1V-03

Evaluating Aspect Category Sentiment Analysis Using Foreign Skiers' Review Dataset

Zhongyue Yi† Yohei Seki‡
†Graduate School of Comprehensive Human Sciences, University of Tsukuba
‡Faculty of Library, Information and Media Science, University of Tsukuba

1 Introduction

Foreign skiers account for a significant percentage of the total number of visitors to Japanese ski resorts. The annual report of HAKUBAVALLEY TOURISM[1] indicates that around 375,000 foreign skiers visited Hakuba valley in 2019. However, facilities managed by Japanese people have been developed to meet the needs of Japanese skiers, so it was likely challenging to meet the needs of foreign skiers [1]. Therefore, this study attempted to analyze the satisfaction level of foreign skiers by utilizing aspect category sentiment analysis (ACSA), which aims to analyze the sentiment of texts regarding specific aspect categories. As there were no suitable datasets available, we collected and annotated English reviews of 10 ski resorts located in Japan from Tripadvisor and created a new dataset for this purpose. Moreover, we employed various data augmentation techniques to increase the number of texts with infrequently mentioned aspect categories and leverage it to improve model performance.

2 Related Work

2.1 Aspect Category Sentiment Analysis

In recent years, a number of deep learning models have been proposed that achieved outstanding performance on the ACSA task, such as the SCAN model [2] which employed two graph attention networks to generate representations of the nodes in sentence constituency parse trees for the aspect category detection and sentiment classification. There are several datasets available for the ACSA task, such as the SemEval-2016 Task-5 Restaurant (SemEval-16-Rest) dataset [3] which includes 12 aspect categories (e.g., "AMBIENCE#GENERAL"), with three sentiment polarity labels (i.e., Positive, Negative and Neutral).

2.2 Data Augmentation

The purpose of data augmentation (DA) is to create additional training data when available data is insufficient [4]. This can be accomplished in a variety of ways, including machine translation, synonym replacement, and model generation, and there are some tools to do this, such as *Helsinki NLP*[2] and *nlpaug*[3].

3 Dataset Construction and Analysis

3.1 Data Collection and Annotation Process

We collected the 2,205 English reviews of 10 ski resorts in Japan on TripAdvisor, and divided them into individual sentences, resulting in 13,203 sentences.

There are numerous aspects, such as snow quality, that could potentially influence the satisfaction of skiers. However, it is challenging to assign an aspect category to each of these potential aspects. Therefore, before annotating, we pre-defined 26 aspect category labels shown in Table 1. Annotation was performed by three annotators (annotators A, B, and C), with annotator A responsible for annotating all of the data, and annotators B and C each responsible for annotating half of the data. The Cohen's κ coefficient between annotators A and B, and annotators A and C were 0.6487 and 0.6583, respectively, indicating substantial agreement among the annotators. Any discrepancies in the annotation results were then discussed and resolved to determine the final annotation results.

3.2 Dataset Analysis

Table 1 shows the distribution of aspect categories in our dataset. For example, "Slope#SnowConditions" is mentioned frequently in the reviews, with 1,741 mentions. On the other hand, some are mentioned infrequently, such as "Accommodation#Price" with only 32 mentions. This data imbalance is also common in SemEval-16-Rest, where "RESTAURANT#PRICES" has only 80 training data, while the aspect category "FOOD#QUALITY" has 1,048. To address these issues, we attempted to increase the amount of data for the aspect categories with infrequently mentioned by using data augmentation techniques.

4 Methodology

The following three methods were employed to augment the low-frequency aspect label data: (1) **Back translation**: a method that translates a text from one language to another and then back again into the original language to generate a new text with similar meaning. (2) **Thesaurus-based DA**: a method that replaces words in the source text with synonyms retrieved from a thesaurus. (3) **GPT-3-based DA**: GPT-3 is trained on a large data set and is capable of generating long and coherent sentences. This GPT-3 is used to generate new texts with opposite sentiment polarity labels to alleviate the problem of lack of negative and neutral sentiment polarity labels.

[1]https://www.vill.hakuba.lg.jp/gyosei/keikan_kankyo_kanko/kanko/hakubavalleytourism.html
[2]https://huggingface.co/Helsinki-NLP
[3]https://github.com/makcedward/nlpaug

Table 1: Pre-defined aspect categories and their distribution in our dataset.

