Analysis of Summarization Evaluation Experiments

Marie-Josée Goulet
CIRAL, Department of Linguistics
Laval University, Quebec City
G1K 7P4, Canada
marie-josee.goulet.1@ulaval.ca

Abstract

The goals of my dissertation are: 1) to propose a French terminology for the presentation of evaluation results of automatic summaries, 2) to identify and describe experimental variables in evaluations of automatic summaries, 3) to highlight the most common tendencies, inconsistencies and methodological problems in summarization evaluation experiments, and 4) to make recommendations for the presentation of evaluation results of automatic summaries. In this paper, I focus on the second objective, i.e. identifying and describing variables in summarization evaluation experiments.

1 Introduction

The general subject of my dissertation is summarization evaluation. As stated in my thesis proposal, my work aims at four goals: 1) proposing a French terminology for the presentation of evaluation results of automatic summaries, 2) identifying and describing experimental variables in evaluations of automatic summaries, 3) highlighting the most common tendencies, inconsistencies and methodological problems in summarization evaluations, and 4) making recommendations for the presentation of evaluation results of automatic summaries. In this paper, I will focus on the second objective.

My ultimate goal is to provide the francophone scientific community with guidelines for the evaluation of automatic summaries of French texts. Evaluation campaigns for NLP applications already exist in France, the EVALDA project[1]. However, no campaign has yet been launched for French automatic summaries, like Document Understanding Conferences for English texts or Text Summarization Challenge for Japanese texts. I hope that such a campaign will begin in the near future and that my thesis work may then serve as a guide for its design.

2 Completed Work

I collected 22 scientific papers about summarization evaluation, published between 1961 and 2005. Each paper has been the subject of an in-depth analysis, where every detail regarding the evaluation has been carefully noted, yielding a quasi-monstrous amount of experimental variables. These variables have been classified into four categories: 1) information about source texts, 2) information about automatic summaries being evaluated, 3) information about other summaries used in the evaluation process, and 4) information about evaluation methods and criteria. At the current stage of my research work, the first three types of variables have been analyzed and will be presented here.

2.1 Variables about source texts

Four types of information about source texts emerged from the analysis: 1) the number of source texts, 2) the length, 3) the type of text, and 4) the language. First, the number of source texts is an indicator of the significance of the evaluation. In my study,

[1] http://www.elda.org/rubrique25.html

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 17–20,
Rochester, April 2007.

all the evaluations used less than 100 source texts, except for Mani and Bloedorn (1999) (300 source texts), Brandow et al. (1995) (250 source texts), Kupiec et al. (1995) (188 source texts) and Teufel and Moens (1999) (123 source texts).

Secondly, regarding source text length, it is expressed in different ways from one evaluation to another. For example, Edmundson (1969) gives the number of words, Klavans et al. (1998) give the number of sentences and Minel et al. (1997) give the number of pages. In some papers, the length of the shortest and of the longest text is provided (Marcu, 1999) while in others it is the average number of words, sentences or pages that is given (Teufel and Moens, 1999). Obviously, it would be wise to standardize the way source texts length is given in evaluation experiments.

In my corpora, there are three main types of source texts: 1) scientific papers, 2) technical reports, and 3) newspapers. Also, Minel et al. (1997) used book extracts and memos, and Farzindar and Lapalme (2005) used judgments of the Canadian federal court. All evaluations used only one type of source texts, except for Kupiec et al. (1995) and for Minel et al. (1997).

Finally, the majority of the evaluations used English texts. Some authors used French texts (Minel et al., 1997; Châar et al., 2004), Korean texts (Myaeng and Jang, 1999) or Japanese texts (Nanba and Okumura, 2000).

2.2 Variables about automatic summaries being evaluated

In this section, I describe variables about automatic summaries being evaluated. The variables have been classified into six categories: 1) the total number of automatic summaries evaluated, 2) the number of automatic summaries produced per source text, 3) if they are multiple document summaries, 4) the length, 5) if they are extracts or abstracts, and 6) their purpose.

