Random Walks for Opinion Summarization on Conversations

Zhongqing Wang, Liyuan Lin, Shoushan Li, and Guodong Zhou

Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, China
{wangzq.antony,scarecrowlly,shoushan.li}@gmail.com, gdzhou@suda.edu.cn

Abstract. Opinion summarization on conversations aims to generate a sentimental summary for a dialogue and is shown to be much more challenging than traditional topic-based summarization and general opinion summarization, due to its specific characteristics. In this study, we propose a graph-based framework to opinion summarization on conversations. In particular, a random walk model is proposed to globally rank the utterances in a conversation. The main advantage of our approach is its ability of integrating various kinds of important information, such as utterance length, opinion, and dialogue structure, into a graph to better represent the utterances in a conversation and the relationship among them. Besides, a global ranking algorithm is proposed to optimize the graph. Empirical evaluation on the Switchboard corpus demonstrates the effectiveness of our approach.

Keywords: Opinion Summarization on Conversations, Graph, Random Walk, Global Ranking.

1 Introduction

Opinion summarization aims to generate a sentimental summary on opinions in a text and has been drawing more and more attention recently in NLP due to its significant contribution to various real applications [5, 14]. However, although there are a few previous studies on extracting opinion summaries, most of them focus on text reviews, such as movie reviews [8, 16] and product reviews [4, 21]. With the increasing amount of conversation recordings, opinion summarization on conversations becomes more and more demanding. In this study, we investigate this novel type of opinion summarization.

Speech summarization is more difficult than well-structured text, because a) speech is always less organized and has recognition errors; b) in conversational speech, information density is low and there are often off topic discussions [24].

As pilots in opinion summarization on conversations, Wang and Liu [24] recast it as an utterance ranking problem, similar to traditional topic-based summarization and general opinion summarization [18, 23]. However, as stressed by Wang and Liu [24], opinion summarization on conversations possesses some unique characteristics and challenges.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 430–437, 2014.

First, it is necessary to consider both the topic relevance of an utterance and the opinion expressions in an utterance. One basic intuition therein is that opinion summarization is prone to containing opinion sentences.

Second, dialogue structures play an important role in utterance selection. One common phenomenon is that if one utterance contains opinions, its adjacent paired utterances much likely contain opinions.

Third, there may be some short utterances in a conversion, e.g. "Uh", "Yeah", "Well", etc. Our preliminary exploration finds that, although most of them are little informative, they are much likely to be selected as "good" utterance candidates in the summary when some frequency-based approaches are employed.

Although above unique characteristics and challenges have been noticed by Wang and Liu [24], they are not well addressed in the literature. This largely limits the performance of opinion summarization on conversations.

In this paper, we re-visit these unique characteristics and challenges, and propose a graph-based framework to opinion summarization on conversations. In particular, a random walk model is proposed to globally rank the utterances in a conversation. The main advantage of our approach is its ability of integrating various kinds of important information, such as utterance length, opinion, and dialogue structure, into a graph to better represent the utterances in a conversation and the relationship among them. Different from Wang and Liu [24], where these factors are separately ranked and combined with a simple weighting strategy, we incorporate them into an utterance graph and perform global graph ranking. Experimental results on the Switchboard corpus show that our approach achieves the performance of 0.5778 in terms of ROUGE-1 measurement, which is 0.034 higher than that reported in Wang and Liu [24].

The rest of this paper is organized as follows. Section 2 overviews the related work on both topic-based summarization and opinion summarization. Section 3 introduces our framework for opinion summarization on conversations. Section 4 reports the experiment results. Finally, Section 5 concludes this paper with future work.

2 Related Work

Some previous studies on opinion summarization focus on to generate aspect-based ratings for an entity [4, 21] which actually consider the opinion summarization as an opinion mining problem. Although such summaries are informative, they lack critical information for a user to understand why an aspect receives a particular rating. Ganesan et al. [5] present a graph-based summarization framework named Opinosis to generate concise abstractive summaries of highly redundant opinions. Nishikawa et al., [14] generates a summary by selecting and ordering sentences taken from multiple review texts according to represent the informative and readability of the sentence order.

