Topic-aware video summarization technique for product reviews exploiting the BERTopic and BART models

Yu-Jin Ha¹, Gun-Woo Kim²*

¹ Department of AI Convergence Engineering, Gyeongsang National University, Jinju, Korea
² School of Computer Science/Department of AI Convergence Engineering, Gyeongsang National University, Jinju, Korea

Abstract
Recently, there has been a growing trend of consumers seeking product information from video platforms such as YouTube. However, when viewing multiple review videos about the same product, viewers often encounter redundant information, resulting in wasted time. To address these issues, we use BERTopic to eliminate repetitive video content and address the problem of missing subtopics, which has been a limitation of traditional video summarization methods. Subsequently, the topic-aware video contents are summarized using the BART model. The ROUGE metric was used to evaluate the model proposed in this paper, and the experimental results showed improved results compared to previous research.

Keywords
Multi video summarization, Product review summarization, BERTopic, BART

1. Introduction

As the use of mobile internet media has increased since the COVID-19 pandemic, the proportion of people who prefer online video platforms as their main source of news and current affairs information has increased. YouTube is one of the most popular online video platforms, and the percentage of users who use it to get news and current affairs information has increased from before the pandemic. [1], [2] Videos provided by video platforms can be as short as a few seconds and as long as several hours or more. It requires time and effort to watch all these long videos. To reduce this waste of time, video summarization is necessary. A video summary is a short summary of the important content of the entire video or what the user requested. [3]

Video summarization can be based on audio, vision, and textual data (i.e., subtitle) [4], [5]. Most traditional video summarization approaches use only low-level visual images as data. [6], [7], [8], [9], [10] Traditional video summarization lacks the understanding of semantic meaning and relationships in videos. This means that they need to explore not only visual images, but also audio and textual information. Focusing on textual information can outperform traditional video summarization in terms of time and space for specific domains such as education (i.e., lectures, tutorials, and information), knowledge (i.e., news and documentaries), sports, entertainment industry (i.e., movies), medicine, and product reviews. [3]

There are several methods for summarizing textual information in video summaries. Summarization of textual information using video images is possible through image capturing [11]. Audio data is summarized by text-based methods using speech-to-text techniques. Video subtitle summaries are summarized using text summarization algorithms using text data. The resulting summaries exist as short videos or text data. [4], [5]
The increase in the use of the YouTube platform as a main source of information mentioned earlier is the same for product reviews. Product reviews are a way for buyers to learn about the pros and cons of a product and its features before making a purchase. If a product review is long, it takes a lot of time to read through the content of the entire review. For this reason, there is a need for product review summaries that minimize the length of product reviews while preserving their content. [12], [13] With the increase of product review videos on the YouTube platform, there has been an increase in the number of multiple product review videos for a single product. Multi-video product review summaries provide product reviews from different perspectives for buyers who want to purchase a product and help sellers or companies to improve their quality of products and services [14].

In this paper, we aim to identify the semantic meanings and relationships of product review videos through subtitle summaries and provide product reviews from different perspectives through multi-video product review summaries.

![Video Summarization](image)

**Figure 1:** The scope of the video summarization

### 2. Related Works

Extractive summaries are summaries that attempt to extract the entire summarized document into sentences, phrases, and words from the source document. Generative summaries are summaries in which the summarized document generates a summary using words, sentences, and phrases that are not present in the original document. [15] In this paper, we use extractive summarization to preserve the review video creator's intent by using sentences extracted from the original document to increase the reliability of the review.

Topic modeling is a statistical model that identifies the topics of a set of documents in the field of machine learning and natural language processing. It analyzes large amounts of textual data to help identify the embedded semantic structure of the text and summarize its key content. [16] In this paper, we use topic modeling to identify various embedded topics and subtopics.

Ansamma et al. (2017) [17] maximized relevance and minimized redundancy by representing them as word vectors, and multi-document extractive summarization using Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF) as objective functions. Alrumiah et al. (2022) [18] is a single summary of lecture video subtitles using LDA. By generating a keyword list using LDA and extracting sentences from the original document that contain at least one word from the keyword list, they improved the summarization performance in terms of length and quality compared to previous studies. Miller et al. (2019) [19] summarized a single lecture video using BERT. This is the first study in which a large language model (LLM) is used for video summarization, and the LLM model, BERT, is used to embed the input sentences. Then, the n most centroid sentences were selected using k-means to summarize the lecture video subtitles. The
performance evaluation was compared with TextRank, but since there was not Golden summary, it was not evaluated using evaluation metrics.

