Generating Fine-Grained Aspect Names from Movie Review Sentences Using Generative Language Model

Tomohiro Ishii[1], Yoshiyuki Shoji[1,2][0000−0002−7405−9270], Takehiro Yamamoto[3][0000−0003−0601−3139], Hiroaki Ohshima[3][0000−0002−9492−2246], Sumio Fujita[4][0000−0002−1282−386X], and Martin J. Dürst[1][0000−0001−7568−0766]

[1] Aoyama Gakuin University, Sagamihara, Kanagawa 252–5258, Japan
{ishii, duerst}@sw.it.aoyama.ac.jp
[2] Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
shojiy@inf.shizuoka.ac.jp
[3] University of Hyogo, Kobe, Hyogo 651–2197, Japan
t.yamamoto@sis.u-hyogo.ac.jp, ohshima@ai.u-hyogo.ac.jp
[4] Yahoo Japan Corporation, Chiyoda, Tokyo 102–8282, Japan
sufujita@yahoo-corp.jp

Abstract. This paper proposes a method for identifying an aspect highlighted in a sentence from a movie review, utilizing a generative language model. For example, the aspect "SFX Techniques" is identified for the sentence "The explosions in cosmic space were realistic." Classically, aspects are commonly estimated in the field of opinion mining within product reviews with classification or extraction approaches. However, because the aspects of movie reviews are diverse and innumerable, they cannot be listed in advance. Thus, we propose a generation-based approach using a generative language model to identify the aspect of a review sentence. We adopt T5 (Text-to-Text Transfer Transformer), a modern generative language model, providing additional pre-training and fine-tuning to reduce the training data. To verify the effectiveness of the learning techniques thus adopted, we conducted an experiment incorporating reviews of Yahoo! movies. Manual labeling of the correctness and diversity of the aspect names generated shows that our method can generates a variety of fine-grained aspect names using little training data.

Keywords: Opinion Mining, Aspect Detection, Generative Language Model

1 Introduction

Even when they are watching the same movie, different people will pay attention to different parts of it. Some will focus on the script and others on the actors' performances. A person who has a specific interest may focus on a specific aspect. For example, a railroad enthusiast may focus on whether the trains shown in a movie are correct in relation to the temporal and spatial setting of the movie.

Thanks to the rapid growth of information and communication technologies, the way that people watch movies and find reviews has changed in recent years. Subscription-based distribution services have brought more encounters with movies. People can decide what to watch from a vast pool of candidates. Movie review services, such as IMDb, have become widespread. Reading review posts to decide what movie to watch has become a part of everyday life.

However, the immense number of posted reviews makes it impossible to read them all. Additionally, as a large and diverse group of users contribute reviews, the aspects highlighted by different reviewers can vary. For instance, one user may want to read reviews that discuss the accuracy of the historical details in a movie, but most reviews concentrate on the actors' performances. Our hypothetical user will not be interested in this review. However, they cannot determine whether a review suits their purpose without reading the reviews.

To address this issue, numerous studies have focused on estimating the aspects of the text of reviews. However, conventional aspect classification and extraction approaches require each description to be categorized into predetermined aspects or extracting aspect names from the sentences. For instance, in common product domains, such as televisions or cameras, the sentence "I can see fine details" would typically be linked to the aspect of "resolution." At the same time, "I can bring it anywhere" would be associated with aspects such as "weight" or "size". For this categorization, a list of aspects must be provided in advance.

Reviews of entertainment content, such as movies and books, are necessarily more specific than reviews of traditional products. Different items may have different aspects; for example, even within the set of movies, the names of the aspects appearing in science fiction and romance movie reviews will be different. The number of such aspects is innumerable, so they cannot be listed in advance. New aspect names may become necessary when a new movie is released.

Therefore, this paper proposes a generative language model method that can generate aspect names from a given review sentence. Generative language models are capable of generating abstract aspect names that traditional classification and extraction approaches cannot address. For instance, take the case of creating an aspect name for the sentence "The explosions in cosmic space were realistic." Here, the review text has not included the phrase "scientific accuracy," and preenumerating specific aspect names for this case would be too fine-grained and laborious. This paper leverages state-of-the-art generative language models to comprehensively handle names of such fine-grained and specific aspects.