To the best of our knowledge, there is only one related work on opinion summarization on conversation, i.e., Wang and Liu [24]. They create a corpus containing both extractive and abstractive summaries of speaker's opinion towards a given topic using telephone conversations. They adopt both sentence-ranking method and graph-based method to perform extractive summarization. However, since they consider the topic,

opinion, dialogue structure and sentence length factors separately, the relations of them are not well integrated. Unlike that, our approach leverages them together in a random walk model and makes them working together to improve the overall performance.

3 Random Walks for Opinion Summarization on Conversations

Formally, a conversation is donated as its contained utterances as $U = [u_1, u_2, ..., u_n]^T$ and the summary using the extracted utterances as $X = [x_1, x_2, ..., x_m]^T$ with $m < n$ and $X \subset U$. Our approach for opinion summarization on conversations consists of three main steps: First, we build a graph G to represent all the utterances with their mutual topic, opinion and dialogue structure information; second, we rank the utterances on graph G with PageRank algorithm. Finally, we select the utterances with top ranking scores to generate a summary.

3.1 Graph Building

Different from traditional topic-based summarization tasks [18, 22], opinion summarization on conversations is encouraged to consider not only the topic relevance, but also the opinion and dialogues structure factors [24]. To integrate them into a uniform graph-based ranking framework, we hope to build a graph which could include all these information.

To achieve that, we build a graph that contains topic relevance, opinion relevance and dialogue structure relevance, which makes the graph a tri-layer model. The first layer contains the topic information; the second one contains the opinion information; the third one contains dialogue structure information.

Representing an Utterance as a Feature Vector
In our approach, an utterance is considered as a node in the graph and it is represented by a feature vector. If two sentences are more related, their feature vectors are supposed to share more features. To represent the relationship between two utterances, three kinds of features are employed to represent topic relevance, opinion relevance and structure relevance respectively.

- **Topic Relevance Features:** If two utterances shares more word unigrams and bigrams, they are thought to be more topic-related, as popularly assumed by many other previous studies [22]. Thus, the word unigrams and bigrams are adopted as the features to representing the topic relevance. The weight of each feature is Boolean which represents the presence or absence of a feature in an utterance. Formally, an utterance u_i can be represented as a feature vector x_i,

$$u_i = \langle bool(t_1), bool(t_2), ..., bool(t_n) \rangle$$

Where n is the number of unique features; $bool(t_i) = 1$ means the occurrence of the feature t_i in the utterance and $bool(t_i) = 0$ means the absence of the feature t_i.

- **Opinion Relevance Features:** If two utterances both contain opinion, they are thought to be opinion-related. To represent such relationship, a new feature named OPINION is added for each utterance. If an utterance contains at least one sentiment word in the pre-given lexicon, the OPINION feature weight is set to be a fixed integer larger than one, i.e., λ ($\lambda > 1$); Otherwise, the feature weight is set to zero. In this study, the sentimental lexicon[1] we used is from MPQA.
- **Structure Relevance Features:** There are two dialogue structures in conversations: One is the *adjacent* relation representing the relation between two adjacent utterances which are said by two speakers; the other is the *turn* relation representing the relation between two utterances which are in the same turn. If two utterances take either the *adjacent* relation or the *turn* relation, they are considered to be more structure-related. To represent such relationships, two kinds of features named ADJ and TURN are added for each utterance. Specifically, 1) we let two adjacent utterances u_i and u_{i+1} share the same feature ADJ-*i*; 2) we let two utterances u_i and u_j in the same turn *k* share the same feature TURN-*k*. Because these two kinds of features are believed to be more important than one unigram word feature, their weights are set to be a fixed integer that lager than one, i.e., ω ($\omega > 1$).

Transition Probability Computation with the Penalization on Short Utterances

The transition probability from the *i*-th node to the *j*-th node, denoted as $p(i \to j)$ is defined as the normalization of the weights of the edges out of the *i*-th node, i.e.,

$$p(i \to j) = \frac{f(i \to j)}{\sum_k f(i \to k)}$$

Where $f(i \to j)$ represents the similarity between u_i and u_j. Here, the cosine similarity is adopted (Baeza-Yates and Ribeiro-Neto, 1999):

$$f(i \to j) = \frac{u_i \cdot u_j}{|u_i||u_j|}$$

Two nodes are connected if their transition probability is larger than zero, i.e., $p(i \to j) > 0$. To avoid self-transition, we set the transition probability of one node to itself as zero.

