Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 255-262

Surrogater: A Simple Yet Efficient Document Condensation System

*Joe Zhou**
LEXIS-NEXIS, A Division of Reed Elsevier

This paper describes Surrogater, a simple yet efficient document condensation system that is applicable for commercial use. The system consists of two components, a preprocessing component for automatic generation of key terms for a predefined topic, and a condensation component that produces the condensed versions of on-line documents. To evaluate its performance, we compared Surrogater with five other summarization technologies, including Searchable LEAD, a commercial product. Twenty topics across four domains were evaluated, totalling 30 documents and 1800 summaries. A two-way ANOVA test suggested that Surrogater performed at least equally well, if not better, compared to other commercial or nearly commercial products. More significantly, the empirical comparison did not show any dramatic differences in performance between Surrogater and Searchable LEAD as reported in an earlier study (Brandow et al. 1995)

1. INTRODUCTION

In the field of information retrieval (IR), megabytes or terabytes of text have become commonplace. This is particularly true for commercial systems. Taking LEXIS-NEXIS, one of the leading providers of electronic research services, as an example, the entire services, at the time of writing, consists of more than 11,000 sources: 7,100 news and business sources, and 4,800 legal sources. There are more than 6,900 databases and over 1,100 billion characters on-line. Each week more than 9.5 million documents are added to the more than one billion documents on-line. To efficiently handle such gigantic text sources, traditional information storage and delivery mechanisms are constantly being challenged. While users are happy having at their finger tips the information they need to do their work, they feel frustrated to be overwhelmed by all sorts of documents, journals, reports, and electronic mail messages. There is an increasing demand that requested information be presented in a clear, succinct and systematic way. For example, it would be helpful if the user could be provided with ability of examining the summary or the condensed version of a document first, then be guided to the full document if he or she chooses to. It would be equally helpful if search and retrieval could be carried out on the condensed form of databases so that the retrieved information could be made more focused on the original need. Undoubtedly, these on-line as well as off-line system capabilities can help alleviating the current information overload.

Up until now two major approaches have been attempted to automatically generate summaries or condensed versions of on-line documents. One is text-based and the other is knowledge-based (Spark Jones and Endres-Niggemeyer 1995). The text-based approach is primarily statistical in nature, either relying on simple heuristics (e.g. Brandow et al. 1995 and Zechner 1996) or on the combination of various heuristics (e.g. Kupiec et al. 1995 and Edmundson 1964). The goal is to identify key terms in the document, assigning weights to individual sentences, and assembling top-scored sentences as the condensed version of the document. According to some empirical studies (e.g. Morris et al.1992), the optimal length of the condensed version is somewhere between 20% and 30% of the original document length. The knowledge-based approach, on the other hand, involves symbolic analysis to extract significant pieces of information from the document (e.g. McKeown and Radev 1995, Ciravegna 1995 and Lin 1995). After a set predefined templates are filled or partially filled, a summary will then be produced via a natural language generation mechanism. Though the text-based approach reveals its own limitations, such as no guarantee for discourse coherence, the general opinion seems to support its capability for real-world applications. The knowledge-based approach

* LEXIS-NEXIS, A Division of Reed Elsevier, Inc., 9555 Springboro Pike, Miamisburg, OH 45342-4400, USA. Email: joez@lexis-nexis.com

can be clean and effective in certain small and limited environments, but since it encompasses both interpretation and generation of natural language, its scalability and generalizability are often questioned. Research performed in this area has been heavily driven by readily available texts, e.g. MUC (Message Understanding Conference) data, and/or restricted to specific topic domains, again those defined by MUC. If needs arise for applications in general domains, a large amount of training and modification would be required. At the present time, such an approach can hardly fit in a commercial environment such as the one at LEXIS-NEXIS.

In this paper we introduce Surrogater, a simple yet efficient document condensation system that is applicable for commercial purposes. Though it follows the general trend of the text-based strategy, i.e. performing automatic document condensation through sentence scoring and selection, Surrogater distinguishes itself from its predecessors in two ways. First, significant terms, ranging from single word terms to four-word terms, are automatically generated from a small and homogeneous dataset which represents a specific topic. Second, when Surrogater produces document summaries, the identified key terms are utilized exactly as they appear in the naturally running text. No term normalization or regulation is needed. Term clustering or grouping is not pursued. The primary goal of such a simple implementation is to achieve maximum efficiency, generalizability and flexibility that the fast-growing information commerce demands.

