A Multi-Granularity Opinion Summarization Method

Anonymous ACL submission

Abstract

Existing opinion mining (OM) is limited to applications on commercial reviews, with aspect and sentiment of the opinions in a coarse-grained form. In this paper, we further explore the definition of OM by extending the concepts of aspect and sentiment, and propose an opinion summarization method based on Multi-granularity Clustering and BERT (Jacob et al., 2018), i.e., MCB for emergent online discussion record in keeping with the further definition. A supporting Chinese corpus, ZH45 comprising 45 groups of discussion, and assorted metrics are also proposed. Experiments based on ZH45 and the metrics demonstrate that MCB produces succinct and insightful opinion summaries.

1 Introduction

Opinion mining (OM), usually interchangeable with the term sentiment analysis, is a typical task in the field of natural language processing. In terms of the opinions obtained, the studies of OM can be divided into two types. The one is aspect-based sentiment analysis (ABSA), which aims at generating opinions in the form of triples like (aspect, opinion, sentiment polarity); the other is opinion summarization (OS) aggregating the opinions in textual form. The advantage of ABSA is the structured output, which is conducive to post-processing, while OS is more informative and readable. OM plays a fantastic role in the big data era. With the mainstream opinions obtained, people can straightforwardly grasp overall cognition, then make decisions about the object of concern without browsing every piece of information.

Despite the considerable advantages, the full development of OM is still far due to two limitations: (1) The premature technical system. In existing researches, aspects of the studied entity are always predictable and concrete attributes (e.g., the hygiene and service for hotel, the wine, flavor, and price for restaurant, etc.). Similarly, the sentiments or opinions are limited to simple description, which can be interpreted as adjectives, or even sentiment polarities. The definition attenuates the difficulty but also the versatility of OM. (2) The supporting evaluation system and corpora have yet to be complemented. The assessment of OS has to mirror that of automatic summarization, adopting metrics such as BLEU (Kishore et al., 2002) and ROUGE (Lin, 2004). The rigorous supervised metrics are not suitable for semi-supervised or unsupervised summarizers, as they tend to underestimate the opinion summaries. Furthermore, the annotation of reference summary is burdensome. Existing quality corpora are all commodity review sets (Chu and Liu, 2019; Bražinskas et al., 2020), preventing OM from penetrating into public opinion monitoring, current affair summary, and big data sentiment analysis.

On a separate note, under the gaze of COVID-19, the face-to-face contact has to be restricted. Consequently, social media like webinars, message boards, micro blogs, etc. have been increasingly spotlighted and used (Jiang et al., 2021), and online discussion record (hereinafter referred to as discussion) has been mushrooming all over the social network. The contents of discussion cover quite a board realm, implying substantial exploring value. However, compared to dialogue, news, and commercial comment, discussion possesses excessive volume, relax structure, and miscellaneous expressions, which really impede the progress of its processing.

Given the task framework and the online discussion record above, we propose to further explore the definition of OM by extending the aspect and sentiment from concrete, knowable concepts to abstractive, implicit concepts.
Subsequently, we introduce an OS method based on Multi-granularity Clustering and BERT (MCB) for summarizing the upstream of complicated textual data resembling discussion in an unsupervised, extractive fashion. Through statistical analysis on item attributes disassembling the text into sentences and terms, gained from customers’ comments helped the MCB bypasses the unstructured problem. At different levels of granularity, MCB employs suitable clustering algorithms in line with the data characteristics and phased needs (e.g., Angelidis et al., 2021). As OM evolved, the dimension of dimensionality or volume reduction, split-flow, and emotion had increased. Conrad and Schilder (2007) aggregation). In sentence level, we leverage fine-tuned BERT models to inject external knowledge when mining opinions from legal blogs. De into the framework and advance the exploitation of subjective analysis and TransfoRank analysis to MCB.

For the sake of assessment, the paper introduces a Chinese corpus, ZH45 comprising 45 groups of surveys (Ku et al., 2006; Meng et al., 2012), discussion from Zhihu. Zhihu is a large-scale Ganesan et al. (2010) were the first to state the Chinese forum, where objects discussed range from deep semantics. With the aid of BERT, we add positive and negative texts as emotion summaries. In the period, some researchers also tried to extract the informative and insightful sentences from the corpus and metrics, experiments including ablation studies and a case study are conducted to prove the practicality and superiority of MCB.