We adopt T5 (Text-to-Text Transfer Transformer), among the best-regarded text-to-text large language models, for this purpose. An overview of our training

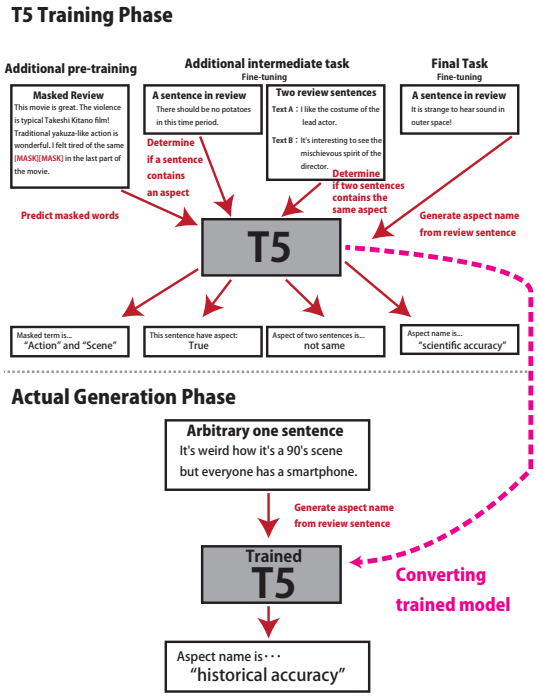

Fig. 1. Overview of our training steps of T5. The model was trained with an additional pre-training and two intermediate tasks before the main task.

method is shown in figure 1. First, using crowdsourcing, we created a dataset consisting of sentences from reviews and the relevant aspect names. We then trained T5 using the review sentences and aspect names. In general, training a language model with an additional intermediate task improves the accuracy of its generation [10]. We interspersed two additional learning tasks that differ from the generation of aspect names: a task to binary classify whether a review contains a statement about an aspect and another to binary classify whether two given sentences refer to the same aspect. This model with additional training was fine-tuned with the task of generating aspect names.

The performance of this generative model was evaluated using real data. Aspect names were generated for review sentences that were collected from real online movie review sites. The experimental participants labeled each aspect name, assessing its correctness and fine grain. Thus, we confirmed the presence of fine-tuning and changes in accuracy depending on the additional tasks.

2 Related Work

This study was undertaken aims to identify the aspects mentioned in online reviews. To provide context and positioning, this section describes related studies

on language models in online reviews, aspect extraction, and summarization, using language models.

2.1 Language Model in Online Review

Traditionally, review data are used as a general source of information for research. In addition, applications used to retrieve reviews have been widely studied. In particular, in opinion mining, it is common to analyze sentiment from product reviews.

For instance, Kim *et al.* [6] propose a sentiment analysis method using online reviews. Singh *et al.* [14] also present a sentiment classification method for movie reviews. Using SentiWordNet, they make classifications for emotional expressions and their polarities with a focus on parts of speech and words in proximity. Xu *et al.* [16] show a method that turns reviews into a source of knowledge that can be used to answer users' questions. Xu *et al.* [15] propose a simple CNN model using two types of pre-trained embeddings.

This research generally uses lexical, probability-based, and classical machine learning approaches. In recent years, as large-scale language models (LLMs) such as Transformers have received more attention, studies have come to focus on them.

For instance, Rietzler *et al.* [13] propose a method using BERT (Bidirectional Encoder Representations from Transformers) to classify the review aspect. Hasib *et al.* also use BERT for classifying sentiments of reviews [2]. Karimi *et al.* [5] propose an architecture called BAT (BERT Adversarial Training). BAT applies adversarial learning to post-training BERT. He *et al.* [3] propose a neural approach for finding coherent aspects. These studies use LLMs as classifiers. They are therefore similar to the traditional approach to estimating aspects.

2.2 Aspect Extraction

Aspect extraction has long been a topic of interest in opinion mining. In review posts, users can freely write their opinions about a product. To use reviews in product search, visualization, and analysis, it is important for shoppers to identify what aspects a specific description refers to [12].

Jo *et al.* [4] propose a method for automatically discovering different viewpoints and combinations of feelings on viewpoints from submission reviews. Peng *et al.* [9] propose a framework for dealing with the aspect of sentiment triplet extraction. Aspect sentiment triplet extraction refers to the task of extracting triples that can indicate what an aspect is, its sentiment polarity, and why it has this polarity. Angelidis *et al.* [1] present a neural network framework for summarizing opinions drawing on online product reviews. In their network, they combine two weak supervised components: an aspect extractor and a sentiment predictor.

