What Are Meeting Summaries? An Analysis of Human Extractive Summaries in Meeting Corpus

Fei Liu, Yang Liu
Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
Richardson, TX, USA
{feiliu,yangl}@hlt.utdallas.edu

Abstract

Significant research efforts have been devoted to speech summarization, including automatic approaches and evaluation metrics. However, a fundamental problem about what summaries are for the speech data and whether humans agree with each other remains unclear. This paper performs an analysis of human annotated extractive summaries using the ICSI meeting corpus with an aim to examine their consistency and the factors impacting human agreement. In addition to using Kappa statistics and ROUGE scores, we also proposed a sentence distance score and divergence distance as a quantitative measure. This study is expected to help better define the speech summarization problem.

1 Introduction

With the fast development of recording and storage techniques in recent years, speech summarization has received more attention. A variety of approaches have been investigated for speech summarization, for example, maximum entropy, conditional random fields, latent semantic analysis, support vector machines, maximum marginal relevance (Maskey and Hirschberg, 2003; Hori et al., 2003; Buist et al., 2005; Galley, 2006; Murray et al., 2005; Zhang et al., 2007; Xie and Liu, 2008). These studies used different domains, such as broadcast news, lectures, and meetings. In these approaches, different information sources have been examined from both text and speech related features (e.g., prosody, speaker activity, turn-taking, discourse).

How to evaluate speech summaries has also been studied recently, but so far there is no consensus on evaluation yet. Often the goal in evaluation is to develop an automatic metric to have a high correlation with human evaluation scores. Different methods have been used in the above summarization research to compare system generated summaries with human annotation, such as F-measure, ROUGE, Pyramid, sumACCY (Lin and Hovy, 2003; Nenkova and Passonneau, 2004; Hori et al., 2003). Typically multiple reference human summaries are used in evaluation in order to account for the inconsistency among human annotations.

While there have been efforts on speech summarization approaches and evaluation, some fundamental problems are still unclear. For example, what are speech summaries? Do humans agree with each other on summary extraction? In this paper, we focus on the meeting domain, one of the most challenging speech genre, to analyze human summary annotation. Meetings often have several participants. Its speech is spontaneous, contains disfluencies, and lacks structure. These all post new challenges to the consensus of human extracted summaries.

Our goal in this study is to investigate the variation of human extractive summaries, and help to better understand the gold standard reference summaries for meeting summarization. This paper aims to answer two key questions: (1) How much variation is there in human extractive meeting summaries? (2) What are the factors that may impact interannotator agreement? We use three different metrics to evaluate the variation among human summaries, including Kappa statistic, ROUGE score, and a new proposed divergence distance score to reflect the coherence and quality of an annotation.

2 Corpus Description

We use the ICSI meeting corpus (Janin et al., 2003) which contains 75 naturally-occurred meetings, each about an hour long. All of them have been transcribed and annotated with dialog acts (DA) (Shriberg et al., 2004), topics, and abstractive and extractive summaries in the AMI project (Murray et al., 2005).

We selected 27 meetings from this corpus. Three annotators (undergraduate students) were recruited to extract summary sentences on a topic basis using the topic segments from the AMI annotation. Each sentence corresponds to one DA annotated in the corpus. The annotators were told to use their own judgment to pick summary sentences that are informative and can preserve discussion flow. The recommended percentages for the selected summary sentences and words were set to 8.0% and 16.0% respectively. Human subjects were provided with both the meeting audio files and an annotation Graphi-

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 80–83,
Columbus, June 2008.

cal User Interface, from which they can browse the manual transcripts and see the percentage of the currently selected summary sentences and words.

We refer to the above 27 meetings **Data set I** in this paper. In addition, some of our studies are performed based on the 6 meeting used in (Murray et al., 2005), for which we have human annotated summaries using 3 different guidelines:

- **Data set II**: summary annotated on a topic basis. This is a subset of the 27 annotated meetings above.
- **Data set III**: annotation is done for the entire meeting without topic segments.
- **Data set IV**: the extractive summaries are from the AMI annotation (Murray et al., 2005).

