Presented at ICDAR '97, pp. 227-232, Ulm, Germany, Aug. 18-20, 1997.

Extraction of Indicative Summary Sentences from Imaged Documents

Francine R. Chen Dan S. Bloomberg
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
{fchen,bloomberg}@parc.xerox.com

Abstract

A system for selecting sentences from an imaged document for presentation as part of a document summary is presented. The extracts are identified without the use of optical character recognition. The sentences are selected based on a set of discrete features characterizing the words within a sentence and the location of the sentence within the imaged document. Each sentence is scored based on the values of the discrete features using a statistically based classifier. The imaged document is processed to identify the word locations, the reading order of words, and the location of sentence and paragraph boundaries in the text. The words are grouped into equivalence classes to mimic the terms in a text document. A sample extract for a technical document is shown, and evaluation against a set of abstracts created by a professional abstracting company is given. These results are compared with text-based abstracts.

1 Introduction

A summary is commonly thought of as a concise interpretation of a document, represented by a small number of sentences. Techniques for creating computer-generated summaries of textual documents have been developed by a number of researchers (an overview is given in [7]). Many of these techniques rely on extraction of phrases and sentences, due in part to the difficulty of natural language understanding and generation. However, a summary of an imaged document can take other forms. These may be extracts, such as sentences, phrases, headings, or figures, that together communicate a sense of the document.

Although many documents are available as ASCII text, many are available only as paper documents. One approach to automatically summarize paper documents is to scan the paper document, convert it to text using optical character recognition (OCR) and then use techniques developed for text. However, character recognition systems are not perfect, and require significantly more processing time than the summarization process. For applications where a quick summary is desired, OCR may be unnecessary. Such applications may include optional summaries provided by reading machines for the blind, or those in which data is primarily transient, such as facsimile transmissions or digital copiers. Another possible application is to scan and store documents digitally, and provide a summary sheet that can be used for later retrieval of the documents.

In contrast to the *text-based* techniques developed by other researchers [5, 6, 7, 9, 10, 11], this paper describes a method for automatically selecting sentences for creating a summary from an *imaged* doc-

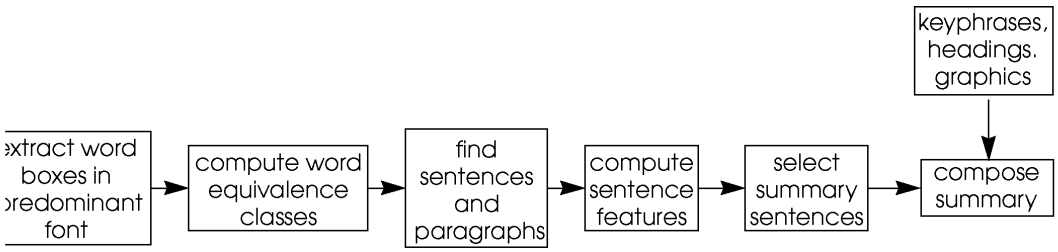

Figure 1: Text image summarization system.

ument without recognition of the characters in each word. Summary sentence selection is performed using a statistical classifier to determine the likelihood of each sentence in a document being a summary sentence. The sentences most likely to be a summary sentence are then selected for extraction. This method parallels the approach to text summary sentence selection described in Kupiec *et al.* [5] and uses the conditional probabilities obtained from training that system. In earlier work [3], we outlined a system for selection of "thematic" sentences and for selection of keyphrases. Our earlier system tended to select sentences throughout the document and these tended to result in informative types of summaries. In contrast, the sentences selected by the current system tend to be more indicative, and therefore are usually more coherent as a group.

Similar to work by others on ASCII text (e.g., [6, 5]) and imaged text [12], our approach to image-based summarization does not rely on language understanding or generation. Many of the text-based characterizations are performed without referring to word meaning, and these underlying text-based concepts can be applied directly to a document that has been scanned to form a set of text images. Some features that are available if the characters in a word are known are unavailable, and therefore not utilized in image-based summarization. For other features, approximations to the information needed to compute them may suffice. In particular, methods for identifying *stop* words without matching character sequences and for identifying *word equivalence classes*, that is words that are instances of the same term, are needed. Our method for handling these differences will be described in this paper.

