Towards Online Speech Summarization

Gabriel Murray , Steve Renals

Centre for Speech Technology Research,
University of Edinburgh, Edinburgh, Scotland
gabriel.murray@ed.ac.uk, s.renals@ed.ac.uk

Abstract

The majority of speech summarization research has focused on extracting the most informative dialogue acts from recorded, archived data. However, a potential use case for speech summarization in the meetings domain is to facilitate a meeting in progress by providing the participants - whether they are attending in-person or remotely - with an indication of the most important parts of the discussion so far. This requires being able to determine whether a dialogue act is extract-worthy before the global meeting context is available. This paper introduces a novel method for weighting dialogue acts using only very limited local context, and shows that high summary precision is possible even when information about the meeting as a whole is lacking. A new evaluation framework consisting of weighted precision, recall and f-score is detailed, and the novel online summarization method is shown to significantly increase recall and f-score compared with a method using no contextual information.

Index Terms: speech summarization, online summarization, multiparty dialogues, meeting assistant, remote monitoring

1. Introduction

When applying speech summarization to the meetings domain, the goal of most research has been to extract and concatenate the most informative dialogue acts from an archived meeting in order to create a concise and informative summary of what transpired. Such summaries are analogous to the traditional manual minutes of a meeting, and are relevant to use cases such as a person wanting an overview of a meeting they missed, or a person wanting to review a meeting they attended, as a mental refresher. However, there are many use cases that go beyond the scenario of a user accessing an archived meeting. For example, someone might join a meeting halfway through and require a method of catching up on the discussion without disturbing the other participants. A second example is a person who is remotely monitoring a meeting with the intention of joining the group discussion when a certain topic is broached. These use cases require the development of online summarization methods that classify dialogue acts based on a much more limited amount of data than previously relied upon.

This paper introduces effective methods for scoring and extracting dialogue acts based on examining each candidate's immediate context. A method of *score-trading* is introduced and described wherein redundancy is reduced while informativeness is maximized, thereby significantly increasing weighted f-scores in our evaluation.

2. Previous Work

McKeown et. al. [1] provide an overview of text summarization approaches and discuss how text-based methods might be extended to speech data. The authors describe the challenges in summarizing differing speech genres such as Broadcast News and meeting speech. In the meetings domain, using the ICSI corpus [2], Murray et al. [3] compared text summarization approaches with feature-based approaches using prosodic features, with human judges favoring the feature-based approaches. In subsequent work, Murray et al. [4] explored speech-specific characteristics beyond prosody. Also using the ICSI corpus, Galley [5] used skip-chain Conditional Random Fields to model pragmatic dependencies such as QUESTION-ANSWER between paired meeting utterances, and used a combination of lexical, prosodic, structural and discourse features to rank utterances by importance. Zechner [6] researched summarization on several genres of speech, including spontaneous meeting speech. Though relevance detection in his work relied largely on *tf.idf* scores, Zechner also explored cross-speaker information linking and question/answer detection, so that utterances could be extracted not only according to high *tf.idf* scores, but also if they were linked to other informative utterances.

3. Weighting Dialogue Acts

This section describes three methods of scoring and extracting dialogue acts, the first of which relies on a simple term-score threshold, and the second two of which rely on a more complex score-trading system within the dialogue act's immediate context.

3.1. Residual IDF

The most common term-weighting scheme in information retrieval is $TF \cdot IDF$, where TF is the frequency of the term in the document at hand, and IDF is given by

$$IDF(t) = -\log(\frac{D_w}{D})$$

where D is the number of total documents in the collection and D_w is the number of documents indexed by the term w. This IDF term is an example of a *collection frequency* measure. Because we are investigating online summarization methods, complete term-frequency information is unavailable and so one option is to rely solely on collection frequency.

One extension of $TF \cdot IDF$ called $TF \cdot RIDF$ [7] has proven effective for automatic summarization [8] and named entity recognition [9]. In $TF{\cdot}RIDF$, the usual IDF component is substituted by the difference between the IDF of a term and its expected IDF according to the poisson model. RIDF can be calculated by the formula

$$expIDF = -\log(1 - e^{(-f_w/D)})$$

$$RIDF = IDF - expIDF$$

where f_w is the frequency of the word across all documents D.

Experiments on creating very brief summaries of archived meetings [10] have shown RIDF to be superior to IDF on this data. Our first method of extraction then is to simply sum RIDF term-scores over each dialogue act and extract a given dialogue act if it exceeds a pre-determined threshold. Based on using various thresholds on a separate development set of meetings, a threshold of 3.0 was decided for the experiments below. RIDF scores were calculated using a collection of documents from the AMI, ICSI, MICASE and Broadcast News corpora, totalling 200 speech documents (AMI test set meetings were excluded).

