Selecting Text Spans for Document Summaries: Heuristics and Metrics

Vibhu Mittal[*] **Mark Kantrowitz**[*] **Jade Goldstein**[†] **Jaime Carbonell**[†]

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

[†]Language Technologies Institute
Carnegie Mellon University
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 an analysis of news-article summaries generated by sentence extraction. Sentences are ranked for potential inclusion in the summary using a weighted combination of linguistic features – derived from an analysis of news-wire summaries. This paper evaluates the relative effectiveness of these features. In order to do so, we discuss the construction of a large corpus of extraction-based summaries, and characterize the underlying degree of difficulty of summarization at different compression levels on articles in this corpus. Results on our feature set are presented after normalization by this degree of difficulty.

Introduction

Summarization is a particularly difficult task for computers because it requires natural language understanding, abstraction and generation. Effective summarization, like effective writing, is neither easy and nor innate; rather, it is a skill that is developed through instruction and practice. Writing a summary requires the summarizer to *select*, *evaluate*, *order* and *aggregate* items of information according to their relevance to a particular subject or purpose.

Most of the previous work in summarization has focused on a related, but simpler, problem: *text-span deletion*. In text-span deletion – also referred to as text-span extraction – the system attempts to delete "less important" spans of text from the original document; the text that remains can be deemed a summary of the original document. Most of the previous work on extraction-based summarization is based on the use of statistical techniques such as frequency or variance analysis applied to linguistic units such as tokens, names, anaphoric or co-reference information (e.g., (Baldwin & Morton 1998; Boguraev & Kennedy 1997; Aone *et al.* 1997; Carbonell & Goldstein 1998; Hovy & Lin 1997; Mitra, Singhal, & Buckley 1997)). More involved approaches have attempted to use discourse structure (Marcu 1997), combinations of information extraction and language generation (Klavans & Shaw 1995; McKeown, Robin, & Kukich 1995), and the use of machine learning to find patterns in text (Teufel & Moens 1997; Barzilay & Elhadad 1997; Strzalkowski, Wang, & Wise 1998). However, it is difficult to compare the relative merits of these various approaches because most of the evaluations reported were conducted on different corpora, of varying sizes at varying levels of compression, and were often informal and subjective.

This paper discusses summarization by sentence extraction and makes the following contributions: (1) based on a corpus of approximately 25,000 news stories, we identified several syntactic and linguistic features for ranking sentences, (2) we evaluated these features – on a held-out test set that was not used for the analysis – at different levels of compression, and (3) finally, we discuss the degree of difficulty inherent in the corpus being used for the task evaluation in an effort to normalize scores obtained across different corpora.

Ranking Text Spans for Selection

The text-span selection 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 problem: *ranking sentences* from the original document according to their salience (or 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 most relevant documents.

Ranking text-spans for importance requires defining at least two parameters: (i) the granularity of the text-spans, and (ii) metrics for ranking span salience. While there are several advantages of using *paragraphs* as the minimal unit of span extraction, we shall conform to the vast majority of previous work and use the *sentence* as our level of choice. However, documents to be summarized can be analyzed at varying levels of detail, and in this paper, each sentence can be ranked by considering the following three levels:

- *Sub-document Level:* Different regions in a document often have very different levels of significance for summarization. These sub-documents become especially im-

portant for text genres that contain either (i) articles on multiple, equally important topics, or (ii) multiple subsections, as often occurs in longer, scientific articles, and books. All sentences within a sub-document are assigned an initial score based on the whole sub-document. Sub-document scores depend both on various properties independent of the content in the sub-document (e.g., length and position), as well as the lexical and syntactic relations that hold between the sub-documents (e.g., discourse relations, co-references, etc.).

- *Sentence Level:* Within a sub-document, individual sentences can be ranked by using both features that are independent of the actual content, such as the length and position, as well as content specific features that are based on number of anaphoric references, function words, punctuation, named-entities, etc.
- *Phrase/Word Level:* Within a sentence, phrases or words can be ranked by using features such as length, focus information, part of speech (POS), co-reference information, definiteness, tense, commonality, etc.