Figure 1 outlines the steps in performing text image summarization. To identify a set of summarizing excerpts, word boxes in the predominant font are extracted from an imaged document, and then word-box equivalence classes are identified based on shape similarity. From the word-box equivalence classes, term frequencies can be estimated. Layout analysis is performed to determine the reading order of textblocks, sentence boundaries and paragraph boundaries. Thematic sentences are identified using algorithms which are based on a statistical characterizations of words in a document, without regard to possible meanings of the words. This information is used to compute features characterizing each sentence. Summary sentences are identified as the best scoring sentences, where the score of each sentence in a document is computed based on the values of a set of discrete features. The resulting sentence excerpts can be composed to form a set of summary sentence images, and combined with other extracts, such as keyphrases, graphics, and/or a table of contents derived from the headings to compose the final summary. In this paper, we focus on identifying and selecting terms, the selection of summary sentences and performance evaluation of summary sentence selection. For a set of 173 imaged documents, the sentences selected by the system versus the sentences selected by professional abstractor are compared and the results are discussed.

2 Term Identification

Word bounding boxes in the predominant font are processed to identify those words that are the same term, based on image matching rather than character matching. This is done by identifying equivalence classes of words, where each equivalence class is assumed to represent a unique term. The statistics characterizing each equivalence class are then used to identify terms that are stop words.

2.1 Word equivalence classes

To identify words corresponding to the same term, an unsupervised classifier is used to place each word box in the predominant font into one of a set of equivalence classes (see [1] for details). The classifier compares word images using a rank blur hit-miss transform [2] which allows some outlier pixels in the word matches. The number of outlier pixels allowed is proportional to the area of the word bounding box. If the number of permitted outliers is set sufficiently large to prevent an equivalence class representing a short word from being split, then in long words, errors occurring in a small number of characters could result in merging of classes. This is prevented by performing a second match with a larger structuring element for the blur and a smaller number of permitted outliers. Both matches must be valid for a word image to be placed in an existing class.

The number of members in an equivalence class serves as an estimate of term frequency for that class. Since term frequencies are used in the identification of stop words, class splitting needs to be minimized for good estimates. Class splitting is further reduced by removing punctuation, such as a comma or period, following a word.

In imaged text, the text may be rendered in different sized fonts. The main body of text is usually printed in the one font, which is generally the predominant font, whereas headings and captions may appear in a variety of fonts. For identifying summarizing sentences using an image-based approach, only text in the predominant font is considered. Since words in the non-predominant font will not match equivalence classes of words in the predominant font, extra computation will not be expended in creating small separate equivalence classes which will not be utilized in later processing. Identification of the predominant font also simplifies the identification of the reading order of the textblocks, because blocks of non-predominant text interspersed on the page are not considered.

2.2 Stop, content and thematic words

Words that are not content words (commonly called *stop* words) are identified from the equivalence classes. Word frequencies, word locations, and word image widths are used to rank equivalence classes as to their likelihood of being a stop word. Generally, a word is more likely to be a stop word if it is of high frequency and of small width or is rarely the first word in a sentence. The list of stop words is chosen by selecting the N highest ranking words, where N is dependent on document length and other characteristics of the document (see [3] for details). All non-stop words are considered to be *content* words. From the content words, a small set of high frequency words is chosen as thematic words or *keywords*. To identify keywords, the content words are sorted by frequency and then by width in the case of ties. Keywords are identified from the sorted list as the most frequent content words in a document.

3 Summary Sentence Excerpt Selection

Sentences are selected for extraction to form a summary using algorithms based on statistical characterizations of words and sentences in a document. This is done without regard to the possible meanings of the words. In constructing a set of summary sentences, word distributions within a document are examined. Since words in different documents are typically in different fonts, comparison of words across documents without performing OCR is not feasible. Thus instead of using corpus-dependent word-level

information, the method relies on simple measures, such as word proximity and statistics on word frequencies within a document. In addition, within document information, such as sentence position and sentence length, are also considered.