Studies related to aspects of reviews are commonly used in both the categorical approach, where the aspect is applied to pre-prepared candidates, and in

extracting aspects that are written directly in the review. This study seeks to address this problem with a generative language model.

2.3 Text Summarization Using Language Model

This study uses a language model to generate names of aspect from review sentences. This approach is similar to summarizing in that it extracts the subject matter from a text, abstracts it, and expresses it in a few words. Summaries of texts can be roughly divided into extractive and abstractive summarization; however, our study relates to abstractive summarization. In recent years, it has been common to use LLMs for abstractive summarization. Significant research has been conducted on summarization using a language model.

Among LLMs used in this area, Liu *et al.* [8] propose BERTSUM, an extractive summarization method using BERT. Pegasus, proposed by Zhang *et al.* [17], is another prominent example of a traditional language model based on the summarization method. Pegasus is also an extension of BERT, characterized the use of Gap Sentence Generation for pre-training. Lewis *et al.* [7] present Bidirectional and Auto Regressive Transformers, a denoising autoencoder for use in pre-training sequence-to-sequence models.

Some methods based on BERT specialize in discrimination by embedding. Generative language models such as GPT (Generative Pre-trained Transformer) and T5 are good at abstractive tasks in particular. Our method adopted T5, a generative LLM.

3 Aspect Name Generation with T5

This section describes a method for generating the names of aspects mentioned in a given movie review text. Our method consists of three steps:

- additional pre-training,
- fine-tuning via an additional intermediate task, and
- fine-tuning via the final task.

The additional pre-training is performed to ensure that the model understands the particular vocabulary and knowledge related to movies. This additional intermediate fine-tuning is intended to focus the model on aspects in particular. The final task is the task of actual generation task. It generates the names of the aspects of any input review text. We collected actual review data and had it labeled it using crowdsourcing to enable this training.

3.1 Creating Training Data Using Crowdsourcing

Initially, we created a dataset for use in machine learning. Actual movie reviews were collected from real online movie review sites. These reviews were divided into individual sentences through splitting at the periods. Crowd workers then

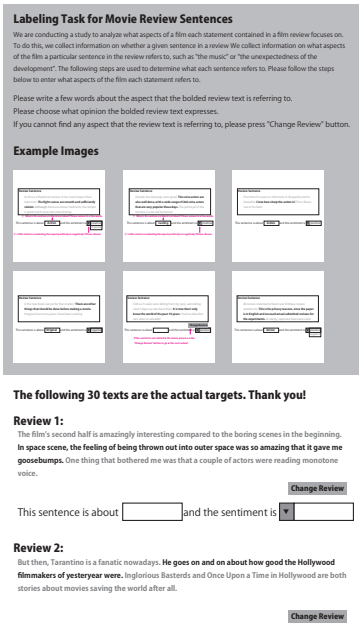

Fig. 2. Screenshot of the system used for labeling. Instructions are given at the gray section at the top. Below are six example images. Below these, the 20 review sentences are listed.

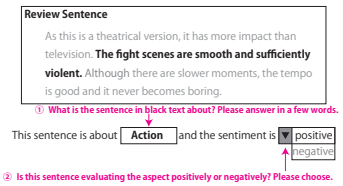

Fig. 3. Example image is presented to the crowd workers. It includes a screenshot of the actual form, instructions in red text, and an inputted example answer. (Translated from Japanese)

identified the aspects that were mentioned in each sentence. In total, the aspect names for twenty sentences were input by each of the 100 crowd workers.

When the crowd workers accessed the system, they were first presented with an instruction. This included the following background statement "This is an experiment for analyzing reviews conducted at a university" and instructions stating, "Please input names of aspects and sentiments (*i.e.*, negative or positive) for the 20 review sentences displayed below." Additional detailed precautions are also provided. Below the instructions, six example images of correct labeling are presented. By clicking on an image, the workers could see examples of aspect names that could be assigned to specific review sentences.

