Extractive Meeting Summarization through speaker zone detection

Mohammad Hadi Bokaei¹, Hossein Sameti¹, Yang Liu²

¹Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
²Department of Computer Science, The University of Texas at Dallas, U.S.A

bokaei@ce.sharif.edu, sameti@sharif.edu, yangl@hlt.utdallas.edu

Abstract

In this paper we investigate the role of discourse analysis in extractive meeting summarization task. Specifically our proposed method comprises of two distinct steps. First we use a meeting segmentation algorithm in order to detect various functional parts of the input meeting. Afterwards, a two level scoring mechanism in a graph-based framework is used to score each dialogue act in order to extract the most valuable ones and include them in the extracted summary. We evaluate our proposed method on AMI and ICSI corpora and compare it with other state-of-the-art graph based algorithms according to various evaluation metrics. The experimental results show that our algorithm outperforms the other state-of-the-art ones according to most of the metrics and on both datasets.

1. Introduction

Improvements of automatic speech recognition systems and increasing amount of audio data (such as broadcast news, voice mail, telephony conversations and meetings) have attracted plenty of research interest in the field of speech summarization. On the other hand, conversation and in its specific form, meetings are an integral part of each organization. Not surprisingly then, there is growing interest in developing automatic methods which summarize meetings in a way that by reading only the extracted summary, the reader will be informed about the important key-notes of the meeting.

Traditionally and based on earlier works on summarization task [1], extractive methods are expressed as a combination of two distinct goals in selecting each sentence: maximizing the information covered by the sentence while minimizing its redundancy according to previously extracted sentences. Different summarization algorithms use different strategies to calculate how much informative a sentence is, but few of them incorporate meeting specific structural information in calculation [2, 3, 4, 5].

In this work, we propose to consider discourse by detecting speaker zones in the meeting flow. We try to segment a meeting transcript into functionally coherent parts such as monologue or discussion. We hypothesize that detecting these parts and summarizing meeting transcript accordingly can improve the accuracy of the extracted output. We compare our proposed algorithm against state-of-the-art algorithms and show improvement over them using standard summarization measures (ROUGE-1, ROUGE-2) and classification measures (Precision, Recall, F-Measure) on AMI and ICSI standard meeting corpora.

This paper is organized as follows. Section 2 describes relevant works. Section 3 illustrates our proposed algorithm. In Section 4 evaluation metrics and experimental results are shown. Finally in Section 5 we describe our future work and conclude the usefulness of our proposed algorithm.

2. Related works

Generally speaking, there are two distinct categories of summarization approaches: supervised methods which need annotated training data to train their models [2, 6, 7, 8, 9] and unsupervised methods which need no labeled training set and use only the information in words of the document to be summarized [10, 11, 12, 13].

Most of the approaches applied for meeting summarization purpose are inherited from text summarization task like Maximal Marginal Relevance (MMR) [1, 14], topic-based [15, 16, 17], graph-based [10, 13, 18], and optimization-based [12, 19, 20]. It has been shown that the graph-based approaches are the most successful ones compared to other methods in the meeting summarization domain [21].

One successful method proposed in graph-based category is ClusterRank [13]. In this method, adjacent utterances are clustered according to their similarity values. A graph is constructed where nodes are these clusters and Random-walk procedure is applied on this graph to score each cluster. Afterwards, each utterance is scored according to its associated cluster and the similarity between the utterance and its corresponding cluster.

The main difference between our work and ClusterRank is the segmentation step. While ClusterRank uses a very simple algorithm which clusters similar adjacent utterances to one segment, our segmentation algorithm tries to segment the meeting into zones in which speakers distribution is steady. We also use weighted combination schema to compute the final score for each utterance.

Another state-of-the-art algorithm proposed in graph-based framework constructs two-layer graph [10]. This algorithm uses speakers information in order to score each utterance. Utterances are represented as nodes in utterance-layer and speakers are represented as nodes in speaker-layer of the graph. Applying Random-walk procedure on this constructed graph, scores from different layers are reinforced so that final utterance scores are influenced by utterances from the same speaker and similar utterances.

