Human evaluation of extraction based summaries

Marcus Johansson, Henrik Danielsson, Arne Jönsson

Santa Anna IT Research Institute AB
Linköping, Sweden
marjo581@student.liu.se, henrik.danielsson@liu.se, arnjo@ida.liu.se

Abstract

We present an evaluation of an extraction based summarizer based on human assessments of the summaries. In the experiment humans read the various summaries and answered questions on the content of the text and filled in a questionnaire with subjective assessments. The it took to read a summary was also measured. The texts were taken from the readability tests from a national test on knowledge and ability to be engaged in university studies (Sw. Högskoleprovet). Our results show that summaries are faster to read, but miss information needed to fully answer questions related to the text and also that human readers consider them harder to read than the original texts.

1. Introduction

Most evaluations of extraction based text summarizations are based on comparisons to gold standards, e.g. (Over et al., 2007; Pitler et al., 2010; Smith and Jönsson, 2011a). Such evaluations are rather straightforward to conduct and are important to assess the performance of a summarizer. The performance of the summarizer is then mainly assessed based on n-gram statistics between the gold standard and the summarization produced by the summarizer. However, such evaluations, termed intrinsic (Hassel, 2004), do not consider how readable a summary is or how much information that is conveyed from the original document.

Extrinsic evaluations, where the usability of a summary is evaluated are less common. Morris et al. (1992), however, present an evaluation of summaries where they have subjects do the American Graduate Management Aptitude test, similar to the Swedish Högskoleprovet based on summaries of varying length, a human produced abstract and no text, i.e. they have to guess when answering the questions. They did found that human produced abstracts were best, but their results were not significant. Mani et al. (1999) found that summaries comprising 17% of the original text were as good as the original text to predict if the information is relevant for a certain subject.

In this paper we present results from an extrinsic evaluation of an extraction based summarizer based on experiments with humans answering test questions after reading the original text, summaries and guessing.

2. Method

We recruited 60 students, mean age 22.6 years, 22 women and 38 men, all at Linköping University. The test was similar to the test by Morris et al. (1992). We used the reading comprehension test from the National test for high school studies (Sw. Högskoleprovet) from 2011 as we assumed that none of them had previous experience with that test set, which they did not have. The experiment used four different test sets. One extra test set was used as training set. The four test sets were summarized using the extraction based summarizer COGSUM (Smith and Jönsson, 2011b) to 30% of the original text length. We also kept the original text for comparisons. The original texts were between 928-1108 words and the summaries between 410-308 words.

Before the test the subjects answered a questionnaire comprising background data such as age, sex, if they have done the test before and esteemed reading ability. They then practiced on a text that was not used in the actual test and after that they did the actual test under three conditions: 1) reading a 30% summarization, 2) reading the original text and, 3) to answer the questions without any text at all, i.e. they had to guess. These three conditions were handed to the subjects in a different and balanced order between the subjects.

For the text conditions the subjects first read the tests' pre-defined questions, then they were handled the text which included answers to the questions and finally the questions were handed back to them and they were asked to answer them. We measured the time it took for the subjects to carry out each sub task, read questions, read text and answer the questions. After the test the subjects had to answer a second questionnaire with Likert scale items (1-7) on their attitudes towards the text such as how easy it was to read, if all relevant information was in the text, if it took long to read and understand etc. We also measured the number of correct answers to the questions.

3. Results

Table 1 shows the number of correct answers on the questions in the test *Högskoleprovet*. The data were analyzed using a within-group ANOVA test which gave the result $F(1.86, 109.77) = 30.735, p < .01, \eta^2 = .34$, Huynh-Feldt corrected. This was followed by a SIDAK post-hoc test to investigate differences between conditions.

Table 1: Number of correct answers and time to read for each text type.

Text type	Correct answers		Time(sec)	
	Mean	StDev	Mean	Stdev
Original text	2.62	1.04	337.6	109.38
Summary	2.2	1.03	153.9	61.34
Guessing	1.3	0.94		

Table 2: Means and standard deviations for questionnaire items for the original text and the summary text.

Item	Original		Summary	
	Mean	StDev	Mean	Stdev
I think the text gives a good conception of the subject	4.63	1.52	3.20	1.37
I experience the text as information rich	4.70	1.37	3.48	1.56
I think the text has a good flow	4.75	1.49	3.63	1.69
I experience that the text misses relevant information in order to answer the questions	3.25	1.44	4.55	1.65
I think the text was easy to comprehend	4.93	1.59	4.12	1.60
I think it took a long time to read the text	3.87	1.33	3.28	1.32
I think the text was easy to read	4.78	1.61	4.08	1.81
I think the text was exhausting to read	3.55	1.67	3.85	1.67

As expected reading the original text gives significantly more correct answers than reading the summary, $p < .05$. Both the summary and the original text give significantly more correct answers, $p < .001$. The difference between the original text and the summary was 10.5% fewer correct answers.

Table 1 also depicts the time it took to read the text. The time to read the summary is 55% shorter than the time it takes to read the original text, a significant difference $t(59) = 17.73, p < .001$.

We did not find any significant difference in the time it took to answer the test questions.

Table 2 depicts the subjective scores on the items in the questionnaire for reading the original text and the summary.

There were statistically significant differences, two-tailed t-test significance level $p < .001$, between the original text and the summary for all items (all $ts > 2.5$, all $ps < .01$), except for *I think the text was exhausting to read* where no significant difference was found.

4. Discussion

We have presented results from an evaluation of extraction based summaries. Sixty subjects read texts from a national readability test and answered questions on the text. In the study we also measured reading time and the subjects answered a questionnaire with items on their perceived text quality.

The time it takes to read the summary is significantly shorter that reading the original text.

When our subjects read the original text they had significant more correct answers to the questions than they had when reading a 30% summary of the text. However, compared to guessing without reading the text the subjects had significantly more correct answers after reading the summary. Furthermore, the amount of information lost in the summary compared to the original text is only 10%, and considering that 70% of the text is lost in the summary this can be considered as an acceptable loss of information, especially as the time it took to read the text was around 50% shorter reading the summary compared to reading the full text.

The importance of losing information depends, of course, on the type of text. Persons reading a news text probably accept losing 10% or even more of the text, especially if it means saving 20% of the time it takes to read. For other texts, such as texts on how to fill in authority forms or the texts we used on taking a test, we can assume that information loss is more problematic.

Overall the original texts were considered better than the summaries. They were easier to read, had a better cohesion, and contained more information. The mean values on the various Likert items for the summaries, as seen in Table 1, are often around 3.5, i.e. the arithmetic mean of the scale with a maximum of 7 and although significantly worse than the originals the difference is only about 1 point on the scale indicating that the summaries are not that bad.

One important target group for automatic text summarization is persons with reading difficulties, such as dyslectics. Conducting studies with such persons is an important future work.

5. References

Martin Hassel. 2004. Evaluation of automatic text summarization. Licentiate Thesis, 3-2, Stockholm University.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of EACL-99*.

Andrew H. Morris, George M. Kasper, and Dennis A. Adams. 1992. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 3(1):17–35.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43:1506–1520, Jan.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality inmulti-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*, pages 544–554.

Christian Smith and Arne Jönsson. 2011a. Automatic summarization as means of simplifying texts, an evaluation for swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010), Riga, Latvia*.

Christian Smith and Arne Jönsson. 2011b. Enhancing extraction based summarization with outside word space. In *Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand*.