Some of these features are harder, or costlier, to compute than others. By analyzing their relative utility for a particular task, users can make informed decisions on the cost-benefit ratios of various combinations in different contexts without having to first build a system with which to experiment. Section lists the features we evaluated and discusses our results in detail.

Data Sets: Properties and Features

A corpus of documents and corresponding summaries at various levels of compression are required for experimenting with various summarization methods because summarizer performance can vary significantly at different compression levels. In our experiments, different algorithms for summarization performed best at different levels of compression. This suggests that experiments evaluating summarizers should be conducted at a variety of compression levels, and should be reported in a manner similar to the 11-point precision-recall scores that are used in information retrieval (Salton & McGill 1983). To conduct our experiments, we collected three corpora. The first data set, *Model Summaries*, was created from four data sets supplied as part of the Tipster (Tipster 1998) evaluation: the training set for the Question and Answer task and three other data sets used in the formal evaluation. This data set is relatively small: it consists of 138 documents and "model" summaries. Each "model" summary contains sentences extracted from the document that answer possible questions for the given document. Because of concerns about the small size of the first data set (and its uniformity of summary compression ratios), we acquired two additional, larger data sets consisting of news-wire summaries from Reuters and the Los Angeles Times. Our analysis covered approximately 24,000 summaries over a six-month period in 1997–1998 on a variety of news topics (international, political, sports, business, health and entertainment news articles). Statistics about the average length of stories and summaries from randomly chosen subsets of all three of these data-sets are shown in Table 1.

However, the summaries in the latter two datasets could not be used directly by us, because these were *not* generated by sentence extraction. Therefore, we first converted the hand-written summaries into their corresponding extracted summaries to conduct an analysis of their discourse, syntactic and lexical properties. This conversion – from hand-written to extracted – was done by matching each sentence in the hand-written summary with the smallest subset of sentences in the full-length story that contained all of the key concepts mentioned in that sentence. 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. This approach is a simplified version of the text alignment problem used to align different languages. 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 will depend upon the lexical distribution patterns between summaries and the articles; it may require knowledge of synonyms and hypernyms, such as those provided by WordNet. More details on this work can be found in (Banko *et al.* 1999). This transformation resulted in a 20% increase in summary length on average, probably because hand-written summaries often employ complex syntactic sentential patterns with multiple clauses. Several story sentences were sometimes necessary to cover a single summary-sentence.

Evaluation Metrics

There have been several recent attempts to define evaluation criteria for summarizers, the most extensive being the one organized by the Tipster (Tipster 1998) program. Tipster was motivated partly by the difficulty of evaluating the relative merits of two summarization systems unless their performance was measured on the same task: summaries generated from identical documents at identical character compression levels. This evaluation recognized the fact that different corpora can yield different results because of the inherent properties of the documents contained in them. For instance, some types of documents can be very structured or focused on the main topic. Extraction of sentences from such a document, even at random, is more likely to form a reasonable summary, than random extraction of sentences from a document that is long and rambling with many digressions. Thus, to be able to compare the performance of a particular heuristic or algorithm for summarization on a corpus, it becomes essential to first understand the underlying degree of difficulty of that corpus. Consider, for instance, the performance of random sentence selection on three randomly selected subsets of approximately 1000 articles each from Reuters, The Los Angeles Times and the Christian Science Monitor: at a 20% compression level, the "score" for the same summarizer was 0.263, 0.202 and 0.353 respectively. If these scores were reported separately, as three dif-

Property	Model Summaries	Reuters Summaries	Los Angeles Times Summaries
task	Q and A	generic summaries	generic summaries
source	Tipster	human ⇒ extracted	human ⇒ 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
minimum sentences per doc	11	5	3
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: Characteristics of data sets used in the summarization experiments

ferent experiments, one might wrongly infer that the 'different' algorithms varied widely in performance Thus, even when testing and evaluation is done on the same document genre, it is important to clearly state the baseline performance expected from that corpus.