First, concerning the total number of automatic summaries, Brandow et al. (1995), Mani and Bloedorn (1999), Kupiec et al. (1995), Salton et al. (1997) and Teufel and Moens (1999) evaluated respectively 750, 300, 188, 150 and 123 automatic summaries. All the other studies for which this information is given evaluated less than 100 automatic summaries. It may appear redundant to give the number of source texts and the number of automatic summaries in an evaluation, but sometimes more than one automatic summary per source text may have been produced. This is the case in Brandow et al. (1995) and Barzilay and Elhadad (1999) where automatic summaries of different lengths have been evaluated.

Automatic summaries can either be produced from one text or more than one text. In my corpora, only Mani and Bloedorn (1999) and Châar et al. (2004) evaluated multiple document summaries.

As for source texts, automatic summary length is expressed in different ways from one evaluation to another. Moreover, it is not always expressed in the same way than source text length, which is inconsistent.

On a different note, most experiments evaluated extracts, except for Maybury (1999) and Saggion and Lapalme (2002) who evaluated abstracts, reflecting the predominance of systems producing extracts in the domain of summarization. Extracts are summaries produced by extracting the most important segments from texts while abstracts are the result of a comprehension process and text generation. Most extracts evaluated are composed of sentences, except for Salton et al. (1997) and Châar et al. (2004) where they are respectively composed of paragraphs and passages. The type of automatic summaries is crucial information because it normally influences the choice of the evaluation method and criteria. Indeed, we do not evaluate extracts and abstracts in the same way since they are not produced in the same way. Also, their purposes generally differ, which can also influence the choice of the evaluation method and criteria.

Last, some papers contain the specific purpose of automatic summaries, not only if they are indicative or informative, which is interesting because it can sometimes explain the choice of the evaluation method. Only 9 experiments out of 22 give this information in my corpora.

2.3 Variables about other summaries used in the evaluation process

One of the most common evaluation methods consists of comparing automatic summaries with other summaries. During my analysis, I identified seven

types of information about these other summaries: 1) the total number of other summaries, 2) the type of summaries, 3) the length, 4) the total number of human summarizers, 5) the number of human summarizers per source text, 6) the instructions given to the human summarizers, and 7) the human summarizers' profile.

The number of other summaries does not necessarily correspond to the number of automatic summaries evaluated, depending on many factors: the use of other summaries of different types or different lengths, the number of persons producing the other summaries, the number of other systems producing the other summaries, and so on.

There are two general types of summaries used for comparison with the automatic summaries being evaluated. First, *gold standard summaries* (or *target summaries*) can be author summaries, professional summaries or summaries produced specifically for the evaluation. Second, *baseline summaries* are generally produced by extracting random sentences from source texts or produced by another system.

In my corpora, gold standard summaries are often produced specifically for the evaluation. In most cases, they are produced by manually extracting the most important passages, sentences or paragraphs, allowing automatic comparison between automatic summaries and gold standard summaries.

On the other hand, many evaluations used baseline summaries. For example, Barzilay and Elhadad (1999) used summaries produced by *Word AutoSummarize*, Hovy and Lin (1999) used summaries produced by automatically extracting random sentences from source texts. In Brandow et al. (1995), Kupiec et al. (1995) and Teufel and Moens (1999), baseline summaries were produced by automatically extracting sentences at the beginning of the texts, and in Myaeng and Jang (1999) by extracting the first five sentences of the conclusion.

Logically, the length of the summaries used for the comparison should be equivalent to the length of the automatic summaries being evaluated. If automatic summaries of different lengths are evaluated, there should be corresponding baselines and/or gold standard summaries for each length, unless the goal of the evaluation is to determine if the length plays a role in the quality of automatic summaries.

Many of the evaluations analyzed do not indicate the number of human summarizers participating in the production of gold standard summaries. A few of them specify the total number of persons involved, but not the number for each source text. This is an important variable because summarizing, either by extracting or abstracting, is a subjective task. The more people involved in the summarization of one text, the more we can consider the final summary to be reliable. From the pieces of information I was able to gather, the number of summarizers per source text ranges from 1 to 13 in my corpora.

