ratios for summaries of news-wire articles: human-generated vs. corresponding extraction based summaries.

It is a common practice for summary evaluations to use a fixed compression ratio. This yields a target number of summary sentences that is a percentage of the length of the document.<sup>4</sup> As noted previously, the empirical analysis of news summaries written by people found that the number of target sentences does not vary with document length, and is approximately constant (see Figures 1 and 2). The theoretical analysis in Section 6 supports the conclusion that a fixed compression ratio is not an effective means for evaluating summarizers.

Consider the impact on  $F_1$  of a fixed compression ratio. The value of  $F_1$  is then equal to  $\frac{2 \cdot M}{M + K}$  multiplied by the compression ratio, a constant. This value does change significantly as  $L$  grows larger. But a longer document has more non-relevant sentences, and so should do significantly worse in an un-informed sentence selection metric. Assuming a fixed value of  $K$ , on the other hand, yields a more plausible result.  $F_1$  is then equal to  $\frac{2 \cdot M}{L \cdot (M + K)}$ , a

<sup>4</sup>Typically ten percent.

quantity that decreases as  $L$  increases. With a fixed value of  $K$ , longer documents yield lower baseline performance for the random sentence selection algorithm.

The theoretical analysis also offers a possible explanation for the popular heuristic that most summarization algorithms work well when they select 1/3 of the document’s sentences for the summary. It suggests that this has more to do with the number of sentences selected and characteristics of the documents used to evaluate the algorithms than the quality of the algorithm. The expected number of summary-relevant sentences for random sentence selection is at least one when  $\frac{K}{L}$ , the compression ratio, is at least  $\frac{1}{M}$ . When reporters write summaries of news articles, they typically write summaries 3 to 5 sentences long. So there is likely to be at least one sentence in common with a human-written summary when the compression ratio is at least 1/3 to 1/5.

A similar analysis can show that for the typical sentence lengths, picking 1/4 to 1/3 of the words in the sentence as keywords yields the “best” summary of the sentence.

It is also worthwhile to examine the shape of the  $F_1$  curve. The ratio of  $F_1$  values at successive values of  $K$  is  $1 + \frac{M}{K \cdot (M + K + 1)}$ . Subtracting 1 from this quantity yields the percentage improvement in  $F_1$  values for each additional summary sentence. Assuming a point of diminishing returns when this quantity falls below a certain value, such as 5 percent or 10 percent, yields a relationship between  $M$  and  $K$ . For typical values of  $M$  for news stories, the point of diminishing returns is reached when  $K$  is between 4.7 and 7.4.

## 8 Experimental Results

Unlike document information retrieval, text summarization evaluation has not extensively addressed the performance of different methodologies by evaluating the contributions of each component. Most summarization systems use linguistic knowledge as well as a statistical component [27]. We are currently exploring the use of both of these features in building effective summarizers. Among statistical methods, we explored the use of query-expansion and pseudo-relevance feedback (prf) techniques.

Our summarizers are capable of using the cosine distance metric (of the SMART search engine [5]) to score sentences with respect to the query. Options such as stemming and stop word removal can be specified. For generic summaries, a “query” can be constructed from the document text (using, for instance, high frequency words). For query-relevant summaries, the query is constructed from key words of the topic for the Relevant Sentence data and from key words of the questions and topic for the Model Summary data. Each TIPSTER set consists a topic-, description-, narrative- and sometimes a concepts-section. For “short queries” we used just the topic and the description and for the standard “queries” we used the topic, description and narrative. For “long queries”, we used the “concepts” section, which only occurred in 5 of the 17 data sets.

Query expansion – which can be effective in monolingual information retrieval – and local context analysis [28] were found to be effective methods for summarization [27]. Our experiments to evaluate the relative

