Using the pyramid method to create gold standards for evaluation of extraction based text summarization techniques

Bertil Carlsson, Arne Jönsson

Department of Computer and Information Science, Santa Anna IT Research Institute AB
Linköping University, SE-581 83, Linköping, SWEDEN
Berca955@student.liu.se, arnjo@ida.liu.se

Abstract

We present results from using a version of the pyramid method to create gold standards for evaluation of automatic text summarization techniques in the domain of governmental texts. Our results show that the pyramid method may be useful to create gold standards for extraction based summarization techniques using only five human summarisers.

1. Introduction

Automatic creation of text summarizations is an area that has gained an increasing interest over the years, for instance in order to allow for skim reading of texts or to facilitate the process of deciding if a text is interesting to read in full. In order to know if the summarization is useful it must be evaluated.

To evaluate automatic text summarization techniques we either need humans to read, and evaluate a number of summarizations, or we can compare it to a gold standard, a "correct" summarization of the text, i.e. extrinsic or intrinsic evaluation of the text. A gold standard is often a compilation of different human created summarizations which is then put together into one.

It is an open question how to assemble such human created summaries into one gold standard. In this paper we present results from using a variant of the pyramid method (Nenkova, 2006) to create gold standards of text summaries. We use the pyramid method on extraction based summaries, i.e. we do not ask our human summarisers to write an abstract summary but to extract a number of whole sentences from the text. The texts are governmental texts. We also present an evaluation of our gold standards.

2. The pyramid method

The pyramid method is a summarization technique used to assemble summary fragments (words, phrases or sentences) from different humans to generate one summarization (Nenkova, 2006). Nenkova used information fragments, brief phrases with the same information content, in her original study in the domain of news texts.

The pyramid method assigns each information fragment a weight, reflected by the number of human summarisers that have highlighted it as an important fragment for the text. Each fragment is then inserted into a pyramid where each layer in the pyramid represents how many summarisers that have suggested the fragment. Consequently, the number of layers in the pyramid is equal to the number of summarisers and the higher up the more likely it is that a fragment is important.

One interesting result from Nenkova (2006) is that pyramids comprising four to five layers produce the best results in evaluations of summaries. Thus, contrary to e.g. Halteren and Teufel (2003), five summaries is all that is needed to produce a gold standard.

3. Creation of the gold standards

We use 5 frequently used fact sheets from the Swedish Social Insurance Administration (Sw. Försäkringskassan) as selected by employees at the Swedish Social Insurance Administration. They comprise 62-91 sentences, each between 1000 and 1300 words. All texts were about allowances and had the same structure.

Our ambition was to create indicative summaries, i.e. they should not replace reading the whole text but rather facilitate deciding if reading the whole text is interesting. A pre-study revealed that 10% is an appropriate length of such a summary (Jönsson et al., 2008).

Five persons created summaries of all five texts, two students, two seniors and one worked in a private company. All had sufficient read and write skills in Swedish and none had ever constructed extraction based summaries before.

The text summarizations were entered into a pyramid, as explained in Section 2., one for each text, and from these the gold standards were created. The variation between the summaries produced by the summarisers versus the produced gold standard were investigated by computing the sentence overlaps for the summaries.

The sentence overlap for the five gold standards created in this study varies between 57,5% and 76,6%, which is in line with previous studies that have found that the sentence overlap normally vary between 61% and 73% where the larger number is achieved by more accustomed summarisers (Hassel and Dalianis, 2005). All but one of the summaries obtain the minimum value which represents a good overlap according to (Hassel and Dalianis, 2005). The 57,5% overlap can be explained by inexperience from the human summarisers part in creating extraction based summaries. Something which has been well documented in earlier work, such as Hassel and Dalianis (2005).

To further investigate the variation obtained by our human summarisers, we calculated the number of new sentences added by each human summariser. These investigations show that the number of new sentences added by the summarisers drops rather quickly. At most the fifth summariser adds three new sentences and at best only one. Thus, we can assume that the summaries comprise the most important sentences from the text. It should be noted that

humans do not agree on what is a good summary of a text (Lin and Hovy, 2002; Hassel and Dalianis, 2005; Jing et al., 1998), which means that there is probably not one single best summary. The results presented here also point towards texts having a limit on important sentences that should be included in summaries. Something that has to be further investigated.

