Exploring Correlation Between ROUGE and Human Evaluation on Meeting Summaries

Feifan Liu and Yang Liu, *Member, IEEE*

***Abstract*—Automatic summarization evaluation is very important to the development of summarization systems. In text summarization, ROUGE has been shown to correlate well with human evaluation when measuring match of content units. However, there are many characteristics of the multiparty meeting domain, which may pose potential problems to ROUGE. The goal of this paper is to examine how well the ROUGE scores correlate with human evaluation for extractive meeting summarization, and explore different meeting domain specific factors that have an impact on the correlation. More analysis than those in our previous work [1] has been conducted in this study. Our experiments show that generally the correlation between ROUGE and human evaluation is not great; however, when accounting for several unique meeting characteristics, such as disfluencies, speaker information, and stopwords in the ROUGE setting, better correlation can be achieved, especially on the system summaries. We also found that these factors have a different impact on human versus system summaries. In addition, we contrast the results using ROUGE with other automatic summarization evaluation metrics, such as Kappa and Pyramid, and show the appropriateness of using ROUGE for this study.**

***Index Terms*—Correlation, disfluencies, evaluation, meeting summarization, ROUGE.**

I. Introduction

SPEECH summarization can provide a convenient interface for browsing the large amount of audio data. Much work in this line has been done in recent years for different domains such as broadcast news speech, lectures, and meetings [2]–[9]. Various automatic evaluation approaches have been adopted in different previous work on speech summarization. Automatic summarization evaluation can help to advance system development and avoid the labor-intensive and sometimes inconsistent human evaluation. While many studies have been conducted on text summarization evaluation, speech summarization, especially meeting summarization, is much more recent and many questions about the automatic evaluation remain to be answered.

ROUGE [10] 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 some specific content units match [10]. 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 grammatical or topic structure in spontaneous utterances. The question of whether ROUGE is a good metric for meeting summarization is unclear. [8] has reported that ROUGE-1 (unigram match) scores have low correlation with human evaluation in meetings. However, they simply used ROUGE as is, without taking into account many characteristics of the meeting domain during evaluation.

In this paper, we investigate the correlation between ROUGE and human evaluation of extractive meeting summaries and focus on a few issues specific to the meeting domain. Both human and system generated summaries are used. We conducted human evaluation of different summaries, calculated ROUGE scores with several variations, and examined their correlation with human evaluation based on Spearman's rho. Our analysis shows that generally the correlation is low, but by integrating meeting characteristics into the ROUGE settings, better correlation can be achieved between the ROUGE scores and human evaluation for the meeting domain, especially for the system generated summaries. In addition, we found different patterns in the effect of these factors on human summaries and system generated ones. This study is a continuation of the experiments conducted in [1], with more detailed analysis of human evaluation and additional impacting factors in the meeting domain. Furthermore, in this study we also compare using the ROUGE scores and other summarization evaluation metrics, such as Kappa and Pyramid.

The rest of this paper is organized as follows. In Section II we describe some related work. Section III discusses two phenomena in the meeting domain. Section IV explains the experimental setup, including data, human evaluation, and variation in ROUGE. We present the detailed analysis results in Sections V and VI. More discussions are provided in Section VII. Conclusions and future work appear in Section VIII.

II. Related Work

Automatic summarization evaluation can be broadly classified into two categories [11]: intrinsic and extrinsic evaluation. Intrinsic evaluation, such as metrics proposed in [10][12], [13]–[15], assesses a summary in itself (for example, informativeness). Extrinsic evaluation tests the effectiveness of a summarization system on other tasks [6], [16], [17]. In this paper, we concentrate on the automatic intrinsic summarization evaluation. It has been extensively studied in text summarization and many different methods have been proposed. Similar to machine translation evaluation, methods based on word N-gram

Manuscript received April 28, 2008; revised October 18, 2008. First published June 10, 2009; current version published October 23, 2009. This work was supported by the National Science Foundation (NSF) under Grant IIS-0714132. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ruhi Sarikaya.

The authors are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2025096

matches between system generated summaries and human summaries have been developed, such as ROUGE [10], which has been used in various summarization tasks (e.g., document understanding conference, DUC [18]). Other approaches have been further proposed in order to consider more meaningful semantic units instead of words, such as factoid analysis [13], the Pyramid method [14], and Basic Element (BE) [15]. Some of these metrics may require human involvement in finding the semantic units, e.g., Pyramid [14].

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 automatic evaluation metrics. Some methods have been developed to take into account recognition errors or to use recognition output and its posterior probability [8], [19], [20]. Inspired by automatic speech recognition (ASR) evaluation, [19] proposed a summarization accuracy metric (SumACCY) based on a word network created by merging manual summaries. [20] compared ASR-oriented evaluation metrics with text summarization based evaluation (e.g., ROUGE), showing considerable differences in the relative rank of different systems between the two different metrics on a subset of the Switchboard data.

For meeting summarization, different evaluation methods have been adopted, including ROUGE, Pyramid (with location constraint), and weighted precision [7]–[9]. Murray *et al.* [8] showed low correlation of ROUGE and human evaluation in meeting summarization evaluation; however, they simply used ROUGE as is and did not take into account the meeting characteristics during evaluation. Thus, it is still unclear whether ROUGE is a good metric for the meeting domain. In our previous work [1], we studied how ROUGE correlates with human evaluation of extractive meeting summaries and whether we can modify ROUGE settings to account for the meeting style for a better correlation with human evaluation. This current study incorporates more analysis of human evaluation results, new human evaluation setup, additional factors on the correlation between human evaluation and ROUGE scores, and comparisons with other automatic evaluation metrics.

