An Opinion Summarization-Evaluation System Based on Pre-trained Models

Han Jiang, Yubin Wang, Songhao Lv, and Zhihua Wei(✉)

Department of Computer Science and Technology, College of Electronics and Information Engineering, Tongji University, Shanghai, China
{1853290,1851731,1852635,zhihua_wei}@tongji.edu.cn

Abstract. As social media appeal more frequently used, the task of extracting the mainstream opinions of the discussions arising from the media, i. e. opinion summarization, has drawn considerable attention. This paper proposes an opinion summarization-evaluation system containing a pipeline and an evaluation module for the task. In our algorithm, the state-of-the-art pre-trained model BERT is fine-tuned for the subjectivity analysis, and the advanced pre-trained models are combined with traditional data mining algorithms to gain the mainstreams. For evaluation, a set of hierarchical metrics is also stated. Experiment result shows that our algorithm produces concise and major opinions. An ablation study is also conducted to prove that each part of the pipeline takes effect significantly.

Keywords: Opinion summarization · Subjectivity analysis · Pre-trained model · Evaluation · Hierarchical metrics

1 Introduction

In the post-pandemic era, social media like webinars, message boards, micro blogs, etc., have been increasingly spotlighted and used. Consequently, a special class of data, discussion, is mushrooming all over the Internet. Compared with other textual data, discussion has features as follows: (1) Single topic & multiple opinions; (2) Numerous participants & big volume; (3) Short lifespan; (4) Low structuredness; (5) Multiform expression. The data shows considerable potential for data mining and natural language processing, especially when real-time public sentiment is in demand.

Given the properties above, we place the emphasis on the angles and sentiments of the opinions in discussion. Hence the general process of opinion summarization is to filter the possible opinions out of a discussion, then refine the opinions in terms of their angles and sentiments to obtain the mainstreams.

The reason why opinion summarization requires a two-stage procedure is that a discussion is too extensive to be processed in one go. Meanwhile, speeches in a discussion vary a lot in length, compromising the traditional methods of treating every speech as an equal document. Another trouble is that there are always miscellaneous but semantically identical expressions, which is severely detrimental to generalization.

H. Jiang and Y. Wang—Equal Contributions.

S. Ramanna et al. (Eds.): IJCRS 2021, LNAI 12872, pp. 225–230, 2021.
https://doi.org/10.1007/978-3-030-87334-9_19

To address the aforesaid problems, we propose an opinion summarization-evaluation system including a pipeline and a set of evaluation metrics. For the pipeline, we adopt pre-trained language models to analyze the discussion semantically, and utilize data mining algorithms to accomplish following generalization. For evaluation, we state hierarchical metrics to assess the summary from the relevance, the sentiment orientation, and the one-to-one correspondence between the generated and reference opinions.

In order to implement the algorithm, we also construct two Chinese corpora: a subjectivity analysis corpus for fine-tuning BERT [3], and an opinion summarization corpus for evaluation. An ablation study is subsequently performed by setting several variants of our pipeline, and the result substantiates the effectiveness of our methods.

2 Related Works

There has been a long history of the research on extractive summarization, opinion mining and metrics for these NLP tasks. In recent years, the tasks of extractive summarization are usually fulfilled through neural network modeling, network graph method and data mining. Neural network modeling is the focus of the field [9]. A summary-level framework using SBERT with superior performance was proposed based on this method [8]. Network graph method is a mainstream [9] which stems from a research result: Human language is also a complex network with the characteristic of small world and is scale-free [2]. One of its most representative examples is TextRank [6]. Another important method is data mining. A typical application of this method is clustering. Opinion mining can be divided into three main levels: the text document level, the sentence level and the subject-part level [5]. An important problem in sentence level opinion mining is to classify sentences into subjective ones and objective ones.

Automatic evaluation metrics mainly include BLEU [7], ROUGE [4], and METEOR [1]. BLEU is a similarity evaluation method based on accuracy, which excels on sentences that are well-matched on corpus-level. ROUGE is based on recall, which calculates the co-occurrence probability of n-grams in the candidate sentences and the reference sentences to evaluate the adequacy and fidelity [4]. METEOR is based on single-precision weighted harmonic mean and the recall of single word, and solves the problem of low correlation between BLEU [7] and manual evaluation results [1].

3 An Opinion Summarization-Evaluation Algorithm

In this section, we first introduce our algorithm for extracting the mainstream opinions (see Fig. 1) in Sect. 3.1. In Sect. 3.2, we state our hierarchical evaluation metrics for opinion summarization.

3.1 Subjective Analysis and Opinion Mining

The subjectivity analysis is applied to ensure that the candidate sentences for the final summary are qualified for opinions. With a fine-tuned BERT [3] model, the process is formulated as a binary classification task, where most subjective statements are retained

for the following steps and others are removed. Since there are usually extensive colloquial or objective speeches in a discussion, the process alleviates the problem of data overload as well.