In a nutshell, our contributions include: (1) we comb through the flux of OM, and put up with a deepened conception of the task in the light of tagging and classification, but significantly emerged data sources, the discussion; (2) we propose MCB, an OM method based on multi-granularity clustering and the SOTA language model, BERT for the further task as a baseline, and prove its effectiveness with abundant experiments; (3) we proffer a Chinese corpus consisting of 45 discussions, and an evaluation system for more unsupervised opinion summarizers to refer to in the future.

2 Exploration of the concept of OM

2.1 Development of Opinion Mining

The idea of OM was raised in 2003 as “processing a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)” (Dave et al., 2003), followed by a technique of classifying the product review.
other summarization tasks, for some second mainstreams would be omitted.

**Excessive volume.** Faced with the extensive text, the methods have to cut it down to a manageable size, or be invulnerable to data overload problem.

**Relax structure.** An individual essay, paragraph or even sentence can become a comment in a discussion, making it difficult to start from the natural comment level as usual.

**Miscellaneous expressions.** Myriad ironies and degressions are blended in a discussion, hoodwinking the summarizers from figuring out true mainstreams. In addition, similar semantics may hide behind distinct expressions, which leads to the sparse problem.

At the beginning of our method, we trim the relax structure by splitting the given discussion into sentences. Through limiting the length of each coping with complicated texts, as they extract sentence and selecting the separators, we try to **aspect and sentiment** in the form of **terms**. A single term can hardly represent an aspect of a whole most, which is different from a linguistic sentence. Then the subjectivity analysis is implemented by a BERT model. The model is simply fine-tuned on 7500 sentences manually labeled with subjectivity (e.g., 1 for subjective, and 0 for objective). Like general classification tasks, with the hidden state $h_0$ of the [CLS] token, the subjectivity is calculated as:

$$y^i = \text{softmax}(W^i h_0 + b^i)$$ (1)

3 **Methodology**

3.1 **Preprocessing**

The further OM should be able to break the limitations. Inspired by “opinion mining analyzes people’s opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes” (Liu, 2011), we suggest selecting such individuals, issues, events, and topics as research target, and generalizing as:

- **ASPECTS**
- **SENTIMENTS**
- **OPINIONS**

![Figure 1: An overview of the OS method based on Multi-grained Clustering and BERT (MCB) in the paper. (a) In preprocessing stage, the fine-tuned BERT models function as filters. (b) In aspect extracting stage, we cluster twice to obtain the extended aspects. (c) In sentiment generalizing stage, affinity propagation and TransfoRank analysis are conducted simultaneously.](image-url)
Apart from alleviating the burden brought by the cumbersome texts, we also insist that sentences with stronger subjectivity are qualified for candidate opinions. The subjective sentences are collected into a set denoted as \( S = \{ s_1, s_2, \ldots, s_m \} \).

Another pre-trained ERNIE model (Xiao et al., 2021) is employed for tokenization and part-of-speech (PoS) tagging. Here we mimic the practice of topic modeling, reserving the tokens with specific part-of-speech tags to eschew non-sense words. A stopword list is also used in the process. Consequently, we get the significant word set \( W = \{ w_1, w_2, \ldots, w_n \} \).

### 3.2 Aspect extracting

The aspect extracting stage identifies several subsets of \( W \) to constitute the aforesaid extended aspects of the complex discussed.

After observing the data, we come up with three conditions where words are likely to be relevant.

- Synonyms and antonyms. That means the words are always interchangeable in the text. From another perspective, they would be especially close in semantic space.

- Logical dependencies. For example, *nurse* and *surgery*, *politics* and *economy*, *bloom* and *fruit*, etc. The various dependencies bring the words together, reflecting global statistical characteristics.

- Relevance under particular circumstance. It means that the words are irrelevant usually but relevant within a certain discussion, which corresponds with local statistical characteristics in task-related corpus.

Firstly, we decide to adopt the Word2Vec representations (Mikolov et al., 2013) and the agglomerative clustering algorithm to aggregate the synonyms and the antonyms into *word-points*.

The plain algorithm can conveniently control the extent of aggregation by set a similarity threshold. In this paper, we set the threshold to 0.6 empirically and get satisfactory word-point set \( P = \{ p_1, p_2, \ldots, p_n \} \). On the other hand, the short texts are always sparse, noisy and ambiguous (Shi et al., 2018), and the operation mitigates the problems by reducing the dimensions.