In this work, sentences are selected only from text in the predominant font. Because our techniques include a feature based on high frequency words, it is unlikely that text in non-predominant fonts would be selected. As mentioned earlier, text in the predominant font generally corresponds to the main body of text, so that the selected sentences are drawn from the main body of text. Text in font sizes significantly larger than the predominant font are often headings that can be used to create a table of contents and serve as complementary summary information. Side information in smaller fonts, such as author descriptions, are generally not desirable for inclusion in a summary and are avoided by selecting only sentences from the predominant font.

3.1 Summary sentence selection

A summary score is computed for each sentence using a statistical model of discrete feature values characterizing each sentence. The best scoring sentences are selected as excerpts. The features used are described in section 3.2. The probabilistic scoring is similar to that described in Kupiec [5]. We review the method here.

The probability that a sentence, s_i, is included in a summary $\mathcal{S}$, given the values of a set of K discrete features, $F_1, F_2, \ldots, F_K$, by Bayes' theorem is:

$$P(s_i \in \mathcal{S}|F_1, F_2, \ldots, F_K) = \frac{P(s_i \in \mathcal{S})\ P(F_1, F_2, \ldots, F_K|s_i \in \mathcal{S})}{P(F_1, F_2, \ldots, F_K)}.$$

Because of a limited amount of data for estimating the probabilities, we assume that the features are statistically independent. Thus $P(s_i \in \mathcal{S}|F_1, F_2, \ldots, F_K)$ can be expressed as:

$$P(s_i \in \mathcal{S}|F_1, F_2, \ldots, F_K) = P(s_i \in \mathcal{S})\ \frac{\prod_{j=1}^{K} P(F_j|s_i \in \mathcal{S})}{\prod_{j=1}^{K} P(F_j)}.$$

$P(s_i \in \mathcal{S})$ is assumed to be uniform for each sentence. Since only the relative scores for each sentence will be compared, $P(s_i \in \mathcal{S})$ can be ignored. We apply Bayes' once more:

$$\begin{aligned} &P(s_i \in \mathcal{S}|F_1, F_2, \ldots, F_K) \\ &\quad = \frac{1}{P(s_i \in \mathcal{S})^{K-1}} \prod_{j=1}^{K} \left(\frac{P(F_j|s_i \in \mathcal{S})P(s_i \in \mathcal{S})}{P(F_j)} \right) \\ &\quad \propto \prod_{j=1}^{K} P(s_i \in \mathcal{S}|F_j). \end{aligned}$$

Since the features are assumed to be independent and $P(s_i \in \mathcal{S})$ is assumed to be uniform, estimation of the conditional feature probabilities, $P(s_i \in \mathcal{S}|F_j)$, may be performed separately for each feature, and the final score of each sentence is computed as the product of the conditional probabilities. The conditional feature probability, $P(s_i \in \mathcal{S}|F_j)$, represents the probability of a sentence being a summary

sentence given the value of a particular feature F_j, and are estimated from the training set. The number of sentences to be selected for a summary, N, is specified by the user. The most probable N sentences are identified as the best scoring N sentences. These sentences are selected as the set of sentence excerpts and can be composed to form a summary image.

3.2 Features used in summary sentence selection

Each sentence is characterized by a set of discrete features: (1) sentence length (2) location of the sentence with respect to both the containing paragraph in the document and the location of the sentence within the containing paragraph and (3) the most thematic sentences. The set of features that we use and how they are computed for the imaged documents are described below:

Sentence Length Feature: This feature tends to remove from consideration sentences that are unlikely to be included in summaries and some "sentences" that are a result of errors in sentence boundary identification. Sentences that are very long tend to be due to segmentation errors since sentence boundaries that are missed result in longer "sentences." Sentences that are very short may be due to undetected abbreviations that are misidentified as a sentence boundary. In addition, true sentences that are very short tend to not be included in summaries. Given a lower and upper threshold on the number of words in a sentence, this feature is true for all sentences for which the number of words is between the two thresholds, and false otherwise. We use a lower threshold of 5 and an upper threshold of 40.

