Weakly Supervised Opinion Summarization with Focus on Implicit Aspects

2310434 NGUYEN, Khanh Vinh

In the era of user-generated content explosion, many social media and e-commerce sites, such as Yelp and Amazon, receive millions of reviews daily. These textual data sources provide valuable insights into customer satisfaction and product quality. However, a huge volume of customer reviews poses a significant challenge for manual analysis by users or enterprises. To tackle this problem, the study of opinion summarization has emerged and has received much attention.

Opinion summarization is a task that aims to automatically generate a concise and informative summary for a given set of reviews, which reflects the overall preference of users. Unlike conventional summarization tasks focusing on factual or event-based content, opinion summarization aims to address subjective expressions and fine-grained aspect-level sentiment, which are denoted by a wide variety of linguistic expressions.

One essential challenge in opinion summarization is the prevalence of implicit opinions where users' sentiments are not directly stated through obvious aspect-opinion pairs but instead emerge from contextual cues. For instance, the sentence "I had to wave three times before getting help" implies dissatisfaction with the quality of the service, even though the word "service" is not explicitly mentioned. Most existing methods in opinion summarization tend not to focus on capturing such implicit expressions. Furthermore, the majority of current approaches rely on supervised learning, which requires large-scale datasets with manually annotated reference summaries. However, the construction of such datasets is time-consuming and expensive. A lack of labeled data severely restricts the scalability and adaptability of opinion summarization systems across new domains or languages.

This study proposes a novel weakly supervised method for opinion summarization. The main idea is to create mixed-structured data that consists of both structured data and unstructured data, which is derived from a set of reviews, and a pseudo-summary of these reviews. The structured data contains opinion-aspect (OA) pairs that explicitly associate an aspect of a product with users' sentiment, while unstructured data consists of implicit sentences that express users' opinions implicitly. In addition, a ground-truth summary is associated with the structured and unstructured data by selecting a representative review from a review set as a pseudo-summary. This approach addresses the challenge of manual annotation of reference summaries while still ensuring that the model is exposed to rich and sentiment-aware training data. The proposed method consists of two main stages: the construction of mixed-structured data and the training of a summarization model.

In the first stage, the mixed-structured is constructed by the five steps. First, OA pairs are extracted using a large language model (LLM), i.e., the LLaMA-2 chat model. We instruct the LLM to extract OA pairs that fulfill strict syntactic requirements (e.g., OA should be a noun-adjective pair). Unlike traditional rule-based or dependency-based extraction methods, which are limited and hard to extract OA pairs from noisy reviews across diverse domains, the LLM-based approach provides greater flexibility, allowing the model to more appropriately handle ungrammatical sentences. Second, opinionated sentences that lack explicit aspect terms are extracted as implicit sentences (ISs). Third, both the OA and IS are annotated with sentiment polarity, i.e., positive or negative, to indicate the emotional orientation of the users' feedback. Fourth, a single review within the set of reviews is chosen as a pseudo-summary by a sentiment-aware selection strategy. For each time, a pseudo-summary is expected to satisfy the following four criteria: (1) it covers commonly mentioned aspects, (2) it maintains a balance of positive and negative sentiment, (3) it does not contain first-person pronouns (e.g., "I", "my"), (4) it does not contain contents other than sentences. This pseudo-summary acts as a substitute for a ground-truth summary, although it is not originally written as a summary. Fifth, the OA pairs are sampled to make a diverse mixed-structured dataset. We divide the OA pairs into two types: popular OA and unpopular OA. The popular OA appears in other reviews within the dataset. The soft matching based on cosine similarity between two embeddings of OAs is employed to check whether an OA appears in other reviews. The unpopular OAs are not included in the pseudo-summary; they are randomly sampled. Meanwhile, all implicit sentences are kept intact. The pseudo-summary selection and OA sampling steps are repeated until no more pseudo-summaries satisfy the aforementioned conditions to construct the final mixed-structured data.

After constructing the mixed-structured data, a summarization model is trained. It is a dual-encoder sequence-to-sequence model based on the pre-trained language model BART. This architecture consists of two encoders for processing opinion-aspect pairs and the implicit sentences. Each input is prefixed with a special token ([OA] or [IS]) and sentiment tag (positive or negative). In the decoder, the representations of OAs and ISs are jointly attended using attention fusion, thereby allowing the model to generate a structurally faithful and sentiment-aware summary. The use of the dual-encoder enables us to consider both explicit and implicit opinions in the reviews, which is essential for handling noisy reviews in real-world settings.

Experiments on the Yelp and Amazon datasets are carried out to evaluate the effectiveness of the proposed method. Despite using only around 43K mixed-structured data in the Yelp dataset and 11K in the Amazon dataset,

which are far fewer than the 100K+ used in the previous study, our model achieves competitive or superior results. Our model does not exceed the baselines in terms of ROUGE-1 and ROUGE-2 scores, which measure strict lexical overlap between generated and reference summaries. However, our method outperforms all baselines in terms of the ROUGE-L metric, which evaluates the longest common subsequence between generated and reference summaries. This finding suggests that our model produces more fluent and coherent summaries. Furthermore, our model achieves the lowest self-BLEU score on the Amazon dataset, indicating that it generates lexically diverse summaries.

To validate the effectiveness of each component, an extensive ablation study is conducted by removing or altering individual components. The removal of ISs results in the most significant drop in ROUGE-L, confirming the vital role of implicit content in ensuring the quality and coverage of generated summaries. The elimination of OAs from mixed-structured data also degrades performance, affirming the efficacy of both structured and unstructured data in the mixed-structured dataset. Finally, the utilization of randomly selected reviews as pseudo-summaries results in a decline in the performance of opinion summarization when compared to the use of sentiment-balanced and aspect-rich pseudo-summaries selected by the proposed criteria.

Some limitations are found during the mixed-structured data construction stage, especially in the extraction of OA pairs. Since OA pairs are generated by prompting an LLM rather than relying on gold-standard annotations, some of the extracted pairs are erroneous or semantically ambiguous. Nevertheless, our model demonstrates robustness in producing coherent summaries, suggesting that the integration of explicit opinion-aspect pairs and implicit sentences, as well as adding the sentiment of OAs and ISs to the input, can contribute to generating higher-quality summaries.