Summarizing Text Documents: Sentence Selection and Evaluation Metrics

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

†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 normalized version of precision-recall curves, with a baseline of random sentence selection, as well as analyze the properties of such a baseline. We illustrate our discussions with empirical results showing the importance of corpus-dependent baseline summarization standards, compression ratios and carefully crafted long queries.

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., [7,23]). 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 1950's [13]. Several researchers continued investigating various approaches to this problem through the seventies and eighties (e.g., [19, 26]). 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., [2, 3, 6, 12, 15, 16, 18, 20]), statistical and information-centric approaches (e.g., [8, 9, 17, 25]), and combinations of the two (e.g., [5, 25]).

Almost all of this work (with the exception of [12,16,20,24]) 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 co-reference, 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 newswire summaries. We used statistical features that have proven effective in standard monolingual information retrieval techniques. Next, we outline an approach to evaluating summarizers that includes: (1) an 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 normalized version of Salton's 11-pt precision/recall method [23]. 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, and 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 *Q&A* section of the TIPSTER formal evaluation with some success [14].) Other linguistic features include quotations, honorifics, and thematic phrases, as discussed in Section 4 [18].

Furthermore, different document genres can be assigned weights to reflect their individual linguistic features, a method used by GE [25]. 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 newswire genre.

Statistical features include several of the standard ones from information retrieval: cosine similarity; TF-IDF weights; pseudo-relevance feedback [22]; query-expansion using techniques such as local context analysis [14, 27] or thesaurus expansion methods (e.g., WordNet); the inclusion of other query vectors such as user interest profiles; and methods that eliminate text-span redundancy such as Maximal Marginal Relevance [8] .

3 Data Sets: Properties and Features

An ideal query-relevant text summary must contain the relevant information to fulfill a user's information seeking goals, as well as eliminate irrelevant and redundant information. A first step in constructing such summaries is to identify how well a summarizer can extract the text that is 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 20 sets of 50 documents from the TIPSTER evaluation sets of articles spanning 1988-1992. For our experiments we eliminated all articles covering more than one subject (news briefs) resulting in 954 documents.

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 not relevant and the final score for the sentence was determined by a majority vote (somewhat relevant was considered not relevant). Of the 954 documents, 176 documents contained no relevant sentences using this scoring method (See Table 1). The evaluators also marked each document as relevant or not relevant to the topic and selected the three most relevant sentences for each article from the sentences that they had marked relevant (yielding a *most relevant sentence* data set of 1-9 sentences per document). This set has an average of 5.6 sentences per document and 58.2% of the relevant sentence summaries contain the first sentence. Note that relevant summaries do not include the first sentence as often as the other sets due to the fact that off topic documents may contain relevant data.

The data set *Q&A Summaries*, was created from the training and evaluation sets for the Question and Answer portion of the TIPSTER evaluation as well as the three sets used in the formal evaluation (See Table 1). Each summary consists of sentences directly extracted (by one person) from the marked sections of the documents that answer a list of questions for the given topic.

To improve generic machine-generated summaries, an analysis of the properties of human-written summaries can be used. 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 (See Table 1).[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 subset of sentences in the full-length story that contained all of the key concepts mentioned in that sentence. Ini-

[1] The Reuters articles covered the period from 11/10/1997 through 11/25/1997 and the Los Angeles Times articles from 1/1/1998 through 7/4/1998