Secondly, to give consideration to both global and local relevancy, we compute the frequency of each word-point in \( S \):

\[
S_{pi} = \{ s \mid w \in p_i, s \text{ contains } w, s \in S \}
\]

\[
freq_i = |S_{pi}|
\]

and the same for co-occurrence frequency \( C_{0ij} \) of every two word-points. Co-occurrence is accessible yet plausible in smaller corpus, especially when the discussion is not large enough to fine-tune the word vectors.

The word-points with low frequencies are removed through a threshold related to the scale of the discussion. Thus, the similarity matrix of remaining word-points can be calculated as:

\[
Sim_{i,j} = \frac{pv_i \cdot pv_j \cdot \ln(C_{0ij} + \epsilon)}{\|pv_i\| \|pv_j\|}
\]

where \( pv_i \) stands for the representation of word-point \( p_i \) obtained by averaging the vectors of the words in the word-point. The similarity matrix is the input for spectral clustering (Ng et al., 2002).

The graph-based algorithm tallies with the organizational form of the word-points in the discussion. We set the number of clusters between 3 and 6, and the best number is assigned according to the silhouette coefficient. The output clusters of word-points \( A = \{ a_1, a_2, \ldots, a_p \} \) are candidates for the extended aspects, among which we will get some clusters of non-sense words. To weed them out thoroughly, we identify these loose clusters by comparing the intra-cluster and inter-cluster co-occurrence:

\[
compact_i = \sum_{j \neq k} C_{0ij} - \sum_{j \neq k} C_{0ik} \]

\[
CR_i = \frac{\sum_{j \neq k} compact_{k,i} \cdot compact_{k,j}}{(p-1) \cdot compact_i} \]

where \( compact_i \) is the sum of \( C_{0ik} \) of every two word-points in the candidate aspect \( a_i \), and \( CR_i \) means the compact degree of \( a_i \). In our method, the loosest cluster will be abandoned if its compact degree is greater than a certain value, and there are enough clusters \( (p > 4 \text{ in our study}) \). Then we get the final aspects \( A' \).

Thirdly, the sentences in \( S \) are categorized into the aspects above. The words involved in \( A' \) can vote for the sentences they belong to:

\[
Count^s_w = \begin{cases} 1, & s \text{ contains } w \\ 0, & \text{otherwise} \end{cases}
\]

\[
Vote^s_i = \sum_{w \in A} Count^s_w \cdot freq_i
\]
In which $Vote^a_s$ is number of votes for sentence $s$ going to aspect $a$. In this way, the word-point frequency act as the voting weight. After voting, most sentences are grouped under one or more aspects, while seldom that contain no voter words will be left out.

### 3.3 Sentiment generalizing

The sentiment generalizing stage is intended to further aggregate the subjective sentences in each group in terms of the emotions they expressed. Given a group under an aspect $S_{ai} = \{s_i^1, s_i^2, \ldots, s_i^k\}$ we first get its embeddings $SV_i = \{sv_i^1, sv_i^2, \ldots, sv_i^q\}$. The encoding model is a Chinese BERT model pre-trained with whole word masking ($\text{Cui et al.}, 2021$) and fine-tuned on a Chinese natural language inference (NLI) dataset. Although the addition of the model enhances the understanding of deep semantics behind the multiform expressions, note that our method is not tied to any BERT model, including the aforementioned ones for subjectivity analysis and PoS tagging.

In the following, $SV_i$ is fed into two algorithms in parallel. Assuming that sentences under the same topic vary in attitude and emotion most, affinity propagation (AP; Frey and Dueck, 2007) is adopted to generalize the extended sentiments. AP excels at clustering multi-class, high-dimensional data, but it has higher complexity than other algorithms. In view of the above, we apply AP to this latter step for the sentiment clusters like $S_{ai,e} = \{s_i^{e1}, s_i^{e2}, \ldots, s_i^{ej}\}$ ($e$ for emotion).

Simultaneously, we put forward TransfoRank by replacing the original similarity function in TextRank (Mihalcea and Tarau, 2004) with a cosine similarity matrix of $SV_i$ to work out the popularity ranking of $S_{ai,e}$.

We mimic the skill factor in multifactorial evolutionary algorithm (MFEA; Gupta et al., 2015) to design the popularity factor of the sentiment clusters:

$$Pop_{ij} = \min_{sv_{ai}, ej} TransfoRank(s)$$  \hspace{1cm} (9)

where $TransfoRank(s)$ is the ranking of the sentence by TransfoRank. A smaller popularity factor signifies higher popularity. The central sentences of the K most popular sentiment clusters are extracted for the mainstreams of the aspect. Ultimately, mainstreams coming from all aspects compose the opinion summary.

