December 2003

Evaluation of an Automatic Text Abstraction System

Wendy Wang
San Jose State University

Sumali Conlon
University of Mississippi

Brian Reithel
University of Mississippi

EVALUATION OF AN AUTOMATIC TEXT ABSTRACTION SYSTEM

Wendy Wang
San Jose State University
wang_w@cob.sjsu.edu

Sumali Conlon
University of Mississippi
sconlon@bus.olemiss.edu

Brian Reithel
University of Mississippi
breithel@bus.olemiss.edu

Abstract

Information overload has been a serious problem to scholars, researchers, and students in academia. There is too much to read and too little time. A good abstract would reduce readers reading time by providing gist of the source. By reading the abstract, readers can decide if it is worthwhile to read the original. Therefore, time could be saved. Since most journal articles do not observe guidelines published by the American National Standards Institute (ANSI) on writing abstracts, they need to be modified by professional abstractors. The huge volume of journals has brought pressure on professional abstractors in various bibliographic abstracting services. A text abstraction system was developed to produce abstracts that meet the ANSI standards. To evaluate the system performance, this study compared its output to two other abstracts: the abstract written by the original author and the manual modification of the abstract system output. All three abstracts are derived from the same source paper. The evaluation results indicated that the author abstract differed significantly from the manually edited system output in terms of meeting the ANSI standards. The abstract was also significantly different from the system output both before and after manual editing in fluency. All three were not significantly different in understandability. No significant difference was identified between the system output and its manual edited version in any of the three dependent variables. Therefore, it is possible to use the system output directly without manual modification. This would, subsequently, save time and resources.

Keywords: Automatic abstraction system, decision support; abstraction evaluation

Introduction

There are many resources available to help researchers stay current in their academic areas of interests. Reading research publications is a very popular and effective tool. Increases in the volume of scholarly publications demands the information about the primary literature in a more timely manner. According to the ANSI standards, an abstract is defined as an abbreviated and accurate representation of the content of a document. An informative abstract should contain four components: purpose (why the study was conducted), methodology (how the study was conducted), results (what was found), and conclusions (interpretation and evaluation of the results). As for an indicative abstract, the ANSI standards suggest that it should be as close to an informative abstract as possible (Borko and Chatman 1962; Milas-Bracovi 1987).

If the author writes an abstract in a manner that the access services can reproduce it with little or no change, the life of professional abstractors would be much easier. The reality is that most abstracts of the research publication need to be modified to meet the ANSI standards on writing abstracts. An automatic text abstraction system was developed to improve the productivity of professional abstractors. By providing high quality abstracts, hopefully the system could help researchers to acquire more information in a shorter period of time. Due to the limitation on length of this paper, details on the system features will not be discussed here.

The emphasis of this paper is on the evaluation of system performance. This paper reported the system performance from three perspectives: satisfying ANSI standards, understandability, and fluency. The evaluation was conducted by comparing three different abstracts out of one source paper. These abstracts are the system output, the manually-edited system output, and the abstract written by the original author(s). The goal of this study is to explore the possibility of automatically producing abstracts that meet ANSI standards.

Literature Review

Evaluation is crucial for identifying the weakness of the systems for further improvement. There are many evaluation methods in abstraction system evaluation literature (Brandow et al. 1995; Morris et al. 1992; Rath et al. 1961). These methods can be roughly categorized into intrinsic and extrinsic evaluation. Intrinsic evaluation focuses more on the quality of the abstract itself. It concerns whether the abstract covers the main content of the source (Paice and Jones 1993*)*, or how readable the output is (Minel et al. 1997). The concept of "main content" is quite elusive although seems easy at first. It could mean different things in different contexts for different users (Jones 1999; Barzilay and Elhadad 1997; Shannon 1951). Accordingly, it is not an easy job, if not an impossible job, to establish an ideal summary that can be used as the golden standards for evaluation purpose (Pollock and Zamora 1975; Edmundson 1968; Kupiec et al. 1995).

Compared to intrinsic evaluation, extrinsic evaluation concentrates more on how the abstract facilitates or serves other purposes, for example, whether the abstract improves precision and recall in an information retrieval environment (Brandow et al. 1995). This kind of evaluation encounters difficulties of formulating questions that are close to the real users.

In this study, an online questionnaire is setup to evaluate whether the system has accomplished what it is designed for: to produce fluent, understandable abstracts that meet the ANSI standards.