III. Characteristics of Meeting Recordings

Meetings differ from other types of speech and written text in many aspects. Below, we briefly describe two of them that we expect to have an impact on summarization evaluation, and will investigate how they affect the correlation between ROUGE and human evaluation of meeting summaries in the following sections.

A. Disfluency

Disfluencies occur frequently in the meeting domain where people talk spontaneously. In this paper, we use the term "disfluencies" to mean a broad range of phenomena, including filled pauses (*uh, um*), discourse markers (e.g., *I mean, you know*), repetitions, corrections, and incomplete sentences. They are problematic to many natural language processing tasks, since most of those techniques have been developed to deal with written text. There have been efforts studying the impact of disfluencies on summarization techniques [8], [21], [22], and human readability [23]. However, it is not clear whether disfluencies impact automatic evaluation of extractive meeting summarization.

B. Multiparty Participation

Unlike monologue speech, generally there are several people participating in one meeting, and participants may have different roles in the meeting. The existence of multiple speakers raises questions to the summarization process as well as the evaluation method. A simple question related to the existence of multiple speakers is, if the same words are spoken by different speakers, should they be treated the same during evaluation? We hypothesize that using speaker information can help the automatic evaluation on meeting summaries. [7] considered some location constraints in meeting summarization evaluation—the units need to be extracted from the same location in the original meeting for them to be equivalent in evaluation, which is a more strict condition than just using speaker identity.

IV. Experimental Setup

A. Data

We use six meetings from the ICSI meeting data [24]. This corpus contains naturally occurring meetings, each about an hour long. The number of speakers in these six meetings ranges from 4 to 9. The speaking style is very informal and the topics are mainly research discussions in the area of natural language processing, artificial intelligence, speech, and networking. All the meetings have been transcribed and annotated with dialog acts (DA) [25], topics, and abstractive and extractive summaries [8]. An excerpt example from a meeting transcript in this corpus is shown in Fig. 1, where each line corresponds to one DA based on the DA annotation in the corpus. We do not show case information because that will not be available when using speech recognition output. Speaker IDs are shown on the left for each DA (underlined in the example). This example clearly illustrates the phenomena in the meeting domain as discussed in Section III (e.g., disfluencies, multiple speakers).

The six meetings used in this study are the same as those in other prior work [7], [8]. All the summaries considered in this paper are extractive meeting summaries, i.e., sentences (DAs to be exact) selected from the meeting transcripts without any rephrasing or compression. We used the extractive summary annotation obtained from the University of Edinburgh, annotated as part of the AMI project.[1] There were three common annotators for the six meetings. We recruited another three human subjects to generate three more human summaries, in order to create more data points for a reliable analysis. The Kappa statistics for those six different annotators varies from 0.11 to 0.35 for different meetings. The human summaries have different length, containing around 5.6% of the selected DAs and 11.1% of the words respectively. We used four different system summaries for each of the six meetings: one based on the MMR method in MEAD [12], [26], the other three are the system output from [7]–[9], respectively. The system generated summaries contain around 4.1% of the DAs and 11.6% of the words. The ROUGE-1

[1]AMI stands for Augmented Multi-Party Interaction. For more information on the project, see http://www.amiproject.org/.

Excerpt of a meeting transcript

me011: i see what you mean
fe008: and that'll be something like i- it's ver- it's interesting
me011: a backchannel or
fe008: once in a while it's a backchannel
mn014: yep [laugh]
fe008: sometimes it seems to be um similar to the ones that are being picked up
me011: [laugh] mm-hmm
fe008: and they're rare events but you can really go through a meeting very quickly
fe008: you just you just you know yo- you s- you scroll from screen to screen looking for blips
fe008: and i think that we are gonna end up with uh better coverage of the backchannels
me013: yeah
fe008: but at the same time we're benefiting tremendously from the pre-segmentation

Fig. 1. Excerpt of a meeting transcript.

F-measures for these system summaries are around 0.7. In total, we have 36 human summaries and 24 system summaries on the six meetings. We will use these to investigate the correlation between ROUGE and human evaluation.

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

B. ROUGE Setup for Automatic Evaluation

Our goal in this study is to investigate how the meeting characteristics we discussed in Section III can be taken into account to adapt ROUGE settings and affect the correlation between human evaluations and ROUGE scores. In most of our experiments, we used the same options in ROUGE as in the DUC summarization evaluation [18], i.e., computing the unigram ($-n$ 1), bigram ($-n$ 2) and skip bigram with the maximum gap length of 4 between two words (-2 4); including the unigram while computing the skip bigram ($-u$); stemming summaries using Porter stemmer before computing various statistics ($-m$); averaging over the sentence unit ROUGE scores ($-t$ 0); assigning equal importance to precision and recall ($-p$ 0.5); computing statistics in the confidence level of 95% ($-c$ 95) based on sampling points of 1000 in bootstrap resampling ($-r$ 1000).

We investigate the following three variations for ROUGE evaluation.

- *Removing disfluencies*: Since we use extractive summarization at the sentence level, summary sentences in the original transcripts may contain difluencies. We hand annotated the transcripts for the six meetings and marked the disfluencies such that we can remove them to obtain cleaned up sentences for both human and system selected summary sentences. To study the impact of disfluencies on the correlation, we run ROUGE using two different inputs: summaries based on the original transcription, and the summaries with disfluencies removed.
- *Using speaker information*: In this paper, we use the data with separate channels for each speaker and thus have speaker information available for each sentence. We associate the speaker ID with each word and treat them together as a new "word" in the input to ROUGE. In this way the same word from different speakers will be counted differently. This is a simple way to incorporate specific speaker information in ROUGE evaluation. Ultimately we would like to capture the discussion flow among different participants in a meeting; however, that is still an open question.
- *Filtering unimportant stopwords*: ROUGE allows using a stopword list to filter some unimportant words during evaluation. We hypothesize that in the meeting domain, different DAs may share a lot of frequently occurring words which will have an impact on the automatic evaluation. The default stopword list provided in the ROUGE package was created based on news wire text. For comparison, in this study we also automatically created another meeting domain dependent stopword list, using the rest of the meetings in the ICSI meeting corpus (69 meetings). We split the 69 training meetings into 463 new documents according to the topic segmentation information annotated in the corpus, and then used these documents to calculate the idf value (inverse document frequency) [27] for each word