Then, workers read and labeled 20 sentences from actual reviews. Each sentence was displayed together with the surrounding text (*i.e.*, the sentence before it and the one after it). The sentence to be labeled was marked in black, and the surrounding sentences were gray. Workers input the aspect names as free keywords into text boxes. Simultaneously, they also input the sentiment related to each aspect name. For example, a sentence like "The lead actress's strange pronunciation was rather refreshing" is a "positive" expression concerning "performance." It should be noted that not every sentence displayed in black men-

tioned any aspect. In this case, the workers could label a different sentence by pressing the button for "Change Review." This skipped statement is recorded and labeled as "no aspect."

3.2 Additional Pre-training: Predicting Masked Term in Movie Review Data

Additional pre-training is performed to prompt the language model to better learn vocabulary on movies and general knowledge about them. Many proper nouns and peculiar expressions appear in movie review data; This includes names of directors, actors, series, *etc.*

The T5 model, which is publicly available, has been pre-trained on public documents (*e.g.*, Wikipedia, OSCAR, and CC-100). Because these documents do not contain sufficient descriptions related to movies, we train the model with movie review data. This model is pre-trained with masked language modeling, in the same way as was done by Raffel *et al.* [11] For additional pre-training, the model should be trained on the same task as that when the original model was created, only using different data.

3.3 1st Intermediate Fine-tuning: Binary Classification of Whether a Sentence Includes an Aspect

Next, we attempt to improve the quality of the model by fine-tuning it by means of intermediate tasks. In general, the accuracy of the final task can be improved by passing it through a task that is different from the original objective. Specifically, when training data for the final task are not large enough for training, it is adequate to fine tune it with different domains and tasks, so that it solves the same task with data from other domains or solves different related tasks using data from the same domain. Thus, here, the model solved tasks related to the goal of making the model understand what an aspect is.

To help the model understand aspects, it solves the most straightforward binary classification problem. In this task, the model receives one sentence and then learns to return 1 if the received sentence mentions an aspect and 0 otherwise. For this training, the sentences that the crowdworkers skipped as having "no aspect," and the sentences that were labeled with an aspect name were used as the answer data.

T5 can be trained using a prefix, which makes the different tasks explicit. Therefore, when fine-tuning this task, the prefix "contains-aspect:" was added.

3.4 2nd Intermediate Fine-tuning: Binary Classification of Whether Two Reviews Have the Same Aspect

Next, we allow the model to solve a slightly more advanced intermediate task. It receives two sentences and determines whether both mention the same aspect. This task is intended to enable the model to learn the differences between aspects.

The review sentences containing aspect names that were collected through crowdsourcing are used for the training. We randomly selected two arbitrary sentences from the dataset and then combined the two sentences with a separator token.

The input is a text with a prefix like "same aspect: typical Takeshi Kitano's violence scene [SEP] it has surprising end part." In this example, the former refers to the director and the latter to the story. Thus, the model was trained to return 0 because the two sentences mention different aspects.

We used this training because we wanted to tune the model more efficiently using fewer training data. It is expensive to have people read sentences and label them with aspect names is expensive. This method, however, can use many combinations of two sentences in a given dataset, creating extensive training data.

3.5 Final Fine-tuning: Aspect Name Generation

The model ultimately solved the same tasks as in the performance task. It input an arbitrary single sentence and outputs an aspect name. The prefix was also used for this training. For instance, the model was trained to generate aspect names, such as "direction," input such as the following is given: "aspect-generation: I was impressed with the way he expressed the main character's feelings by making it rain."

The same task was performed for the actual aspect name generation. A model trained in this way outputs aspect names even if they are input with review sentences that are not included in the training data. In such a case, the common linguistic sense obtained from training on public documents and knowledge of the movie studied in additional pre-training should be used. For example, the generated aspect name may be extracted from the review text, an abstraction of an expression in the review, or a completely new aspect name.

4 Evaluation

The evaluation tests were conducted to confirm the accuracy and effectiveness of the proposed aspect name generation and the effectiveness of the fine-tuning through the intermediate tasks. The experiment was conducted in two parts: first, a preliminary experiment was performed with automatic accuracy evaluation, followed by a main experiment that included subject evaluation. We created a dataset through the collection of reviews from real review sites and labeled them using crowdsourcing. Multiple comparative models were created for training using different fine-tuning methods. For each of these models, we evaluated the reproducibility (automatic evaluation) of the generated aspect names, their correctness, their fine granularity, and their novelty.