Short utterances, e.g., backchannels, appear frequently in dialogues while they typically contain little important content [24]. For example, as shown in Fig 3, we can see that the short utterances such as u_2 (*and*) and u_5 (*Oh*) contains little information for opinion expression.

Since the short utterances are sometimes highly frequency words, and thus the transfer probability between these utterances are larger than many other utterances, which makes the PageRank algorithm more likely to select long utterances. To avoid this happening, we propose a novel formula of similarity function $f(i \to j)$ to penalize the

[1] http://www.cs.pitt.edu/mpqa/subj_lexicon.html

short utterances using the information of the utterance length. The basic idea is to make the transition probability to be proportion to the length of the utterance. In this way, the shorter the utterance, the lower transition probability to it will be assigned. The revised formula for computing transition probability is given as follows:

$$f(i \to j) = \frac{u_i \cdot u_j}{|u_i||u_j|} \log(|u_j|)$$

3.2 Ranking the Utterances with PageRank

Given the graph G, the saliency score $s(u_i)$ for utterance u_i can be deduced from those of all other utterances linked with it and it can be formulated in a recursive form as in the standard PageRank algorithm.

$$s(u_i) = \mu \sum_{j \neq i} s(u_j) \cdot p(j \to i) + (1 - \mu)$$

In the implementation, μ is the damping factor and usually set to be 0.85 (Page et al., 1998). The initial scores of all utterances are set to one, and the iteration algorithm is adopted until convergence [22].

As long as the saliency scores of utterances are obtained, the utterances are ranked with the sores. The utterances with largest ranking scores form the summary. In the implementation, the utterances from both speakers in the conversion are emerged for ranking with our PageRank algorithm.

4 Experimental Evaluation

4.1 Evaluation Setup

In the experiment, we use the Opinion Conversion Corpus which is drawn from the Switchboard corpus [24]. The corpus contains 88 conversations from 6 topics, among which 18 conversations are annotated by three annotators. We use these 18 annotated conversations as the testing set and perform our ranking approach on it.

We use the ROUGE toolkit [10] which has been widely adopted for automatic summarization evaluation. We choose three automatic evaluation methods ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-W in our experiment.

4.2 Experimental Results

In this subsection, we evaluate the performance of our PageRank approach and compare it with three baseline approaches for opinion summarization on conversions, together with the human summarization:

- **Max-length:** select the longest utterances for each speaker, this has been shown to be a very strong baseline for summarization on conversations (Gillick et al., 2009).
- **Sentence-Ranking:** the topic, opinion and dialogue structure score are separately calculated and then combined via a linear combination (Wang and Liu, 2011). We report the best performance of ROUGE-1 measurement for comparison.
- **PageRank:** the PageRank approach which only considers the topic relevance, which is a popular approach in topic-based text summarization (Wan and Yang, 2008).
- **Human:** calculate ROUGE scores between each reference and the other references, and average them. This can be considered as the upper bound of the performance.

Followed by Wang and Liu [24], the average compression ratio of the extractive summary of each conversation is set to be 0.25. In our approach, we set the parameters as $\lambda = 6$ and $\omega = 5$.

Table 1 shows the results of different approaches. From this table, we can see that the approach by Wang and Liu [24] is more effective than the basic PageRank approach, i.e., Topic-PageRank. This is because it takes the specific characteristics in conversations, such as utterance length and opinion information, into account. Our approach outperforms all the other approaches and improves the performances from 0.5448 to 0.5778 compared to the approach by Wang and Liu [24].

Table 1. Comparison Results with Baseline Approaches

Method	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W
Max-length	0.5279	0.3600	0.5113	0.2495
Sentence-Ranking	0.5448	-	-	-
PageRank	0.4959	0.3105	0.4821	0.2293
Our Approach	0.5778	0.4202	0.5670	0.2786
Human	0.6167	0.5200	0.6100	0.3349

It is interesting to find that the simplest baseline Max-length is able to get a decent performance of 0.5279, which is even much better than Topic-PageRank. This result reveals that the length of the utterances is an important factor for selecting "*good*" utterances in summarization on conversations.

5 Conclusion and Future Work

In this paper, we propose a graph-based framework to opinion summarization on conversations by first representing a conversation as an utterance graph and then performing global ranking via a PageRank algorithm. Besides topic relevance, both opinion relevance and structure relevance are incorporated systematically to meet the

specific characteristics and challenges in the task. Empirical studies demonstrate that our approach performs much better than other alternatives.