3.2. Score-Trading

The previously described method uses no knowledge of dialogue act context, and therefore does not address redundancy or importance relative to neighboring dialogue acts. A dialogue act was simply extracted if it scored above a given threshold. In contrast, the following two methods use a limited amount of context in order to maximize informativeness in a given region and to reduce redundancy, via a simple score-trading scheme.

For each dialogue act, we examine the ten preceding and ten subsequent dialogue acts. For each unique word in that 21-dialogue-act window, we total its overall score (its RIDF score times its number of occurrences in that window) and reapportion that overall score according to the relative informativeness of the dialogue acts containing the term. For example, if the word 'scroll' has an RIDF score of 1.2 and it occurs twice in that window, in two different dialogue acts, it has a total score of 2.4. If one of the dialogue acts containing the term has a dialogue act score of 5.0 and the other has a dialogue act score of 3.0, the overall term score is apportioned in favor of the former dialogue act, so that is receives a revised term score of 1.5 and the latter receives a revised term score of 0.9. As a result, the dialogue act score for the former has increased while it has decreased for the latter. This method of score-trading places the burden of carrying that term's information content onto the more generally informative dialogue acts, which also has the effect of reducing redundancy.

More formally, the revised term-score for word W in dialogue act D is given by

$$Sc_b(W,D) = (Sc_a * N_W) * (Ascore_D / \sum_{d_W} Ascore_{d_W})$$

where Sc_a is the original RIDF score for the word, N_W is the number of times that the word appears in all of the dialogue acts examined, $Ascore_D$ is the original dialogue act score for D, i.e. its summed RIDF scores, and d_W is a dialogue act in the examined context that contains the word W.

A dialogue act's Bscore is then the sum of its revised term-scores. After deriving the Bscore score, the dialogue act in question is extracted if it satisfies the formula

$$Bscore >= 3.0$$

The second score-trading method is similar to the first, but a dialogue act is extracted if it satisfies the formula

$$Bscore - (Ascore - Bscore) >= 3.0$$

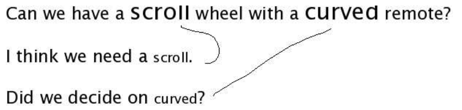
Can we have a scroll wheel with a curved remote?
I think we need a scroll.
Did we decide on curved?

Figure 1: *Score-Trading Between Dialogue Acts*

where Ascore is the original score and Bscore is the adjusted score. The reasons motivating this latter method are twofold. First, a dialogue act's adjusted score (i.e. Bscore) may still be below the 3.0 threshold, but if it has increased significantly compared to the Ascore, that indicates its importance in the local context and we want to increase its chances of being extracted. Second, a dialogue act's adjusted score may be above 3.0 but it is well below its original Ascore, indicating that it has lost informativeness and may well be redundant in the local context. As a result, we want to reduce its chance of being extracted.

4. Experimental Setup

4.1. Data

The data used for these experiments is the AMI meeting corpus [11], a corpus of 100 hours of spontaneous multiparty spoken dialogues. While the corpus contains both scenario and non-scenario meetings, these experiments utilized solely the scenario-based portion. In these scenario meetings, four participants take part in each meeting and play roles within a fictional company. The scenario given to them is that they are part of a company called Real Reactions, which designs remote controls. Their assignment is to design and market a new remote control, and the members play the roles of project manager (the meeting leader), industrial designer, user-interface designer, and marketing expert. Through a series of four meetings, the team must bring the product from inception to market. The participants are also given real-time information from the company during the meetings, such as information about user preferences and design studies, as well as updates about the time remaining in each meeting. While the scenario given to them is artificial, the speech and the actions are completely spontaneous and natural.

The AMI test set consists of 19 meetings, or 4 sequences of 4 meetings each and 1 sequence of 3 meetings.

For each of the meetings, a manual abstract was created summarizing the meeting in terms of decisions, goals and problems from the meeting. Multiple human annotators then worked through the meeting transcript and linked dialogue acts to the abstract if they believed that the dialogue act supported a specific abstract sentence. A given dialogue act is able to be linked to multiple abstract sentences, and vice-verse, so that we end up with a many-to-many mapping between dialogue acts and abstract sentences. The unit of extraction in these summarization experiments is the dialogue act.