Aspect Categories		Pos.	Neu.	Neg.	Aspect Categories		Pos.	Neu.	Neg.
Entity (pct.)	Attribute (pct.)				Entity (pct.)	Attribute (pct.)			
Accommodation (4.9)	AccessConvenience (2.3)	150	22	20	Lift (11.6)	OpenStatus (2.9)	38	50	157
	Price (0.4)	15	4	13		Price (1.3)	68	33	8
	Quality (1.2)	75	6	24		Quality (4.2)	107	19	228
	Service (1.0)	37	1	50		WaitingTime (3.2)	184	19	66
RentalShop (2.1)	Equipment (0.9)	42	18	17	Restaurant (9.0)	F&B (6.8)	469	56	46
	Price (0.6)	22	13	14		Price (1.9)	95	14	52
	Service (0.6)	27	2	19		Service (0.3)	16	2	12
SkiResort (21.2)	AccessConvenience (6.7)	279	160	124	Slope (51.2)	Crowd (6.3)	333	40	159
	EnglishAvailability (2.0)	113	11	47		Preparation (2.2)	118	12	58
	InformationService (2.0)	58	12	97		Size (4.3)	268	22	70
	Price (2.6)	137	42	44		SnowConditions (20.6)	1498	89	154
	Scenery (3.5)	290	6	2		Variety (17.8)	1155	163	181
	Service (4.4)	297	15	60	Out-of-Scope		-	-	-

5 Experiment

5.1 DA Implementation Details

We utilized the Helsinki NLP to translate the original text into Chinese and then back-translate it into English, resulting in a back-translated text. The *nlpaug* and *text-davinci-003* models were employed to perform the thesaurus-based and GPT-3 based DA for the original text and back-translated text.

5.2 DA Process

We performed three data augmentation methods described above to generate more training data in the aspect categories of "Accommodation#Price," "RentalShop#Price," "Lift#Price," and "Restaurant#Price" for our dataset, and "RESTAURANT#PRICES" and "DRINKS#PRICES" aspect categories for SemEval-16-Rest dataset. By combining back translation and thesaurus-based DA, we obtained 500 and 300 new training data for our dataset and the SemEval-16-Rest dataset, respectively. We also obtained the same number of data by combining back-translation and GPT-3-based DA. The back-translation texts in the above two DA methods are identical. Therefore, the combination of the above two methods yielded 900 and 500 new training data for our dataset and the SemEval-16-Rest dataset, respectively. Using three methods, we generated new training data for the aspect categories described above. For example, using the three DA methods, we obtained 135 new training data for "Accommodation#Price" in our dataset and 390 new training data for "RESTAURANT#PRICES" in the SemEval-16-Rest dataset.

5.3 Results and Analysis

The SCAN model was employed to evaluate the effectiveness of our DA methods. We adopted Micro-F1 and accuracy as evaluation metrics. The experimental results indicated that our DA methods can improve the performance of the SCAN model, as shown in Table 2. We were also able to improve the prediction accuracy of the aspect categories with less original training data. For example, by leveraging the three DA methods, we observed a 23.99% accuracy improvement for "Accomodation#Price" and a 9.52% improvement for "RESTAURANT#PRICES" on the SemEval-16-Rest dataset.

Table 2: Experimental results on our dataset and the SemEval-16-Rest dataset with SCAN model. Best scores are boldfaced.

	Our Dataset		**SemEval-16-Rest**	
	Acc.	F1	Acc.	F1
Original dataset	59.93	58.38	87.07	53.77
Original dataset + Back Translation + Thesaurus	**61.47**	58.99	87.48	51.58
Original dataset + Back Translation + GPT-3	61.20	**59.50**	**88.97**	56.12
Original dataset + Back Translation + Thesaurus + GPT-3	61.10	59.21	87.82	**57.93**

6. Conclusion

This paper presents a foreign skiers' review dataset for the ACSA task, which can be used to investigate foreign skiers' satisfaction. Furthermore, data augmentation methods were used to resolve label imbalances in our dataset and the SemEval-16-Rest dataset.

Acknowledgements

This work was partially supported by the Japanese Society for the Promotion of Science Grant-in-Aid for Challenging Exploratory Research (#22K19822), Grant-in-Aid for Scientific Research (B) (#19H04420), and by ROIS NII Open Collaborative Research 2022 (22S0103).

References

[1] 呉羽 正昭. "日本のスキーリゾートにおけるインバウンド・ツーリズムの発展," 日本地理学会発表要旨集, 2017a(0),100065, 2017.

[2] Yuncong Li. et al. "Sentence Constituent-Aware Aspect-Category Sentiment Analysis with Graph Attention Networks," In CCF Int'l Conf. on Natural Language Processing and Chinese Computing, pages 815-827, 2020.

[3] Maria Pontiki. et al. "SemEval-2016 Task 5: Aspect Based Sentiment Analysis", In Proc. of the 10th Int'l Workshop on Semantic Evaluation (SemEval-2016), pages 19-30, 2016.

[4] Bohan Li. et al. "Data Augmentation Approaches in Natural Language Processing: A Survey", AI Open, Volume 3, pages 71-90, 2022.