4.1 Dataset

We collected reviews from Yahoo! Movies, among the most extensive review sites in Japan. We extracted 176,970 of the reviews, avoiding those that were extremely short or too minor (this number was constrained by the graphic memory on the video card that was used for the training). All of these review data were first used for the additional pre-training of T5.

Next, sentences that were neither long nor short were extracted. We asked the crowdworkers to label the review sentences with the name of the aspect that they reviewed and to indicate whether they mentioned any aspect. Crowdsourcing continued until 1,500 sentences were labeled with the aspect name and collected.

The labeled review sentences thus obtained were used to create data for intermediate tasks. In the first intermediate task (the binary classification of whether an aspect was included), 3,759 reviews were prepared. For the second, (binary classification of whether two sentences refer to the same aspect), 1,000 pairs of review sentences were prepared.

During the creation of a dataset for the second intermediate task, we addressed the distortion of the notation of the aspect name, *e.g.*, "story" and "scenario," were essentially the same aspect. For this purpose, the similarity of two aspect names was calculated using Sentence-BERT, and aspect names with a semantic similarity of 0.9 or higher were considered identical.

In addition to the training dataset, a dataset for the subject experiments was also generated. Review sentences not used for either training that have a standard length were extracted. We only used reviews that seemed to mention an aspect that was used for evaluation. We classified such reviews using a simple BERT classifier (accuracy 0.63).

4.2 Comparison Methods

We prepared different methods, and only some of the training was given to verify the effectiveness of each of the trainings described in Section 3. Specifically, one set was given no additional training at all, one set only had additional pre-training, and one set had only a part of the intermediates task. For the intermediate tasks, we also compared which task was solved most rapidly.

Table 1 shows the 10 methods compared in the evaluation. The methods of **Pre-training+Include>Same** and **Pre-training+Same>Include** are the proposed methods (that is, these are the models that successfully completed all training tasks).

4.3 Implementation

Hugging Face Transformers[5], a library of Transformer-based models was used to implement T5. We used the Japanese T5 pre-trained model[6]. Additional training

[5] Hugging Face Transformer: https://huggingface.co/docs/transformers/index
[6] sonoisa/t5-base-japanese: https://huggingface.co/sonoisa/t5-base-japanese

Table 1. Comparison of methods for evaluation and their accuracy during preliminary experiments. The precision assessment indicates whether the method generated the same aspect name as the test data during cross-validation (using BERT for distortion of the notation). The bottom two are the proposed methods (the ones that successfully completed all training).

Method	Additional Pre-training	Aarlier Task	Later Task	Final Task	# aspect name generated	# aspect names not in training data	Precision in Auto Evaluation	F1 Score in Auto Evaluation
FinalTask Only	None	None	None	Done	76	25	0.800	0.795
FinalTask+Include	None	Include	None	Done	74	28	0.795	0.791
FinalTask+Same	None	Same	None	Done	74	23	0.794	0.791
FinalTask+Include>Same	None	Include	Same	Done	69	27	0.798	0.795
FinalTask+Same>Include	None	Same	Include	Done	77	28	0.797	0.794
Pre-training Only	Done	None	None	Done	84	31	0.800	0.796
Pre-training+Include	Done	Include	None	Done	**92**	**35**	0.799	0.795
Pre-training+Same	Done	Same	None	Done	79	23	0.800	0.796
Pre-training+Include>Same	Done	Include	Same	Done	56	11	**0.802**	**0.797**
Pre-training+Same>Include	Done	Same	Include	Done	11	1	0.798	0.790

parameters were set as follows: Maximum sequence length, 512; batch size, 16; learning rate, 0.005; weight decay, 0.001; and warmup steps, 2,000.

The learning rate was 0.0003 (determined empirically). The other parameters were set to the default values for Hugging Face Transformers. SentenceTransformers[7] was used to evaluate whether the aspect name was correct. We used a pre-trained multilingual model[8] for it.

4.4 Preliminary Experiment: Automated Accuracy Evaluation

We first performed an automated cross-validation evaluation to roughly assess the generation accuracy. In this evaluation, we split the dataset, consisting of crowdsourced labeled review sentences and aspect names, into training and test sets.