4. Evaluation

Evaluation of the gold standards was conducted by having subjects read the summaries and answer a questionnaire on the quality of the summary. The questionnaires used six-point Likert items and comprised the following items on the summary: [Q1] ... has a good length to give an idea on the content in the original text, [Q2] ... is experienced to be information rich, [Q3] ... is experienced as strenuous to read, [Q4] ... gives a good idea on what is written in the original document, [Q5] ... gives a good understanding of the content of the original document. [Q6] ... is experienced as missing relevant information from the original document, and [Q7] ... is experienced as a good complement to the original document.

The subjects for our evaluation where 10 students and 6 professional administrators at the Swedish Social Insurance Administration.

All subjects read the summary but did not have the original text at hand, to more resemble future use of the system. Discourse coherence for extraction based summaries is, of course, a problem. Our evaluators were not instructed to disregard discourse coherence since this is a factor which has to be accounted for when creating texts of this sort.

The results from the student evaluations are presented in Table 1. Note that, as the items are stated, a high score is considered positive on Q1, Q2, Q4, Q5 and Q7 whereas as low score on Q3 and Q6 is considered positive. Note also that the questions themselves are intertwined and hence act as some sort of control questions to each other in order to assure that the data given by the participants in the questionnaire is correct.

Table 1: Mean from the students' responses

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
S1	4,5	4,5	2,8	4,0	3,8	2,5	4,2
S2	4,7	4,8	1,5	4,2	4,6	2,2	4,5
S3	5,2	5,1	2,0	4,4	4,6	1,9	4,7
S4	4,9	5,3	2,2	4,7	4,9	2,1	4,7
S5	4,5	4,2	1,9	4,3	4,4	2,8	4,5

As can be noted from Table 1 the evaluators give positive opinions on all items.

Table 2: Mean from the professionals' responses

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
S1	4,0	4,2	4,0	4,2	4,2	2,5	4,2
S2	4,7	4,5	2,8	4,3	4,2	2,3	4,3
S3	4,5	4,5	3,0	4,5	4,7	2.2	4,8
S4	4,5	4,7	2,2	4,7	4,7	1,7	5,0
S5	4,5	4,0	3,5	4,3	4,5	1,8	4,0

The results from the professional administrators' answers to the questionnaires, Table 2, also demonstrate positive opinions on all items, but Q3. The professional administrators are indifferent regarding how hard the texts are to read. In fact, two subjects rank them as rather hard to read.

Notable is that the students and professional administrators provide very similar answers to most of the questionnaires. They all consider the text to be informative, Q2, and having an appropriate length, Q1. They also, all think that the texts provide a good idea on what was in the original text, Q4 and Q5. Furthermore, the subjects do not think that the texts miss relevant information.

5. Summary

We have used the pyramid method to create extraction based summaries of governmental texts.The summaries are evaluated by both novices (students) and professionals (administrators at the local governmental agency) and the evaluations show that the summaries are informative and easy to read.

Our results are in line with previous research (Nenkova, 2006) which states that five human summarisers are enough to produce a gold standard. It can be further stated that the pyramid method then not only can be used in order to create gold standards from abstract summaries but also from extraction based summaries.

Acknowledgements

This research is financed by Santa Anna IT Research Institute AB. We are grateful to our evaluators and especially the staff at the Swedish Social Insurance Administration.

6. References

Hans Van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: Initial. In *In HLT-NAACL DUC Workshop*, pages 57–64.

Martin Hassel and Hercules Dalianis. 2005. Generation of Reference Summaries. In *Proceedings of 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland, April 21-23.

H Jing, R Barzilay, K McKeown, and M Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. *AAAI Symposium on Intelligent Summarization*, Jan.

Arne Jönsson, Mimi Axelsson, Erica Bergenholm, Bertil Carlsson, Gro Dahlbom, Pär Gustavsson, Jonas Rybing, and Christian Smith. 2008. Skim reading of audio information. In *Proceedings of the The second Swedish Language Technology Conference (SLTC-08), Stockholm, Sweden*.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Morristown, NJ, USA. Association for Computational Linguistics.

Ani Nenkova. 2006. *Understanding the process of multi-document summarization: Content selection, rewriting and evaluation*. Ph.D. thesis, DigitalCommons@Columbia, January 01.