$$\mathrm{idf}(w_i) = \log(N/n_i) \tag{1}$$

where n_i denotes the number of the documents containing word w_i, and N is the total number of the documents in the collection. A higher idf value of a word means that this word appears in fewer documents and thus is more indicative of a specific topic for a document, whereas a low idf value indicates a word appears in many documents and is not topic indicative. Examples of stopwords are "a, the, so, ok, ...". We selected 600 words with the lowest idf scores to use as our meeting domain stopwords, a size equivalent to the one used in the ROUGE package.

C. Human Evaluation

Five human subjects (all undergraduate students in Computer Science) participated in human evaluation. There are 20 different summaries for each of the six test meetings: six human-generated and four system-generated using the original transcripts, and their corresponding ones with disfluencies removed. We assigned four summaries from a meeting with different configurations to each human subject: human versus system-generated summaries, with and without disfluencies. Each human subject evaluated 24 summaries in total, for the six 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 avoids redundancy.
- 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 is discussed in the meeting.

These statements were an extension of those used in [8] for human evaluation of meeting summaries. We added three more statements (S1, S6, and S9) in the hope of accounting for the discussion flow in meetings. In addition, during human evaluation, we conducted a survey on how much disfluencies in the summaries affect the subjective evaluation using the following two statements:

- S10: The disfluencies affect the readability and your comprehension of the summary very much in your evaluation.
- S11: The disfluencies affect very much the scores you give to the evaluation statements.

V. Results: Analysis of Human Evaluation and Baseline Correlation Between Human Evaluation and ROUGE

In this section, we first analyze the validity of the human evaluation used in our study, and then study the baseline correlation between human evaluation and the ROUGE score. We will focus on the effect on the correlation when incorporating different meeting characteristics in Section VI.

A. Analysis of Human Evaluation Results

Since the focus of this study is to measure the correlation of ROUGE score and human evaluation, we need to ensure that the human evaluation criterion we used is sound and reliable. Then we can proceed to study if automatic metrics correlate well with human evaluation and thus approximate such kind of human evaluation. Deciding whether the questions we use in the human evaluation (in Section IV-C) are proper is not straightforward. It may need discussions from the research community. Additionally, for the summarization task, selecting an evaluation metric may depend on the application of the summaries. The study in this paper is only a first step along these directions.

For the purpose of making sure our human evaluation approach is reasonable, we performed a sanity check using the following criteria. Intuitively, we would expect human summaries to have better quality than system generated ones. Therefore, we first verify if our human evaluation metric is able to distinguish between human and automatic summaries in a relatively reliable way. We selected four human summaries for each meeting, such that we have the same number of human and system summaries for the six test meetings. For these 24 pairs of summaries, we measure the human evaluation scores for each individual statement $(S_n, \mathrm{n} = 1, 2, \ldots 9)$ and their average (AVG). For the negative statements, such as S7 and S8, their scores were the maximum score value (i.e., 5) minus the original score from human subject. Then we applied a paired statistical t-test to see if human summaries can obtain statistically better scores than system ones.

TABLE I
Statistical T-Test Results for Human Evaluation Scores Using 24 Pairs of Human Summaries and System Summaries

	$score_h - score_s$	p-value
S1	0.875	0.003
S2	0.5	0.028
S3	0.083	0.403
S4	-0.083	0.615
S5	0.375	0.133
S6	0.292	0.198
S7	-0.292	0.841
S8	-0.167	0.68
S9	0.542	0.048
AVG	0.236	0.146

Table I shows the statistical test results, where $\mathrm{score\_}h - \mathrm{score\_}s$ is the difference of the human evaluation scores for the human summaries and the system summaries, averaged over the 24 summary pairs. Our null hypothesis $H0$ in this test is that the mean scores for the human and system summaries are equal. The corresponding p-value in Table I indicates the probability of obtaining the difference if the null hypothesis is true. The lower the p-value is, the more likely that $H0$ is wrong, suggesting more likely the human evaluation metrics can distinguish human versus system summaries. We can see that the scores are significantly different for statement S1, S2 and S9 (where $p < 0.05$), but for statements S4, S7, and S8, the p-value is rather high $(p > 0.5)$. There are a few possible reasons for the negative results for statement S4, S7, and S8: these statements themselves are not able to distinguish human and system summaries; the human subjects have problems making the judgment properly in those aspects; or the system summaries and the human summaries are not very different in terms of information relevance and redundancy evaluated in those statements.