The test set was then fed into the model trained on the training set. We determined whether the generated aspect names matched the original manually assigned aspect names. However, to deal with the notation distortion, we used BERT to determine the identity of the aspect name (as in subsection 4.1).

4.5 Experiment: Subject Evaluation of Generated Aspect Names

The participants manually assessed whether the aspect names generated for the unknown sentences were correct. They read one review sentence and the aspect names generated by each method and labeled them as correct or incorrect.

The three evaluation factors were as follows:

- **format**: Whether the aspect name generated is a likely aspect name (0 or 1),
- **correctness**: How well the aspect name matches the content of the review sentence (scale of 1 to 5), and

[7] SentenceTransformers: https://www.sbert.net/

[8] Hugging Face sentence-transformers https://huggingface.co/sentence-transformers/

Table 2. Ratings for each method in the subject experiment (normalized to 0 – 1).

Method	Format Correctness	Granularity
FinalTask Only	0.964	0.691	0.380
FinalTask+Include	0.969	0.677	0.399
FinalTask+Same	0.971	0.670	0.384
FinalTask+Include>Same	0.969	0.666	0.384
FinalTask+Same>Include	0.971	0.648	0.387
Pre-training Only	**0.973**	**0.700**	0.396
Pre-training+Include	0.962	0.677	**0.407**
Pre-training+Same	0.971	0.663	0.406
Pre-training+Include>Same	**0.973**	0.671	0.386
Pre-training+Same>Include	0.969	0.640	0.336

– **granularity**: Whether the aspect name is sufficiently fine-grained (scale of 1 to 5).

Three participants read 100 review sentences. A maximum of 10 aspect names (with duplicates removed) were appended to each review text in random order. Participants labeled each aspect name.

4.6 Result

In this section, the results of the preliminary experiment and the subject experiment are described. Table 1 presents the results for the preliminary experiments. The model that had the highest accuracy had performed additional pre-training; it determined whether the aspect was included first and determined the same aspect name second. The models that had high accuracy had a lower probability of generating a new aspect name, often fitting an existing aspect names drawn from the training data.

Table 2 presents the results of the subject experiment. The method of **Pre-training Only** was evaluated as having the highest score. However, the differences were minor, and the models' performances were not significantly different from each other. The granularity of the generated aspect names was determined to be too rough in many cases.

5 Discussion

This section discusses the usefulness of the generative approach, the generation of aspect names by T5, and the effectiveness of every additional training according to the experimental results. Both the additional pre-training and the intermediate task showed increased accuracy. However, the additional pre-training was extremely effective, although the intermediate task had a limited effect.

First, we discuss the results of the automatic evaluation. To clarify the accuracy, we focus on the F_1 scores. As shown in Table 1, the model trained with all of the intermediate tasks achieved the highest accuracy. This model, developed with additional pre-training, first estimated the presence or absence of each

Table 3. Example of a review text and the aspect name generated from it. The judgment is the evaluation of the appropriateness of the aspect name.

Review Sentences	Aspect Name	Type	Judge	Method
I must miss it if it kept its original title.	Original Title	Extraction	1	Many methods
It is from an era I am unfamiliar with, but I recognized most of the songs.	Song	Extraction	1	Many methods
I would have felt satisfied if they ended with a in the final scene :)	Last Scene	Classification	1	FinalTask Only
Regardless, it is undeniably a high-quality work that keeps you engaged till the end.	Story	Classification	0	Many methods
All the other actors fit their comic roles perfectly and delivered high-quality performances.	Casting	Classification	1	Pre-training Only, Pre-training+Include
I think that it is a black comedy that cleverly satirizes current social issues.	Category	Classification	1	Pre-training Only
I think that it is a black comedy that cleverly satirizes current social issues.	Story	Classification	0	Many methods
It is good for amateurs, but the previous anime version was better.	Difference from Original	Generation	1	FinalTask+Include
The strangeness of this movie must came out from a director's taste.	Director's Personality	Generation	1	Pre-training+Include
The dancing has improved and the singing is moving.	Singing ability	Generation	1	Pre-training+Include
All the other actors fit their comic roles perfectly and delivered high-quality performances.	Difference from Original	Generation	1	FinalTask+Same, Pre-training+Same

Table 4. The ten aspect names most frequently generated by each method (translated from Japanese).