In analyzing the evaluations of my corpora, I realized that some authors gave clear instructions to the human summarizers, for example Edmundson (1969). In other cases, authors asked the summarizers to extract the most "important" sentences. The term "important" includes other terms like representative, informative, relevant, and eligible. It is rarely mentioned however if those words were explained to the summarizers.

I also noticed that some evaluations used people coming from different backgrounds, for example in Salton et al. (1997), while others used more homogeneous groups, for example in Barzilay and Elhadad (1999) and Kupiec et al. (1995).

3 Future Directions

In the next couple of months, I plan to analyze evaluation methods identified in my corpora, for example comparing automatic summaries with gold standard or baseline summaries, and asking judges to give their opinion on the quality of automatic summaries. I will also describe evaluation criteria used to assess the quality of the automatic summaries, for example informativeness and readability. Next, I will make recommendations for the presentation of summarization evaluation results, based on the knowledge acquired from my analysis of 22 scientific papers, and from previous evaluation campaigns.

4 Conclusion

In this paper, I described variables about source texts, about automatic summaries being evaluated and about other summaries used in summarization evaluation experiments. These variables provide important information for the understanding of the

evaluation results presented in a scientific paper. My analysis is based on 22 scientific papers on summarization evaluation, which is to my knowledge the largest study on the variables found in evaluation experiments. This constitutes a notable contribution in the domain of summarization. In another paper (in French) to appear, I propose a French terminology for the presentation of evaluation results in the domain of summarization, which is also a major contribution.

To conclude, the analysis presented in this paper gave an overview of summarization evaluation habits since 1961. Also, it showed that there is no common agreement as to how evaluation results should be presented in a scientific paper about automatic summaries.

Acknowledgements

I would like to thank the SSHRC and the FQRSC for granting me doctoral scholarships. I would also like to thank Joël Bourgeoys, Neil Cruickshank, Lorraine Couture and the anonymous reviewer for their useful comments.

References

R. Barzilay and M. Elhadad. 1999. Using lexical chains for text summarization. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 111–121, Cambridge, Massachusetts. MIT Press.

R. Brandow, K. Mitze, and L. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing Management*, 31(5):675–685.

S. L. Châar, O. Ferret, and C. Fluhr. 2004. Filtrage pour la construction de résumés multidocuments guidée par un profil. *Traitement automatique des langues*, 45(1):65–93.

H. P. Edmundson. 1969. New methods in automatic abstracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

A. Farzindar and G. Lapalme. 2005. Production automatique de résumé de textes juridiques : évaluation de qualité et d'acceptabilité. In *TALN*, pages 183–192, Dourdan.

E. Hovy and C.-Y. Lin. 1999. Automated text summarization in SUMMARIST. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94, Cambridge, Massachusetts. MIT Press.

J. L. Klavans, K. R. McKeown, M.-Y. Kan, and S. Lee. 1998. Resources for the evaluation of summarization techniques. In Antonio Zampolli, editor, *LREC*, pages 899–902, Granada, Spain.

J. Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *SIGIR*, pages 68–73, Seattle.

I. Mani and E. Bloedorn. 1999. Summarizing similarities and differences among related documents. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 357–379, Cambridge, Massachusetts. MIT Press.

D. Marcu. 1999. Discourse trees are good indicators of importance in text. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–136, Cambridge, Massachusetts. MIT Press.

M. Maybury. 1999. Generating summaries from event data. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 265–281, Cambridge, Massachusetts. MIT Press.

J.-L. Minel, S. Nugier, and G. Piat. 1997. How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE protocols and their results on SERAPHIN. In *EACL*, pages 25–31, Madrid.

S. H. Myaeng and D.-H. Jang. 1999. Development and evaluation of a statistically-based document summarization system. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 61–70, Cambridge, Massachusetts. MIT Press.

H. Nanba and M. Okumura. 2000. Producing more readable extracts by revising them. In *18th International Conference on Computational Linguistics*, pages 1071–1075, Saarbrucker.

H. Saggion and G. Lapalme. 2002. Generating indicative-informative summaries with SumUM. *Computational Linguistics*, 28(4):497–526.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 33(2):193–207.

S. Teufel and M. Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 155–171, Cambridge, Massachusetts. MIT Press.