Overall, the p-value for the average score of the nine statements is 0.146 (last row in Table I), which seems reasonable but not significant at $p < 0.05$ level to distinguish between human and system summaries. Hence, we performed a few other statistical tests in order to obtain a more reliable measurement for human evaluation before studying its correlation with ROUGE. Various results are shown in Table II. (A) is the same average of the nine statements as shown in Table I. (B) is the average of the six statements, excluding S4, S7, and S8, based on the analysis from Table I. (C) uses only those three statements with $p < 0.05$. (D) is a slightly modified average, that is, it first calculates the average for the pairs (S3, S8) and (S4, S7), and then the average of all the scores. We expect this can avoid extra weights on the statements that measure the same aspect. Rows E-G correspond to using the six statements used in previous work [8], i.e., S2, S3, S4, S5, S7, and S8 in our work. (E) is the average of these six statements. (F) removes S4, S7, and S8, thus leaving

TABLE II
STATISTICAL T-TEST ANALYSIS ON DIFFERENT HUMAN EVALUATION MEASUREMENTS

	statements	$score_h - score_s$	p-value
(A)	1 2 3 4 5 6 7 8 9	0.236	0.146
(B)	1 2 3 5 6 9	0.444	0.042
(C)	1 2 9	0.639	0.008
(D)	1 2 3 4 5 6 7 8 9 modified average	0.336	0.073
Murray et al. [8] i.e., S2, S3, S4, S5, S7, S8			
(E)	2 3 4 5 7 8	0.069	0.384
(F)	2 3 5	0.319	0.119
(G)	2 3 4 5 7 8 modified average	0.162	0.245

only the other three statements. (G) is the modified version of the average over the six statements, similar to (D).

We observe from Table II that the six statements after removing S4, S7, and S8 (row B) show a much more reliable measurement based on the statistical test analysis ($p < 0.05$) than using all the nine statements. As expected, using only those three statements with an individual low p-value (row C) yields better distinguishing ability between human and system summaries according to the t-test results. The modified version of the average scores also improves the significance test. Regarding the six statements used in previous work [8], we can see that using those did not provide a reasonable measurement of human evaluation, and that using the modified method for calculating the average scores and removing the three statements that perform poorly in the t-test helped better differentiate between human and system summaries. Overall, the average score is significantly better using the statements used in our study compared to those used in prior work, indicating that adding the three new statements we proposed is helpful for human evaluation of meeting summarization. Murry *et al.* [8] did not conduct such an analysis for human evaluation; however, we noticed that the three statements with negative correlation with ROUGE-1 based on their results (i.e., informative-2, informative-5, and informative-6) correspond to the three statements in our experiments (S4, S7, and S8) which are less effective based on our significance test. This suggests there is some consistency between our human evaluation results and theirs.

Based on the above results, we will use the six statements excluding S4, S7, and S8, referred as "H_AVG" hereafter, and the three statements (S1, S2, and S9), referred as "H_AVG-3", to explore correlation of human evaluation with ROUGE in the following experiments. In addition, based on the aspects those human evaluation statements capture, we further group the remaining six statements into the following categories, in order to measure the correlation for different aspects.

- Informative Structure (IS): S1, S5 and S6.
- Informative Coverage (IC): S2 and S9.
- Informative Relevance (IRV): S3.

B. Analysis of ROUGE Scores

Since ROUGE can generate different measurements, we first analyze their relationship to choose appropriate measurements for the following study on its correlation with human evaluation.

TABLE III
CORRELATION (SPEARMAN'S RHO) BETWEEN DIFFERENT ROUGE SCORES

spearman's rho	R-1	R-2	R-L	R-SU4
R-1	-	0.719	0.999	0.838
R-2	0.719	-	0.716	0.95
R-L	0.999	0.716	-	0.834
R-SU4	0.838	0.95	0.834	-

TABLE IV
CORRELATION (SPEARMAN'S RHO) BETWEEN HUMAN EVALUATION (H) AND ROUGE (R) WITH BASIC SETTING

Correlation for Human Summaries					
	H_AVG	H_AVG-3	H_IS	H_IC	H_IRV
R-1	0.225	0.206	0.222	0.201	0.027
R-2	0.386	0.45	0.344	0.419	0.095
R-SU4	0.355	0.404	0.326	0.375	0.042
Correlation for System Summaries					
R-1	-0.125	-0.149	-0.016	-0.17	-0.272
R-2	-0.021	-0.048	-0.025	-0.021	-0.186
R-SU4	0.03	-0.008	0.045	0.012	-0.146

We calculated the F-measures for R-1 (unigram), R-2 (bigram), R-L (longest sequence match), and R-SU4 (skip-bigram with maximum gap length of 4, including unigram as well) with the basic setting as described in Section IV-B. Table III shows the correlation (measured using Spearman's rho) between different measurements using human summaries. Overall, R-1 and R-L are highly correlated, and R-2 and R-SU4 also show good correlation. The same trend is also observed using the 24 system summaries. Therefore, we choose to use R-1, R-2 and R-SU4 for the following correlation experiments. These are also the measurements frequently used in DUC. In addition, due to the different length of the human and system summaries, we will use F-measures, instead of recall or precision values.

C. Correlation Between Human Evaluation and Original ROUGE Scores

Similar to [8], we use Spearman's rank coefficient (rho) to investigate the correlation between ROUGE and human evaluation. Results are presented in Table IV using 36 human summaries and 24 system summaries for the six test meetings in this study. We show the results for the average score over the six statements (H_AVG) and three statements (H_AVG-3) for human evaluation (see Section IV-C), as well as for each evaluation category. We compute the correlation separately for the human and system summaries in order to avoid the bias due to the inherent difference between the two kinds of different summaries. When calculating the ROUGE scores, for a human-generated summary, we used each of the other five human summaries as a reference and computed the average. For the system-generated summaries, it is the average based on the six human references.