Pre-training Only	Pre-training +Same	Pre-training +Include	Pre-training +Same>Include	Pre-training +Include>Same	FinalTask Only	FinalTask +Same	FinalTask +Include	FinalTask +Same >Include	FinalTask +Include >Same
Cinema	Scale	Nature	Acting, Acting,...	Dubbing	Tools	Translation	Drama	Entertainment	Tools
Culture	Love	Scale		Fans' Expectations	Landscape	Love	Love	Expression	Spoiler
Tears	Attraction	C		The Idea	Fatigue	Lighting	Spoilers	Love	Disappointment
Drama	Last Scene	Love		Passion	Love	Brainwashing	Sound Effects	Schedule	Original
Screenplay Award	Difference from Drama	Description		Commentary	Evaluation Criteria	Doraemon's tools	Love	Spoilers	Inclusion
Kissing Scene	Compilation	Lacrimal Gland		Generation	Spoiler	Crying Scene	Cooking	Talk of Cast	Dubbing
Disappointment	Dubbing	Language Difference		Scale	Last scene	Overseas travel	Scenery	Cooking	Travel Expenses
Special Makeup	Homage	Story		Cosmology	Brainwashing	Snow	Trailer	Disappointment	Animation
Target	Original Title	Brainwashing		Snow	Atmosphere of the Original	Travel Expenses	Dubbing	For Kids	Feeling of Support
Dubbing	Reproduction	Target		Travel Expenses	Overloaded	Original Title	Doraemon's Tools	Snow	Meet Expectations

aspect and then determined whether the aspect of the two sentences was identical. This suggested that the accuracy of the model could be improved through training it by solving aspect-related tasks.

In particular, fine-tuning with the intermediate task reduced accuracy in some cases. The model named **Pre-Training+Same>Include**, which trained on all tasks, was the least accurate. However, fine-tuning in general tended to lead to higher accuracy. In all cases except **Same>Include** models, the pre-trained models were more accurate than those models trained using the same task. For instance, F_1 value increased by 0.005 relative to **FinalTask+Same** to **Pre-training+Same**. From these results, it appears that increasing the amount of data and focusing on additional pre-training may be a more efficient approach than performing fine-tuning through increasing the number of intermediate tasks.

Next, we discuss the novelty, granularity, and quality of the aspects generated and not their accuracy. Each model differed in terms of the number of unknown aspect names that were generated. Some models did not generate novel aspect

names but forcibly classified review sentences with aspect names that were contained in the training data labels.

For example, the model **Pre-training+Same>Include** generated only one new aspect name. It also tied all review sentences to a total of only 11 different aspect names. The novelty and the diversity of the aspect names that were generated tend to weaken with the amount of training. Models that were trained on both additional pre-training and two intermediate tasks output fewer aspect names.

This may be due to an overfitting to the training data. It is possible that the model was over-trained to consider the definition of the aspect names should refer only to the labels defined in the given dataset. There is room to investigate this effect in the future by increasing or decreasing the data for training and changing the ratio of fine-tuning.

Table 3 represents an example of the aspect names generated by the models[9]. Throughout this, reasonable aspect names are generated. These names included those derived from extraction, classification, and generation. The aspect names designated by extraction are those in which the model outputs words in the review text as they are. That is, the model extracts words based on the inference that these words are likely to be used as the aspect name. The aspect names according to classification are those that are not directly included in the input text but are included in the training data labels. Here, the model abstracts the review text to choose a word; however, these aspect names simply entail a classification as the pre-prepared aspect names. The aspect names according to generation are the output terms that are not included in the review text or in the training data labels. In these cases, the aspect names generated are completely abstracted from the meaning of the review sentences, and the model infers them from nothing.

Here, our discussion focuses only on the aspect names generated. Table 4 shows the top ten most frequent aspect names included in the output of the models. Many models generate valid aspect names that are fine-grained and are not included in the data set, excluding **Pre-train+Same>Include**. The only new aspect name generated by the **Pre-train+Same>Include** model was an incorrect name, simply repeating the word "acting" several times. At other times, this model either extracted a term in the review or applied a sentence to the label name in the training data.