Location Feature: This feature jointly captures position of a sentence within a paragraph and paragraph position of a sentence within a document. Sentences in the introductory or concluding sections of an article tend to be included as summary sentences. The first and last sentences in a paragraph also tend to be included more often as summary sentences. Sentences in the introductory and concluding sections are approximated by noting the "paragraph number" of the sentences in the first ten or last five paragraphs. All other paragraphs are given the same feature value. With imaged text, it may be easier to identify section headings by font changes, and this could be used to identify sentences only within introductory or concluding sections. However, we chose to keep our feature definition as consistent as possible to that used for text, in order to use the parameter values obtained for text. The location of a sentence within a paragraph is characterized by noting whether a sentence is paragraph initial, final (for paragraphs composed of more than one sentence), or medial (for paragraphs composed of more than two sentences).

Thematic Feature: The most frequent content words are defined as thematic words. Sentences that contain at least one thematic word, or keyword, are referred to as "thematic" sentences. Using only the keywords, $\{k\}$, each sentence j is assigned a score, s_j, based on the number of occurrences of each keyword in the sentence, $c_j(k)$, and the frequency with which each keyword occurs relative to other keywords in the document, $f(k)$:

$$s_j = \sum_{k \in keywords} c_j(k) \cdot (1 + f(k)).$$

The sentence counts are weighted more heavily towards more frequent words in an effort to bias the selected sentences to include those which are more likely to be related to other selected sentences. The weighting by $f(k)$ is generally used to break ties when several sentences contain the same number of keywords. The N best-scoring sentences are identified and assigned the value "true," indicating these sentences are the more thematic sentences in the document; the remaining sentences are assigned the value "false." Here we use N equal to 10.

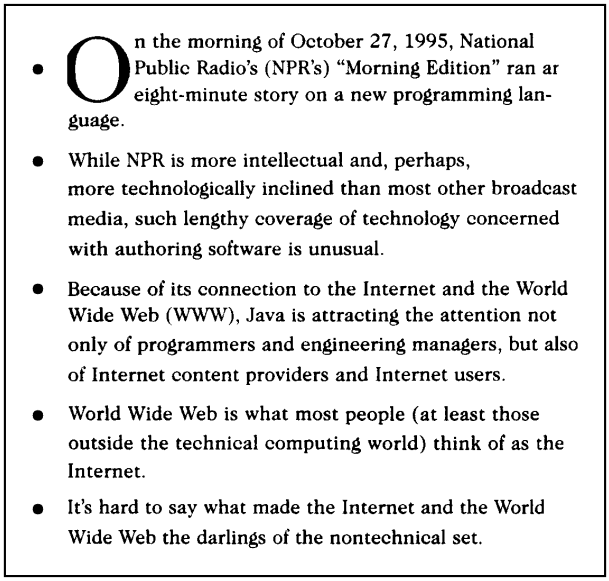

Figure 2: Sample summary sentences.

4 Experiments

In this section we describe the training and test corpora that was used for our experiments. Evaluation of the results of selecting summary sentences by comparison to sentences selected by professional abstractors is also described.

4.1 Corpus

The corpus was derived from 175 paper articles without an abstract from 21 scientific/technical journals provided by Engineering Information. Professional abstractors created an abstract for each of these articles. Many of the sentences in the created abstracts were observed to closely correspond to sentences in the text [5].

A version of the corpus was used by Kupiec *et al.* [5] for evaluating a text-based summarizer. In their work, the paper articles were scanned at 600 pixels/inch (ppi) and then OCR was performed to extract the text in the documents. The text was then hand-corrected, and normalized so that the first line of each text file contained the document title. A set of conditional probabilities, $P(s \in \mathcal{S}|F_j)$, was estimated for the hand-corrected ASCII text version of the original documents. We used these conditional probabilities for our experiments. Ideally, the conditional probabilities, $P(s \in \mathcal{S}|F_j)$, should be estimated from a corpus of imaged text. The errors in extracting the feature values would then be modeled together with the feature values, providing a better overall model. However, labeling the images to indicate the position of the correct summary sentences would have been tedious.

