LANGUAGE IN INDIA
Strength for Today and Bright Hope for Tomorrow
Volume 12 : 5 May 2012
ISSN 1930-2940

Managing Editor: M. S. Thirumalai, Ph.D.
Editors: B. Mallikarjun, Ph.D.
Sam Mohanlal, Ph.D.
B. A. Sharada, Ph.D.
A. R. Fatihi, Ph.D.
Lakhan Gusain, Ph.D.
Jennifer Marie Bayer, Ph.D.
S. M. Ravichandran, Ph.D.
G. Baskaran, Ph.D.
L. Ramamoorthy, Ph.D.
Assistant Managing Editor: Swarna Thirumalai, M.A.

Creating Summaries with an Automated Tool

Renu Gupta, Ph.D.

==

Abstract

Students and researchers are frequently required to write summaries, either during examinations or as part of their ongoing work. This paper describes some features of a summary, derived from work in Natural Language Processing, briefly examines the AutoSummarize tool in Microsoft Word, and then proposes that such tools can be used in teaching.

1. Introduction

I've always been puzzled by the activity called summary writing. In school, I wrote summaries for tests and examinations without knowing why or how to write one. As a teacher, I once inflicted summary writing on my students in a writing class merely because other teachers were doing so. And as an examiner, I found that evaluators could never agree on what we were looking for on the summary/précis item.

This lack of clarity is curious because summaries are used extensively in everyday life as well as in professional work and at the university. In everyday conversations, we relate the plot of a movie or the gist of a conversation without describing every digression, ‘umm’ and ‘er’. Students know this intuitively; when writing a laboratory report, no student would launch into a description of the laboratory or the acid that fell on their clothes, because they know that this information is irrelevant in a report. Yet, in language classes, students are taught to write summaries without reference to the purpose; instead, summary writing is taught as an idealized or abstract skill.

Life after school relies heavily on the ability to summarize documents, even if these are not viewed as summary writing. At the university, the conventional closed book time-bound examination requires answers that summarize ‘all you know about X’. In the government and corporate sector, reports have to be summarized for superiors or the general public. Independent research involves summarizing data, reading and summarizing the research literature, and even summarizing one’s own document for the abstract. Like other writing, summaries are driven by their purpose; for example, an entire article can either be summarized in a paragraph or as a single reference in a research paper with “see X (2005)”.

So, it seems that summary writing is an indispensible skill for professional life. The volume of information on the Internet has exacerbated the problem; students and junior researchers download hundreds of pages from the Internet and then cannot find their way through it.

Despite the importance of this skill, there seems to be little clarity on the features of a summary and how to teach summary writing. One teaching technique is to give students a text and instruct them to summarize it in a given number of words (usually one-third of the original). One student joked that he was told to copy every third sentence from the original to create a summary. There is a fuzzy notion that a summary contains only the main points of a text, and omits the details. But what is the difference between main points and details, and how does one find them?

Although written discourse consists of sentences that are written and read linearly, the ideas are organized in a hierarchy, with the most general information at the top level subsuming details below (Hinds, 1979). Although this is primarily true of expository texts, narratives too have a structure (or story grammar, from Propp, 1928/1968) as do argumentative texts. Experiments have found that after a few days, readers remember only the top-level information, which would be the main points, and cannot recall information at lower levels of the hierarchy, namely, the specific details (Meyer and Freedle, 1984; Meyer and Rice, 1984). It is this logic that underlies textbook advice on writing summaries: "Read the original, put it aside, and write down what you remember". This works as a procedure, but does not help students identify the features of a summary.

This paper identifies some features of summaries that are given in the literature. It then shows what this means using one tool in Microsoft Word—AutoSummarize—and concludes with suggestions for teaching.

2. Features of a Summary

Clarity about the features of a summary can be found in the literature on Natural Language Processing (NLP). When information on the Internet began growing exponentially in the 1990s, computer scientists recognized the need to search for and retrieve information. To help users extract information and generate summaries, they worked with linguists to formalize the features of summaries and incorporate them in their programs. In the literature, a frequently cited definition is given in Radev (2002; cited in Das and Martins, 2007): summaries can be of one document or of several documents; they should be short; and they should preserve important information. As Das and Martins (2007) point out, "a more elaborate definition for the task would result in disagreement within the community" (p.1), which we see in moderation meetings for examinations.