We can see that R-2 and R-SU4 obtain a higher correlation with human evaluation than R-1 on the whole. This is expected since R-2 and R-SU4 can capture long-distance relation so that more useful information is considered when comparing two summaries. The best result on human summaries is up to 0.386 between R-2 and H_AVG. The correlation on the system summaries is very low on the whole, which is consistent with the previous conclusion from [8], suggesting that it is more

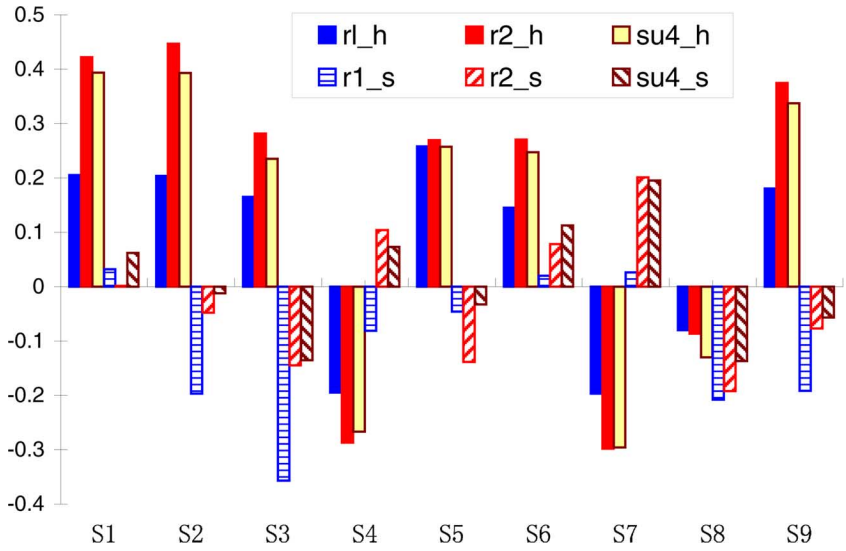


Fig. 2. Correlation results between ROUGE (R-1, R-2 and R-SU4) and the nine individual evaluation statements using human and system summaries.

challenging to evaluate those system summaries by humans as well as automatic metrics. The results indicate that it is not a good choice to use ROUGE as is to get a reasonable evaluation, at least on the meeting data. This seems to be different from previous results on other data sets [10]. We will further discuss this in Section VII.

Among the three categories, better correlation is achieved for information structure (IS) and information coverage (IC) compared to the information relevancy category (IRV). This indicates that ROUGE can model IS and IC well by n-gram and skip-bigram matching but not relevancy (IRV). This is also consistent with the earlier results in Table I, where we have seen that human evaluation of the IRV category does not show significant difference between human and system summaries. For the two average human evaluation scores, H_AVG and H_AVG-3, in general they show similar correlation with various ROUGE scores.

For a more detailed picture of the correlation between ROUGE and human evaluation, we calculated Spearman's rho between ROUGE scores and each individual statement. Fig. 2 shows these results. Consistent with the analysis in Section IV-C, the correlation between ROUGE scores and human evaluation on S4, S7, and S8 is very low. In addition, we observe some mixed results for S2, S3, S5, and S9, for example, much better correlation in human summaries than system summaries. This supports the general better correlation results on human summaries than on the system ones in Table IV. Same as for the average results, we notice that R-2 and R-SU4 show more reliable correlation with most evaluation statements than R-1, and for most evaluation statements there is a better correlation on human summaries than system ones.

VI. Effects of Incorporating Meeting Characteristics on Correlation Between ROUGE and Human Evaluation

We hypothesize that one of the reasons for the low correlation between human evaluation and ROUGE scores is the meeting style, and thus we believe that some modification is needed for a better correlation in the meeting domain. In the following experiments, for human evaluation, we use H_AVG since its corresponding p value is good and it covers more aspects than using three statements (H_AVG-3). For ROUGE scores, we use F-measure results for R-1 (unigram match) and R-2 (bigram match).

TABLE V
Correlation Results After Removing Disfluencies

Correlation for Human Summaries					
ROUGE	Summary	H_AVG	H_IS	H_IC	H_IRV
R1	Original	0.225	0.222	0.201	0.027
	Cleaned	0.121	0.002	0.135	0.096
R2	Original	0.386	0.344	0.419	0.095
	Cleaned	0.29	0.211	0.313	0.151
Correlation for System Summaries					
R1	Original	-0.126	-0.016	-0.17	-0.272
	Cleaned	0.11	0.178	0.16	-0.075
R2	Original	-0.021	-0.025	-0.021	-0.186
	Cleaned	0.049	0.136	0.124	0.009

A. Impacts of Disfluencies on Correlation

Table V shows the correlation results between ROUGE scores (R-1 and R-2) and human evaluation on the original and cleaned up summaries respectively. For human summaries, after removing disfluencies, the correlation between ROUGE and human evaluation degrades on the whole as well as for individual categories. However, for system summaries, there is a significant gain of correlation between ROUGE and average human scores, as well as all the other categories. Our hypothesis for the different patterns between human and system summaries is that removing disfluencies helps more to remove the noise in the system generated summaries and makes them easier 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.

As mentioned in Section IV-C, during human evaluation we also asked human subjects for their opinions about how disfluencies affect their evaluation of the human and system sum-

TABLE VI
EFFECTS OF SPEAKER INFORMATION ON THE CORRELATION BETWEEN ROUGE AND HUMAN EVALUATION OF CLEANED UP MEETING SUMMARIES

Using cleaned transcripts					
Correlation for Human Summaries					
ROUGE	W/O Spkr	H_AVG	H_IS	H_IC	H_IRV
R1	No	0.121	0.002	0.135	0.096
	Yes	0.197	0.118	0.183	0.112
R2	No	0.29	0.211	0.313	0.151
	Yes	0.251	0.187	0.268	0.11
Correlation for System Summaries					
R1	No	0.11	0.178	0.16	-0.075
	Yes	0.167	0.271	0.216	-0.014
R2	No	0.049	0.136	0.124	0.009
	Yes	0.104	0.195	0.154	0.005

maries. Using the original transcripts (which contain disfluencies), the average human subject score for Statement 10 is 2.39 for human summaries and 2.83 for system summaries; and the average score of S11 is 1.81 for human summaries and 2.13 for system summaries. This suggests that disfluencies impact human judgment more for system summaries than human summaries. In addition, as expected, disfluencies have a greater effect on readability of summaries and the difficulty to evaluate them than the actual score given to the summaries by the human evaluators (higher scores for S10 than S11).