The test corpus is composed of the document images that were scanned from the paper articles provided by Engineering Information for 175 papers, composed of a total of 879 pages. The original scanned article images were reduced to 300 ppi. For testing, cross-validation was performed. That is, summaries for articles from a given journal were created using the conditional probabilities estimated from the articles from all other journals.

Table 1: Comparison of Text and Imaged Text Summarizers

		text	*image*	
			all	subset
test corpus	# of test sentences	451	445	96
	# of articles	175	173	39
	# of journals	21	21	11
segmentation	hand corrected text	yes	no	no
	1st line in doc id'd	yes	no	yes
features	sentence length	yes	yes	yes
	paragraph location	yes	yes	yes
	thematic	yes	yes	yes
	fixed phrase	yes	no	no
	upper-case word	yes	no	no
results	% correct sentences	42	23	27

4.2 Results

Informal evaluation of a subset of the selected summary sentences indicates that this image-based method produces indicative summaries. A sample set of summary sentences for a six page document[1] is shown in Figure 2. For evaluation, the system was run to obtain a similarly composed image summary for each document in the test corpus. The number of summary sentences was specified to be the same as the number of matchable sentences in the manual summary produced by the professional abstractors. That is, only those sentences in the professional abstracts which roughly matched a sentence in the document were included. (These sentences composed 79% of the sentences in the professional abstracts; see "direct matches" in Kupiec *et al.* [5].) There were a total of 445 direct sentence matches. OCR using ScanWorX was performed on each composed image summary. The sentences in the OCR text were compared with the sentences identified as a direct match. A summary sentence was considered to be correct if at least 70% of the words in the OCR summary sentence corresponded to words in a direct match sentence from the professional abstract for the document.

A comparison of the features, test conditions and performance of the text summarizer described in Kupiec *et al.* [5] and our imaged text summarizer is shown in Table 1. We tested our system with the full set of articles (except for 2 articles for which the scanned image files were missing) and on a subset of the articles that did not have major segmentation errors on the first page. Articles without such errors were defined to be those for which the first line of the document and the majority of the paragraph and sentence boundaries on the first page were correctly identified. On the full set of test articles, the percentage of the 445 matchable sentences that were correctly identified by the image summarizer is 23%. For the subset of test articles, the percentage of the 96 matchable sentences that were correctly identified by the image summarizer is 27%. In the experiments reported by Kupiec *et al.* [5], 42% of the matchable summary sentences were correctly identified. Our poorer performance is due to several factors. In Kupiec *et al.* [5] the ASCII data was hand-corrected so that the first line of the document corresponded to the document

[1] Jim Waldo, "Java and the internet," *Unix Review,* pp. S1-S10, Feb 1996.

title; there was no hand-correction in image-based sentence selection for the full test set. Segmentation errors and errors in identification of the predominant font may all contribute to errors in identifying the beginning of text in imaged text. The increase in performance for our subset of test articles indicates that correct identification of the beginning of the document is important. Our feature set was poorer because in image-based summarization, character-based features, such as fixed-phrases and upper-case-words cannot be used. And finally, there may be a mismatch between the text and image probabilities, because the definition of some of the features was modified to accommodate imaged text and we did not retrain the probabilities for these features.

It should be noted that although we are evaluating against a set of sentences selected by professional abstractors, a set of sentences that is definitively the "best" summary sentences does not always exist. It has been observed that when subjects are given the task of selecting sentences to form a summary, the results are not consistent. This was observed between different subjects [8, 4], as well as between the same subject at one time and at a later date [8], where the same sentences were selected only 55% of the time.

The experiments in selecting summary sentences from imaged text documents were performed on a 75 mHz Sun SPARCstation10. On a per page basis, the processing time was approximately 6 sec per page, of which nearly 2 sec is spent reading and deskewing each page image.

5 Summary

We have described a system for selecting and extracting sentences from imaged documents to create a summary. This system relies only on image processing and statistically based techniques; OCR is not performed. The summary sentences were evaluated by comparison with professional abstracts. When the number of summary sentences is specified to be the same as the number of direct matching sentences in a professional summary, 23% of the summary sentences match those in the professional summary. These summary sentences can be combined with other extracts, such as keyphrases, graphics, and/or a table of contents derived from the headings to compose the final summary.