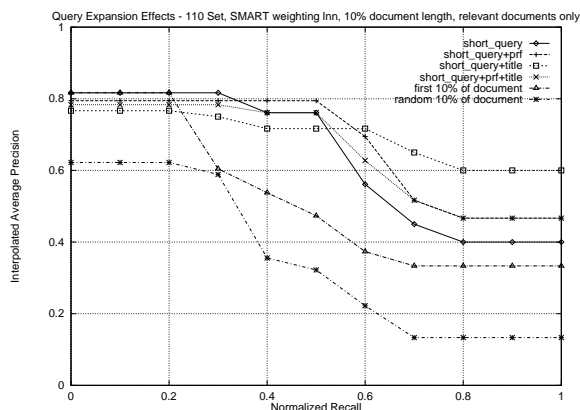


Figure 6: Query expansion effects at 10% document length

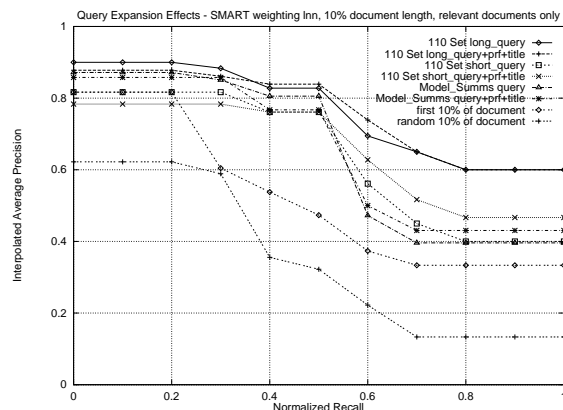


Figure 8: Query expansion effects at 10% document length (2 data sets)

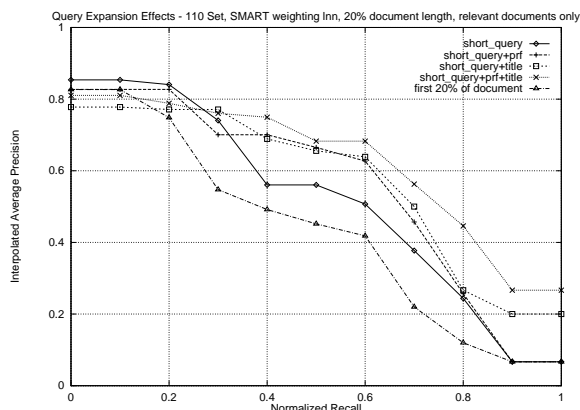


Figure 7: Query expansion effects at 20% document length

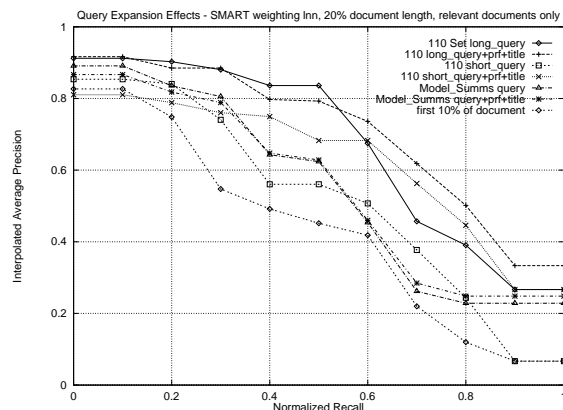


Figure 9: Query expansion effects at 20% document length (2 data sets)

benefits of query expansion for summarization consisted of comparing the standard query to the “short query” version. The results of this experiment are shown in Figure 6 and 7 for 10% and 20% document character compression (calculated by rounding up to the nearest sentence).

Pseudo-relevance feedback has been shown to improve performance in monolingual information retrieval [20]. In our experiments on summarization, we took the top ranked sentence to the query and added it to the query, and then used the result for our new summary. We also examined a variation of pseudo-relevance feedback which added the title to the query. The results are shown in Figures 8 and 9.

The most significant score improvements occur for short queries. For the longer queries, the effect was less. For 20% document length (characters rounded up to the sentence boundary) adding the highest ranked sentence (prf) and title to the query helps performance for the 110 set relevant summary judgments (figures 7 and 9). For 10% document length, for short queries just adding the title performed better than prf and the title (figures 6, 8). These results are similar to those obtained for document information retrieval.

While these statistical techniques work well across the board, they can often be supplemented by using com-

plementary features based on exploiting characteristics specific to either the document type or language being used. For instance, English documents often begin with an introductory sentence that can be used in a generic summary. Less often, the last sentence of a document can also repeat the same information. Intuitions such as these – that positional aspects of document structure – can be exploited by system designers. Since not all of these features are equally probable in all situations, it is also important to gain an understanding of the cost-benefit ratio for these feature-sets in different situations. Linguistic features occur at many levels of abstraction: document level, paragraph level, sentence level and word levels. Section 4 discusses some of the sentence and word-level features that can help select summary sentences in news-wire articles. Our efforts have focused on trying to discover as many of these linguistic features as possible for specific document genres (news-wire articles, emails, scientific documents, etc.). Figure 10 shows the normalized  $F_1$  scores at different levels of compression for some of the other sentence level linguistic features for a dataset of approximately 1200 articles from Reuters.