The research of opinion summarization on conversations is still in its early stage since the pilot work by Wang and Liu [24]. In the future work, we will explore more factors in a conversation and better ways of representing a conversation.

Acknowledgments. This research work is supported by the National Natural Science Foundation of China (No. 61273320, No. 61331011, and No. 61375073), National High-tech Research and Development Program of China (No. 2012AA011102). We thank Dr. Dong Wang for providing their corpus and useful suggestions. We also thank anonymous reviewers for their valuable suggestions and comments.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrival. ACM Press and Addison Wesley (1999)
2. Celikyilmaz, A., Hakkani-Tur, D.: Discovery of Topically Coherent Sentences for Extractive Summarization. In: Proceeding of ACL 2011 (2011)
3. Erkan, G., Radev, D.: LexPageRank: Prestige in Multi-document Text Summarization. In: Proceedings of EMNLP 2004 (2004)
4. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of SIGKDD 2004 (2004)
5. Ganesan, K., Zhai, C., Han, J.: Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions. In: Proceeding of COLING 2008 (2008)
6. Koumpis, K., Renals, S.: Automatic Summarization of Voicemail Messages using Lexical and Prosodic Features. ACM-Transactions on Speech and Language Processing (2005)
7. Li, F., Tang, Y., Huang, M., Zhu, X.: Answering Opinion Questions with Random Walks on Graphs. Proceeding of ACL 2010 (2010)
8. Li, S., Huang, C., Zhou, G., Lee, S.: Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In: Proceedings of ACL 2010 (2010)
9. Lin, C.: Training a Selection Function for Extraction. In: Proceedings of CIKM 1999 (1999)
10. Lin, C.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of ACL 2004, Workshop on Text Summarization Branches Out (2004)
11. Lin, S., Chen, B., Wang, H.: A Comparative Study of Probabilistic Ranking Models for Chinese Spoken Document Summarization. ACM Transactions on Asian Language Information Processing 8(1) (2009)
12. Mckeown, K., Hirschberg, J., Galley, M., Maskey, S.: From Text to Speech Summarization. In: Proceedings of ICASSP 2005 (2005)
13. Murray, G., Carenini, G.: Detecting Subjectivity in Multiparty Speech. In: Proceedings of Interspeech 2009 (2009)
14. Nishikawa, H., Hasegawa, T., Matsuoand, Y., Kikui, G.: Optimizing Informativeness and Readability for Sentiment Summarization. In: Proceeding of ACL 2010 (2010)
15. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Libraries (1998)
16. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP 2002 (2002)

17. Radev, D., McKeown, K.: Generating Natural Language Summaries from Multiple Online Sources. Computational Linguistics 24(3), 469–500 (1998)
18. Radev, D., Jing, H., Stys, M., Tam, D.: Centroid-based Summarization of Multiple Documents. Information Processing and Management 40, 919–938 (2004)
19. Raaijmakers, S., Truong, K., Wilson, T.: Multimodal Subjectivity Analysis of Multiparty Conversation. In: Proceedings of EMNLP 2008 (2008)
20. Ryang, S., Abekawa, T.: Framework of Automatic Text Summarization Using Reinforcement Learning. In: Proceeding of EMNLP 2012 (2012)
21. Titov, I., Mc-donald, R.: A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In: Proceedings of ACL 2008 (2008)
22. Wan, X., Yang, J.: Multi-document Summarization using Cluster-based Link Analysis. In: Proceedings of SIGIR 2008 (2008)
23. Wan, X.: Using Bilingual Information for Cross-Language Document Summarization. In: Proceedings of ACL 2011 (2011)
24. Wang, D., Liu, Y.: A Pilot Study of Opinion Summarization in Conversations. In: Proceeding of ACL 2011 (2011)
25. Xie, S., Liu, Y.: Improving Supervised Learning for Meeting Summarization using Sampling and Regression. Computer Speech and Language 24, 495–514 (2010)
26. Zhang, J., Chan, H., Fung, P.: Improving Lecture Speech Summarization using Rhetorical Information. In: Proceedings of Biannual IEEE Workshop on ASRU (2007)
27. Zhu, X., Penn, G.: Summarization of Spontaneous Conversations. In: Proceedings of Interspeech 2006 (2006)