4.2. Evaluation

The evaluation method is an extension of the *weighted precision* metric introduced by Murray et al [4], and relies on the many-to-many mapping between dialogue acts and abstract sentences described in the previous section. The work described in [4] involved the creation of very short summaries of 700-words, and the evaluation was therefore limited to weighted precision due to the very low recall scores of all approaches. In the present experiments, we extend the evaluation metric to weighted pre-

sys	man-prec	man-rec	man-fsc	asr-prec	asr-rec	asr-fsc
ridf	0.608	0.286	0.382	0.612	0.276	0.374
trade	0.611	0.295	0.391	0.610	0.285	0.383
tdiff	0.603	**0.305**	0.399	0.605	**0.295**	**0.392**

Table 1: Weighted Precision, Recall and F-Scores
ridf=DA extracted if Ascore $>=$ 3.0, **trade**=DA extracted if Bscore $>=$ 3.0, **tdiff**=DA extracted if Bscore - (Ascore-Bscore) $>=$ 3.0

cision, recall and f-score, as our new summaries tend to be much longer and are of varying lengths.

To calculate weighted precision, we count the number of times that each extractive summary dialogue act was linked by each annotator, averaging these scores to get a single dialogue act score, then averaging all of the dialogue acts scores in the summary to get the weighted precision score for the entire summary. To calculate weighted recall, the total number of links in our extractive summary is divided by the total number of links to the abstract as a whole. A difference between weighted precision and weighted recall is that weighted recall has a maximum score of 1, in the case that all linked dialogue acts are included in the extractive summary, whereas there is no theoretical maximum for weighted precision since annotators were able to link a given dialogue act as many times as they saw fit.

More formally, both weighted precision and recall share the same numerator

$$num = \sum_{d} L_s/N$$

where L_s is the number of links for a dialogue act d in the extractive summary, and N is the number of annotators.

Weighted precision is equal to

$$precision = num/D_s$$

where D_s is the number of dialogue acts in the extractive summary. Weighted recall is given by

$$recall = num/(L_t/N)$$

where L_t is the total number of links made between dialogue acts and abstract sentences by all annotators, and N is the number of annotators.

The f-score is calculated as

$$(2 * precison * recall)/(precision + recall)$$

The summaries range between 600 and 3000 words in length, as the meetings themselves greatly vary in length.

4.3. Results

One of the most surprising results is that the weighted precision in general is not drastically lower than the scores found when creating very brief summaries of archived meetings. For example, in [10], creating 700-word summaries of the same test set using RIDF yielded an average weighted precision of 0.66. All three online approaches presented here have average weighted precision around 0.61. This is particularly surprising and encouraging given that these summaries are on average much longer than 700 words.

The third approach, labeled **tdiff** in Table 1, is superior in terms of f-score on both manual and asr transcripts. RIDF performs the worst on both sets of transcripts, and the second approach labeled **trade** is in-between. Significant results in the table are presented in boldface. The method **tdiff** achieves significantly higher recall than the other two methods on manual transcripts, and both recall and f-score are significantly higher on ASR (paired t-test, $p<0.05$). The most encouraging result of this third approach is that it is able to significantly increase recall without significantly reducing precision.

Having determined the effectiveness of the third approach, we subsequently ran this score-trading method at multiple thresholds of 2.0, 3.0 and 4.0 to gauge the effect on weighted precision, recall and f-score. The results are displayed in Figure 2. A threshold between 2 and 3 results in a good balance between recall and precision, while a threshold of 4 results in drastically lower recall and only slightly higher precision.

The score-trading results reported so far stem from an implementation of the method that has an algorithmic delay of 10 dialogue acts. We were interested in what benefit, if any, could be gained by increasing the algorithmic delay and thereby increasing the amount of context used. The two score-trading approaches were therefore run fully offline, so that the context for each dialogue act is the entire meeting (the first approach, based simply on RIDF results, is the same online versus offline since it does not use context). Because there is a larger amount of score-trading when using all meeting dialogue acts for comparison, a given dialogue act would have to be very informative in order to have its overall Ascore increase. The expectation was that running this method offline would therefore result in higher precision and perhaps lower recall. The third approach, labeled **tdiff** in Table 2, is again superior to the second approach, labeled **trade**, with significant differences between the two in terms of recall and f-score on both manual and ASR transcripts. However, neither approach was significantly different when run offline versus online. The trend was for precision to be slightly lower when run offline and recall to be slightly higher, the opposite of what was expected.

5. Discussion

The results above show that the score-trading scheme is able to significantly increase recall and f-score with no significant decrease in precision. More specifically, it allows us to reject dialogue acts that may have scored high but were redundant compared with similar and more informative neighboring dialogue acts, and allows us to retrieve dialogue acts that may have scored below the threshold originally but subsequently had their scores adjusted based on local context.