As discussed in Section 7, the level of compression being achieved can have an important effect on the quality of the summarization. Our analysis in Section 6 also illustrated the connection between the baseline performance from random sentence selection and compression ratios.

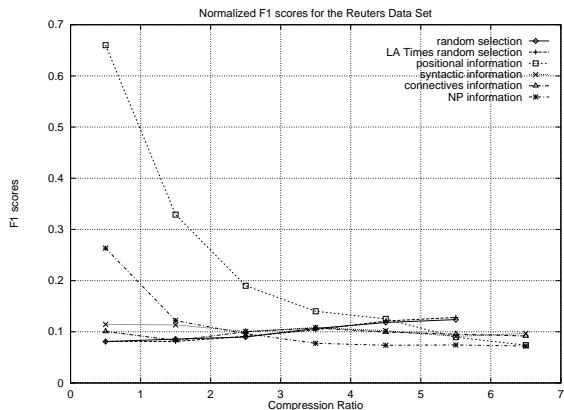


Figure 10: Normalized  $F_1$  scores for sentence level linguistic features.

We investigated the quality of our summaries (in terms of the  $F_1$ ) for different compression ratios. In one analysis, we determined how compression affected performance (Figure 11). We used a document compression factor based on the number of characters in the document. If this cutoff fell in the middle of a sentence the rest of the sentence was allowed, thus the output summary could sometimes be slightly longer than the compression factor. As can be seen from the figure, statistical approaches did significantly better than the baseline performance on random selection and the *first-n-sentences* on these data sets. These figures suggest that performance drops off with compression.

Figure 11 indicates that the normalized  $F_1$  score is helped by having the pseudo-relevance feedback and title in the query thereby extracting relevant sentences that would otherwise be missed, clearly demonstrating the impact of good queries. For 10% compression, the long query has a 15% improvement in the raw  $F_1$  score over the short query (or 42% improvement taking the baseline random selection into account based on equation 7). Furthermore, using the first 10% of the document as a summary, the long query has a 41% improvement in the raw  $F_1$  score (or 277% improvement taking the baseline random selection).

A graph of normalized  $F_1$  versus the baseline random recall value looks almost identical to Figure 11, empirically confirming that the baseline random recall value is the compression ratio. A graph of the normalized  $F_1$  scores adjusted relative to the random baseline using Equation 6 looks similar to Figure 11, but tilts downward, showing worse performance as the compression ratio increases.

If we calculate the normalized  $F_1$  score for the first sentence retrieved in the summary, we obtain a score of .80 for 110 Set standard query, .67 for 110 Set short query and .79 for the Model Summaries. This indicates that even for the short query we obtain a relevant sentence two thirds of the time. However, ideally this first sentence retrieval score would be 1.0 and we will explore methods to increase this score as well as select a highly relevant first retrieved sentence for the document.

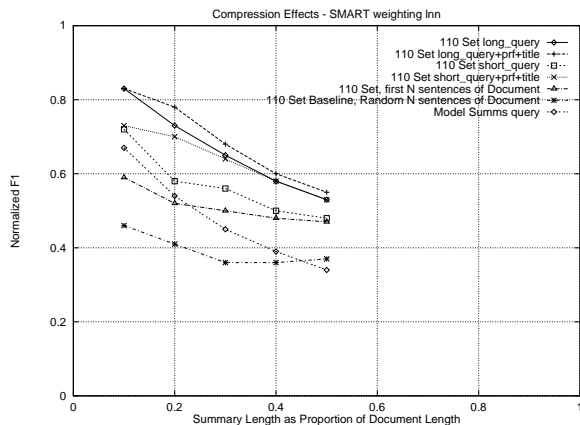


Figure 11: Normalized  $F_1$  versus compression ratios.

## 9 Conclusions and Future Work

This paper presents our analysis of news-article summaries generated by sentence selection. Sentences are ranked for potential inclusion in the summary using a weighted combination of statistical and linguistic features. The statistical features were adapted from standard IR methods. Potential linguistic ones were derived from an analysis of news-wire summaries. To evaluate these features, we use a modified version of precision-recall curves, with a baseline derived from a theoretical analysis of text-span overlap based on random selection. This paper demonstrates that an evaluation of summarization systems must take into account *both the compression ratios and the characteristics of the document set* being used. This work has shown the importance of baseline summarization standards and the need to discuss summarizer effectiveness in this context. This work has also demonstrated the importance of query formation in summarization results.