B. Incorporating Speaker Information

We first examined the individual contribution of adding speaker information on the original transcripts; however, it decreased correlation for most of the conditions, except some slight improvement for R-1 using human summaries. Hence the next question we investigated is whether it helps on the cleaned transcripts. Table VI presents the resulting correlation values between ROUGE score and human evaluation.

For human summaries, adding speaker information slightly degrades the correlation in term of R-2, but still yields some improvement on the correlation of R-1 both on average and for individual categories. On system summaries, consistent improvement is observed for both R-2 and R-1, suggesting 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. Therefore, on the whole the improvement of correlation yielded by adding speaker information on the human summaries is not as significant as on system summaries. Overall, even though using speaker information itself does not help, combining with disfluency removal results in better correlation, especially for the system summaries.

C. Effect of Different Stopword Lists

First we employed the two stopword lists discussed in Section IV-B (one in ROUGE and another one generated using the meeting corpus) on the original transcripts to evaluate the impact of the stopwords factor by itself. We found that it did not improve the correlation for the system summaries, but yielded better correlation on the human summaries—correlation with H_AVG improved from 0.225 to 0.283 for R-1 (using ROUGE stopwords), and 0.386 to 0.393 for R-2 (using meeting stopwords). This is different from the two previous factors, where they both degraded correlation for human summaries.

TABLE VII
IMPACT OF DIFFERENT STOPWORD (SW) LISTS ON THE CORRELATION BETWEEN ROUGE AND HUMAN EVALUATION OF CLEANED UP MEETING SUMMARIES

Using cleaned transcripts					
Correlation for Human Summaries					
ROUGE	StopWord	H_AVG	H_IS	H_IC	H_IRV
R1	No	0.121	0.002	0.135	0.096
	R-SW	0.298	0.179	0.294	0.212
	Mtg-SW	0.268	0.169	0.241	0.141
R2	No	0.29	0.211	0.313	0.151
	R-SW	0.392	0.314	0.436	0.189
	Mtg-SW	0.395	0.323	0.404	0.212
Correlation for System Summaries					
R1	No	0.11	0.178	0.16	-0.075
	R-SW	0.14	0.294	0.16	-0.129
	Mtg-SW	0.018	0.175	0.026	-0.221
R2	No	0.049	0.136	0.124	0.009
	R-SW	-0.013	0.112	0.036	-0.024
	Mtg-SW	-0.045	0.082	-0.048	-0.017

Second we evaluated joint effect of disfluencies and stopwords. The reason we chose to combine stopwords and disfluencies rather than with speaker information is because using speaker information is not helpful by itself, whereas removing disfluencies has positive effect on system summaries, a pattern that is different from stopwords. We used the stopwords on the cleaned up summaries and computed their correlation with human evaluation of the corresponding cleaned up summaries. Results of this experiment are shown in Table VII. "R-SW" and "Mtg-SW" indicates the standard stopwords in ROUGE and the domain specific stopwords for meeting transcripts, respectively.

As we can see, there seems to be some difference between the two different stopword lists. In general, using domain specific stopwords did not yield any gain compared to using the ROUGE stopwords, contrary to our expectation. We also varied the number of stopwords for the meeting domain, but found no significant difference when using a different size of stopword list. From Table VII, we observe that combining the removal of disfluencies and using stopwords has better results than each individual approach and we achieve the best correlation on human summaries: the correlation between R-1/R-2 and human evaluation increase to 0.298/0.395 compared to the baseline results in Table IV. Using stopwords on the cleaned up transcripts did not consistently help with the correlation on the system summaries. When using the ROUGE stopwords, it increases correlation for R-1, however, neither stopword lists improves result for R-2 scores. The effect of stopwords is different from the patterns using the other two adaptations: adding speaker information and removing disfluencies, which helps on system summaries but decreases correlation on the human condition.

Finally, we investigate if the three factors are complementary and can yield further improvement on the correlation between ROUGE and human evaluation. Table VIII shows the comparison between the two stopword lists, combining with

TABLE VIII
CORRELATION RESULTS WHEN COMBINING STOPWORD LISTS WITH ADDING SPEAKER AND REMOVING DISFLUENCIES

Using cleaned transcripts, with speaker info					
Correlation for Human Summaries					
ROUGE	StopWord	H_AVG	H_IS	H_IC	H_IRV
R1	No	0.197	0.118	0.183	0.112
	R-SW	0.166	0.085	0.16	0.073
	Mtg-SW	0.183	0.1	0.166	0.11
R2	No	0.251	0.187	0.268	0.11
	R-SW	0.221	0.155	0.246	0.101
	Mtg-SW	0.238	0.171	0.241	0.167
Correlation for System Summaries					
R1	No	0.167	0.271	0.216	-0.014
	R-SW	0.213	0.326	0.238	-0.04
	Mtg-SW	0.18	0.296	0.198	-0.035
R2	No	0.104	0.195	0.154	0.005
	R-SW	0.243	0.345	0.269	0.011
	Mtg-SW	0.242	0.359	0.255	-0.058

the other two factors: adding speaker and removing disfluencies. We can see that adding stopwords degrades correlation on the human summaries. Therefore, the best result so far for the human summaries is achieved by using stopwords and removing disfluencies (i.e., 0.298 for R-1 and 0.395 for R-2, as shown in Table VII). However, we found that the correlation on the system summaries can be further improved by involving those modifications in the ROUGE setting that we investigated in this paper. This suggests that there is an effective interaction among the three factors. The combination of the three approaches yields the best results for the system summaries, i.e., 0.213 for R1 and 0.243 for R2, compared to the baseline results of -0.125 for R1 and -0.021 for R2 (Table IV).