Research Methodology

Concept Definitions

Three independent variables can influence the quality of abstracts: methods used to produce abstracts, the original source text quality, and the level of difficulty of the abstracts. In this study, however, the impact of only one independent variable, the method by which the abstract is produced, was the target of interest. Based on the interested independent variable, abstracts are categorized into three groups: abstract written by the original author (author abstract), abstract generated by abstraction system (system abstract), and abstract generated by the system and manually edited (hybrid abstract).

Three dependent variables are examined in this study, *fluency* (presentation style), *satisfaction of ANSI standards* (content), and *understandability* (meaning). Each covers one aspect of an abstract. *Fluency* refers to the presentation aspect of an abstract. It concerns issues such as is the text coherent, does it flow naturally, does the text contain dangling anaphors, are there any obvious gaps between sentences, and is structured environment preserved etc (Mani 1999). It is possible that an abstract has a perfect flow of logic and grammar, and misses the main content. *Satisfaction of ANSI standards* focuses more on the content of the abstract. The *ANSI standards* suggest that an abstract should be as informative as it is permitted by the type and the style of the document. An informative abstract should state the purpose, methods, results, and conclusions presented in the original document, either in that order, or with an initial emphasis on results and conclusions. An indicative abstract should be like an informative abstract as much as possible. The ANSI standards also suggest avoiding background information, or citing the work of others in the abstract, unless the study is a replication or evaluation of their work. *Understandability* focuses on the meaning of the abstract. It is possible that the abstract meets the ANSI standards, has good flow of logic, yet is hard to understand. It might be that the abstract itself is hard to read. For example, given the same idea, some texts are much easier to understand since they contain less complicated vocabulary with shorter sentence.

Hypothesis

The abstraction system was designed to produce abstracts that meet the ANSI standards. To measure if the system has achieved this goal, the following hypothesis was examined:

$H1_0$: There is no difference in terms of meeting the ANSI standards among the author abstract, the computer abstract, and the hybrid abstract.

Humans have a certain level of tolerance for incoherent text due to the world knowledge that they possess. People can derive meanings out of an incoherent sentence from the context. Since the output of the Abstraction Assistant is an extract, not an abstract, it is not as fluent. It would be interesting to examine whether the problem of incoherence in the output of Abstraction Assistant is so serious that readers would notice. Therefore, the following hypothesis was proposed:

$H2_0$: There is no difference in terms of the fluency of the text among the author abstract, the computer abstract, and the hybrid abstract.

Finally, the understandability of these three abstracts was examined. Understandability focuses more on meaning, and fluency, more on the presentation aspect of texts. It is quite possible that a not so fluent text is understandable and a perfectly smooth text is hard to understand. To examine the understandability of these three kinds of abstracts, the following hypothesis was investigated:

$H3_0$: There is no difference in terms of the text understandability among the author abstract, the computer abstract, and the hybrid abstract.

Research Design

One pilot study and three major pre-tests have been carried out to ensure the quality of the questionnaire designed for evaluation. In the pilot study, six graduate students and faculty members participated. The pilot study aimed to assess whether the question items were clearly written and whether each item was measuring what it was designed to measure. Changes were made according to the feedback.

Next, the questionnaire was put online through a school server hosted by the Mississippi Center for Supercomputing Research (MCSR). Three pre-tests were conducted with the participation of second year master level students, Ph.D students, faculty members, and information system professionals such as web developers and web masters in the University of Mississippi. Altogether 20 graduate students and faculty members participated the first two pretests. The first two pre-tests provided information on how long it took to finish the questionnaire etc. More changes were made to the wording and the design of the questionnaire. The 3rd pre-test lasted over two weeks. Altogether 48 subjects participated. The ∂ score was 0.699 for ANSI, 0.78 for fluency, and 0.85 for understandability. After these pretests, the questionnaire was considered reliable and ready to be used (see attachment A for questionnaire).

Data Collection Procedure

Due to the characteristics of the evaluation task, this study needs subjects who are accustomed to reading journal articles. ISWORLD is a mailing list that serves the international information system research community. Currently there are over 2000 subscribers. Most subscribers are Ph.D. students, faculty members, and information system practitioners. The characteristics of ISWORLD subscribers satisfy the participation requirements for this study.

To recruit participants for this study, two study participation request emails were sent to ISWORLD one week apart from each other. The emails to ISWORLD stated the purpose of this study and provided the address of evaluation web site. Interested readers were linked to the evaluation web site by an address provided at the end of the emails. The evaluation website randomly assign participants to read one version of abstracts. In the end, participants were required to vote whether the abstract was written by the original author, produced by a computer program or was produced by a computer program and edited manually. Only one answer was allowed. After readers filled in all the required questions, their input were saved to a text file, and the data was ready for further data analysis.

