Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries

Feifan Liu, Yang Liu
The University of Texas at Dallas
Richardson, TX 75080, USA
ffliu,yangl@hlt.utdallas.edu

Abstract

Automatic summarization evaluation is critical to the development of summarization systems. While ROUGE has been shown to correlate well with human evaluation for content match in text summarization, there are many characteristics in multiparty meeting domain, which may pose potential problems to ROUGE. In this paper, we carefully examine how well the ROUGE scores correlate with human evaluation for extractive meeting summarization. Our experiments show that generally the correlation is rather low, but a significantly better correlation can be obtained by accounting for several unique meeting characteristics, such as disfluencies and speaker information, especially when evaluating system-generated summaries.

1 Introduction

Meeting summarization has drawn an increasing attention recently; therefore a study on the automatic evaluation metrics for this task is timely. Automatic evaluation helps to advance system development and avoids the labor-intensive and potentially inconsistent human evaluation. ROUGE (Lin, 2004) has been widely used for summarization evaluation. In the news article domain, ROUGE scores have been shown to be generally highly correlated with human evaluation in content match (Lin, 2004). However, there are many differences between written texts (e.g., news wire) and spoken documents, especially in the meeting domain, for example, the presence of disfluencies and multiple speakers, and the lack of structure in spontaneous utterances. The question of whether ROUGE is a good metric for meeting summarization is unclear. (Murray et al., 2005) have reported that ROUGE-1 (unigram match) scores have low correlation with human evaluation in meetings.

In this paper we investigate the correlation between ROUGE and human evaluation of extractive meeting summaries and focus on two issues specific to the meeting domain: disfluencies and multiple speakers. Both human and system generated summaries are used. Our analysis shows that by integrating meeting characteristics into ROUGE settings, better correlation can be achieved between the ROUGE scores and human evaluation based on Spearman's rho in the meeting domain.

2 Related work

Automatic summarization evaluation can be broadly classified into two categories (Jones and Galliers, 1996): intrinsic and extrinsic evaluation. Intrinsic evaluation, such as relative utility based metric proposed in (Radev et al., 2004), assesses a summarization system in itself (for example, informativeness, redundancy, and coherence). Extrinsic evaluation (Mani et al., 1998) tests the effectiveness of a summarization system on other tasks. In this study, we concentrate on the automatic intrinsic summarization evaluation. It has been extensively studied in text summarization. Different approaches have been proposed to measure matches using words or more meaningful semantic units, for example, ROUGE (Lin, 2004), factoid analysis (Teufel and Halteren, 2004), pyramid method (Nenkova and Passonneau, 2004), and Basic Element (BE) (Hovy et al., 2006).

With the increasing recent research of summarization moving into speech, especially meeting recordings, issues related to spoken language are yet to be explored for their impact on the evaluation metrics. Inspired by automatic speech recognition (ASR) evaluation, (Hori et al., 2003) proposed the summarization accuracy metric (SumACCY) based on a word network created by merging manual summaries. However (Zhu and Penn, 2005) found a statistically significant difference between the ASR-inspired metrics and those taken from text summarization (e.g., RU, ROUGE) on a subset of the Switchboard data. ROUGE has been used in meeting summarization evaluation (Murray et al., 2005; Galley, 2006), yet the question remained whether ROUGE is a good metric for the meeting domain. (Murray et al., 2005) showed low correlation of ROUGE and human evaluation in meeting summarization evaluation; however, they

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 201–204,
Columbus, Ohio, USA, June 2008.

simply used ROUGE as is and did not take into account the meeting characteristics during evaluation.

In this paper, we ask the question of whether ROUGE correlates with human evaluation of extractive meeting summaries and whether we can modify ROUGE to account for the meeting style for a better correlation with human evaluation.

3 Experimental Setup

3.1 Data

We used the ICSI meeting data (Janin et al., 2003) that contains naturally-occurring research meetings. All the meetings have been transcribed and annotated with dialog acts (DA) (Shriberg et al., 2004), topics, and extractive summaries (Murray et al., 2005).