This definition covers user inputs—the number of documents and the length of the summary. However, it is the final feature, namely, preserve important information, which is the central concern for the student, the writer and researchers working on text summarization techniques. What features of a text help identify 'importance'?

Early work in text summarization identified three features of a summary that still hold good. Note that Features 2 and 3 draw on the concept of text structure.

1. The *frequency* of content words (Luhn, 1958; cited in Lloret, 2008). Discourse consists of connected sentences. One device used to achieve cohesion is the repeated use of lexical items or their synonyms (Hoey, 1991). This approximates the keyword approach. So, sentences that contain the most frequent content words are considered important for a summary.

2. The *position* of sentences in the text. Baxendale (1958; cited in Das and Martins, 2007) found that the topic sentence of a paragraph occurred 85% of the time as the first sentence and 7% of the time the last sentence of the paragraph. This feature is used in text analysis programs such as *Criterion* to automatically score student essays (Burstein, Chodorow, and Leacock, 2003). However, this feature has to be treated with some caution because of differences across texts. First, genre determines the structure of the text; for example, news items are structured with the most important information at the beginning, which automatic summarizers confirmed (Das and Martins, 2007). Second, the disciplinary area affects the structure of texts, with texts in the sciences and social science conforming more to this canonical structure than texts in the humanities. Third, there are regional differences between texts written in English (Biber, 1987; Gupta, 2009; Hall *et al*., 2007).

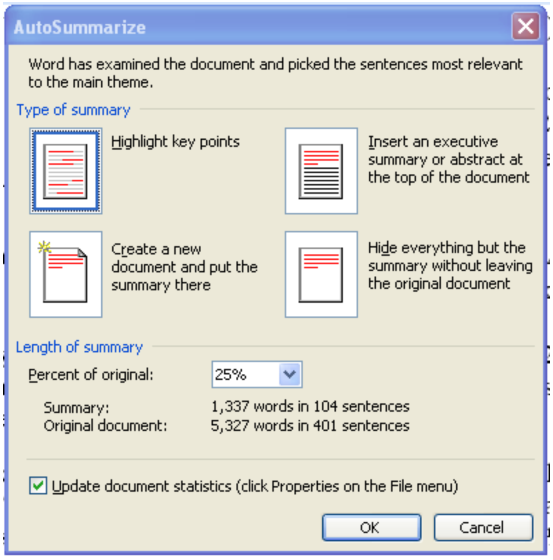
3. C*ue* words, such as *in conclusion*, *important*, *in this paper*, and *hardly,* signal the relevance of the sentence (Baxendale, 1958; cited in Das and Martins, 2007).

Work on automatic text summarization now deals with more complex problems, such as summarizing multiple documents, customizing summaries, reducing sentences (Knight and Marcu, 2002; Jing, 2000) and evaluating the quality of summaries, and employs techniques that go beyond surface textual features to capture semantic relationships in the texts. However, for our purposes these three features provide us with some tangible tools for teaching our students how to summarize a text.

3. AutoSummarize

Several summarization tools are available (see Das and Martins for an overview of current research), but Microsoft Word offers a basic tool to summarize a single document. It extracts sentences but does not paraphrase them (see Hovy and Lin, 1999).

The summarization tool can be found under the Tools menu. With the document open, you can select AutoSummarize and choose to highlight the important points or to create an abstract. You can also specify the length of the summary—from 1% to 50%.



AutoSummarize

Word has examined the document and picked the sentences most relevant to the main theme.