Evaluation Result

One hundred and fifty six responses were collected shortly after two request for participation emails were sent to ISWORLD. Most of the responses came back within two days after both emails were sent. The response rate was about 7.5%. 30% of the participants were randomly assigned to read the author abstract by the system, about 40% read the computer abstract, and 30% read the hybrid version. As for the composition of the participants, 48% of the subjects identified themselves as Ph.D. students, 28% as junior faculty, 17% as senior faculty, about 6% as Information System researchers, and about 1% as Information System practitioners. More than 70% stated that they were familiar with the area that this paper covered, about 4% reported that they were not familiar with the area at all.

Before the study, it was predicted that most people would identify the computer abstract correctly since its sentences might not be as fluent. The result showed that overall, 36% of the participants identified correctly the kind of abstracts they read. Among these participants, more than 50% identified correctly they read an author abstract, 44% identified that they read an hybrid abstract, only 19% voted correctly that they read a computer abstract. The low correct identification rate of the computer abstract suggested that it was difficult for readers to distinguish between the abstract written manually and the one produced automatically. This finding suggested that the computer abstract could be used directly without the need for manual modification.

The reliability and validity tests showed that the alpha value for ANSI conformity was 0.61, 0.79 for fluency and 0.66 for understandability. Once the requirements for measurement reliability and discriminant validity were reasonably established (Nunnally 1967), further analysis were conducted. Three variables were created to represent the three dependent variables: *ansi*, *fluency*, and *understandability*.

To test if there were differences in the ANSI, fluency, and understandability among these three groups, the Multivariate Analysis of Variance (MANOVA) was run. The MANOVA test indicated that there was a significant difference among these three groups (see table 1).

MANOVA only indicated whether there were significant differences among the groups in terms of the vectors of dependent variables, but did not indicate which dependent variable(s) created this difference. To test hypothesis one: *There is no difference in terms of meeting the ANSI standards among an author abstract, a computer abstract, and an hybrid abstract*, ANOVA was run. The result indicated that there was a significant difference in terms of meeting the ANSI standards ($F_2 = 3.611$, $P = 0.029$). Scheffe's test was conducted to identify which group or groups was/were different from others.

No significant difference was identified on meeting the ANSI standards between the author abstract and the computer abstract, yet there was a significant difference between the author abstract and the hybrid abstract. No significant difference was discovered between the hybrid abstract and the computer abstract. So the hypothesis: *There is no difference in terms of meeting the ANSI standards among an author abstract, a computer abstract, and an hybrid abstract,* was partially supported.

ANOVA test was run to test hypothesis two. The result indicated that there was a significant difference among the author abstract, the computer abstract, and the hybrid abstract ($F_2 = 9.212$, $P < 0.000$). Scheffe's test showed that there was a significant difference between the author abstract and the other two kinds of abstracts. No difference was identified between the computer abstract and the hybrid abstract on fluency. So the second hypothesis: *There is no difference in terms of the fluency of the text among an author abstract, a computer abstract, and an hybrid abstract,* was partially supported.

The ANOVA test was used again to uncover whether differences among the abstracts were caused by difference in understandability. The result did not indicate any significant difference among these three groups in term of abstract understandability. So the third working hypothesis: *There is no difference in terms of the text understandability among the author abstract, the computer abstract, and hybrid abstract,* was supported.

To compare the performance of the three abstracts on individual question items, more descriptive analysis was carried out. Results showed that the computer abstract was perceived to contain more redundant sentences than the author abstract, and the hybrid abstract contained more redundant sentences than the computer abstract. The hybrid abstract was considered to be more likely to contain the methodology component than the computer abstract. The computer abstract was considered more likely to contain the methodology component than the author abstract. All three abstracts were likely to include the conclusion component.

Discussion

The evaluation result is quite inconclusive. More studies need to be done before any generalized comments can be made. *First of all*, the result show that in meeting the *ANSI* standards, the hybrid abstract performed better than the author abstract. Before claiming that the system has accomplished what it is designed for, three abstracts were examined closely. First of all, table 2 showed that the hybrid abstract contained 267 words whereas the author abstract had 107. Since the more words an abstract contained, the more information it could provide. It was possible that the difference in the number of words that the abstract contained, not how it was produced, created the difference in meeting the ANSI standards between the author abstract and the hybrid abstract. The fact that the computer abstract contained 254 words (quite close to the number of words that the hybrid abstract had), might explain why no significant difference was identified between the computer abstract and the hybrid abstract in terms of meeting ANSI standards.