2. SYSTEM DESCRIPTION

The Surrogater system consists of two components: a preprocessing component for the automatic construction of key terms and a document condensation component that produces document summaries. Section 2.1 offers a description of the unique technology we have developed for generating a set of meaningful terms. Section 2.2 gives a detailed account of the algorithm that uses these key terms for document condensation.

2.1 Automatic Generation of Key Terms

Automatically identifying meaningful terms from naturally running texts has been an important task for information technologist. It is widely believed that a set of good terms can be used to express the content of the document. By capturing a set of good terms, for example, relevant documents can be searched and retrieved from a large document collection. Though what constitutes a good term still remains to be answered, we know that a good term can be a word stem, a single word, a multiple word term (a phrase), or simply a syntactic unit.

Various existing and workable term extraction tools are either statistically driven, or linguistically oriented, or some hybrid of the two. They all target frequently co-occurring words in a running text. The earlier work of Choueka (1988) proposed a pure frequency approach in which only quantitative selection criteria were established and applied. Church and Hanks (1990) introduced a statistical measurement called mutual information for extracting strongly associated or collocated words. Tools like Xtract (Smadja 1993) were based on the work of Church and others, but made a step forward by incorporating various statistical measurements like z-score and variance of distribution, as well as shallow linguistic techniques like part-of-speech tagging and lemmatization of input data and partial parsing of raw output. Exemplary linguistic approaches can be found in the work by Strzalkowsky (1994) where a so-called fast and accurate syntactic parser is the prerequisite for the selection of significant phrasal terms.

Different applications aim at different types of key terms. To generate key terms for the Surrogater system, we have adopted a "learn data from data" approach. The novelty of this approach lies in the automatic comparison of two sample datasets, a topic-focused dataset based on a predefined topic and a larger and more general base dataset. The focused dataset is created by a domain expert either through a submission of an on-line search or through a compilation of documents from specific sources. The construction of the corresponding base dataset is performed by pulling documents out of a number of sources, such as news wires, newspapers, magazines and legal databases. The intention is to make the resulted corpora cover a much greater variety of topics or domain subjects than the focused dataset.

To identify interesting word patterns in both samples, a set of statistical measures are applied. The identification of single word terms is based on the variation of a t-test. Two-word terms are captured through the computation of mutual information (Church et al. 1991), and an extension of mutual information assists in extracting three-word and four-word terms. Once the significant terms of these

four types are identified, a comparison algorithm is applied to differentiate terms across the two samples. If significant changes in the values of certain statistical variables are detected, associated terms are selected from the focused sample and included in the final generated lists. (For a complete description of the algorithm and preliminary experiments, please refer to Zhou and Dapkus 1995.)

2.2 Automatic Condensation of Document

The design goal for this component of Surrogater was to use as simple an algorithm as possible, avoiding sophisticated or advanced linguistic features employed by other summarization algorithms in the literature. Implemented in the Unix platform using C, this component performs a series of tasks to generate various lengths of condensed versions of input documents.

First, this component uses the significant terms the preprocessing component has generated out of a specific focused file. The focused file is prepared by the domain expert in advance based on a predefined topic. Its size is small, normally about 20 documents. But, there is one prerequisite. At least half of the documents in the focused file must be relevant to the predefined topic. As described in Section 2.1, the terms generated are four kinds: single word terms, two-word terms, three-word terms, and four-word terms. They, as a whole, are supposed to represent the content of the predefined topic. (For the simplicity of explanation, we will hereafter refer this component simply as Surrogater.)

To prepare the input document collection for Surrogater, sentential boundaries in each document are programmatically detected and marked. We use a simple and efficient sentence boundary algorithm to perform this task (Humphrey and Zhou 1989). While processing the input dataset, Surrogater scans each document and builds a term frequency and weight matrix across all sentences using the key terms generated by the preprocessing component.