There are a few previous works which incorporate discourse information in the summarization task. [2] adds some very simple discourse features (such as existence of specific keywords in the utterance and the position of that utterance in the meeting) to calculate informative score for each sentence. In [3] it is shown that adding structural features can improve the effectiveness of the summarization algorithms. [5] studies the effect of turn-taking and participant involvement on the task of finding more informative segments in a meeting. The authors in [4] study the usefulness of discourse more profoundly. They use Conditional Random Fields to extract rhetorical structure and summary in a single step. However the main difference between this work and ours is that this work is supervised in the
Our algorithm needs a very small development set (one meeting in our study) to tune its parameters. However its performance is not very sensitive on the changes of these parameters.
normalize the scores of outgoing edges of a node, so the graph can be seen as a Markov chain. We apply random walk process on $G$ to compute salience score for each segment. Specifically Eqn. 2 is used iteratively to compute segments scores.

$$P(u) = \frac{d}{d+1} + \frac{1-d}{\sum_{v \in \text{adj}(u)} \text{weight}_{\text{adj}}(u,v) \ast P(v)}$$

In Eqn. 2, $P$ is a vector whose elements are salience scores of each segment. $d$ is damping factor which is typically chosen in the interval $[0.1, 0.2]$ and ensures convergence [18]. $u$ and $v$ are two nodes of $G$ and $\text{adj}(u)$ denotes nodes that are adjacent to $u$. Vector $P$ can be chosen randomly at first and Eqn. 2 is then iteratively applied on all nodes of $G$ until changes on vector $P$ becomes lower than a predefined threshold.

At this point, scores for each segment is computed. Now utterance ($U$) in the meeting must be scored according to its associated segment ($S$) score and the degree of importance of $U$ in $S$. As previous works, the degree of importance of $U$ in $S$ can be computed according to their similarity. Both $U$ and $S$ are considered as bag of words and cosine similarity measure is used to compute the similarity. These two scores are combined in a weighted manner to compute the final score for each utterance according to Eqn. 3.

$$\text{score}(U) = \text{cosine}(U,S) + \omega \ast P(S)$$

In the above equation, $P$ is computed according to Eqn. 2. $\omega$ and other parameters of the algorithm are tuned according to a small development set as shown in the experimental result. Using the procedure described above, we obtain the salience score for each utterance. Now all utterances are sorted according to their scores and the best ones are greedily selected until summary length exceeds predefined threshold.

4. Experimental results

4.1. Setup

The AMI meeting corpus [25] is a collection of 100 hours of meeting data that includes annotations in various layers such as speech audio, transcripts, focus of attention, and etc. In order to evaluate the effectiveness of our proposed algorithm, a subset of 11 meetings in the AMI corpus were manually annotated and used as test set\(^3\). We employed graduate students as annotators and asked them to segment the meetings according to different events. They were given a guideline which included the task definition and various examples used to clarify the concept of events and function segmentation of a meeting. Each meeting was annotated by one annotator. One special trained annotator then checked and finalized the annotations. The average number of segments in this set is 19.6. One of these meetings was used as our development set on which we tuned the parameter $TH_{cu}$ of the segmentation algorithm.

We used $P_k$ [27], which is the most widely used metric for segmentation evaluation. Given two points in a sequence, $P_k$ specifies the probability of segmentation error, which is the average probability that the segmenter’s decision is incorrect. Note that $P_k$ is a measure of error and thus a lower score means better segmentation performance.

To evaluate the whole summarization algorithm, we used 20 meetings in the same AMI corpus\(^3\). Each meeting is prepared with one reference summary. These summaries have no unique compression ratio. The average compression ratio for our test set is 0.38 and its variance is 0.0091.

We also tested our summarization algorithm on ICSI meeting corpus [26] which contains 75 recordings from natural meetings. Each meeting is about an hour long. We evaluated the results of our summarization algorithm according to the whole meetings in this corpus. For our evaluation purpose we used human transcription of the meeting and also assumed that the speaker of each utterance is specified in advance.

According to the fact that extractive meeting summarization is indeed a classification task where important utterances must be distinguished from not-important ones, we can use common evaluation metrics in classification tasks such as precision, recall and F-measure. From another viewpoint, ROUGE [28] evaluates a summarization system based on the number of overlapping units such as n-grams, between the system generated summary and the ideal summary created by human annotator. We show results of our summarization algorithm using ROUGE-1 (unigram overlap) and ROUGE-2 (bigram overlap) along with classification measures to compare the result with the other state-of-the-art algorithms\(^4\).