The second issue, directly related to the previous one, is the desired compression level. Clearly, generating summaries at a 50% compression level should be much easier than generating summaries at a 10% compression level.

Current methods of evaluating summarizers often measure summary properties on absolute scales, such as precision, recall, and F_1 (Salton & McGill 1983). Although such measures can be used to compare summarization performance on a common corpus, 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

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

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.

If D_t is the total number of sentences in a document, D_r the number of summary-relevant sentences in the document, and S_r the target number of sentences to be selected for inclusion in the summary, then let $P_i(D_t, D_r, S_r)$ denote the probability of selecting S_r sentences such that i of them are from the set of D_r relevant sentences. Then $P_i(D_t, D_r, S_r)$ is the product of the number of ways to select i sentences from the D_r relevant sentences, multiplied by the number of ways to select the remaining $S_r - i$ sentences from the $D_t - D_r$ non-relevant sentences, and divided by the number of ways to select S_r sentences from the D_t sentences in the document. Thus

$$P_i(D_t, D_r, S_r) = \frac{\begin{pmatrix} D_r \\ i \end{pmatrix} \begin{pmatrix} D_t - D_r \\ S_r - i \end{pmatrix}}{\begin{pmatrix} D_t \\ S_r \end{pmatrix}} \tag{2}$$

Let $E(D_t, D_r, S_r)$ be the expected number of relevant sentences. Then

$$E(D_t, D_r, S_r) \;=\; \sum_{i=0}^{D_r} i \cdot P_i(D_t, D_r, S_r) \;=\; \frac{D_r \cdot S_r}{D_t}$$

From this it can be derived that

$$F_1 \;=\; \frac{2 \cdot D_r \cdot S_r}{D_t \cdot (D_r + S_r)} \tag{3}$$

This formula relates F_1, D_t, D_r, and S_r. Given three of the values, the fourth can be easily calculated. In particular, the value of a baseline F_1 can be calculated once the average corpus statistics are known (lengths of an average document and summary and the number of relevant sentences per document). For instance, for the Reuters articles in our case, the values of the relevant parameters D_t, D_r and S_r are 23.10, 10, 4.3. This yields a baseline F_1 score of approximately 0.260 at a compression level of 20%. In similar fashion, one can compute the baseline F_1 scores for any desired compression level (or vice versa).

Experiments

Summary lengths in our corpus seemed to be mostly independent of document length; they were narrowly distributed around 85–90 words, or approximately five sentences. Thus, compression ratios decrease with document length. This

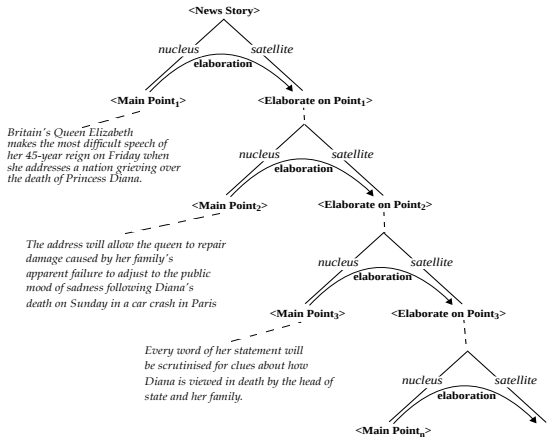

Figure 1: Discourse structure of a stereotypical news story

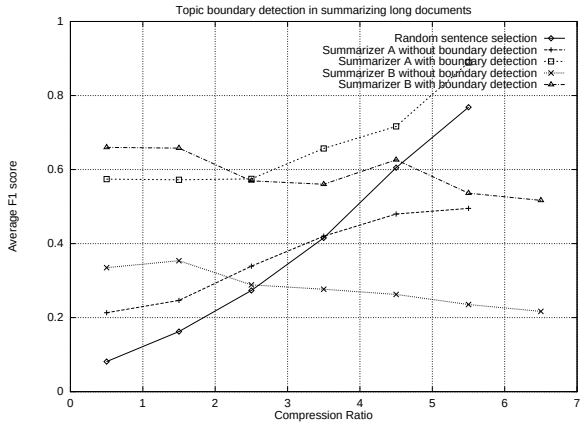

Figure 2: Effect of topic detection.

suggests that the common practice of using a fixed compression ratio in evaluation may not be appropriate, and that using a genre-specific constant summary length may be more natural.