Since there is no feature such as term clustering or word stemming associated with the generated key terms, the term frequency count and weight assignment are determined purely by taking the term as it is. We find that this is not only computationally inexpensive but also quite effective. For example, when handling the predefined topic Medical Malpractice, the generated term lists may include "medical", "medical malpractice", "medical malpractice action", and "medical malpractice actions". Here, the first term is embedded in the second term and the second in the third. They are nested collocations or phrasal terms, i.e. those being part of longer collocations (Frantzi and Ananiadou 1996). Without the additional step of normalizing these variants, our algorithm would assign more weight to the sentence that contains "medical malpractice action(s)" than the sentence that only has "medical" in it because the formal nested phrase is counted three times, whereas the latter non-nested term only receives one count. Conceptually, the multi-word "medical malpractice action(s)" is unambiguous and highly relevant to the topic in comparison to the single term "medical." The same is true for the topic Campaign. The single word "dole" can refer to any person whose last name is "dole," but the multi-word term "majority leader senator dole" has to be the "dole" who campaigned for the presidency.

After the sentence weights are determined for all the sentences in the document, the sentences are sorted according to the weights they have received. The condensed version of the document can then be produced by extracting and assembling a certain number of the highest scored sentences. It is important to note that the produced condensed form would always include the first sentence of the document regardless of its weight, especially for a general news story (Bradow et al. 1995). The extracted sentences mimic the original text order so that maximum possible coherence is maintained. Surrogater keeps the word counts for each document, as well as for all the sentences in it; therefore, the produced condensation can be flexible in terms of its lengths in proportion to the total size of the corresponding document. Naturally, one cannot expect such a simple technique to resolve complex problems such as anaphora. Co-referential linking might be lost when the extracted sentences contain pronouns such as "he" and "us," and word compounds such as "the company" and "that person".

3. EXPERIMENT

3.1 Experimental Environment

In order to be able to evaluate the quality of the condensed documents produced by our simple and straightforward algorithm, we conducted an empirical experiment by comparing the results of Surrogater with five other text summarization technologies. Among these five technologies, two are

from LEXIS-NEXIS. Search LEAD, a commercial product for LEXIS-NEXIS services, takes the beginning segment of a document as its extract, where the amount of text included in the extract is based on overall document length. The other technology, still in its prototype stage, statistically selects and ranks key terms in a given document and then extracts the sentences that contain these terms. The other three technologies are external to LEXIS-NEXIS with two already in commercial use and one in near production mode. Contract restrictions prevent us from revealing the names and technical details of these three external products. However, there is one feature that makes Surrogater different from the other five participants. Surrogater, as described above, uses statistical data (i.e. a set of key terms) derived from a group of documents for a predefined topic, whereas Searchable LEAD does not use word statistics and the other four technologies obtain word statistics at the document level rather than the corpus level. In addition, Surrogater is primarily statistically oriented with some very shallow language heuristics such as sentence boundary detection. Among the three external technologies, two resort to deep-level NLP techniques. As a result, their processing of large volume of text is relatively slow.

3.2 Experimental Data

The test data for our comparison experiment consists of 15 pre-determined topics in four groups. Four groups represent four subject domains - Company, Tax, Environment, and Insurance. For each topic, there are ten relevant documents and ten irrelevant documents. Both these pre-determined topics and their associated documents were prepared by a domain expert who was not among the human evaluators in the later stage. The specific topics in each of the four groups are illustrated in Figure 1.

Group 1: Company

01: Amgen Inc

04: Netscape Communication

05: Philip Morris

18: Columbia/HCA healthcare

Group 2: Tax

07: Capital gains tax

08: Property taxes

09: Tax cuts

10: Tax evasion

Group 3: Environment

13: Landfills & dumps

14: Radioactive wastes

16: Superfund

Group 4: Insurance

21: Malpractice liability insurance

22: Life insurance fraud

24: Automobile insurance

30: Workers' compensation

Figure 1: Experimental Data: 15 Topics in 4 Groups
(Note: the first 2 digits represent the topic ID, assigned as such to ensure randomness)

3.3 Experimental Procedure and Criteria

All six participating technologies processed the same test data individually to produce summaries or condensed versions of input documents. The result totaled 1,800 summaries, i.e., 6 participants x 15 topics x 20 documents. The size of a summary was required to be about 25% of the original document.