	Summary Data Set Comparison			Relevant Sentence Data Comparison		
Property	Q&A Summaries	Reuters Summaries	Los Angeles Times Summaries	All Docs with Rel. Sent.	Rel. Docs with Rel. Sent.	Non-Rel. Docs with Rel. Sent.
task	Q&A	generic summaries	generic summaries	relevance	relevance	relevance
source	TIPSTER	human ⇒ extracted	human ⇒ extracted	user study	user study	user study
Document Features						
number of docs	128	1000	1250	778	641	137
avg sent/doc	32.1	23.10	27.9	29.8	29.5	31.9
median sent/doc	26	22	26	26	26	26
max sent/doc	122	89	87	142	142	107
min sent/doc	11	5	3	5	5	7
query formation	topic+Q's	–	–	topic	topic	topic
Summary Features						
% of doc length	19.6%	20.1%	20.0%	23.4%	27.1%	6.4%
incl. 1st sentence	61.7%	70.5%	68.3%	43.4%	52.7%	0%
avg size (sent)	5.8	4.3	3.7	5.7	6.5	1.6
median size (sent)	4	4	4	5	5	1
size (75% of docs)	2–9	3–6	3–5	1–11	2–12	1–2

Table 1: Data Set Comparison: For relevant sentence data the summary consists of majority vote relevant sentences.

tially, 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 [4]. Detailed inspections of the two sets of sentences indicate that the transformations are highly accurate, especially in this document genre of newswire articles.[2] We found that this transformation resulted in a 20% increase in summary length on average (see Table 6), presumably because document sentences include extraneous clauses.

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 *Q&A 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 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 Figure 1, document compression ratio decreases as document word length increases.[3] The graphs are approximately hyperbolic, suggesting that the product of the compression and the document length (i.e., summary length) is roughly constant.

Table 1 contains information about characteristics of sentence distributions in the articles and the summaries. Figure 2 shows that the summary length in words is narrowly distributed around 85–90 words per summary, or approximately three to five sentences.

[2]The success of this technique depends on consistent vocabulary usage between the articles and the summaries, which, fortunately for us, is true for newswire articles. Application of this technique to other document genres would require knowledge of synonyms, hypernyms, and other word variants.

[3]Graphs for the LA Times data appeared similar, though slightly more diffuse.

Table 2: Frequency of word occurrence in summary sentences vs 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%

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 2 shows that the token “A” appeared 62% more frequently in the summaries.

In the Reuters articles, the word “Reuters” appeared 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) re-

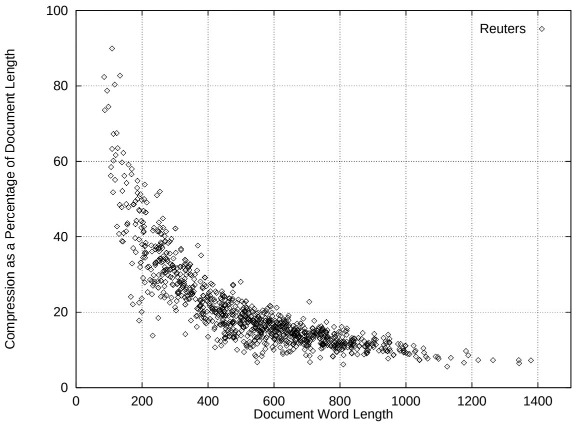

Figure 1: Compression Ratio versus Document Word Length (Reuters)

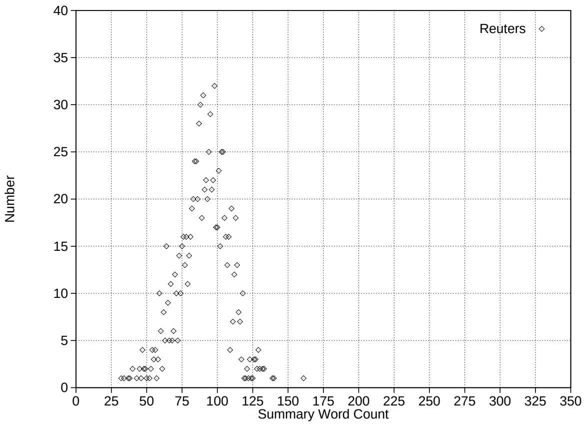

Figure 2: Distribution of Summary Word Length (Reuters)

lated 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", and "those", 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", and "never".
- Auxiliary verbs, such as "was", "could", and "did".
- 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 that qualify a statement, such as "often", "about", "significant", "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 3, 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, i.e, a readable, useful, intelligible, appropriate length summaries from which the information that the user is seeking can be extracted.