Type of summary

Length of summary

Percent of original: 25%

Summary: 1,337 words in 104 sentences

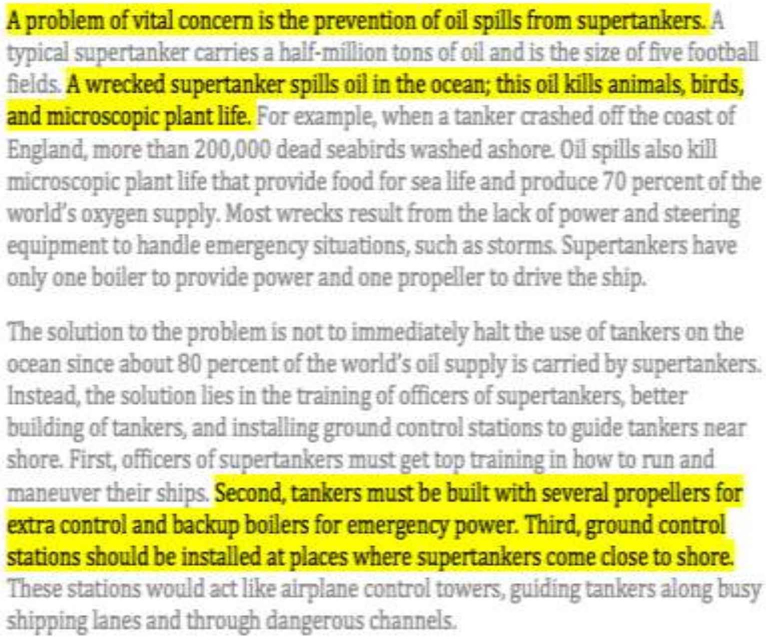
Original document: 5,327 words in 401 sentences

☑ Update document statistics (click Properties on the File menu)

Highlight key points
Insert an executive summary or abstract at the top of the document
Create a new document and put the summary there
Hide everything but the summary without leaving the original document
OK
Cancel

Figure 1. AutoSummarize window in Microsoft Word

You can choose whether to create the summary in a separate document or to have the sentences highlighted in the text. In the latter case, the output appears as in Figure 2, which uses the *Supertanker* passage from Meyer, Brandt, and Bluth (1980).



A problem of vital concern is the prevention of oil spills from supertankers. A typical supertanker carries a half-million tons of oil and is the size of five football fields. A wrecked supertanker spills oil in the ocean; this oil kills animals, birds, and microscopic plant life. For example, when a tanker crashed off the coast of England, more than 200,000 dead seabirds washed ashore. Oil spills also kill microscopic plant life that provide food for sea life and produce 70 percent of the world's oxygen supply. Most wrecks result from the lack of power and steering equipment to handle emergency situations, such as storms. Supertankers have only one boiler to provide power and one propeller to drive the ship.

The solution to the problem is not to immediately halt the use of tankers on the ocean since about 80 percent of the world's oil supply is carried by supertankers. Instead, the solution lies in the training of officers of supertankers, better building of tankers, and installing ground control stations to guide tankers near shore. First, officers of supertankers must get top training in how to run and maneuver their ships. Second, tankers must be built with several propellers for extra control and backup boilers for emergency power. Third, ground control stations should be installed at places where supertankers come close to shore. These stations would act like airplane control towers, guiding tankers along busy shipping lanes and through dangerous channels.

Figure 2. Sample output from AutoSummarize

4. Using an Automated Tool

According to the documentation on the Microsoft Office website (n.d.), the AutoSummarize tool is based on only one of the three parameters listed above, namely, word frequency, so it picks up sentences with the word *supertanker*; however, the output shows that cue words, such as *second* and *third*, are also used in the analysis.

On the Internet, several similar summarizing tools are available. Below is a comparable summary of the same length from a site called FreeSummarizer (http://freesummarizer.com/).

Summary of Supertanker passage from FreeSummarizer

A wrecked supertanker spills oil in the ocean; this oil kills animals, birds, and microscopic plant life.

The solution to the problem is not to immediately halt the use of tankers on the ocean since about 80 percent of the world's oil supply is carried by supertankers.

Instead, the solution lies in the training of officers of supertankers, better building of tankers, and installing ground control stations to guide tankers near shore.

FreeSummarizer does not provide the criteria on which the summary tool is based, but the output is a more coherent summary. One possible criterion that it uses is text structure; notice that sentence with the words *problem* and *solution* have been selected—the *Supertanker* text was in fact designed to illustrate a problem-solution text structure.