3 Analysis Results

3.1 Kappa Statistic

Kappa coefficient (Carletta, 1996) is commonly used as a standard to reflect inter-annotator agreement. Table 1 shows the average Kappa results, calculated for each meeting using the data sets described in Section 2. Compared to Kappa score on text summarization, which is reported to be 0.38 by (Mani et al., 2002) on a set of TREC documents, the inter-annotator agreement on meeting corpus is lower. This is likely due to the difference between the meeting style and written text.

Data Set	I	II	III	IV
Avg-Kappa	0.261	0.245	0.335	0.290

Table 1: Average Kappa scores on different data sets.

There are several other observations from Table 1. First, comparing the results for Data Set (II) and (III), both containing six meetings, the agreement is higher for Data Set (III). Originally, we expected that by dividing the transcript into several topics, human subjects can focus better on each topic discussed during the meeting. However, the result does not support this hypothesis. Moreover, the Kappa result of Data Set (III) also outperforms that of Data Set (IV). The latter data set is from the AMI annotation, where they utilized a different annotation scheme: the annotators were asked to extract dialog acts that are highly relevant to the given abstractive meeting summary. Contrary to our expectation, the Kappa score in this data set is still lower than that of Data Set (III), which used a direct sentence extraction scheme on the whole transcript. This suggests that even using the abstracts as a guidance, people still have a high variation in extracting summary sentences. We also calculated the pairwise Kappa score between annotations in different data sets. The inter-group Kappa score is much lower than those of the intragroup agreement, most likely due to the different annotation specifications used in the two different data sets.

3.2 Impacting Factors

We further analyze inter-annotator agreement with respect to two factors: **topic length** and **meeting participants**. All of the following experiments are based on Data Set (I) with 27 meetings.

We computed Kappa statistic for each topic instead of the entire meeting. The distribution of Kappa score with respect to the topic length (measured using the number of DAs) is shown in Figure 1. When the topic length is less than 100, Kappa scores vary greatly, from -0.065 to 1. Among the entire range of different topic lengths, there seems no obvious relationship between the Kappa score and the topic length (a regression from the data points does not suggest a fit with an interpretable trend).

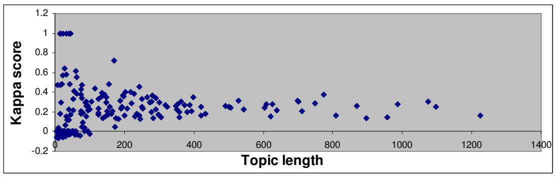

Figure 1: Relationship between Kappa score and topic length.

Using the same Kappa score for each topic, we also investigated its relationship with the number of speakers in that topic. Here we focused on the topic segments longer than a threshold (with more than 60 DAs) as there seems to be a wide range of Kappa results when the topic is short (in Figure 1). Table 2 shows the average Kappa score for these long topics, using the number of speakers in the topic as the variable. We notice that when the speaker number varies from 4 to 7, kappa scores gradually decrease with the increasing of speaker numbers. This phenomenon is consistent with our intuition. Generally the more participants are involved in a conversation, the more discussions can take place. Human annotators feel more ambiguity in selecting summary sentences for the discussion part. The pattern does not hold for other speaker numbers, namely, 2, 3, and 8. This might be due to a lack of enough data points, and we will further analyze this in the future research.

# of speakers	# of topics	Avg Kappa score
2	2	0.204
3	6	0.182
4	26	0.29
5	26	0.249
6	33	0.226
7	19	0.221
8	7	0.3

Table 2: Average Kappa score with respect to the number of speakers after removing short topics.

3.3 ROUGE Score

ROUGE (Lin and Hovy, 2003) has been adopted as a standard evaluation metric in various summarization tasks. It is computed based on the n-gram overlap between a summary and a set of reference summaries. Though the Kappa statistics can measure human agreement on sentence selection, it does not account for the fact that different annotators choose different sentences that are similar in content. ROUGE measures the word match and thus can compensate this problem of Kappa.