To proceed with the pipeline, we choose *distiluse-base-multilingual-cased-v2* [10] instead of BERT to calculate the semantic representations of the subjective sentences, as BERT is not expert in capturing the semantic meaning of the sentences. Next, the sentences are grouped with spectral clustering algorithm. Spectral clustering relies heavily on the similarity matrix, and the encoder above is verified to work well in extracting semantic information, therefore the two methods are complementary to each other. To balance the integrity and conciseness of the generated summary, we recommend the number of clusters between 3 and 6. Within the interval, we refer to silhouette coefficient, a reasonable and reliable measure to select the optimal clustering result. Moreover, it is necessary to abandon some excessively small clusters.

For each cluster, the vector closest to the geometric center is extracted to be the representation of the cluster, and its corresponding sentence will be the candidate opinion for the summary. Since it may appear colloquial, we just simply remove some irrelevant function words from the sentence to get a mainstream opinion. Finally, all the mainstream opinions acquired constitute the generated summary of our algorithm (Fig. 1).

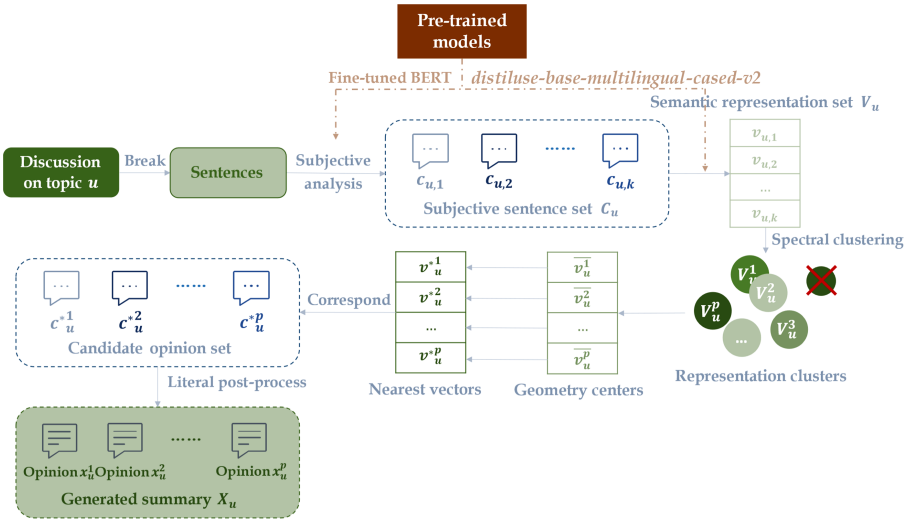

Fig. 1. A schematic diagram of the opinion summarization algorithm

3.2 Hierarchical Metrics

To perfect the opinion summarization algorithm, we state a set of hierarchical metrics, combining automatic and artificial methods to evaluate the generate summary from three aspects progressively.

Evaluate the Relevance on Summary Level Automatically. While assessing a summary, it is most basic to ensure whether it is relevant to the topic, and whether it involves most significant content of the discussion. When design the metric, we refer to the method of collecting the word pairs between the generated and reference summary in ROUGE [4]. For any discussion on topic u, given the generated summary $X_u = \{x_u^1, x_u^2, \ldots, x_u^p\}$ and the reference summary $Y_u = \{y_u^1, y_u^2, \ldots, y_u^q\}$, the relevance between X_u and Y_u can be defined as

$$Relev_u = \frac{1}{q}\sum_{i=1}^{q} Relev\left(X_u, y_u^i\right) \tag{1}$$

$Relev\left(X_u, y_u^i\right)$ denoting the relevance between X_u and opinion y_u^i, is the average value of the cosine similarity between the terms in y_u^i and their most similar terms in X_u. The cosine similarity is computed in the semantic space induced by the model used while clustering. $Relev_u \in [0, 1]$, and the larger $Relev_u$ implies higher relevance.

Evaluate the Sentiment Orientation on Summary Level Automatically. We take the evaluation a step further by examining how the emotion tendency of the generated summary match expectations. With fine-tuned BERT [3], opinions in the summaries can be classified as positive or negative. Then we compare the proportions of the positive opinions in generated summary and reference summary like

$$Senti_u = 1 - \text{abc}\left(\frac{\sum_i Count_{pos}\left(x_u^i\right)}{p} - \frac{\sum_i Count_{pos}\left(y_u^i\right)}{q}\right) \tag{2}$$

It is knowable that $Senti_u \in [0, 1]$. When $Senti_u = 1$, the generated summary captures the sentiment orientation of the discussion perfectly.