4.2. Results

Using the single meeting in our development set, we tune the parameters of our algorithm. Specifically we choose $TH_{cu} = 0.2$ and $T = 10$. We use $\epsilon = 10^{-4}$ for the threshold used in utterance ranking step. We also test various values for $\omega$ and found that the best result is achieved on $\omega = 0.25$.

First we analyze the performance of our proposed segmentation algorithm. The result of applying this algorithm on our test set is shown in Table 2. In order to further analyze the algorithm, the segmentation found in a sample meeting is shown in Figure 2.a. This figure shows the boundaries found by our proposed algorithm (dashed lines) and the reference boundaries (solid lines). Figure 2.b shows the speaker distribution for the first 6 segments in the segmentation found for this meeting. As this figure shows, we have some segments in which one person lectures all others and thus it will be a monologue segment (segments 1, 4 and 6). We also have segments which contain discussions between specific participants in the meeting (segments 2, 3 and 5).

We also compare our proposed summarization algorithm against state-of-the-art algorithms proposed in literature. Specifically these algorithms are:

- ClusterRank [13]: This is our base algorithm. As stated, our utterance ranking step is inspired by this algorithm.
- MRRW-WBP [10]: This is the state-of-the-art algorithm, recently proposed. It uses Probabilistic Latent Semantic Indexing in a graphical framework in order to score each utterance.

To compare results fairly, we consider different compression ratios for meetings. For each meeting we compute the compression ratio according to the one used to generate reference extractive summary in the corpus, which is the number of words in the reference extracted summary divided by the total number of the

\(^3\) The ids of annotated meetings are: es2008a, is1000a, is1001a, is1001b, is1001c, is1003b, is1006b, is1008a, is1008b, is1008c and ts3005a. We chose these meetings since they have more annotations in the AMI corpus, which can be useful for our future studies. The reference segmentation as well as the annotation guide can be found here: http://iie.sharif.edu/bokaei/resources/funseg/

\(^4\) The ids of meetings are: ES2004, ES2014, IS1009, TS3003 and TS3007.

\(^5\) We use ROUGE package (http://www.berouge.com/) to evaluate our proposed summarization algorithm.
Table 1: Results of our proposed algorithm compared to ClusterRank [13] and MRRW-WBP [10]

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>Classification measure (%)</th>
<th>ROUGE1 (%)</th>
<th>ROUGE2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>AMI</td>
<td>ClusterRank</td>
<td>39.07</td>
<td>42.26</td>
<td>39.74</td>
</tr>
<tr>
<td></td>
<td>MRRW-WBP</td>
<td>54.43</td>
<td>37.53</td>
<td>43.67</td>
</tr>
<tr>
<td></td>
<td>Proposed Method</td>
<td>53.46</td>
<td>37.87</td>
<td>43.58</td>
</tr>
<tr>
<td>TCSI</td>
<td>ClusterRank</td>
<td>18.70</td>
<td>16.21</td>
<td>17.07</td>
</tr>
<tr>
<td></td>
<td>MRRW-WBP</td>
<td>25.67</td>
<td>17.97</td>
<td>20.87</td>
</tr>
<tr>
<td></td>
<td>Proposed Method</td>
<td>26.36</td>
<td>23.36</td>
<td>24.48</td>
</tr>
</tbody>
</table>

5. Conclusion and future works

In this work we investigated the effect of considering discourse structure in the performance of summarization systems. Specifically we tried to segment meeting discourse into functionally coherent segments and then scored each utterance according to this segmentation. Results showed improvement over other state-of-the-art algorithms.

We believe that the performance of this algorithm can be further improved using other levels of discourse analysis such as extracting relations between utterances. From another viewpoint, literature proves that using other weighting schema other than tfidf, which we used in this study, can improve the accuracy of summarization system in multi-party conversations [29].

However, the main track of our future works will concentrate on designing a system which summarizes each segment separately. According to this fact that demands of users change according to the type of the segments. While in a monologue segment, readers are interested in finding the key notes of the lecture, in a discussion segment, they are interested in finding the topic of the discussion, its outcome and maybe sentiments of participants to that result. According to these needs, the way a summary for monologue segment is generated must be different from the way a summary is generated for discussion segment.

6. References