These automatic summarizers could help junior researchers reduce their reading material to manageable proportions, but they need to be used with caution because, as we see above, their output varies and the summaries may omit important information.

A more important application is in teaching students about summaries, not as a procedure but in terms of their features. Although AutoSummarize is a basic tool, it is available to students who have access to Microsoft Word, and if they have access to the Internet they could use other text summarization tools.. Using a short text (less than 250 words), ask students to use the tool and then discuss the output in class in terms of the three criteria listed above—word frequency, sentence position, and cue words. This gives students something concrete to work with. As a follow-up, students could use another summarization tool on the same passage and discuss the differences. The purpose of such an exercise is not to give students hard-and-fast rules for creating summaries, but to raise their awareness about the features of summaries, and help them understand and use the concept of a text structure.

The use of an automated tool in teaching provides students with an objective measure beyond the teacher's notion of a summary. And although we like to think that students have faith in what their teachers tell them, studies like Schmitt and Christianson (1998) find that university students are more attentive to computer-generated feedback than to teacher feedback.

==

References

Baxendale, P. (1958). Machine-made index for technical literature - an experiment. *IBM Journal of Research Development*, 2, 4, 354–361.

Biber, D. (1987). A textual comparison of British and American writing, *American Speech*, 62, 99-119.

Burstein, J., Chodorow, M. and Leacock, C. (2003). Criterion: Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence.*

Das, D. and Martins, A.F.T. (2007). *A Survey on Automatic Text Summarization. Literature Survey for the Language and Statistics II course at CMU, November 2007*. Available at http://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf

Gupta, R. (2009). Separated by a common language: Asian students writing in English, *Language in India*, 9, 43-62.

Hall, C., McCarthy, P.M., Lewis, G.A., Lee, D.S., and McNamara, D.S. (2007). Using Coh-Metrix to assess differences between English language varieties, *Coyote Papers: Working Papers in Linguistics*, 15. 40-54. Available at *http://coyotepapers.sbs.arizona.edu/CPXV/hall_mccarthy_lewis_lee_mcnamara 40-54.pdf*

Hinds, J. (1979). Organizational patterns in discourse. In T. Givon (Ed.), *Syntax and Semantics* (Vol. 12. Discourse and Syntax). New York: Academic Press.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.

Hovy, E. and Lin, C-Y. (1999). Automated Text Summarization in SUMMARIST. In I. Mani and M. Maybury (Eds.), *Advances in Automated Text Summarization* (pp. 81-97). MIT Press.

Jing, H. (2000). Sentence reduction for automatic text summarization**,** *Proceedings of the sixth conference on Applied Natural Language Processing*.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression, *Artificial Intelligence*, 139, 91–107.

LLoret, E. (2008). *Text summarization: An overview*. Available at http://www.dlsi.ua.es/~elloret/publications/TextSummarization.pdf

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2, 2, 159-165.

Meyer, B. J. F., Brandt, D. M., & Bluth, G. J. (1980). Use of the top-level structure in text: Key for reading comprehension of ninth-grade students, *Reading Research Quarterly*, 16, 72–103.

Meyer, B.J.F. and Freedle, R.O. (1984). Effects of discourse type on recall, *American Educational Research Journal*, 21, 1, 121-143.

Meyer, B.J.F. and Rice, E. (1984). The structure of text. In P.D. Pearson, R. Barr, M.L. Kamil, and P.B. Mosenthal (Eds.). *Handbook of Reading Research: Volume I* (pp. 319-350). White Plains, NY: Longman.

Microsoft Office (n.d.). http://office.microsoft.com/en-us/word-help/automatically-summarize-a-document-HA010255206.aspx. Accessed March 25, 2012.

Propp, V. I. ([1928] 1968), Morphology of the Folktale. Austin: University of Texas Press.

Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization, *Computational Linguistics*, 28, 4, 399–408.

Schmitt, L.M. and Christianson, K. T. (1998). Pedagogical aspects of a UNIX-based management system for English instruction, *System*, 26, 567-589.

===

Renu Gupta, Ph.D.
renu@stanfordalumni.org