One example of an incorrect aspect name was extraction using an incorrect word tokenization. The model **Pre-training+Include** output the aspect name "C." This indicates a failure to tokenize when reading the review text. The review text included statements on visual expressions, such as, "This visual effect is the kind of expression I see in TVCM (Television Commercial Message)." Because the social review includes data with many colloquial expressions, it may be necessary to proofread the text using conventional methods before performing the language model.

[9] For the sake of translation and anonymization, the reviews are fictitious, as the experiment was in Japanese and uses real review sentences prepared by an individual.

In summary, T5 showed good performance, even without any fine-tuning or additional training. We had thought that generating aspect names in a generative language model would be more difficult. However, performing the same fine-tuning on plain T5 as in the production task achieved an F_1 value of 0.79. In addition, many of the aspect names generated were correct and were not found in the training data. A larger dataset that includes more fine-grained aspect names and that uses T5 in a straightforward manner could produce a more accurate and practical language model.

6 Conclusion

This paper proposed a method for generating aspect names for arbitrary sentences from reviews by training the generative language model T5. Evaluation experiments showed that the generative approach can also generate aspect names with high accuracy. We found that fine-tuning using intermediate tasks was less effective, but additional pre-training was highly effective.

The contributions of this paper are as follows:

- we revealed that a generative approach could be used for aspect name inferences, which is conventionally done using classification and extraction, and
- we revealed that additional pre-training is effective for training generative language models for the aspect name generation task.

In future work, we plan to enhance data cleansing and applied evaluation. Specifically, we plan to construct a system for searching for review sentences using generated aspect names. Recently, methods involving more extensive and more innovative language models have emerged (*e.g.*, GPT-4 and Bard). It may be possible to generate aspect names by prompts rather than using methods based on fine-tuning. Significant room remains for improvement in aspect name generation by generative AI.

Acknowledgements

This work was supported by JSPS KAKENHI Grants Number 21H03775, 21H03774, and 22H03905.

References

1. Angelidis, S., Lapata, M.: Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In: Proc. of the Conference on Empirical Methods in Natural Language Processing. pp. 3675–3686 (2018). https://doi.org/10.18653/v1/D18-1403
2. Hasib, K.M., Towhid, N.A., Alam, M.G.R.: Online review based sentiment classification on bangladesh airline service using supervised learning. In: 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). pp. 1–6. IEEE (2021)

3. He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: An unsupervised neural attention model for aspect extraction. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2017)
4. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proc. of the Fourth ACM International Conference on Web Search and Data Mining. p. 815–824 (2011). https://doi.org/10.1145/1935826.1935932
5. Karimi, A., Rossi, L., Prati, A.: Adversarial training for aspect-based sentiment analysis with bert. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8797–8803 (2021). https://doi.org/10.1109/ICPR48806.2021.9412167
6. Kim, R.Y.: Using online reviews for customer sentiment analysis. IEEE Engineering Management Review **49**(4), 162–168 (2021)
7. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.703
8. Liu, Y.: Fine-tune bert for extractive summarization. arXiv preprint arXiv:1903.10318 p. arXiv:1903.10318 (2019)
9. Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., Si, L.: Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. Proc. of the AAAI Conference on Artificial Intelligence **34**(05), 8600–8607 (2020). https://doi.org/10.1609/aaai.v34i05.6383
10. Pruksachatkun, Y., Phang, J., Liu, H., Htut, P.M., Zhang, X., Pang, R.Y., Vania, C., Kann, K., Bowman, S.R.: Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5231 – 5247 (2020)
11. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1 – 67 (2020)
12. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems **89**, 14–46 (2015). https://doi.org/https://doi.org/10.1016/j.knosys.2015.06.015
13. Rietzler, A., Stabinger, S., Opitz, P., Engl, S.: Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In: Proc. of the Twelfth Language Resources and Evaluation Conference. pp. 4933–4941 (2020)
14. Singh, V.K., Piryani, R., Uddin, A., Waila, P.: Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In: 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s). pp. 712–717 (2013). https://doi.org/10.1109/iMac4s.2013.6526500
15. Xu, H., Liu, B., Shu, L., Yu, P.S.: Double embeddings and cnn-based sequence labeling for aspect extraction. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (2018)
16. Xu, H., Liu, B., Shu, L., Yu, P.S.: Bert post-training for review reading comprehension and aspect-based sentiment analysis. In: Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (2019)
17. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning. pp. 11328–11339 (2020)