To make sure that the experimental design would provide valid and objective information, six domain experts were assigned to evaluate the results. The 1,800 summaries were presented to them in random sequence and in one chunk of text. Two error measurements were used for the manual judgment. A "miss" means that the domain expert believes the summary of a relevant document is irrelevant. A "false positive" means that the domain expert believes the summary of an irrelevant document is relevant. For the empirical study, our experimental hypothesis is that there are no differences among the six summarization technologies.

4. RESULTS AND EVALUATION

After the six domain experts completed the manual evaluation, we performed a two-way ANOVA on the six technologies against the four topic groups. Table 1 reports the "miss" errors. For the clarity of description, the other five technologies are named respectively as Internal-1 (i.e.Searchable LEAD), Internal-2, External-1, External-2 and External-3.

Table 1: Grouping and ranking in terms of mean "miss" errors

Grouping	mean	Technology
A	3.4222	External-3
A B	2.8000	External-1
A B	2.7000	External-2
B	2.4111	Internal-1
B	2.4000	Internal-2
B	1.9556	Surrogater

As indicated, the range of mean is from 1.9555 (Surrogater) to 3.4222 (External-3). The pairing test suggests that Surrogater, together with Internal-1 and Internal-2, made significantly less "miss" errors than External-3.

In terms of "false positive" errors, however, the 2-way statistical test did not detect any substantial difference among the six technologies (see Table 2). The mean difference, ranging from 0.9222 to 1.3000, is too small to be significant.

Table 2: Grouping and ranking in terms of mean "false positive" errors

Grouping	mean	Technology
A	1.3000	Surrogater
A	1.1222	Internal-1
A	1.0667	Internal-2
A	1.0111	External-1
A	0.9889	External-2
A	0.9222	External-3

It is interesting to note that Surrogater exchanges its mean ranking position with External-3. In other words, it seems that Surrogater made the least "miss" errors and the most "false positive" errors, while External-3 made the most "miss" errors and the least "false positive" errors. But, since the range of the mean "false positive" errors is so small (0.9222 to 1.3000), the distinction becomes negligible. All these pieces of data indicate that it is difficult to draw any conclusion regarding one technology as being better or worse than the others. Our original assumption that there are no statistical differences among the six technologies is supported.

Statistically however, it is clear that Surrogater performs at least equally well, if not better, compared to other commercial or nearly commercial products. Although some other factors have to be taken into account, such as processing speed, the ease of maintenance, etc. when one is forced to select, two unique features would make Surrogater stand out. First, it embodies a fairly simple yet robust condensation algorithm. No deep linguistic analysis which is normally slow and computationally expensive is required. Second, Surrogater derives statistically significant data from a small batch of documents, unlike other summarization technologies that process one document at a time.

Another observation well worth noticing is that our comparison experiment has not shown dramatic difference in performance between Surrogater and Searchable LEAD (Internal-1). When evaluating their statistical/heuristic summarization system called ANES, Brandow et al. (1995) conducted a careful study comparing ANES output to comparable amounts of leading text. They reported that "the lead-based summaries outperformed the intelligent summaries significantly, achieving

acceptability ratings of over 90%, compared to 74.4%" (Brandow et al. 1995). In our comparison, however, such a big gap was not observed. In fact, Surrogater made slightly less "miss" errors than Searchable LEAD (Internal-1) and slightly more "false positive" errors, but the differences were by no means significant. One would assume that assembling relevant sentences from different locations of the document night reveal more about the content of the document than simply taking its beginning text.