In future work, we plan to investigate machine learning techniques to discover additional features, both linguistic (such as discourse structure, anaphoric chains, etc.) and other information (including presentational features, such as formatting information) for a variety of document genres, and learn optimal weights for the feature combinations.

**Acknowledgements:** We would like to acknowledge the help of Michele Banko, who implemented the text alignment code to match text spans from the handwritten summaries to the sentences in the original documents, allowing us to construct large corpora of "summaries by sentence extraction".

## References

- [1] Aone, C., Okunowski, M. E., Gorlinsky, J., and Larsen, B. A scalable summarization system using robust NLP. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, 1997), pp. 66–73.
- [2] Baldwin, B., and Morton, T. S. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3)* (Granada, Spain, June 1998).

- [3] Barzilay, R., and Elhadad, M. Using lexical chains for text summarization. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, 1997), pp. 10–17.
- [4] Boguraev, B., and Kennedy, C. Saliency based content characterization of text documents. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, 1997), pp. 2–9.
- [5] Buckley, C. Implementation of the SMART information retrieval system. Tech. Rep. TR 85-686, Cornell University, 1985.
- [6] Carbonell, J. G., and Goldstein, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR-98* (Melbourne, Australia, Aug. 1998).
- [7] Fellbaum, C., Ed. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.
- [8] Hand, T. F. A proposal for task-based evaluation of text summarization systems. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, July 1997), pp. 31–36.
- [9] Hovy, E., and Lin, C. Y. Automated text summarization in SUMMARIST. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, July 1997), pp. 18–24.
- [10] Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. Summarization evaluation methods experiments and analysis. In *AAAI Intelligent Text Summarization Workshop* (Stanford, CA, Mar. 1998), pp. 60–68.
- [11] Jones, K. S., and Galliers, J. R. *Evaluating Natural Language Processing Systems: an Analysis and Review*. Springer, New York, 1996.
- [12] Klavans, J. L., and Shaw, J. Lexical semantics in summarization. In *Proceedings of the First Annual Workshop of the IFIP Working Group FOR NLP and KR* (Nantes, France, Apr. 1995).
- [13] Luhn, P. H. Automatic creation of literature abstracts. *IBM Journal* (1958), 159–165.
- [14] Marcu, D. From discourse structures to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, 1997), pp. 82–88.
- [15] McKeown, K., Robin, J., and Kukich, K. Designing and evaluating a new revision-based model for summary generation. *Info. Proc. and Management* 31, 5 (1995).
- [16] Mitra, M., Singhal, A., and Buckley, C. Automatic text summarization by paragraph extraction. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, July 1997), pp. 31–36.
- [17] Paice, C. D. Constructing literature abstracts by computer: Techniques and prospects. *Info. Proc. and Management* 26 (1990), 171–186.
- [18] Radev, D., and McKeown, K. Generating natural language summaries from multiple online sources. *Computational Linguistics* (1998).
- [19] Salton, G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.
- [20] Salton, G., Allan, J., Buckley, C., and Singhal, A. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science* 264 (1994), 1421–1426.
- [21] Salton, G., and Buckley, C. Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences* 41 (1990), 288–297.
- [22] Salton, G., and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill Computer Science Series. McGraw-Hill, New York, 1983.
- [23] Shaw, J. Conciseness through aggregation in text generation. In *Proceedings of 33rd Association for Computational Linguistics* (1995), pp. 329–331.
- [24] Strzalkowski, T., Wang, J., and Wise, B. A robust practical text summarization system. In *AAAI Intelligent Text Summarization Workshop* (Stanford, CA, Mar. 1998), pp. 26–30.
- [25] Tait, J. I. *Automatic Summarizing of English Texts*. PhD thesis, University of Cambridge, Cambridge, UK, 1983.
- [26] Teufel, S., and Moens, M. Sentence extraction as a classification task. In *ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization* (Madrid, Spain, July 1997), pp. 58–65.
- [27] Tipster text phase III 18-month workshop notes, May 1998. Fairfax, VA.
- [28] Xu, J., and Croft, B. Query expansion using local and global document analysis. In *Proceedings of the 19th ACM/SIGIR (SIGIR-96)* (1996), ACM, pp. 4–11.