VII. DISCUSSION

A. Remarks on ROUGE Scores

Our experimental results so far have shown that when considering the meeting characteristics, ROUGE can get a better correlation with human evaluation. For R-2, the best results are 0.395 and 0.243 on the human and system summaries, respectively. These correlations are not very high, but we think they are acceptable. Note that ROUGE was developed in order to approximate a particular kind of evaluation used in DUC, where annotators took a human generated summary, broke down into clauses, and matched the occurrence of each clause with content of peer summaries. The insight in ROUGE was that this matching can be done on word level and it will still work. ROUGE has been shown to work well only by the means of correlating ROUGE scores with this matching of clauses [10]. No claim has been made that ROUGE correlates well with human evaluation in general. Therefore, we cannot expect that ROUGE can approximate what a human subject does in the general evaluation, especially in meeting domain, which we believe is even more challenging than evaluating summarization for written text.

The correlation results in this paper are computed based on each single summary for each test document (i.e., meetings in this study). This is also different from what was used in [10], where each summarization system was used as a sample for the correlation study. For a comparison, we also calculated the correlation that way, first taking an average of the scores for the six test meetings, and then using the four systems and six human annotators as the samples. We used the average of the six human evaluation statements for this experiment. For the baseline setup (using the original transcripts as is), we obtained a correlation of 0.543 and 0.657 for R1 and R2, respectively, for the human summaries, and 0.4 and -0.4 for the system summaries. When incorporating different factors we considered in this paper, we also observed some improvement over the baseline. The best correlation results were yielded by different configurations, for example, 0.714 for R1 for the human summaries when stopwords are removed (no change for R2), 0.8 for R1 for system summaries when disfluencies are removed and speaker information is used, and 0.4 for R2 for system summaries when using the original transcripts with speaker information. These are significantly higher than the results we have shown earlier in this paper. However, as pointed out in [10], the correlation scores measured this way may not be stable when the number of test documents for each system is small, which is the case in our evaluation. We thought that it is more reasonable to investigate the correlation between ROUGE and human evaluation using individual test documents as we did in this paper, even though this will naturally lead to a lower correlation score as there are more data points compared with that computed when averaging over all the test documents for each system. When there are more test documents, it may be a better solution to take the average of the system and human evaluation scores over all the samples and then measure correlation. This is also more likely to yield a better correlation score than those reported in this paper.

B. Comparing to Other Evaluation Metrics

There exist other metrics for meeting summarization (or speech summarization in general). Since our task is focused on the extractive summaries, one possible metric is recall, precision, and F-measure, or Kappa coefficient based on sentence level match. However, these measurements have some weakness, as some sentences may contain similar information but do not necessarily match exactly with reference summary sentences. Another popular automatic summarization evaluation, Pyramid, has also been used in meeting summarization. In [7], an intuitive adaptation of Pyramid is proposed, resulting in better discriminative power, but the correlation of that metric with human evaluation is not clear.

The question we ask here is if ROUGE is a reasonable metric for this study. Using our human evaluation results, we calculated their correlation with other automatic evaluation methods, i.e., Kappa and Pyramid. Table IX shows the correlation results. We used the adapted Pyramid method of [7]. The summary content unit (SCU) is defined as a word and its document location (represented using sentence ID). To some extent, this is similar to our setting when adding speaker information in ROUGE. Therefore, for a better comparison with Pyramid, in Table IX we present the correlation of ROUGE (R-2) by integrating speaker information. For both ROUGE and Pyramid, we show results using the original transcripts and the cleaned transcripts condition (removing disfluencies). For Kappa measurement, there is

TABLE IX
COMPARISON AMONG DIFFERENT AUTOMATIC SUMMARIZATION EVALUATION METHODS ON THE CORRELATION WITH HUMAN EVALUATION

Correlation for Human Summaries					
		H_AVG	H_IS	H_IC	H_IRV
original	R-2	0.361	0.345	0.382	0.129
	Pyramid	0.102	0.134	0.037	-0.032
cleaned	R-2	0.251	0.187	0.268	0.11
	Pyramid	0.257	0.253	0.247	0.083
Kappa		0.35	0.38	0.33	0.23
Correlation for System Summaries					
original	R-2	-0.083	-0.042	-0.071	-0.264
	Pyramid	-0.177	-0.225	-0.108	-0.221
cleaned	R-2	0.104	0.195	0.154	0.005
	Pyramid	0.072	0.115	0.106	0.144
Kappa		-0.12	-0.12	-0.08	-0.23

only one result as removing disfluencies does not change the sentence-level match.

Table IX shows a better correlation for ROUGE than Pyramid on the human summaries with the basic setting and on the system summaries with the disfluencies removed. In addition, we see that removing disfluencies yields a better correlation for Pyramid on both human and system summaries. Even our study is for extractive summarization evaluation, the comparison between Kappa and Pyramid or ROUGE indicates that by considering other aspects of meetings, Pyramid or ROUGE can achieve a higher correlation with human evaluation than Kappa (that does sentence-level matching), especially on the system generated summaries. We also examined the correlation using R-SU4 and H_AVG-3 as the ROUGE score and human evaluation score respectively under all the settings. The results show similar trends as R-2 and H_AVG and thus are not presented here. These comparison results suggest that it is a reasonable choice to use ROUGE to investigate its correlation with human evaluation in this study. However, based on the results in the table, it seems that none of the original versions of the three metrics above has a high correlation, suggesting that any summarization approach optimizing on those metrics may not generate good summaries based on human evaluations. On the other hand, as mentioned in Section VII-A, taking an average over the test documents yields better correlation scores, which is often how correlation is measured between automatic metrics and human evaluation. Therefore, like in text summarization, these metrics may still be used for speech summarization in practice, but we may need to pay more attention to what a summary is for speech input and the characteristics in the speech genre.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we have systematically investigated the correlation of automatic ROUGE scores with human evaluation for extractive meeting summarization using a subset of the ICSI meeting corpus. 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-2 and ROUGE-SU4 scores showing better correlation than ROUGE-1 score. We have evaluated the impact of three factors individually and by combination, disfluencies, speaker information, and stopwords. Even though using one adaptation does not often improve correlation, there is complementary effect from them and their combination yields significant improvement. We also observe difference between using human summaries and system generated summaries. The best correlation for system summaries is achieved when the three factors are all accounted for, and the best result on the human summaries is obtained when removing disfluencies and using stopwords. There is no significant difference between different stopword lists.