TIPSTER has recently focused on evaluating summaries [14]. The evaluation consisted of three tasks (1) determining 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 a human generated "model" 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 (for 40 sentence articles) 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 (a TIPSTER task) 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 [17, 25] 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, but 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 the data sets described in Section 3 to examine how performance varied for different features of our summarization systems. To evaluate performance, we selected a baseline measure of random sentences. An analysis of the performance of random sentences reveals interesting properties about summaries (Section 6).

We used interpolated 11-point precision recall curves [23] to evaluate performance results. 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 be the number of relevant sentences in document, J be the number of relevant sentences in summary, and K be the number of sentences in summary. The standard definitions of precision, recall, and F_1 are $P = \frac{J}{K}$, $R = \frac{J}{M}$, and $F_1 = \frac{2 \cdot P \cdot R}{(P+R)}$. We define the normalized versions as:

$$R' = \frac{J}{min(M,K)} \quad (1)$$

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

6 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

$$p' = \frac{(p-b)}{(1-b)} \quad (3)$$

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

$$\frac{(s'-g')}{g'} = \frac{(s-g)}{(g-b)} \quad (4)$$

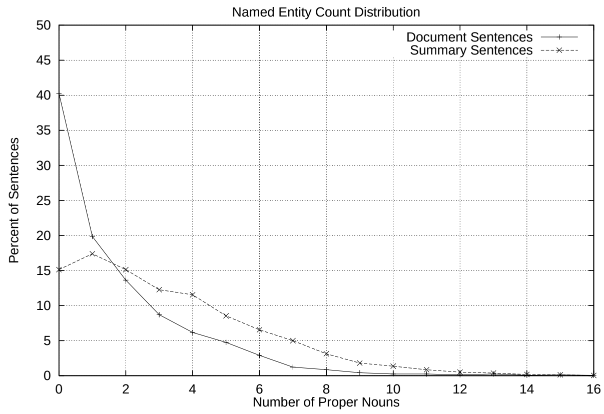

Figure 3: Number of Proper Nouns per Sentence

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.

Assuming a uniform likelihood of relevance, the probability that a sentence is relevant is $\frac{M}{L}$. The expected precision is also $\frac{M}{L}$ since the same proportion should be relevant no matter how many sentences are selected. Then $E(L, M, K)$, the expected number of relevant sentences, is the product of the probability a sentence is relevant with the number of sentences selected, so $E(L, M, K) = \frac{M \cdot K}{L}$. Then recall is $\frac{E(L,M,K)}{M} = \frac{K}{L}$.

From these values for recall and precision it follows that

$$F_1 = \frac{2 \cdot M \cdot K}{L \cdot (M+K)} \quad (5)$$

This formula relates F_1, M, K, and L. Given three of the values, the fourth can be easily calculated. For example, 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 3 and 4, is independent of the evaluation method and can also apply to evaluation of document information retrieval algorithms.

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. As noted previously, the empirical analysis of news summaries written by people found that the number of tar-

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

Table 3: Compression ratios for summaries of newswire articles: human-generated vs. corresponding extraction based summaries.

get sentences does not vary with document length, and is approximately constant (see Figures 1 and 2). Our previous derivation supports our 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 not change significantly as L grows larger. But a longer document has more non-relevant sentences, and so should do significantly worse in an uninformed 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 quantity that decreases as L increases. With a fixed value of K, longer documents yield lower baseline performance for the random sentence selection algorithm.

Our 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 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 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.

7 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. Since most summarization systems use linguistic knowledge as well as a statistical component [14], we are currently exploring the contributions of both types of features.