Our experiments to evaluate features were conducted as follows: first, we identified article-summary pairs at various levels of compression ranging from 0–10%,10–20% … 40-50%. Our dataset contained summaries upto seven-tenths of the length of the original article, but at these relatively large ratios, the summaries begin approaching the original article, so we restricted the comparison to summaries that were at most half the length of the original article. For each of these compression ratios, we randomly selected 3000 of these article-summary pairs. These 15,000 pairs were then split up into 10,000 training articles for analysis and 5000 were held out for testing. During testing, the summarizers were invoked for each article with a parameter specifying the length of the summary to be generated. This length, in sentences, was also the length of the original summary for that article. The overlap between the generated summary and the original summary was then measured using the well-known F_1 measure from information retrieval. This enabled us to measure the summarizer performance at various levels of compression. To reduce the five values to one number for reporting purposes, we computed the average F_1 score for all five intervals. Some of the tables towards the end of this section contain this "five point average" score.

The rest of this section discusses some of the features we looked at, starting from the largest granularity level (the document structure) and moving to the smallest (individual words).

Document Level Features in Summarization

It has been argued previously that discourse structure can be a very useful source of information for summarization (Marcu 1997; Sparck-Jones 1993). This argument is based on theories of discourse which recursively indicate "core" and "contributor" spans of text in the document (Mann & Thompson 1988). The problem of summarizing a document can be addressed by selecting the most abstract "core" spans until the desired summary length is reached. The utility of this approach is very effectively illustrated in the news-wire document genre, where selecting the first few sentences of the story results in better-than-average summaries (Brandow, Mitze, & Rau 1995). This is because news-wire stories are written with a rigid, right-branching discourse structure. (This stereotypical structure is partly due to the fact that the page and column layouts of the newspaper are not known at the time the article is written; thus, the article can be chopped off at any point, and must still be comprehensible.) (Figure 1 shows one of these structures.) Thus selecting the first n sentences of a news-wire story approximates the selection of the top "core" spans in the text. (This approach is not perfect because news-wire stories do not always employ a perfectly right-branching discourse structure.)

Unfortunately, finding the underlying discourse structure is almost as difficult as generating a good summary (Marcu 1997). In this paper, we considered a simpler version of the problem: rather than finding the underlying discourse structure, we segmented the text into sub-documents according to topic boundaries using the TextTiling system (Hearst 1994). (Our experiments with other segmentation systems resulted approximately equivalent bopundaries.) The segmentation represents an approximation to the top level discourse structure, minus knowledge of the relations between nodes, which represent the sub-documents. Note that in theory, sub-document segmentation can be carried out recursively, yielding approximate boundaries for a hierarchical approximation to the discourse structure. At this point, information at the sub-document level, such as the position and length of a sub-document relative to other sub-documents, can be used to augment information available at finer levels of detail. Segmentation becomes increasingly important as documents grow in length and complexity. For instance, the average conference paper at 5,000 words is over 10 times longer than the average news-wire article in the Reuters corpus. In such cases, making use of information at this level can yield significant advantages.