Acknowledgments

We would like to thank Julian Kupiec for providing the conditional feature statistics derived from textual training data, and Engineering Information for providing the articles and abstracts which formed the basis for the corpus.

References

[1] D.S. Bloomberg and F.R. Chen, "Extraction of text-related features for condensing image documents," *SPIE Conf. 2660, Document Recognition III*, San Jose, CA, Jan 29-30, 1996, pp. 72-88.

[2] D.S. Bloomberg and L. Vincent, "Blur Hit-Miss transform and its use in document image pattern detection," *SPIE Conf. 2422, Document Recognition II*, San Jose, CA, pp. 278-292, Feb 6-7, 1995.

[3] F.R. Chen and D.S. Bloomberg, "Extraction of Thematically Relevant Text from Images," Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, pp. 163-178, April 1996.

[4] F.R. Chen and M.M. Withgott, "The use of emphasis to automatically summarize a spoken discourse," *Proceedings of the IEEE Intl. Conf. on Acoust., Speech and Signal Proc.,* vol. 1, pp. 229-232, March 1992.

[5] J. Kupiec, J. Pedersen and F. Chen, "A trainable document summarizer," *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, pp. 68-73, 1995.

[6] H.P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Develop.,* vol. 2, pp. 159-165, 1959.

[7] C. Paice, "Constructing literature abstracts by computer: Techniques and prospects," *Information Processing and Management,* vol. 26, pp. 171-186, 1990.

[8] A. Resnick, "The formation of abstracts by the selection of sentences: Part II. The reliability of people in selecting sentences," *American Documentation,* 12(2): 141-143, April 1961.

[9] G. Salton, J. Allan, C. Buckley and A. Singhal, "Automatic analysis, theme generation, and summarization of machine-readable texts," *Science,* 264(3), pp. 1421-1426, June 1994.

[10] K. Sparck Jones, "Discourse modelling for automatic summarising," Technical Report 29D, Computer Laboratory, University of Cambridge, 1993.

[11] L.C. Tong and S.L. Tan, "A statistical approach to automatic text extraction," *Asian Library Journal,* 3(1), pp. 46-54, Mar 1993.

[12] M.M. Withgott, D.P. Huttenlocher, S.C. Bagley, P.-K. Halvorsen, D.S. Bloomberg, R. M. Kaplan, T. A. Cass, R. R. Rao, "Method and apparatus for document processing," European Patent Application 0 544 432 A2.

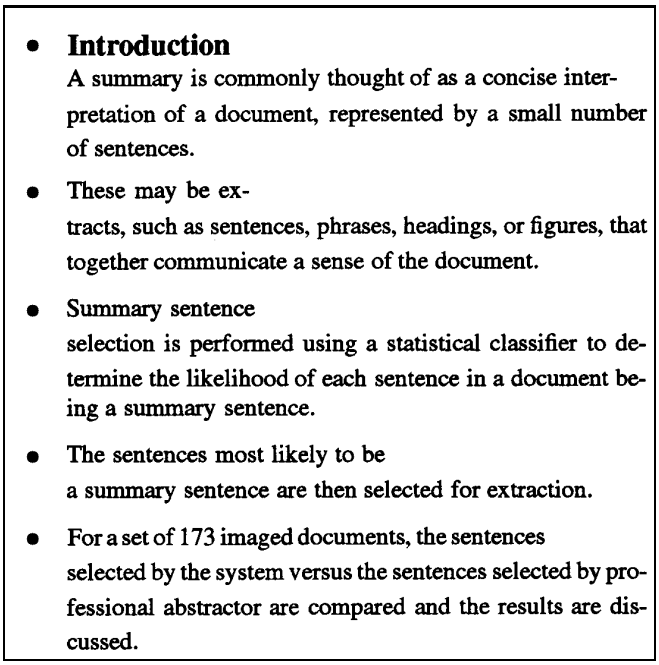

Figure 3: Summary sentences generated by the image-based summarizer. A version of this paper was printed without this figure and accompanying text, and the Abstract was blanked out. Then the paper was scanned and the image summarizer run on the scanned images.