In our future work, we will consider different information needs when using meeting summaries and investigate how to evaluate summaries accordingly. Further studies are still needed to use a larger data set than the one used in this study and consider more meeting aspects (such as adjacency pairs in conversations). Finally, we plan to investigate meeting summarization evaluation using speech recognition output.

ACKNOWLEDGMENT

The authors would like to thank the University of Edinburgh for providing the annotated ICSI meeting corpus and M. Galley for sharing his tool to process the annotated data. They would also like to thank G. Murray and M. Galley for letting them use their automatic summarization system output for this study.

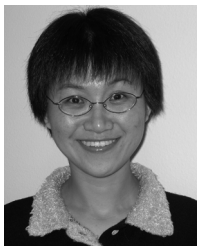
REFERENCES

[1] F. Liu and Y. Liu, "Correlation between rouge and human evaluation of extractive meeting summaries," in *Proc. ACL*, 2008, pp. 201–204.
[2] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 401–408, Jul. 2004.
[3] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Proc. Interspeech*, 2005, pp. 621–624.
[4] X. Zhu and G. Penn, "Evaluation of sentence selection for speech summarization," in *Proc. RANLP, Workshop Crossing Barriers in Text Summarization Research*, 2005, pp. 39–45.
[5] L. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 42–60, Sep. 2005.
[6] K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg, "Do summaries help?," in *Proc. SIGIR*, 2005, pp. 210–217.
[7] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. EMNLP*, 2006, pp. 364–372.
[8] G. Murray, S. Renals, J. Carletta, and J. Moore, "Evaluating automatic summaries of meeting recordings," in *Proc. ACL MTSE Workshop*, 2005, pp. 33–40.
[9] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in *Proc. ICASSP*, 2008, pp. 4985–4988.
[10] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out ACL*, 2004, pp. 74–81.
[11] K. S. Jones and J. Galliers, "Evaluating natural language processing systems: An analysis and review," *Lecture Notes in Artificial Intelligence*, vol. 1083, 1996.
[12] D. R. Radev, H. Jing, M. Stys, and T. Daniel, "Centroid-based summarization of multiple documents," *Inf. Process. Manage.*, vol. 40, pp. 919–938, 2004.
[13] S. Teufel and H. Halteren, "Evaluating information content by factoid analysis: Human annotation and stability," in *Proc. EMNLP*, 2004, pp. 419–426.
[14] A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The pyramid method," in *Proc. HLT/NAACL*, 2004, pp. 145–152.
[15] E. Hovy, C. Lin, L. Zhou, and J. Fukumoto, "Automated summarization evaluation with basic elements," in *Proc. LREC*, 2006, pp. 899–902.
[16] I. Mani, T. Firmin, D. House, M. Chrzanowski, G. Klein, L. Hirschman, B. Sundheim, and L. Obrst, "The Tipster Summac Text Summarization Evaluation: Final Report," 1998, The MITRE Corp., Tech. Rep.

[17] S. P. Hobson, B. J. Dorr, C. Monz, and R. Schwartz, "Task-based evaluation of text summarization using relevance prediction," *Inf. Process. Manage. Special Iss. Summarization*, vol. 43, pp. 1482–1499, 2007.
[18] NIST, "DUC2007: Task, Documents, and Measures," in *Proc. Document Understanding Conf. (DUG)*, 2007 [Online]. Available: http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html, NIST
[19] C. Hori, T. Hori, and S. Furui, "Evaluation methods for automatic speech summarization," *Eurospeech*, pp. 2825–2828, 2003.
[20] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proc. Interspeech'05*, 2005, pp. 2441–2444.
[21] Y. Liu, F. Liu, B. Li, and S. Xie, "Do disfluencies affect meeting summarization? a pilot study on the impact of disfluencies," in *Proc. MLMI Workshop, Poster Session*, 2007.
[22] X. Zhu and G. Penn, "Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization," in *Proc. HLT/NAACL*, 2006, pp. 197–200.
[23] D. Jones, F. Wlof, E. Gilbson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. Eurospeech*, 2003, pp. 1585–1588.
[24] A. Janin, D. Baron, J. Edwards, D. Ellis, G. Gelbart, N. Norgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, 2003, pp. 364–367.
[25] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SIGDIAL Workshop*, 2004, pp. 97–100.
[26] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," *Proc. SIGIR*, pp. 335–336, 1998.
[27] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 28, no. 1, pp. 132–142, 1972.

Feifan Liu received the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2006.

He has since been working as a Postdoctoral Research Fellow in the Computer Science Department, University of Texas at Dallas, Richardson, (2006-2009). His research interests are in information extraction, spoken language processing, sentimental analysis, and statistical machine learning.



Yang Liu (M'05) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2004.

She was a Researcher at the International Computer Science Institute, Berkeley, CA, from 2002 to 2005. She has been an Assistant Professor in Computer Science at the University of Texas at Dallas, Richardson, since 2005. Her research interests are in the area of speech and language processing.