Table 3 shows the ROUGE-2 and ROUGE-SU4 F-measure results. For each annotator, we computed ROUGE scores using other annotators' summaries as references. For Data Set (I), we present results for each annotator, since one of our goals is to evaluate the quality of different annotator's summary annotation. The low ROUGE scores suggest the large variation among human annotations. We can see from the table that annotator 1 has the lowest ROUGE score and thus lowest agreement with the other two annotators in Data Set (I). The ROUGE score for Data Set (III) is higher than the others. This is consistent with the result using Kappa statistic: the more sentences two summaries have in common, the more overlapped n-grams they tend to share.

		ROUGE-2	ROUGE-SU4
data (I)	Annotator 1	0.407	0.457
	Annotator 2	0.421	0.471
	Annotator 3	0.433	0.483
data (III)	2 annotators	0.532	0.564
data (IV)	3 annotators	0.447	0.484

Table 3: ROUGE F-measure scores for different data sets.

3.4 Sentence Distance and Divergence Scores

From the annotation, we notice that the summary sentences are not uniformly distributed in the transcript, but rather with a clustering or coherence property. However, neither Kappa coefficient nor ROUGE score can represent such clustering tendency of meeting summaries. This paper attempts to develop an evaluation metric to measure this property among different human annotators.

For a sentence i selected by one annotator, we define a distance score d_i to measure its minimal distance to summary sentences selected by other annotators (distance between two sentences is represented using the difference of their sentence indexes). d_i is 0 if more than one annotator have extracted the same sentence as summary sentence. Using the annotated summaries for the 27 meetings in Data Set (I), we computed the sentence distance scores for each annotator. Figure 2 shows the distribution of the distance score for the 3 annotators. We can see that the distance score distributions for the three annotators differ. Intuitively, small distance scores mean better coherence and more consistency with other annotators' results. We thus propose a mechanism to quantify each annotator's summary annotation by using a random variable (RV) to represent an annotator's sentence distance scores.

When all the annotators agree with each other, the RV d will take a value of 0 with probability 1. In general, when the annotators select sentences close to each other, the RV d will have small values with high probabilities. Therefore we create a probability distribution Q for the ideal situation where the annotators have high agreement, and use this to quantify the quality of each annotation. Q is defined as:

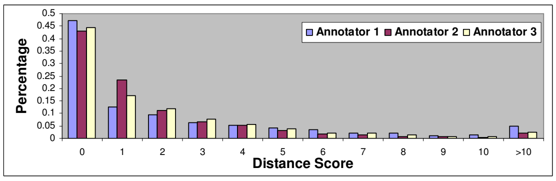

Figure 2: Percentage distribution of the summary sentence distance scores for the 3 annotators in Data Set (I).

$$Q(i) = \begin{cases} (d_{max} - i + 1) \times q & i \neq 0 \\ 1 - \sum_{i=1}^{d_{max}} Q(i) & \\ = 1 - \frac{d_{max} \times (d_{max}+1)}{2} \times q & i = 0 \end{cases}$$

where d_{max} denotes the maximum distance score based on the selected summary sentences from all the annotators. We assign linearly decreasing probabilities $Q(i)$ for different distance values i ($i > 0$) in order to give more credit to sentences with small distance scores. The rest of the probability mass is given to $Q(0)$. The parameter q is small, such that the probability distribution Q can approximate the ideal situation.

For each annotator, the probability distribution P is defined as:

$$P(i) = \begin{cases} \frac{w_i \times f_i}{\sum_i w_i \times f_i} & i \in D_p \\ 0 & \text{otherwise} \end{cases}$$

where D_p is the set of the possible distance values for this annotator, f_i is the frequency for a distance score i, and w_i is the weight assigned to that distance (w_i is i when $i \neq 0$; w_0 is p). We use parameter p to vary the weighting scale for the distance scores in order to penalize more for the large distance values.

Using the distribution P for each annotator and the ideal distribution Q, we compute their KL-divergence, called the Divergence Distance score (DD-score):

$$DD = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

We expect that the smaller the score is, the better the summary is. In the extreme case, if an annotator's DD-score is equal to 0, it means that all of this annotator's extracted sentences are selected by other annotators.