Table 1. MANOVA

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	0.961	1253.12[a]	3.000	151.000	.000
	Wilks' Lambda	0.039	1253.12[a]	3.000	151.000	.000
	Hotelling's Trace	24.897	1253.12[a]	3.000	151.000	.000
	Roy's Largest Root	24.897	1253.12[a]	3.000	151.000	.000
TYPE	Pillai's Trace	0.325	9.824	6.000	304.000	.000
	Wilk's Lambda	0.681	10.643 [a]	6.000	302.000	.000
	Hotelling's Trace	0.459	11.463	6.000	300.000	.000
	Roy's Largest Root	0.438	22.182[b]	3.000	152.000	.000

[a]Exact statistic
[b]The statistic is an upper bound on F that yields a lower bound on the significance level.
[c]Design: Intercept + TYPE

Table 2. Length Comparison of Three Abstracts

Type	Paragraph Number	Words	Number of Sentences
Author abstract	1	107	5
Computer abstract	2	254	10
Hybrid abstract	3	267	9

Secondly, on fluency, the evaluation result indicated that the author abstract performed better than both the computer and the hybrid abstract. Since the computer abstract was produced by a program, it was not surprising that the author abstract was perceived as more fluent. Yet it was not clear why the author abstract was considered to be more fluent than the hybrid abstract given that both abstracts were smoothed manually. It was possible that the fluency of the abstract might relate to its length and the number of paragraphs the abstract had. Table 2 indicated that the author abstract had one paragraph, the computer abstract had two, and the hybrid abstract had three. It could be that the more paragraphs the abstract has, the less fluent it is considered to be. The hybrid abstract had fewer sentences than the computer abstract did although it had three paragraphs, and the computer abstract had two. All these could be interpreted that the longer the abstracts were, the less fluent the abstracts were perceived to be. The reason that the author abstract was considered to be more fluent might be because the author abstract was shorter, making it easier for readers to finish reading. Therefore it is considered to be more fluent than the ones which took longer to read.

One interesting thing to notice about abstract fluency was that, although the author abstract was rated more fluent than the computer abstract, only 19% of the subjects voted correctly that what they read was produced by a computer program. This could be that since the final output of the Abstraction Assistant combined the author abstract and sentences extracted from the main text body, although sentences in later part of the output were not as coherent, readers simply failed to notice that.

Thirdly, in terms of understandability of the abstracts, no significant difference was identified among the three groups. No significant difference was found between the computer abstract and the hybrid abstract in all aspects. The mean scores on understandability for these three abstracts were quite close to each other. It implied that three abstracts were equally

understandable to the readers. Before making this claim, when examining the characteristics of the participants, we found out that over 70% of the subjects reported that they were familiar with the specific area that this abstract covered. This suggests that maybe a balanced population that has various backgrounds could be used for testing in the future studies.

Since the statistical analysis did not identify any differences in all three dependent variables between the computer and the hybrid abstract, it implied that it was possible to use the computer abstract directly without the need for manual modification. Therefore, time and resources could be saved. If that is the case, it is good news for professional abstractors: Abstraction Assistant could be used as a tool for professional abstractors to improve their productivity.

Future Work

This study leaves many issues that need to be addressed in the future.

First of all, none of the alpha values for the three factors is over 0.8. It suggests that more refinement and fine-tuning of the measurements need to be done.

Secondly, since most of the subjects are information systems researchers who are familiar with the specific area covered in the abstract used in this study, a population with subjects having a more diversified background needs to be used to generalize the findings.

Thirdly, since the hybrid abstract contains more words and sentences than the other two, and more words could mean more information, it is possible that the number of words might confound the result. In the future studies, however, a system that produces a similar number of words as the author abstract could be developed for evaluation purpose. Similarly, the length of the abstract could play a role upon examining its fluency. Statistical tests show that the author's original abstract is more fluent than both the computer and hybrid abstracts, while the length of the author abstract is shortest. Therefore, it is possible that the length of the author's abstract affecting the statistical result.

To summarize, with increasing availability of online articles, the demand for an automated production of abstracts is getting greater. Providing high quality abstracts at a very low cost is becoming increasingly important to help the research community as a whole. This study gained valuable information for further research in this area.

References

American National Standards Institute, Inc., New York, NY. (*1979)* ANSI Z39.14.

Barzilay, R., and Elhadad, M. "Using Lexical Chains for Text Summarization," ACL/EACL-97 summarization workshop. pp.10-17.

Borko, H., and Chatman, S. "Criteria for Acceptable Abstracts: A Survey of Abstracters' Instructions," American Documentation. April 1962.