One summarizer uses the cosine distance metric (of the SMART search engine [7]) to score sentences with respect to a query. For query-relevant summaries, the query is constructed from terms of the TIPSTER topic description, which consists of a topic, description, narrative, and sometimes a concepts section. "Short queries" consist of the terms in the topic section, averaging 3.9 words for the 20 sets. The full query consists of the entire description (averaging 53 words) and often contains duplicate words, which increase the weighting of that word in the query.

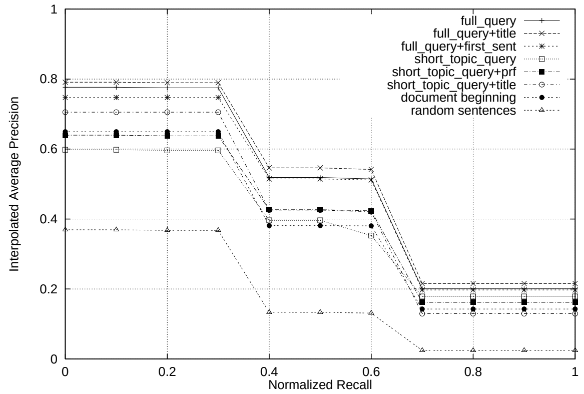

Figure 4: Query expansion effects for fixed summarizer output of 3 sentences (most relevant sentences data).

Query expansion methods have been shown to improve performance in monolingual information retrieval [22, 23, 27]. Previous results suggest that they are also effective for summarization [14]. We evaluated the relative benefits of various forms of query expansion for summarization by forming a new query through adding: (1) the top ranked sentence of the document (*pseudo-relevance feedback - prf*) (2) the title, and (3) the document's first sentence. The results (relevant documents only) are shown in Figures 4, 5, and 6.

Figure 4 examines the output of the summarizer when fixed at 3 sentences using the most relevant sentence data selected by the evaluators (see Section 3). Figures 5 and 6 show the summary performance of 20% document character compression (rounded up to the nearest sentence) using the majority vote relevant sentences data (for all relevant documents, all relevant sentences). [4] Figures 5 and 6 compare the effect of query length and expansion. Figure 6 compares short queries to full queries and medium queries for the five sets of data that include a concept section in the topic description. In this case, the full queries (average 90 words) contain all terms, the medium query eliminates the terms from the concept section (average 46.2 words) and the short queries just include the topic header (average 5.4 words).

Short query summaries show slight score improvements using query expansion techniques (*prf*, the title, and the combination) for the initial retrieved sentences and then decreased performance. This decrease is due to the small size of the query and the use of R' (Equation 1) - a small query often returns only a few ranked sentences and adding additional document related terms can cause the summary to include additional sentences which may be irrelevant. For the longer queries, the effects of *prf* and title addition appear effectively negligible and the first sentence of the document slightly decreased performance. In the case of the most relevant sentence data (Figure 4), in which the summarizer output was fixed at 3 sentences, the summary containing

[4] 20% compression was used as reflecting the average document compression for our data (refer to Table 1).

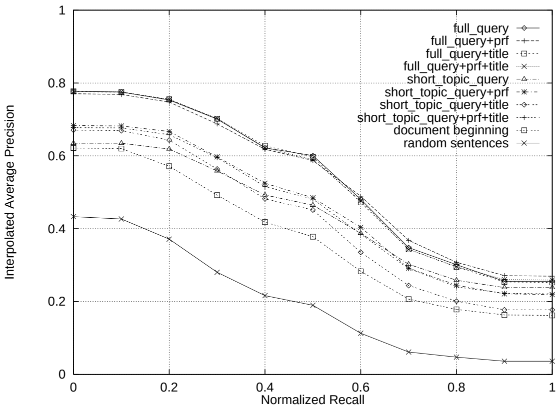

Figure 5: Query expansion effects at 20% document length: all relevant sentences, all relevant documents (641).