For this study, we used the same 6 test meetings as in (Murray et al., 2005; Galley, 2006). Each meeting already has 3 human summaries from 3 common annotators. We recruited another 3 human subjects to generate 3 more human summaries, in order to create more data points for a reliable analysis. The Kappa statistics for those 6 different annotators varies from 0.11 to 0.35 for different meetings. The human summaries have different length, containing around 6.5% of the selected DAs and 13.5% of the words respectively. We used four different system summaries for each of the 6 meetings: one based on the MMR method in MEAD (Carbonell and Goldstein, 1998; et al., 2003), the other three are the system output from (Galley, 2006; Murray et al., 2005; Xie and Liu, 2008). All the system generated summaries contain around 5% of the DAs and 16% of the words of the entire meeting. Thus, in total we have 36 human summaries and 24 system summaries on the 6 test meetings, on which the correlation between ROUGE and human evaluation is calculated and investigated.

All the experiments in this paper are based on human transcriptions, with a central interest on whether some characteristics of the meeting recordings affect the correlation between ROUGE and human evaluations, without the effect from speech recognition or automatic sentence segmentation errors.

3.2 Automatic ROUGE Evaluation

ROUGE (Lin, 2004) measures the n-gram match between system generated summaries and human summaries. In most of this study, we used the same options in ROUGE as in the DUC summarization evaluation (NIST, 2007), and modify the input to ROUGE to account for the following two phenomena.

- Disfluencies

  Meetings contain spontaneous speech with many disfluencies, such as filled pauses (uh, um), discourse markers (e.g., I mean, you know), repetitions, corrections, and incomplete sentences. There have been efforts on the study of the impact of disfluencies on summarization techniques (Liu et al., 2007; Zhu and Penn, 2006) and human readability (Jones et al., 2003). However, it is not clear whether disfluencies impact automatic evaluation of extractive meeting summarization.

  Since we use extractive summarization, summary sentences may contain difluencies. We hand annotated the transcripts for the 6 meetings and marked the disfluencies such that we can remove them to obtain cleaned up sentences for those selected summary sentences. To study the impact of disfluencies, we run ROUGE using two different inputs: summaries based on the original transcription, and the summaries with disfluencies removed.

- Speaker information

  The existence of multiple speakers in meetings raises questions about the evaluation method. (Galley, 2006) considered some location constrains in meeting summarization evaluation, which utilizes speaker information to some extent. In this study we use the data in separate channels for each speaker and thus have the speaker information available for each sentence. We associate the speaker ID with each word, treat them together as a new 'word' in the input to ROUGE.

3.3 Human Evaluation

Five human subjects (all undergraduate students in Computer Science) participated in human evaluation. In total, there are 20 different summaries for each of the 6 test meetings: 6 human-generated, 4 system-generated, and their corresponding ones with disfluencies removed. We assigned 4 summaries with different configurations to each human subject: human vs. system generated summaries, with or without disfluencies. Each human evaluated 24 summaries in total, for the 6 test meetings.

For each summary, the human subjects were asked to rate the following statements using a scale of 1-5 according to the extent of their agreement with them.

- S1: The summary reflects the discussion flow in the meeting very well.
- S2: Almost all the important topic points of the meeting are represented.
- S3: Most of the sentences in the summary are relevant to the original meeting.
- S4: The information in the summary is not redundant.
- S5: The relationship between the importance of each topic in the meeting and the amount of summary space given to that topic seems appropriate.
- S6: The relationship between the role of each speaker and the amount of summary speech selected for that speaker seems appropriate.
- S7: Some sentences in the summary convey the same meaning.
- S8: Some sentences are not necessary (e.g., in terms of importance) to be included in the summary.
- S9: The summary is helpful to someone who wants to know what are discussed in the meeting.

These statements are an extension of those used in (Murray et al., 2005) for human evaluation of meeting summaries. The additional ones we added were designed to account for the discussion flow in the meetings. Some of the statements above are used to measure similar aspects, but from different perspectives, such as S5 and S6, S4 and S7. This may reduce some accidental noise in human evaluation. We grouped these statements into 4 categories: Informative Structure (IS): S1, S5 and S6; Informative Coverage (IC): S2 and S9; Informative Relevance (IRV): S3 and S8; and Informative Redundancy (IRD): S4 and S7.