Figure 3 shows the DD-score for each annotator calculated using Data Set (I), with varying q parameters. Our experiments showed that the scale parameter p in the annotator's probability distribution only affects the absolute value of the DD-score for the annotators, but does not change the ranking of each annotator. Therefore we simply set $p = 10$ when reporting DD-scores. Figure 3 shows that different weight scale q does not impact the ranking of the annotators either. We observe in Figure 3, annotator 1 has the highest DD score to the desirable distribution. We found this is consistent with the cumulative distance score obtained from the distance score distribution, where annotator 1 has the least cumulative frequencies for all the distance values greater than 0. This is also consistent with the ROUGE scores, where annotator 1 has the lowest ROUGE score. These suggest that the DD-score can be used to quantify the consistency of an annotator with others.

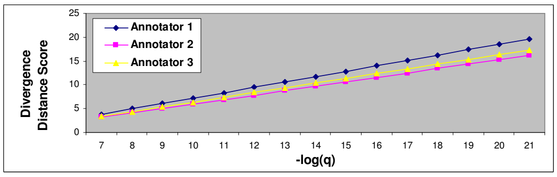

Figure 3: Divergence distance score when varying parameter q in the ideal distribution Q.

We also investigated using the sentence distance scores to improve the human annotation quality. Our hypothesis is that those selected summary sentences with high distance scores do not contain crucial information of the meeting content and thus can be removed from the reference summary. To verify this, for each annotator, we removed the summary sentences with distance scores greater than some threshold, and then computed the ROUGE score for the newly generated summary by comparing to other two summary annotations that are kept unchanged. The ROUGE-2 scores when varying the threshold is shown in Figure 4. No threshold in the X-axis means that no sentence is taken out from the human summary. We can see from the figure that the removal of sentences with high distance scores can result in even better F-measure scores. This suggests that we can delete the incoherent human selected sentences while maintaining the content information in the summary.

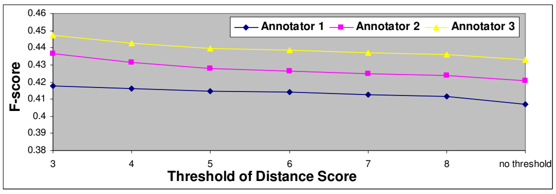

Figure 4: ROUGE-2 score after removing summary sentences with a distance score greater than a threshold.

4 Conclusion

In this paper we conducted an analysis about human annotated extractive summaries using a subset of the ICSI meeting corpus. Different measurements have been used to examine interannotator agreement, including Kappa coefficient, which requires exact same sentence selection; ROUGE, which measures the content similarity using n-gram match; and our proposed sentence distance scores and divergence, which evaluate the annotation consistency based on the sentence position. We find that the topic length does not have an impact on the human agreement using Kappa, but the number of speakers seems to be correlated with the agreement. The ROUGE score and the divergence distance scores show some consistency in terms of evaluating human annotation agreement. In addition, using the sentence distance score, we demonstrated that we can remove some poorly chosen sentences from the summary to improve human annotation agreement and preserve the information in the summary. In our future work, we will explore other factors, such as summary length, and the speaker information for the select summaries. We will also use a bigger data set for a more reliable conclusion.

Acknowledgments

The authors thank University of Edinburgh for sharing the annotation on the ICSI meeting corpus. This research is supported by NSF award IIS-0714132. The views in this paper are those of the authors and do not represent the funding agencies.

References

A. H. Buist, W. Kraaij, and S. Raaijmakers. 2005. Automatic summarization of meeting data: A feasibility study. In *Proc. of the 15th CLIN conference*.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

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

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

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

C. Y. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of HLT–NAACL*.

I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. 2002. Summac: a text summarization evaluation. *Natural Language Engineering*, 8:43–68.

S. Maskey and J. Hirschberg. 2003. Automatic summarization of broadcast news using structural features. In *Proc. of EUROSPEECH*, pages 1173–1176.

G. Murray, S. Renals, J. Carletta, and J. Moore. 2005. Evaluating automatic summaries of meeting recordings. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proc. of HLT-NAACL*.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. of 5th SIGDial Workshop*, pages 97–100.

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

J. Zhang, H. Chan, P. Fung, and L. Cao. 2007. A comparative study on speech summarization of broadcast news and lecture speech. In *Proc. of Interspeech*.