To evaluate the cost-benefit tradeoff of pre-processing

documents to find topic boundaries, we created a synthetic data-set for our experiments. This is because, there does not yet exist a corpus of longer documents with "gold standard summaries" to test on.[1] We created composite documents by concatenating several news-wire articles of various lengths. The number of such sub-documents in a larger document was normally distributed between 2 and 14. The summary for this composite document was assumed to be the collection of sentences in the summaries for the individual sub-documents. Our experiments found that in all cases, pre-processing for topic boundaries can significantly improve the quality of a summarization system, often by a factor of 2. (This factor is based on average 5-point F_1 score averaged across compression levels and normalized with the random-sentence-selection baseline. The actual improvement will depend on the length and complexity of the original document and the summarization algorithm being used. Clearly this approach has maximum utility when a sentence-position based approach is used for selecting sentences.) This approach is analogous to related work on summarization aimed at reducing redundancy (Carbonell & Goldstein 1998). As in this case, that approach attempts to select summary sentences from different topic areas in the document, but without any explicit topic segmentation steps. Figure 2 shows the effectiveness of being able to identify topic boundaries on two different summarization algorithms. (Algorithm-A was based on a TF·IDF approach to ranking sentences; Algorithm-B was based on a combination of syntactic complexity and named-entity relationships between sentences in the document; neither of the two used positional information.)

Sentence Level Patterns in Summary Sentences

Most of the heuristics that have been used in selecting text-spans for summarization have been at the sentence level. These include sentence-level features such as the length of the span, its complexity, the presence of certain punctuation, thematic phrases, anaphora/co-occurrence density, etc. It is essential to understand the relative advantages of these features over one another for a specific corpus, especially since these features vary widely in computational cost and some of them subsume one another. We looked at the following sentence level characteristics as possible cues for summary selection:

- *Syntactic Complexity:* Our analysis of summary sentences found that sentences included in summaries differed from non-summary sentences in several characteristics related to complexity. Two of these are:
  - NP Complexity: We found that the average length of complex noun phrases in summary sentences was more than twice as long than those in non-summary sentences.
  - Coordinate Conjunctions: On the negative side, our dataset possessed a higher density of coordinate conjunctions in story sentences than summary sentences.
- *Density of Named-Entities:* Named entities 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.
- *Punctuation:* Punctuation symbols (not counting periods) tend to also appear more frequently in story sentences as compared to summary sentences.
- *Given vs. New Information:* In a simplified analysis of "given" vs. "new" information (Werth 1984), as indicated by the presence of the definite or indefinite articles, we found that summaries included new information 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.
- *Pronominal References:* Anaphoric references at sentence beginnings, such as "these", "this", "those", etc. are a good source of negative evidence, possibly because such sentences cannot introduce a topic. Personal pronouns such as "us", "our" and "we" are also a good source of

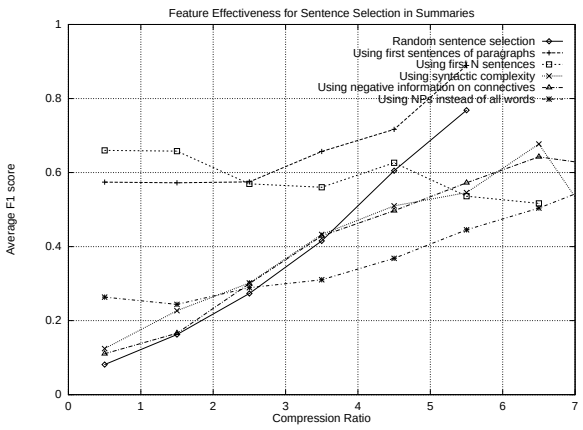

Figure 3: Feature effectiveness for Reuters dataset.

[1] Some researchers have reported experiments using conference papers from the CMP-LG archive, but there are two problems with that approach: (i) abstracts in scientific papers are not generic summaries, and (ii) scientific papers and their abstracts are written by technical people who are not necessarily skilled abstractors. We have conducted experiments with non-synthetic datasets, but we do not have sufficiently large numbers of such documents at various compression levels for us to be able to report on them here.

Table 2: A comparison of word occurrences in summary sentences to 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%

Table 3: Effectiveness of sentence level heuristics by raw score and normalized score (Equation 1).