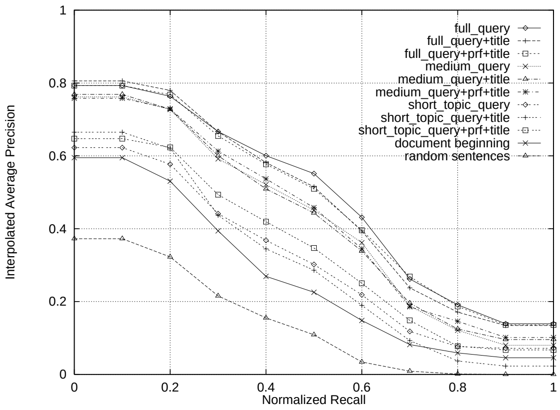

Figure 6: Query expansion effects at 20% document length: all relevant sentences, 5 data sets with "concept section" in topic, all relevant documents (208).

the inital sentences of the document, *"document beginning"* had a higher accuracy than the short query's summary for the initial sentence reflecting the fact that the first sentence has a high probability of being relevant (Table 1).

While these statistical techniques can work well, they can often be supplemented by using complementary features that exploit 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 (positional bias) 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 newswire articles. Our efforts have focused on trying to discover as many of these linguistic features as possible for specific document genres (newswire articles, email, scientific documents, etc.). Figure 7 shows the F_1' scores (Equation 2) at different levels of compression for sentence level linguistic features for a data-set of approximately 1200 articles from Reuters. The output summary size is fixed at the size of the provided generic summary, whose proportion to the document length determines the actual compression factor.

As discussed in Section 6, the level of compression has an effect on summarization quality. Our analysis also illustrated the connection between the baseline performance from random sentence selection and compression ratios. We investigated the quality of our summaries for different features and data sets (in terms of F_1') at different compression ratios (setting the summarizer to output a certain percentage of the document size). Figure 8 suggests that performance drops as document length increases, reflecting the decrease in precision that often occurs as the summarizer selects sentences. For low compression (10-30%), the statistical approach of adding *prf* and title improved results for all data sets (albeit miniscule for long queries). Queries with or without expansion did significantly better than the baseline performance of random selection and document beginning. For 10% of the document length, the long query summary has a 24% improvement in the raw F_1 score over the short query (or 52% improvement taking the baseline random selection into account based on equation 4). This indicates the importance of query formation in summarization results.

A graph of F_1' versus the baseline random recall value looks almost identical to Figure 8, empirically confirming that the baseline random recall value is the compression ratio. A graph of the F_1' scores adjusted relative to the random baseline using Equation 3 looks similar to Figure 8, but tilts downward, showing worse performance as the compression ratio increases.

If we calculate the F_1' score for the relevant sentence data for the first sentence retrieved in the summary, we obtain a score of .65 for the full query and .53 for the short topic query. Ideally, the highest ranked sentence of the summarizer would be among the most relevant, although at least relevant might be satisfactory. We are investigating methods to increase this likelihood for both query-relevant and generic summaries.

8 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 newswire summaries. To evaluate these features, we use a normalized version of precision-recall curves and compared our improvements to a random sentence selection baseline. Our analysis of the properties of such a baseline indicates 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

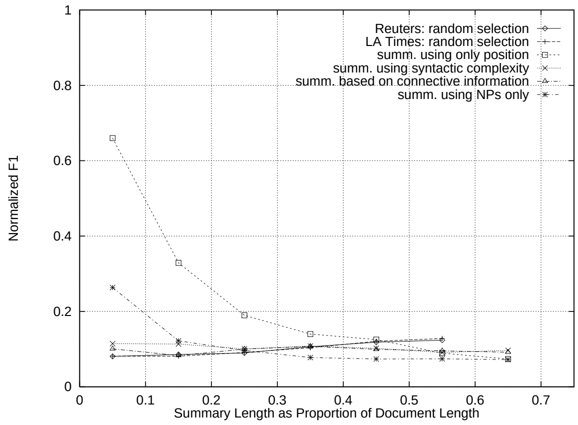

Figure 7: Compression effects for sentence level linguistic features.