Brandow, R.; Mitze, K.; and Rau, L. F. (1995) Automatic Condensation of Electronic Publications by Sentence Selection, Information Processing and Management (31:5), 1995, pp. 675-685.

Edmundson, H.P. (1968) "New Methods in Automatic Extracting," Journal of the Association for Computing Machinery (16:2), pp. 264-285.

Jones, K.S. Automatic Summarizing: Factors and Directions. "Advances in Automatic Text Summarization," Mani and Maybury, MIT Press, Cambridge, Massachusetts, London, England, 1999, pp. 1-12.

Kupiec, J.; Pedersen, J.; and Chen, F. "A Trainable Document Summarizer," In Proceedings of ACM-SIGIR'95, Seattle, WA. pp. 68-73.

Mani, I. (1999) *Automatic Summarization*. John Benjamins Publishing Company. Amsterdam/Philadelphia. 70.

Milas-Bracovi, M. "The structure of Scientific Papers and Their Author Abstracts," Informatologia Yugoslavica, (19: 1-2), 1987, pp. 51-67.

Minel, J., Nugier, S., and Piat, G. "How to Appreciate the Quality of Automatic Text Summarization," In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, July 1997, pp. 25-30.

Morris, A. H.; Kasper, G. M.; and Adams, K. A. "The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance," Information Systems Research (3:1), pp. 17-35.

Nunnally, J. C. Psychometric Theory, New York: McGraw-Hill. 1967.

National Information Standards Organization, Guidelines for abstracts: An American National Standard, 1997.

Paice, C., and Jones, P. "The Identification of Important Concepts in Highly Structured Technical Papers," In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM-SIGIR'93), 1993, pp. 69-78.

Pollock, J.J., and Zamora, A, "Automatic Abstracting Research at Chemical Abstracts Service," Journal of the American Society for Information Science (46:3), 1975, pp. 225-234.

Rath, G.J.; Resnick, A.; and Savage, T. R. (1961) "The Formation of Abstracts by the Selection of Sentences," Part 1. Sentence Selection by Men and Machines. Part 2. The Reliability of People in Selecting Sentences. American Documentation, (12:2), 1961. Reprinted in *Advances in Automatic Text Summarization*, I. Mani and M.T. MayBury(eds.), 287-292. Cambridge, Massachusetts: MIT Press.

Shannon, C.E. "Prediction and Entropy of Printed English," Bell System Technical Journal (30), 1951, pp. 50-64.

Appendix A. Evaluation Questionnaire

Section A

(1) You are a
- a) Ph.D. student
- b) junior faculty
- c) senior faculty
- d) IS researcher
- e) IS practitioner
- f) Other (please specify) _____

2) You are a
- a) Male
- b) Female

3) How familiar are you with the topic discussed in this abstract?

Very familiar						Not familiar
7	6	5	4	3	2	1

Section B

		Strongly agree						Strongly disagree
1)	The purpose of the research project is stated in this abstract.	7	6	5	4	3	2	1
2)	This astract flows very well from one sentence to the next.	7	6	5	4	3	2	1
3)	This abstract adequately summarizes the study.	7	6	5	4	3	2	1
4)	This abstract does not contain redundant sentences.	7	6	5	4	3	2	1
5)	The methodology and techniques used in thie research project are stated in this abstract.	7	6	5	4	3	2	1
6)	This abstract contains pronouns – or similar elements – that cannot be understood due to the lack of proper nouns or other references.	7	6	5	4	3	2	1
7)	The research implications of the results in the study are stated in this abstract.	7	6	5	4	3	2	1
8)	This abstract is informative.	7	6	5	4	3	2	1
9)	The findings (results) of the study are stated in this abstract.	7	6	5	4	3	2	1
10)	This abstract is well organized.	7	6	5	4	3	2	1
11)	This abstract is coherent.	7	6	5	4	3	2	1

12)	All of the sentences in this abstract are complete	7	6	5	4	3	2	1
13)	This abstract does a good job of explaining the purpose and nature of the study.	7	6	5	4	3	2	1
14)	After reading this abstract, I could easily describe to someone else what the study was all about.	7	6	5	4	3	2	1

Section C (plese indicate your judgment on who wrote this abstract)

1)	The abstract you just read was produced manually by the original author of the article.	7	6	5	4	3	2	1
2)	The abstract you just read was produced by a computer program.	7	6	5	4	3	2	1
3)	The abstract you just read was produced by a computer program and then edited by a human.	7	6	5	4	3	2	1

4) This abstract was produced by (pick one)
   a) the original author of the article.
   b) a computer program.
   c) a computer program and then edited by a human.