3. Proposed Method

![Figure 2: product multi-video summary architecture](image)

3.1. Preprocessing

In the data preprocessing stage, we remove emoticons, onomatopoeia, and onomatopoeia, which are unnecessary information in this summary. For onomatopoeia and onomatopoeia containing more than one consonant and vowel, we removed all but one word. If there were missing values and subtitles other than Korean, the columns were deleted. In addition, we removed stop words, which are information that does not add meaningful information to the sentence. We used the list of Korean stop words provided by NLTK.

3.2. Topic Modeling

BERTopic is a popular BERT-based algorithm for topic modeling. It utilizes Transformer and c-TF-IDF to generate dense clusters while maintaining important words in the topic. The process of BERTopic includes three steps: Document Embedding, Document Clustering, and Topic representation. Each step is designed to be independent and can be customized according to the purpose [20].

In this paper, when comparing topic modeling algorithms, we found that BERTopic shows equally good results in Topic Coherence and Topic Diversity. This means that it generates topics with diversity while maintaining the coherence of the document’s topics. In addition, BERTopic has the advantage of requiring relatively little time even as the size of the vocabulary increases. For these reasons, we chose BERTopic for this paper.

3.3. Summarization

BART is a denoising autoencoder based on the Transformer architecture.[21] The corrupted text is input to the encoder and the text representation learned by the encoder is sent to the decoder, which recovers the original undamaged text. BART is a model that can be trained by combining natural language understanding and generation, and performs well on summarization tasks.
3.4. Video Making

This is the process of converting summarized text to video. The process extracts the number of the video where the text is located and the timeline of the video where the text is located. The video is then cut to include the summarized content. The process is iterated over the number of summarized texts, and then the summarized videos are joined together into a single video. Figure 3 is an image that captures the summarized video created through this process.

![Figure 3: Captures of created video](image)

4. Experiments

4.1. Dataset

For Korean data, there is no existing product review data for multi-video summarization, thus we created a dataset in this paper. The dataset is composed of 271 Korean product review videos of 22 different products such as cell phones, game consoles, and robot vacuum cleaners. Shown in Figure 4 below is the shape of the dataset. Each column of the data is in the format of product name, video title, video link, video subtitle, time of video matching with subtitle, and answer data. The answer data was generated randomly.

![Figure 4: Structure of Dataset](image)

4.2. Evaluation metric

For evaluation metrics, we used Topic Coherence and Topic Diversity to evaluate Topic modeling, and Rouge for summary evaluation.

4.2.1. Topic Coherence

Topic Coherence evaluates Normalized Pointwise Mutual Information (NPMI) [22], a method proposed by Röder et al. (2015) [23]. The evaluated result has a value between -1 and 1. The closer it is to 1, the more the coherence of the topic.
\[ \vec{v}(W') = \left\{ \sum_{w_i \in W'} NPMI(w_i, w_j) \right\}_{j=1, \ldots, |W'|} \]

Let \( W' \) be a vector representing the similarity between words in the corpus. \( \vec{v}(W') \) computes the similarity of each word in \( W' \) to the other words and represents it as a vector.

\[ NPMI(w_i, w_j)^\gamma = \left( \frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} - \log (P(w_i, w_j) + \epsilon)}{\epsilon} \right)^\gamma \]

NPMI is a metric that measures the relevance of two words \( w_i, w_j \). \( P(w_i, w_j) \) is the probability that two words \( w_i, w_j \) co-occur. \( P(w_i) \) is the probability of \( w_i \) occurrence and \( P(w_j) \) is the probability of \( w_j \) occurrence. \( \gamma \) represents an exponent and serves to weight the NPMI values exponentially.

\[ \Phi_S(i, \vec{w}) = \sum_{i=1}^{|W|} u_i \cdot w_i \]

\( i \) represents the dimension of the vector. \( \|\vec{u}\|_2 \) represents the L2 norm of vector \( \vec{u} \), \( \|\vec{w}\|_2 \) represents the L2 norm of vector \( \vec{w} \), and \( \Phi_S(i, \vec{w}) \) computes the similarity between two vectors \( \vec{u}, \vec{w} \) by dividing the dot product of the vectors by the L2 norm of the vectors.

