Measuring the quality of multi-document cluster headlines

Frank van Kesteren
University of Twente
Enschede, The Netherlands
Email: f.vankesteren@student.utwente.nl

Wessel Kraaij
TNO
P.O. Box 5050, 2600 GB Delft, The Netherlands
Email: kraaijw@acm.org

***Abstract*— Headline summaries of multi-document clusters enable efficient navigation and selection of content, provided headlines are of sufficient quality. This study compares several methods for automated headline extraction, with considerable variation in length. The reliability of the automated evaluation is validated by a comparison with human produced headlines, taking into consideration the variability in manually created headlines and inter-human agreement in quality judgements. Results suggest that, ROUGE precision is a suitable measure for automatic evaluation of headlines of differing lengths. Also ROUGE Recall can be used after applying a length penalty.**

I. Introduction

This research is part of a long-term research project aiming at a system for automatic clustering and indexing news stories from multiple sources and multiple media types [1]. The system generates clustered representations of documents from various sources, automatically annotated with several types of metadata. One of the research projects concerned improving the method for generating headlines for these clusters, which play a crucial role for browsing and search.

Evaluation of system variants was done manually, and automatically using ROUGE, modeled after the headline task in DUC 2004. However, the best system variant according to ROUGE evaluation did not produce satisfactory headlines. The manual evaluation revealed that the automatic evaluation cannot be used as a replacement of manual evaluation. Two hypotheses why correlation between both methods is so low will be tested and discussed.

II. headline extraction

Background: Headline construction can roughly be split into two groups: headline extraction and headline generation. Headline extraction is based on the idea that the headline is present in the available text, it is only a matter of extracting the right sentence. Novalist works with this principle, also Angheluta and Mira [3] follow the same method. They differ in that they try to compress the most salient sentence, instead of extracting a NP from it. Other experimented with using different features, such as Erkan and Radev [4] with there MEAD system. They included a feature called LexPageRank, similar to the Google page rank system. Rong Jin compared several methods of word extraction [6]. Headline generation tries to build a headline from extracted keywords from a text. These method were not included in this research.

Methods: Experimented were carried out with several methods of headline extraction, most of them based on the original algorithm used in Novalist [2]. But we also used methods that picked the best fitting headline from the documents in the cluster (in fact this produced the best headline according to manual evaluation). Note that this approach often led to headlines with various lengths. Lengths vary from 16 characters to about 150 characters! To evaluate the headlines, we manually created a headline for 100 clusters, and manually evaluated the generated headlines using these headlines as a reference. We also tried to use ROUGE for evaluation. ROUGE was picked because it was the major tool for automatic evaluation used at DUC. We both evaluated for quality of the headline: how fit is the headline for the given cluster, and for readability: is the generated headline grammatically correct? The measure chosen for quality of the headline was ROUGE-1 Recall, for readability ROUGE-2 Recall was chosen.

ROUGE indicated different "best" methods than human evaluation did. Comparing ROUGE and the manual evaluation resulted in a weak correlation. At this moment ROUGE could not be trusted as a good evaluator for headlines. This gave rise to the question of how/if ROUGE could be used to automatic evaluate headlines.

The hypothesis about the origin of the problem is that ROUGE does not work well with headlines of different lengths. To find a solution to this problem, two approaches were investigated: The possibility to measure the quality of headlines by use of ROUGE-1 precision, or by implementing a penalty for the length of a headline and use ROUGE-1 Recall.

Evaluation: Evaluation is always an important part of developing any summarization system. Until recently this used to be a human task of comparing generated summaries with one or more human created summaries. Problems with this human evaluation are that it is difficult to reproduce scores. Inter annotator agreement is often low. Therefore the need for an objective evaluation tool rise.

The first step in this direction was the Summary Evaluation Environment (SEE). This environment aids human evaluators, e.g. by highlighting overlap with the model summary and specifying a classification for every category [5]. Research in the field of Machine Translation evaluation led to the development of BLUE (BiLanguage Evaluation Understudy) [7] based on n-gram co-occurrence. Requirements for the system were that the results strongly correlated with human evaluators. This method works well for (translated) summaries and longer texts. Building on the same idea a system called ROUGE was developed [8]. Main focus again, was the strong correlation with human evaluators. In both cases several reference models need to be created. The shorter the summary the more important this is. Inter annotator agreement between summaries becomes smaller when a summary gets smaller.

III. Experiments

Comparison of headline extraction methods: The dataset consisted of clusters produced from Dutch newspaper articles and transcripts of a Dutch TV news show. A test set was created by selecting 100 medium-sized rather homogenous clusters, after a qualitative analysis. Note that this is an important step, since headline quality is dependent on the cluster quality. For each cluster one reference headline was created manually. Each cluster was pre-processed using a PoS tagger and a chunker. Several methods for headline extraction were run on the clusters, resulting in a headline for every cluster, for every method tried. Finally the headlines were evaluated using ROUGE. Results are listed in table I. Several methods are tried. From each category the best method (manually evaluated) was picked for this paper. The methods used are:

1) Original method used in Novalist. The method used thus far was based on a Naive Bayes system, that ranked the sentences from the cluster on a feature vector that was extracted from the cluster [2]. From the best, most salient sentence a noun phrase (NP) was then extracted that contained a trigger word (also extracted from the cluster).
2) Use first x ranked sentences to extract noun phrases from. Select longest or shortest. In the original system only the most salient sentence was used to extract the NP. Now use more sentences and pick the longest or shortest of all NP's found. Empirically was found that 8 sentences gave the best results.
3) Select headline from existing titles in document. It was found that, given a coherent cluster, an existing headline of a document of the cluster often gave a good headline for the entire cluster.
4) Compress most salient sentence. Instead of picking an NP from the most salient sentence, in this method we tried to compress the sentence to reach a final satisfactory headline.
5) Changes in clustering. It was found that clusters are not always as coherent as we hoped they would be. The idea of this method was to re-cluster the given cluster, so only the most important documents in the cluster were included in finding the headline.