4 Results

4.1 Correlation between Human Evaluation and Original ROUGE Score

Similar to (Murray et al., 2005), we also use Spearman's rank coefficient (rho) to investigate the correlation between ROUGE and human evaluation. We have 36 human summaries and 24 system summaries for the 6 meetings in our study. For each of the human summaries, the ROUGE scores are generated using the other 5 human summaries as references. For system generated summaries, we calculate the ROUGE score using 5 human references, and then obtain the average from 6 such setups. The correlation results are presented in Table 1. In addition to the overall average for human evaluation (H_AVG), we calculated the average score for each evaluation category (see Section 3.3). For ROUGE evaluation, we chose the F-measure for R-1 (unigram) and R-SU4 (skip-bigram with maximum gap length of 4), which is based on our observation that other scores in ROUGE are always highly correlated (rho$>$0.9) to either of them for this task. We compute the correlation separately for the human and system summaries in order to avoid the impact due to the inherent difference between the two different summaries.

Correlation on Human Summaries					
	H_AVG	H_IS	H_IC	H_IRV	H_IRD
R-1	0.09	0.22	0.21	0.03	-0.20
R-SU4	0.18	0.33	0.38	0.04	-0.30
Correlation on System Summaries					
R-1	-0.07	-0.02	-0.17	-0.27	-0.02
R-SU4	0.08	0.05	0.01	-0.15	0.14

Table 1: Spearman's rho between human evaluation (H) and ROUGE (R) with basic setting.

We can see that R-SU4 obtains a higher correlation with human evaluation than R-1 on the whole, but still very low, which is consistent with the previous conclusion from (Murray et al., 2005). Among the four categories, better correlation is achieved for information structure (IS) and information coverage (IC) compared to the other two categories. This is consistent with what ROUGE is designed for, "recall oriented understudy gisting evaluation" — we expect it to model IS and IC well by ngram and skip-bigram matching but not relevancy (IRV) and redundancy (IRD) effectively. In addition, we found low correlation on system generated summaries, suggesting it is more challenging to evaluate those summaries both by humans and the automatic metrics.

4.2 Impacts of Disfluencies on Correlation

Table 2 shows the correlation results between ROUGE (R-SU4) and human evaluation on the original and cleaned up summaries respectively. For human summaries, after removing disfluencies, the correlation between ROUGE and human evaluation improves on the whole, but degrades on information structure (IS) and information coverage (IC) categories. However, for system summaries, there is a significant gain of correlation on those two evaluation categories, even though no improvement on the overall average score. Our hypothesis for this is that removing disfluencies helps remove the noise in the system generated summaries and make them more easily to be evaluated by human and machines. In contrast, the human created summaries have better quality in terms of the information content and may not suffer as much from the disfluencies contained in the summary.

Correlation on Human Summaries					
	H_AVG	H_IS	H_IC	H_IRV	H_IRD
Original	0.18	0.33	0.38	0.04	-0.30
Disfluencies removed	0.21	0.21	0.31	0.19	-0.16
Correlation on System Summaries					
Original	0.08	0.05	0.01	-0.15	0.14)
Disfluencies removed	0.08	0.22	0.19	-0.02	-0.07

Table 2: Effect of disfluencies on the correlation between R-SU4 and human evaluation.

4.3 Incorporating Speaker Information

We further incorporated speaker information in ROUGE setting using the summaries with disfluencies removed. Table 3 presents the resulting correlation values between ROUGE SU4 score and human evaluation. For human summaries, adding speaker information slightly degraded the correlation, but it is still better compared to using the original transcripts (results in Table 1). For the system summaries, the overall correlation is significantly improved, with some significant improvement in the information redundancy (IRD) category. This suggests that by leveraging speaker information, ROUGE can assign better credits or penalties to system generated summaries (same words from different speakers will not be counted as a match), and thus yield better correlation with human evaluation; whereas for human summaries, this may not happen often. For similar sentences from different speakers, human annotators are more likely to agree with each

other in their selection compared to automatic summarization.