**4.2.2. Topic Diversity**

Dieng et al. (2020) [24] proposed a metric to measure the diversity of vocabulary or words in the topics of a certain topic model. The metric is the proportion of unique words, excluding duplicate words, that exist in the topics for all topics, and the measure ranges from 0 to 1. 0 represents duplicate topics and 1 represents various topics.

\[ \text{Topic Diversity} = \frac{|\{\text{unique words}\}|}{\text{topk} \cdot |\text{topics}|} \]

**4.2.3. Rouge**

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) is a co-occurrence statistical measure of N-grams and is to be defined as follows (5). \( \text{Count}_{\text{match}}(N \text{- Gram}) \) is the number of N-grams that are included in both summarized results and reference summary. \( \text{Count}(N \text{- Gram}) \) is the number of N-grams included in the reference summary. These ROUGE metrics can generate three scores: Recall, Precision, and F-measure [25].

\[ \text{ROUGE} - N = \frac{\sum_{S \in \sum_{\text{ref}}} \sum_{N \text{- grams} \in S} \text{Count}_{\text{match}}(N \text{- Gram})}{\sum_{S \in \sum_{\text{ref}}} \sum_{N \text{- grams} \in S} \text{Count}(N \text{- Gram})} \]
4.3. BERTopic

We compared the performance difference between embedding models using TC (Topic Coherence) and TD (Topic Diversity), as the performance difference exists depending on the embedding model of BERTopic. As shown in Table 1, the multilingual model of SBERT, distilbase-multilingual-cased-v1, recorded the best score, thus we compared the topic model with this embedding model.

We calculated the TC and TD of three models, the BERTopic model with distilbase-multilingual-cased-v1 as the embedding model, the Combined Topic Models (CTM) model, and the LDA model, and found that BERTopic shows the best performance, as shown in Table 2. This shows that the BERTopic model generates various topics and coherent topics in one topic.

<table>
<thead>
<tr>
<th>Table 1 Comparison of embedding model performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Distilbase-multilingual-cased-v1</td>
</tr>
<tr>
<td>KR-SBERT-V40K-klueNLI-augSTS</td>
</tr>
<tr>
<td>Ko-sbert-multitask</td>
</tr>
<tr>
<td>ko-sroberta-multitask</td>
</tr>
</tbody>
</table>

We calculated the TC and TD of three models, the BERTopic model with distilbase-multilingual-cased-v1 as the embedding model, the Combined Topic Models (CTM) model, and the LDA model, and found that BERTopic shows the best performance, as shown in Table 2. This shows that the BERTopic model generates various topics and coherent topics in one topic.

<table>
<thead>
<tr>
<th>Table 2 Comparison of Topic modeling model performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>BERTopic</td>
</tr>
<tr>
<td>CTM</td>
</tr>
<tr>
<td>LDA</td>
</tr>
</tbody>
</table>

4.4. Bart

Summarization step uses BART to generate a summary for each topic generated by BERTopic. After sorting the sentences with the highest cosine similarity, we extracted the top n sentences to summarize each topic. In this paper, we set n to 3 arbitrarily. The performance comparison was performed with TextRank [26] and Bert-extractive-summarizer [19]. The results can be seen in Table 3. In precision, TextRank and Bert-extractive-summarizer showed good performance. But, in recall and f1-measure, we can show that the proposed model has the best performance. This means that the proposed model performs the most equally well compared to the other models.

<table>
<thead>
<tr>
<th>Table 3 ROUGE score of different models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Proposed model</td>
</tr>
<tr>
<td>TextRank</td>
</tr>
<tr>
<td>Bert-extractive-summarizer</td>
</tr>
</tbody>
</table>
5. Conclusion

In this paper, we reduce the duplication of data by grouping the same or similar information into one topic through Topic Modeling. In addition, we compared models such as CTM and LDA, and selected BERTopic with the best topic diversity and topic cohesion to generate coherent and diverse topics. This process reduces wasted time searching for information, provides users with multiple perspectives on the product, and helps eliminate information loss in long videos. We used BART for summarization. In comparison to the previous research such as Text Rank and Bert-extractive-summarizer, ROUGE showed the best performance in recall and F1 measures.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (NRF-2021R1G1A1006381).

References