Evaluate the One-to-One Correspondence on Opinion Level Artificially. Since automatic approaches may be coarse-grained and inexact, we suggest grading the one-to-one correspondence between the generated and reference opinions manually. Considering an opinion x in X_u and y in Y_u, they can compose a matching pair (x, y) if they show similarity in semantics. Thus, the one-to-one correspondence can be quantified as

$$Corre_u = \min\left\{\frac{\theta_m \sum_{x \in X_u, y \in Y_u} Scr_u(x, y)}{\sqrt{pq}}\right\} \tag{3}$$

Here θ_m is a bonus parameter to improve the score when all the opinions are matched. $Scr_u(x, y) \in [0, 1]$ is determined by the graders, and a higher value implies higher similarity.

4 Experiments and Analysis

4.1 Experimental Settings

With discussions from a large-scale Q&A forum named Zhihu, we build two Chinese corpora. To support the subjectivity analysis, we provide a corpus containing 7500

sentences from 15 discussions, annotated by three annotators as subjective or not, for fine-tuning the BERT model. For the sake of evaluation, we generate reference summaries for another 45 discussions to construct an opinion summarization dataset. Considering a discussion appears as one topic or question with numerous answers in Zhihu, each summary is made up of several thesis statements of the most popular answers.

4.2 Experiment Results and Analysis

With the hierarchical evaluation metrics in the proposal, we assess our algorithm on the opinion summarization corpus. An ablation study is performed over our pipeline and its two variants, using the same corpus and metrics. The results listed in Table 1 illustrate how the critical modules mentioned above take effects.

Table 1. Results of ablation study

Pipeline	*Relev*	*Senti*	*Corre*
SA & CE (Ours)	0.715	**0.730**	**0.428**
No SA & CE	**0.718**	0.702	0.257
SA & TextRank	0.685	0.729	0.252

Corresponding to the above two modules of our pipeline, here **SA** represents the subjectivity analysis, and **CE** stands for center extraction, i. e. the method of extracting the mainstream opinions from the centers of the clusters. The results prove that the algorithm brings fantastic sentiment orientation and one-to-one correspondence, also acceptable relevance.

First, we demonstrate the importance of the subjectivity analysis. In Table 1, **No SA & CE** gets a markedly low *Senti* score, which indicates that removing the subjectivity analysis critically hurts performance in capturing the sentiment orientation. Without the subjectivity analysis, the algorithm tends to be misled by salient but overwhelming contents and produce summaries with biased emotional perception.

Second, we observe the necessity of center extraction. As listed, **SA & TextRank** is defeated by our **SA & CE** with especially large drops on the *Relev* and *Corre* score. A noteworthy fact is that the center extraction gets the central sentence of each viewpoint cluster, this way the mainstreams are guaranteed to be juxtaposed, and semantic overlaps between opinions extracted would be minimized.

Besides, note that the *Corre* score of our pipeline is prominently higher than the other two. That is because the two variants can be misled by crucial and overlapping contents easily, and the rule we use to grading the correspondence severely punishes overlaps. Maybe there are still some unknown benefits brought by our algorithm.

5 Conclusion and Future Works

The contributions of our paper are as follows:

First, we observe a class of recently prevalent textual data, namely discussion, analyze its features and value, and conceptualize the task of opinion summarization.

Second, we propose an opinion summarization-evaluation system with two matching Chinese corpora, and accomplish the task well.

Third, we conduct an extra ablation study to substantiate the effectiveness of our peculiar methods, the subjectivity analysis and the center extraction.

Our opinion summarization-evaluation system paves a new way for automatic summarization, while it still requires further research. In our algorithm, a more flexible measure for clustering result shall be introduced to replace the silhouette coefficient, and more semantic information should be taken into account when locate the centers of the clusters. Also, we will try migrate our system to other languages by adjusting the pre-trained model, the corpora, and some strategies accordingly.

Acknowledgement. The work is partially supported by the National Nature Science Foundation of China (Grant No. 61976160, 61906137) and the Technology Research Plan Project of Ministry of Public and Security (Grant No. 2020JSYJD01).

References

1. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005)
2. Cancho, R.F., Solé, R.: The small world of human language. In: Proceedings of the Royal Society of London. Series B: Biological Sciences, pp. 2261–2265. The Royal Society (2001)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). https://arxiv.org/abs/1810.04805
4. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches out, pp. 74–81 (2004)
5. Liu, B.: Sentiment analysis and opinion mining. Morgan and Claypool Publishers (2012)
6. Mihalcea, R., Tarau, P.: Textrank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
8. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019)
9. Zhao, J.S., Zhu, Q.M., Zhou, G.D., Zhang, L.: Review of research in automatic keyword extraction. J. Softw. (2017)
10. Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., Huang, X.: Extractive Summarization as text matching. In: ACL 2020 (2020)