	Reuters		LA Times	
Feature Set	Raw	$p'_{0.26}$	Raw	$p'_{0.20}$
synt. complexity	0.32	0.08	0.24	0.05
named-entity density	0.30	0.05	0.26	0.07
synonym density	0.31	0.07	0.25	0.06
first n (FN)	0.61	0.47	0.58	0.47
FN + syntax	0.61	0.47	0.61	0.51
FN + named-entity	0.65	0.53	0.64	0.55
FN + given/new	0.67	0.55	0.64	0.55
FN + SD	0.64	0.51	0.62	0.52
ALL (weighted comb.)	0.82	0.75	0.72	0.65

negative evidence for summary sentences, perhaps because they frequently occur in quoted statements.

- *Density of Related Words:* Words that have multiple related terms in the document – synonyms, hypernyms and antonyms – are much more likely to be in the summary than not.

Figure 3 illustrates the effectiveness of some these features at different compression levels from the Reuters dataset.[2] Note that some of the features may not appear as important as others in the 5-point average scheme used here because these features, such as for instance, named-entities, may not occur in more than a small percentage of the sentences. As the summaries get larger, their individual effect on the summary decreases. Table 3 shows the average performance of various features for summarization. Since this was a news-wire corpus, where the selection of the first n sentences has been shown to be a very effective heuristic, the table also includes, for illustration, the performance of the summarizers when other features are combined with the heuristic. The tables indicate both the raw scores, as well as normalized scores. While the raw scores between the two datasets vary widely in some cases, the normalized scores are much closer and are a better reflection on the effectiveness of the features being used.

[2]Note that we did not include *punctuation* or *pronominal information* information for our table, since these features are mostly used diminish the likelihood of summary selection; thus these features are best used in combination with other positive features. Similarly, in Table 4, certain features have been left off.

Phrase/Word Level Patterns in Summary Sentences

In addition to features at the document and sentence level, there are certain characteristics at the phrase/word level that can be used to rank sentences as well. These include:

- *Word Length:* 63% of the Reuters summaries and 66% of the Los Angeles Times summaries had a greater average word length than the average word length of the article's non-summary sentences.
- *Communicative Actions:* Words and phrases common in direct or indirect quotations also suggest 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.
- *Thematic Phrases:* Phrases such as "finally," "in conclusion," etc. also occur more frequently in summary sentences. We found a total of 22 such terms in our data-set that occur at least 75% more frequently in summary sentences than non-summary sentences.
- *Miscellaneous:* Furthermore, informal or imprecise terms such as "got", "really" and "use" also appear significantly more frequently in non-summary sentences. Other sources of negative lexical significance we found are:
  - Honorifics: Honorifics such as "Dr.", "Mr.", and "Mrs.", are also negatively indicated. This may be due to the fact that news articles often introduce people by name, (e.g., "John Smith") and subsequently refer to them either formally (e.g., "Mr. Smith") or by pronominally.
  - Auxillary verbs: such as "was", "could", "did", etc.
  - Negations: such as "no", "don't", "never", etc.
  - Integers, whether written using digits (e.g., 1, 2) or words (e.g., "one", "two").
  - Evaluative and qualifying words, such as "often", "about", "significant", "lot", "some" and "several".
  - Prepositions, such as "at", "by", "for" "of", "in", "to", and "with".

Our analysis found that each of these features, individually, can be helpful as a cue in selecting summary sentences. Figure 3 shows the effectiveness of some of these features in relationship to the random baseline. At very small summary lengths, all of these features are better than the baseline. As summaries get longer, the effectiveness of an individual feature starts to fall. This is understandable, since features such as anaphoric references or dangling connectives are unlikely to occur in more than a small percentage of the sentences. The outstanding line in this case is the positional feature, which, as discussed earlier, implicitly takes advantage of the underlying discourse structure of the news story. As discussed earlier, it is essential to clearly understand the relative costs and benefits of each of these features. It is also important to understand the relative effects of using *combinations* of these features. Results based on combinations of

Table 4: Effectiveness of word level heuristics.