### 4 Evaluation of furthered OS

#### 4.1 The ZH45 Corpus

In order to add fuel to the research of further OM, we introduce ZH45, a medium-scale Chinese OM corpus. The ZH45 is constructed on the well-known Q&A community on the Chinese Internet, Zhihu. In the community, the users can pose questions, and discuss other users’ questions in turn. Zhihu has a column named “How do you view / evaluate X”, where X symbolizes the object discussed, covering social phenomena, news, particular communities, interpersonal problems, and celebrities, etc. The discussion taking place in the column meets the definition of discussion in the paper. We selected and crawled 45 of them. After filtering out the non-text comments, 165K comments were collected in total. The number of comments under each question ranges from hundreds to more than 10K, and the comments vary in length from a few characters to thousands of characters. The corpus contains no reference summaries, which is helpful for unsupervised methods. Actually, the crowdsourcing is unfeasible, as it is unrealistic for people to digest then summary discussions with thousands of comments. The high-quality OS corpus SPACE (Angelidis et al., 2021), where every human-annotated summary is based on 100 reviews, has reached the largest crowdsourcing in the field.

#### 4.2 Metrics

Referring to the work of Angelidis et al. (2021), we evaluate the method from two angles: aspect and opinion summary.

In this paper, the change of the concept of aspect is noteworthy. We believe that there are different appropriate observation angles for different objects, resulting in different aspects. Therefore, we avoid commenting on the right and wrong of the aspects, but focus on their role in the diversion of the subjective sentences.

**Aspect Variance** is for watching the uniformity of the size of the $S_{ai}$. The variance of the sizes of them is taken for the metrics. Considering that between the opinions exist the popularity gaps, a moderate Aspect Variance is acceptable.

**Aspect attraction** aims at measuring the capacity of the aspects to gather the sentences in a given discussion. We divide the number of subjective sentences collected by $A^+$ by the number
of word-points (or words in the case of other methods without word-points) in $A$ for the metric.

Having difficulties in gaining ground-truth annotation in OS research, we are confronted with the challenge of evaluating the existing results without any reference summaries. Human evaluation is indispensable in the situation.

**Summary Richness** aims at quantizing the information retrieved by the opinion summary. It is computed by preprocessing the summary to get $W$, and divide the size of $W$ by the number of opinions in the summary.

**Summary Coverage** examines the ability of the mainstreams in the summary to represent other comments within the discussion. For the metric, we recruited 6 annotators, including undergraduates, graduate students and white-collar workers to carry out a human study. 150 sentences sampled from 5 arbitrarily were offered with the opinion summary of the discussion, and the annotators had to decide if the sentences were represented by the summary respectively. Each sentence was annotated for 3 times, and there is an average cover rate. We divide the rate also by the number of opinions in the summary, to eliminate the influence of the summary size.

## 5 Experiments

### 5.1 Experimental Setup

**Dataset.** As far as we know, ZH45 is the first corpus to serve further OM, and our experiments take it as the testbed.

**Implementation.** All of the discussions in ZH45 were involved in evaluation. In the preprocessing stage, the model for subjectivity analysis was BERT-base-Chinese (Jacob et al., 2018). As for fine-tuning, we collected 7500 sentences randomly from another 15 posts in the “How do you view / evaluate X” column, and the sentences were annotated by 3 annotators independently as subjective or not. Tokenization and PoS tagging were implemented by a pre-trained multi-task ERNIE model (Xiao et al., 2021) offer in HanLP project (He and Choi, 2021). In the sentiment generalizing stage, we borrowed RoBERTa-wwm-ext (Cui et al., 2021), and fine-tuned it in the framework of Siamese-BERT (Reimers and Gurevych, 2019) using multiple negatives ranking loss. The NLI dataset for fine-tuning was a combination of OCNLI (Hu et al., 2020) and a Chinese NLI corpus built on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). During training, we used the Adam optimizer, with initial learning rate of $3 \times 10^{-5}$. We warmed up the model for the first 10% steps, and ran 5 training epochs in total. The preference in AP clustering was simply set to -1. We let $K = 5$ while choosing the most popular sentiment clusters. More implementation details agree with the method explained in Section 3.