Method	ROUGE-1 Recall	ROUGE-2 Recall
1	0.21831	0.07791
2	0.38883	0.15944
3	0.32039	0.12834
4	0.38693	0.08758
5	0.36326	0.14423

TABLE I
ROUGE scores for headline extraction methods

In order to assess the validity of automatic ROUGE evaluation versus manual evaluation, all automatically extracted headlines were also judged by a human annotator. Human evaluation scored every headline on a scale of 1 to 5. Results are listed in table II. We compared the rank ordering based on ROUGE with the rank order based on the manual judgements using the Spearman correlation coefficient. This resulted in a low and insignificant correlation (row 1, Table III).

Method	quality	readability
1	2.67	3.63
2	3.18	3.57
3	3.37	4.31
4	2.92	3.21
5	2.79	2.23

TABLE II
Human scores for headline extraction methods

	Length	Quality
Recall-1 no penalty	0,44524	0,20925
Recall-1 + BP	0,40607	0,18189
Recall-1 + LP	-0,02728	0,39056
Precision-1 no penalty	-0,47319	0,44303
Precision-1 + BP	-0,46689	0,46093
Precision-1 + LP	-0,54152	0,48633

TABLE III
Spearman's rho correlation between ROUGE and generated headline length and manual peer quality

Precision as a measure for automatic evaluation: To make our results more reliable, we first created additional models. Since the variability in the formulation of a headline of a cluster without a guiding context is high, we hoped for better results this way. For a randomly chosen subset of 10 clusters of the 100 test clusters, four additional reference headlines were created by four individuals. Also we asked them to manually judge the generated headlines for these clusters for several methods. The judgments of these 4 manual evaluators were averaged, and used to determine the correlation between the ROUGE scores. Inter annotator agreement between the quality judgment of the generated headlines was high and significant (>,675 using

Spearman's rho). Table III lists the results of correlation of the several (altered) ROUGE measurements with the averaged quality. We can conclude that ROUGE-1 Precision is at least slightly better than ROUGE-1 Recall. Also included in this table are the ROUGE scores correlated with the length of the generated headline. We can see that precision is negative correlated with the length which could be expected due to the nature of ROUGE Precision (longer sentences result in less overlap with model).

ROUGE with a penalty: Using the same averaged quality explained above, we also experimented with adding a penalty to the ROUGE measurement. Two penalties were tried, modeled after the penalties in BLEU [7]. A brevity penalty (BP), which should increase ROUGE Precision, and a length penalty (LP) for ROUGE Recall scores. The BP was applied as follows: if a headline is of the same length or longer than the shortest model, no penalty is applied, otherwise multiply original ROUGE score with BP. The same structure was used for the LP: if the headline was of the same length or shorter than the longest model, no LP was applied, otherwise the ROUGE score was multiplied by the LP. Results can be found in Table III. It can be concluded that BP slightly increases the score of ROUGE Precision, and slightly decreases the score of ROUGE Recall. Results with the BP are not very significant because it turned out that the BP was not often applied, since some of the manually created models were already very short. The LP increases both the ROUGE Recall as the ROUGE Precision. The correlation of the Recall score is almost doubled. This result seems promising. Using this score, ROUGE-1 Recall can be applied in a more reliable way when length of headlines differ. More important is also the fact that ROUGE-1 Recall with the LP is no longer correlated with the generated headline length. The penalties are calculated as follows.

BP = exp(1-r/c)
r: sum (over all clusters) of the lengths of shortest models
c: sum of the lengths of generated headlines (of one method)

LP = exp(1-c/r)
r: sum (over all clusters) of the lengths of longest models
c: sum of the lengths of generated headlines (of one method)

IV. Conclusion

Our small scale experiment suggests that both hypothesis 1 and hypothesis 2 are more or less correct. Based on former experiments the basic ROUGE-1 Recall is not a good measurement for the evaluation of variable length headlines. Using ROUGE-1 Precision improved the correlation with human annotators, as did the use of a length penalty in combination with the ROUGE-1 Recall score. Although not a high correlation can be obtained, the results were significant. Further research could improve the use of a penalty score.

V. Acknowledgements

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multiparty Interaction, FP6-506811, publication AMI-180).

References

[1] F. de Jong and W. Kraaij. Novalist: Content Reduction for Cross-media Browsing. In: *Proceedings of the RANLP workshop Crossing barriers in Text Summarization Research*, 2005.

[2] W. Kraaij, M. Spitters et al. Headline extraction based on a combination of uni- and multidocument summarization techniques. In: *Proceedings of DUC-2003*, 2003.

[3] R. Angheluta, R. Mitra et al. K.U.Leuven summarization system at DUC 2004. In: *DUC Workshop Papers and Agenda*, pp. 53-60, Boston, 2004.

[4] G. Erkan, D.R. Radev. The University of Michigan at DUC 2004. In: *Proceedings of the Document Understanding Conferences Boston*, MA. 2004

[5] C.Y.Lin, E. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: *Human Technology Conference. Edmonton-Canada.* p. 150-157. 2003.

[6] R. Jin. Statistical approaches towards title generation. Carnegie Mello University, 2003.

[7] K. Papineni et al. Bleu: a method for automatic evaluation of machine translation, IBM Research Report, RC22176, September 2001.

[8] C.Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL.* Barcelona, Spain. 2004