In general, it is interesting that high precision is attained via methods that use either no context or only local context. As mentioned earlier, previous experiments on creating very concise summaries using global information about the meeting achieved weighted precision of only a few points higher. It turns out that restrictions such as the inability to create an overall ranking of dialogue acts in a meeting or to rely on term-frequency information are not severely detrimental to the ultimate results.

sys	man-prec	man-rec	man-fsc	asr-prec	asr-rec	asr-fsc
trade	0.599	0.291	0.386	0.608	0.291	0.388
tdiff	0.589	**0.306**	**0.398**	0.593	**0.304**	**0.398**

Table 2: Weighted Precision, Recall and F-Scores (Offline)
trade=DA extracted if Bscore >= 3.0, **tdiff**=DA extracted if Bscore - (Ascore-Bscore) >= 3.0

A related finding is that there is no benefit to running the score-trading methods completely offline, using the entirety of the meeting's dialogue acts as context. In fact, precision results were slightly better when examining only the limited context. It may be that dialogue acts sharing some of the same terms and existing within proximity to each other tend to be more similar than dialogue acts sharing some of the same terms but existing at various locations spread throughout the meeting. In that case, score-trading between ostensibly similar dialogue acts would not always be beneficial if the examined context is too great.

While the score-trading methods outperform the simple RIDF threshold method, with the third summarization system performing the best, it would seem that the methods are complementary. Because the RIDF method requires no contextual information, a dialogue act can be immediately extracted or rejected on a preliminary basis. Once the subsequent context for a dialogue act becomes available, that decision can be revised based on score-trading. User feedback could provide a further source of input for such dynamic summary creation.

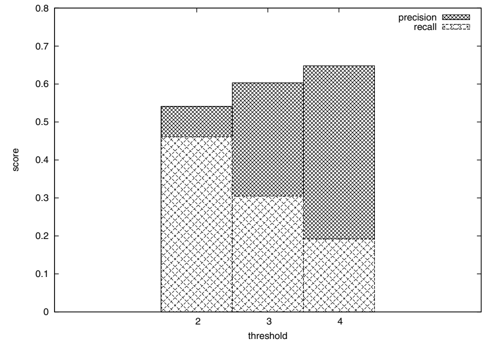

Figure 2: *Score-Trading at Multiple Thresholds*

6. Conclusion

This paper has introduced a novel method for the online summarization of spoken dialogues, using a score-trading scheme intended to reduce redundancy and to develop a more subtle view of informativeness. By looking at informativeness beyond the level of the dialogue act and examining local context around the candidate dialogue act, we are able to locate words that are generally informative in a local region of the meeting transcript and to place the burden of carrying those words' informativeness onto the most informative dialogue acts in that region. An encouraging finding for the prospect of online meeting analysis is that weighted precision scores are not drastically lower than the precision scores found in previous work on very concise summariation of archived meetings, even when the recall of the summaries contained herein is much higher. Running the score-trading methods offline did not result in any added benefit compared with using only a small amount of context and executing the method online.

7. Acknowledgements

Thanks to Thomas Hain and the AMI-ASR group for speech recognition output. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication)

8. References

[1] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey, "From text to speech summarization," in *Proc. of ICASSP 2005, Philadelphia, USA*, 2005.

[2] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. of IEEE ICASSP 2003, Hong Kong, China*, 2003.

[3] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. of Interspeech 2005, Lisbon, Portugal*, 2005.

[4] G. Murray, S. Renals, J. Moore, and J. Carletta, "Incorporating speaker and discourse features into speech summarization," in *Proc. of HLT/NAACL 2006, New York City, USA*, 2006, p. to appear.

[5] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. of EMNLP-06, Sydney, Australia*, 2006.

[6] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, 2002.

[7] K. Church and W. Gale, "Inverse document frequency IDF: A measure of deviation from poisson," in *Proc. of the Third Workshop on Very Large Corpora*, 1995, pp. 121–130.

[8] C. Orasan, V. Pekar, and L. Hasler, "A comparison of summarisation methods based on term specificity estimation," in *Proc. of LREC 2004, Lisbon, Portugal*, 2007, pp. 1037–041.

[9] J. Rennie and T. Jaakkola, "Using term informativeness for named entity recognition," in *Proc. of SIGIR 2005, Salvador, Brazil*, 2005, pp. 353–360.

[10] G. Murray and S. Renals, "Term-weighting for summarization of multi-party spoken dialogues," in *Proc. of MLMI, Brno, Czech Republic*, to appear.

[11] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. of MLMI 2005, Edinburgh, UK*, 2005.