### 5.2 Results

As for aspect evaluation, noticing the aspect is a collection of words, which is consistent with the concept of the topic, we select LDA (Blei et al., 2003) and LSI (Deerwester et al., 1990) as the baseline models. The upper part of Table 1 shows the Variance and Attraction scores of our method (MCB), LDA, and LSI, which is an encouraging result. Aspects from MCB evidently outperform that from the general topic models in helping produce the opinion summary. From the outputs, LDA and LSI tend to generate highly overlapped word sets, and the sentences are likely to amass around the one with the highest average weight. We conjecture that it is because the discussion has already targeted at a certain object, and the aspects may function as sub-topics under the overall topic, while the typical topic model may concentrate on the latter. Apparently, our method has the ability to discover and distinguish more fine-grained relations.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Methods</th>
<th>Variance</th>
<th>Attraction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCB</td>
<td>92270</td>
<td>2.9754</td>
</tr>
<tr>
<td></td>
<td>LDA</td>
<td>608132</td>
<td>0.7096</td>
</tr>
<tr>
<td></td>
<td>LSI</td>
<td>645968</td>
<td>0.6663</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary</th>
<th>Methods</th>
<th>Richness</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCB</td>
<td>5.7339</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>BERT-Spec</td>
<td>5.1497</td>
<td>0.0214</td>
</tr>
</tbody>
</table>

Table 1: MCB compared with our baselines.
MCB and the baseline (BERT-Spec) is listed in the Table 1 (lower). It can be seen that MCB gains the upper hand with more informative and representative opinion summaries.

5.3 Ablation Study

Since few baseline models can be found, we conducted ablative experiments to confirm the effectiveness of MCB as an extended OS method. The variants considered are as follows:

- **MCB-Nopnt**: The procedure of aggregating the words into word-points is removed.
- **MCB-Noco**: The co-occurrence between the word-points is dismissed in the aspect-extracting stage.
- **MCB-FA**: While clustering the word-points, spectral clustering is replaced by agglomerative clustering.
- **MCB-Noemb**: No BERT models are involved in the sentiment generalizing stage. The top-K popular sentences are worked out by TextRank.
- **MCB-Spec**: The AP clustering is replaced by spectral clustering in the sentiment generalizing stage.

The MCB-Nopnt, MCB-Noco, and MCB-FA are variants making changes in aspect extracting stage. They are compared with the full method in the aspect evaluation in Table 2 (upper). Among the methods, MCB has a moderate Variance and the highest Attraction. The result of MCB-FA is particularly irrational, and we expelled it from the summary evaluation. In the lower part of Table 2 are some interesting findings. MCB-Noemb wins the highest Richness, followed by our methods. However, MCB maintains the best Coverage far beyond the MCB-Noemb. The reason lies in the summaries themselves: TextRank, the word-frequency-based algorithm, is prone to be misled by sentences that are tediously long and not suitable for mainstreams.

From the above results, it can be concluded that the data dimension reduction, exploitation of statistical characteristics in the task-related data, appropriate clustering algorithms, and BERT models for grasping the deep semantics are all imperative for our method.

5.4 Case Study

To provide deeper insight into the advantages of our method, we adopt the discussion on the American "fake smile boy" Gavin for a case study. The aspects generated by MCB and the variant with the lowest Variance, MCB-Noco, are shown in Table 3. MCB produces 4 aspects, and we can summarize them as behavioral intention, Gavin himself, social effects, and public reactions. MCB-Noco offers 6 aspects, among which 3 aspects are nearly the same with the social effects (the first), public reactions (the fourth), and Gavin himself (the last). The remaining 3 aspects are ambiguous, though.

In Figure 2, we also present the opinion summaries created by MCB and BERT-Spec.
through-synthesis techniques are not a long-term solution. I feel I am not the first person to have this feeling, and I feel like some body would say that I don’t know the difficulties.

I can’t work hard. However, that’s people’s helplessness and sadness.

Having compassion without enough strength to reverse the reality, that’s people’s helplessness and sadness. (Having compassion without enough strength to reverse the reality, that’s people’s helplessness and sadness.)

In my opinion it really makes sense.

The way of getting a return is not worth pursuing.

As long as he is responsible for his choice, it isn’t honorable or rewarding for me.

I feel like some day, I will be happy.

If he feels happy and loves such life, it’s wo.

Some other children have an unusual tone, it’s not cute because of himself rather than the fake smile, however that’s attitudes are inappropriate.

I will be happy.

I hope that the rest worrying about...


Kavita Ganesan, Chenguang Zai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. https://www.ideals.illinois.edu/handle/2142/16949.


Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. TaxoGen: unsupervised topic taxonomy construction by adaptive term embedding and clustering. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2701-2709. https://doi.org/10.1145/3219819.3220064.