There may be a number of reasons for the performance difference between Surrogater and ANES (Brandow et al. 1995) in contract with different leading text. First, thanks to the lessons drawn by ANES developers, Surrogater always includes the first sentence of the document in its condensed format regardless its assigned weight. (But, note that "the news genre by its very nature is arguably biased toward lead-based summarization, whereas other genres might be less amenable to such a technique" (Brandow et al. 1995). As we have already been aware, for example, legal document does not follow such a "bias".) Second, evaluation criteria for ANES appear to be more stringent than Surrogater. When comparing with different leading text, both content as well as readability were used as evaluation guidelines in the studies by Brandow et al. (1995), whereas our experiment mainly used relevance as the sole criterion. Leading text extracts are generally judged as good summaries. Brandow et al. (1995) reported that ANES extracts generally had more content but were judged to be less readable than leading text extracts. In a relevance judgment task such as the one reported here, Surrogater's more complete content apparently offsets Searchable LEAD's readability advantage. Lastly, the key terms, ranging from single words to four-word terms, generated by the Surrogater system, as a whole, may be of higher quality or more content oriented than the individual "signature" words generated by ANES. Phrasal terms are less ambiguous than individual words.

5. CONCLUSION

In this paper we have introduced Surrogater, a simple yet efficient document condensation system. It consists of two main components, a preprocessing component for automatic generation of key terms from naturally running texts, and a condensation component that produces document summaries. The methodology we have adopted is innovative, flexible, and most importantly suitable for large-scale commercial environment. Primarily statistical in nature, the key term component learns to select candidate terms through a meaningful comparison of a focused sample with a large and diverse base sample. The terms generated are restricted to one- to four-word terms based on our empirical observation that terms within such an range are of better value for IR applications. The resulting term lists may contain lexical, syntactic or proximity-based terms, as well as their variants. We made no attempt to differentiate these types or normalize their variants. When applying the acquired lexical resources for on-line document condensation, again a simple yet efficient algorithm was implemented. No advanced natural language processing techniques were utilized except shallow linguistic heuristics, such as sentence boundary detection. The generated terms were used exactly as they appear in raw data. One may argue that these terms, if taken individually, are partial or weak representations of document content. But, if taken as a coherent whole, they are sufficient enough for IR applications. As we have demonstrated, their usage can lead to viable and practical commercial products at the present time when genuine natural language understanding is still so elusive.

ACKNOWLEDGEMENTS

The author would like to express his sincere thanks to Dan Pliske, Johnyoung Lee and Mark Wasson for their helpful comments on the early draft.

REFERENCES

1 R. Brandow, K. Mitze and L. Rao. Automatic condensation of electronic publications by sentence selection. *In Information Processing & Management*, 31(5), pp.675-685, 1995.

2 K. Church and P. Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1), March 1990.

3 K. Church, et al. Using statistics in lexical analysis. *In U. Zernik, editor, Lexical Acquisition: Exploring On-line Resources to Build a Lexicon*, Lawrence Erlbaum Association, 1991.

4 Y. Choueka. Looking for needles in a haystack. *In Proceedings, RIAO, Conference on User-Oriented Context Based Text and Image Handling.* Cambridge, MA. 1988.

5 F. Ciravegna. Understanding messages in a diagnostic domain. *In Information Processing & Management*, 31(5), pp.687-701, 1995.

6 H. Edmundson. New methods in automatics abstracting. *Journal of the ACM*, 16(2):264-285, 1969.

7 K. Frantzi and A Ananiadou. Extracting nested collocations. *In the Proceedings of COLING, pp41-46*, 1996.

8 T. Humphrey and J. Zhou. Period Disambiguation Using a Neural Network. *In the proceedings of International Joint Conference on Neural Network*, 1989.

9 J. Kupiec, J. Pedersen and F. Chen. A Trainable Document Summarizer. *In Proceedings of the 18th ACM-SIGIR Conference*, pp68-73, 1995.

10 C. Lin. Knowledge-based automatic topic identification. *In Proceedings of the 33rd ACL Conference*, pp308-310, 1995.

11 K. Mckeown, and D. Radev. Generating summaries of multiple news articles. *In Proceedings of the 18th ACM-SIGIR Conference*, pp74-79, 1995.

12 Morris, G. Kasper and D. Adams. The effect and limitations of automated text condensing and reading comprehension performance. *Information System Research*, 3(1), pp17-35, 1992.

13 F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), March 1993.

14 K. Sparck Joens and B. Endres-Niggermeyer. Automatic summarization. *In Information Processing & Management*, 31(5), 1995.

15 T. Strzalkowski. Document Indexing and Retrieval Using Natural Language Processing. *In Proceedings, RIAO*, New York, NY. 1994.