Feature Set	Reuters		LA Times	
	Raw	$p'_{0.26}$	Raw	$p'_{0.20}$
Phrase Complexity	0.28	0.03	0.22	0.04
thematic phrases	0.29	0.04	0.25	0.06
misc ftrs	0.25	-0.01	0.19	-0.01

features can be useful for inferences about the orthogonality and the interdependence between these features. The space in which a hill-climbing technique must search for appropriate weights in a linear combination of these features is quite large; experiments to understand these approaches are currently under way.

Conclusions and Future Work

Human-quality text summarization systems are difficult to design, and even more difficult to evaluate; results reported by different research projects are also difficult to compare because the reported results often do not discuss the characteristics of the corpora on which the experiments were conducted – specifically, characteristics such as redundancy, which can have significant effects on any evaluation measures. We have argued that for specific datasets, characterizing (1) sentence redundancy, and (2) results at a variety of compression levels, are necessary if these results are to be useful to other researchers.

This paper has presented a discussion of sentence selection heuristics at three different granularity levels. We conducted experiments to evaluate the effectiveness of these heuristics on the largest corpus of news-article summaries reported so far. Our experiments emphasized the need for sub-document segmentation on longer, more complex documents. This paper also shows that there are significant advantages in using fine grained lexical features for ranking sentences. Results in our work are reported using a new metric that combines scores at various compression levels taking into account the corpus difficulty on the task.

In future work, we plan to characterize different document genres, and attempt to achieve a better understanding of why certain phenomena play a greater/lesser role for sentence selection in those genres – phenomena such as the role of coreference chains, the given/new distinction and others.

References

Aone, C.; Okurowski, M. E.; Gorlinsky, J.; and Larsen, B. 1997. A scalable summarization system using robust NLP. In Mani and Maybury (1997), 66–73.

Baldwin, B., and Morton, T. S. 1998. Dynamic coreference-based summarization. In *Proceedings of EMNLP-3 Conference.*

Banko, M.; Mittal, V.; Kantrowitz, M.; and Goldstein, J. 1999. Generating Extraction-Based Summaries from Hand-Written One by Text Alignment. In *Submitted to the 1999 Pac. Rim Conf. on Comp. Linguistics.*

Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In Mani and Maybury (1997), 10–17.

Boguraev, B., and Kennedy, C. 1997. Salience based content characterization of text documents. In Mani and Maybury (1997), 2–9.

Brandow, R.; Mitze, K.; and Rau, L. F. 1995. Automatic condensation of electronic publications by sentence selection. *Info. Proc. and Management* 31(5):675–685.

Carbonell, J. G., and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR-98*.

Hearst, M. A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the ACL*.

Hovy, E., and Lin, C.-Y. 1997. Automated text summarization in SUMMARIST. In Mani and Maybury (1997), 18–24.

Klavans, J. L., and Shaw, J. 1995. Lexical semantics in summarization. In *Proceedings of the First Annual Workshop of the IFIP Working Group FOR NLP and KR.*

Mani, I., and Maybury, M., eds. 1997. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*.

Mann, W. C., and Thompson, S. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text* 8(3):243–281.

Marcu, D. 1997. From discourse structures to text summaries. In Mani and Maybury (1997), 82–88.

McKeown, K.; Robin, J.; and Kukich, K. 1995. Designing and evaluating a new revision-based model for summary generation. *Info. Proc. and Management* 31(5).

Mitra, M.; Singhal, A.; and Buckley, C. 1997. Automatic text summarization by paragraph extraction. In Mani and Maybury (1997), 31–36.

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

Sparck-Jones, K. 1993. Discourse modelling for automatic summarising. Technical report, Cambridge University, Cambridge, England.

Strzalkowski, T.; Wang, J.; and Wise, B. 1998. A robust practical text summarization system. In *AAAI Intell. Text Summarization Wkshp*, 26–30.

Teufel, S., and Moens, M. 1997. Sentence extraction as a classification task. In Mani and Maybury (1997), 58–65.

Tipster. 1998. Tipster text phase III 18-month workshop notes. Fairfax, VA.

Werth, P. 1984. *Focus, Coherence and Emphasis*. London, England: Croom Helm.