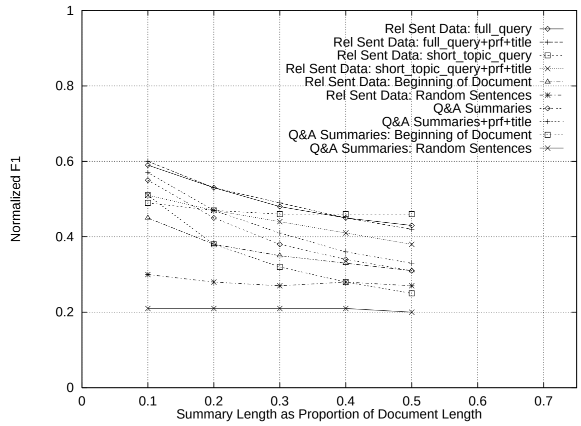

Figure 8: Compression effects for query expansion using relevant sentence data and Q&A summaries.

learn optimal weights for the feature combinations.

Acknowledgements: We would like to acknowledge the help of Michele Banko. This work was partially funded by DoD and performed in conjunction with Carnegie Group, Inc. The views and conclusions do not necessarily reflect that of the aforementioned groups.

References

[1] *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization.* Madrid, Spain, 1997.

[2] Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. A scalable summarization system using robust NLP. [1], pp. 66–73.

[3] 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).

[4] Banko, M., Mittal, V., Kantrowitz, M., and Goldstein, J. Generating extraction based summaries from handwritten summaries by aligning text spans. In *Proceedings of PACLING-99 (to appear)* (Waterloo, Ontario, July 1999).

[5] Barzilay, R., and Elhadad, M. Using lexical chains for text summarization. [1], pp. 10–17.

[6] Boguraev, B., and Kennedy, C. Salience based content characterization of text documents. [1], pp. 2–9.

[7] Buckley, C. Implementation of the SMART information retrieval system. Tech. Rep. TR 85-686, Cornell University, 1985.

[8] 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).

[9] Hovy, E., and Lin, C.-Y. Automated text summarization in SUMMARIST. [1], 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] Mani, I., House, D., Klain, G., Hirschman, L., Obrst, L., Firmin, T., Chrzanowski, M., and Sundheim, B. The tipster summac text summarization evaluation. Tech. Rep. MTR 98W0000138, Mitre, October 1998.

[15] Marcu, D. From discourse structures to text summaries. [1], pp. 82–88.

[16] 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).

[17] Mitra, M., Singhal, A., and Buckley, C. Automatic text summarization by paragraph extraction. [1].

[18] Mittal, V. O., Kantrowitz, M., Goldstein, J., and Carbonell, J. Selecting Text Spans for Document Summaries: Heuristics and Metrics. In *Proceedings of AAAI-99* (Orlando, FL, July 1999).

[19] Paice, C. D. Constructing literature abstracts by computer: Techniques and prospects. *Info. Proc. and Management 26* (1990), 171–186.

[20] Radev, D., and McKeown, K. Generating natural language summaries from multiple online sources. *Computational Linguistics 24*, 3 (September 1998), 469–501.

[21] Salton, G., Allan, J., Buckley, C., and Singhal, A. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science 264* (1994), 1421–1426.

[22] Salton, G., and Buckley, C. Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences 41* (1990), 288–297.

[23] Salton, G., and McGill, M. J. *Introduction to Modern Information Retrieval.* McGraw-Hill Computer Science Series. McGraw-Hill, New York, 1983.

[24] Shaw, J. Conciseness through aggregation in text generation. In *Proceedings of 33rd Association for Computational Linguistics* (1995), pp. 329–331.

[25] 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.

[26] Tait, J. I. *Automatic Summarizing of English Texts.* PhD thesis, University of Cambridge, Cambridge, UK, 1983.

[27] 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.