Correlation on Human Summaries					
Speaker Info.	H_AVG	H_IS	H_IC	H_IRV	H_IRD
NO	0.21	0.21	0.31	0.19	-0.16
YES	0.20	0.20	0.27	0.12	-0.09
Correlation on System Summaries					
NO	0.08	0.22	0.19	-0.02	-0.07
YES	0.14	0.20	0.16	0.02	0.21

Table 3: Effect of speaker information on the correlation between R-SU4 and human evaluation.

5 Conclusion and Future Work

In this paper, we have made a first attempt to systematically investigate the correlation of automatic ROUGE scores with human evaluation for meeting summarization. Adaptations on ROUGE setting based on meeting characteristics are proposed and evaluated using Spearman's rank coefficient. Our experimental results show that in general the correlation between ROUGE scores and human evaluation is low, with ROUGE SU4 score showing better correlation than ROUGE-1 score. There is significant improvement in correlation when disfluencies are removed and speaker information is leveraged, especially for evaluating system-generated summaries. In addition, we observe that the correlation is affected differently by those factors for human summaries and system-generated summaries.

In our future work we will examine the correlation between each statement and ROUGE scores to better represent human evaluation results instead of using simply the average over all the statements. Further studies are also needed using a larger data set. Finally, we plan to investigate meeting summarization evaluation using speech recognition output.

Acknowledgments

The authors thank University of Edinburgh for providing the annotated ICSI meeting corpus and Michel Galley for sharing his tool to process the annotated data. We also thank Gabriel Murray and Michel Galley for letting us use their automatic summarization system output for this study. This work is supported by NSF grant IIS-0714132. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

References

J. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336.

M. Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *EMNLP*, pages 364–372.

C. Hori, T. Hori, and S. Furui. 2003. Evaluation methods for automatic speech summarization. In *EUROSPEECH*, pages 2825–2828.

E. Hovy, C. Lin, L. Zhou, and J. Fukumoto. 2006. Automated summarization evaluation with basic elements. In *LREC*.

A. Janin, D. Baron, J. Edwards, D. Ellis, G. Gelbart, N. Norgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *ICASSP*.

K. S. Jones and J. Galliers. 1996. Evaluating natural language processing systems: An analysis and review. *Lecture Notes in Artificial Intelligence*.

D. Jones, F. Wlof, E. Gilbson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman. 2003. Measuring the readability of automatic speech-to-text transcripts. In *EUROSPEECH*, pages 1585–1588.

C. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out at ACL*, pages 74–81.

Y. Liu, F. Liu, B. Li, and S. Xie. 2007. Do disfluencies affect meeting summarization? a pilot study on the impact of disfluencies. In *MLMI Workshop, Poster Session.*

I. Mani, T. Firmin, D. House, M. Chrzanowski, G. Klein, L. Hirschman, B. Sundheim, and L. Obrst. 1998. The tipster summac text summarization evaluation: Final report. Technical report, The MITRE Corporation.

G. Murray, S. Renals, J. Carletta, and J. Moore. 2005. Evaluating automatic summaries of meeting recordings. In *ACL 2005 MTSE Workshop*, pages 33–40.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: the pyramid method. In *HLT/NAACL.*

NIST. 2007. Document understanding conference (DUC). *http://duc.nist.gov/.*

D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, E. Drabek, W. Lam, D. Liu, H. Qi, H. Saggion, S. Teufel, M. Topper, and A. Winkel. 2003. The MEAD Multidocument Summarizer. http://www.summarization.com/mead/.

D. R. Radev, H. Jing, M. Stys, and T. Daniel. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *SIGDAL Workshop*, pages 97–100.

S. Teufel and H. Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *EMNLP*.

S. Xie and Y. Liu. 2008. Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *ICASSP*.

X. Zhu and G. Penn. 2005. Evaluation of sentence selection for speech summarization. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

X. Zhu and G. Penn. 2006. Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. In *HLT/NAACL.*