16 K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. *COLING*. 1996.

17 J. Zhou and P. Dapkus. Automatic Suggestion of Significant Terms for a Predefined Topic. *In Proceedings of the 3rd Workshop on Very Large Corpora, Association for Computational Linguistics*, MIT, Boston. 1995.

APPENDIX: Sample input and Output for Group 4: *Insurance;* Topic 30: *Workers' compensation.*

A total of 186 Key terms was generated for this topic. The sample key terms are:

31 single word terms: *preexisting, claimant, wcb, pataki, portability, unicar, managed-care, fee-for-service...*

107 two-word terms: *workers compensation, health insurance, group health, health plan, murder rate, experience rating, health plans, health care, cobra continuation, preexisting condition, wc claims...*

40 three-word terms: *group health plan, health and safety, health insurance coverage, group health insurance, cobra continuation coverage, costs and frequency, employer medical expert, workers compensation package...*

8 four-word terms: *occupational health and safety, accident costs and frequency, killed on the job, specializing in workers compensation, health plans and health, occupational safety and health, penalties and special assessments, plan or group health*

Sample input document:

800311

ALBANY -- Legislation reforming New York's workers' compensation system, the cornerstone of Gov. George Pataki's 1996 legislative session, has yet to become law because of a dispute between the governor and the state comptroller.

Because of that disagreement, the Assembly hasn't released the workers' compensation package, approved by the Legislature July 13, for Pataki's signature. Under the state constitution, whichever house of the Legislature passes a bill first controls it when it's sent to the governor for his signature. Meanwhile, a number of provisions that were set to kick in immediately, such as those limiting law suits filed against employers by equipment manufacturers, are on hold.

At the center of the feud is Comptroller H. Carl McCall's insistence that the legislation is unconstitutional. McCall objects to a provision removing his authority to audit payments by the state Insurance Fund before the money is spent. The fund is a public entity that writes many of the workers' compensation insurance policies in the state.

McCall said his agency's past audits of the fund have saved millions by uncovering such mistakes as payments made to dead people. In an audit last year, McCall said the fund failed to do basic investigations to uncover fraud.

"If a state agency makes a payment we have to pre-audit, so that portion of the bill is unconstitutional," McCall said. "We just don't know if it was inadvertent or if they attempted to do this."

McCall said he will insist that his objections be dealt with.

Pataki administration spokesman Michael McKeon, however, dismissed McCall's concerns. "We don't believe there are any constitutional issues," he said, adding that Assembly officials said last week the bill would be sent to the governor in the next few weeks. "We don't think there's an issue here," he added. He would not elaborate.

Assembly Speaker Sheldon Silver, in Chicago for the Democratic Party's national convention, could not be reached for comment Monday. Late last week, however, he acknowledged that McCall's concerns had delayed the sending of the bill to Pataki.

"I believe it will be fixed," is all Silver would say.

Precisely how remains unclear. McCall said he would support Pataki signing the bill so long as he ignores the provision that removes the comptroller's pre-audit powers. In a recent letter from McCall's counsel, Paula Chester, to Pataki counsel Michael Finnegan, the comptroller also insisted Pataki must agree to then amend the package to remove the provision.

In that letter, Chester noted that Finnegan already had acknowledged that Attorney General Dennis Vacco "expressed the view that we have the better side of the argument."

Sample output document (condensed by Surrogater):

800311

ALBANY -- Legislation reforming New York's workers' compensation system, the cornerstone of Gov. George Pataki's 1996 legislative session, has yet to become law because of a dispute between the governor and the state comptroller.

Because of that disagreement, the Assembly hasn't released the workers' compensation package, approved by the Legislature July 13, for Pataki's signature. The fund is a public entity that writes many of the workers' compensation insurance policies in the state. How long Silver is willing to hold onto the bill remains uncertain, because the disagreement has done more than just hold up the workers' compensation package. The workers' compensation package was the most heatedly contested piece of legislation in this year's session. But Ronald Kermani, a spokesman for the New York State Trial Lawyers Association, dismissed the notion of lawyers lining up at court clerk offices trying to file last-minute cases.