Towards a New Investigation Method for Tourists' Needs: Dictionary-based Aspect-level Review Classification

SHUANG SONG †1 TOMOHISA YAMASHITA †1
HIDENORI KAWAMURA †1 HAJIME SAITO †2

Abstract: Many travel surveys have been carried out in tourist destinations to investigate tourist's needs. Meanwhile, the analysis of massive and up-to-date travel reviews may be able to provide a low-cost and real-time substitute. This paper presents a method for machines to related text in the reviews with the questions used in traditional surveys. This method includes two steps: 1) co-occurrence based keywords extraction and 2) dictionary-based text classification. The 11 questions from the Hokkaido Survey and travel reviews from TripAdvisor are used as an example; classification results from a previous manual analysis are used for the evaluation.

Keywords: text mining, survey method, keyword extraction, fixed aspect, tf-idf

1. Introduction

Inbound tourism can bring huge economic contributions to the destinations. To attract more international tourists, many travel surveys have been carried out by destination marketing organizations to investigate tourists' expectations, satisfactions and etc. Traditional survey methods such as interview, mail or Internet survey usually takes a lot of time (e.g. seasonally to yearly) and money (e.g. personnel expense, distribution expense and incentives) to acquire a large amount of samples. Nowadays, travel-related data are constantly generated and can be conveniently collected via the Internet. Recent studies have shown the possibility of extracting valuable information such as tourists' preference, complaints or destination images from those online data [1][2][3]. Using those online data, the cost of sample collecting could be significantly reduced. However, data analysis introduces many external factors into needs investigation [4][5][6]; thus, whether the results from data analysis can represent tourists' needs in general in unclear.

Building towards a low-cost and real-time needs investigation method, this research aims to find out whether it is possible to find tourists' needs through text data mining. During a previous manual analysis [7], positive correlations ($p<0.05$) are found between the attitudes in 1,158 travel reviews collected from TripAdvisor and results of a satisfaction survey [8] carried out by the Hokkaido government (n=1,709). This finding suggests the possibility of using review analysis for the prediction of tourists' satisfaction. However, this assumption needs further validation with more review samples and results from larger-scaled surveys, which urges the development of an automated analysis method.

To automatically extract the attitudes from travel reviews to be compared to the results of a traditional survey, we need to extract text related to each question (i.e. aspect) used in the survey and then judge the underlying attitudes towards each question. This paper will focus on the first step: aspect-level text extraction.

†1 Hokkaido University
†2 Hokkaido Information University

There are two often discussed approaches: dictionary-based approach and machine learning approach (supervised and unsupervised) [9]. This research will focus on the first approach for the following reason: Supervised machine learning usually requires a large amount of labeled data. However, each destination has its own local specialties and characteristics, causing the fact that different surveys have different question settings. Also, due to limited context, it is difficult for human to label data on phrase or sentence level [10]. Therefore, the creation of aspect-level labeled data can be costly and difficult.

This paper presents a dictionary-based text classification method where the dictionary is automatically created using co-occurrence based keywords extraction. We use the 11 questions from the Hokkaido Satisfaction Survey as 11 aspects and travel reviews posted during the survey period in Hokkaido from TripAdvisor for keywords extraction. We apply both aspect-level and sentence-level classification to the 1,158 reviews used in the previous manual analysis.

The rest of the paper is organized as follows. Sections 2 reviews related works. Section 3 explains the methodology. Section 4 shows the results. Section 5 discusses the findings, limitations and future works. And conclusions are presented in Section 6.

2. Related works

2.1 Aspect-level Text classification

Aspect-level text classification is used when a document or a sentence contains multiple topics to enable the analysis of individual attitude towards each topic. Aspects can be either generated from text data or pre-defined. The 11 questions from the Hokkaido survey, for example, are 11 pre-defined aspects of tourist satisfaction. When generating aspects from given text, the procedure is called aspect detection, where the extraction of keywords is needed but the interpretation of the meaning of each keyword can be left to human objects. Meanwhile, when aspects are pre-defined, not only the extraction, but also the classification of keywords is necessary.

Methods of Aspect-level text classification can be divided into dictionary-based, machine learning and hybrid approach [9]. In dictionary-based approach, a set of keywords, or a set of

syntax (patterns of keywords) are usually assigned to each aspect, simplifying the procedure to the identification of keywords. Meanwhile, in machine learning approach, aspect detection can be considered as a labeling problem.

Keywords are usually single nouns and compound nouns with high occurrence frequency. However, not all frequent nouns are suitable as keywords; those words are referred to as stopwords. Stopwords can be generated from materials outside of the current research domain. For example, frequent words in news, daily conversational text or etc. can provide a baseline for the extraction of tourism terms. When using only nouns, one limitation is that only explicit words can be detected, which makes it difficult for extracting words that are alternative to each other. Meanwhile, the use of adjectives or verbs helps the detective of implicit association between words [11].

To prepare a set of keywords to each aspect, we can use thesaurus such as WordNet [12] [13], or use keywords derived from the data themselves [14]. Previously, we compared the words in the 1,158 reviews with the words in WordNet2.1 and found that 1935 / 7432 = 26% kinds of words do not exist in WordNet [7]. Most of these words are proper nouns or Romaji such *onsen* or *ramen*, which suggests that a method is needed to automatically extract and classify unknown words into a set of given aspects.

The classification of unknown words is frequently used in document or sentence-level classification. On document-level classification, for example, unknown words can be extracted from document by the calculation of tf-idf or etc. An unknown document can then be classified using the similarity between unknown keywords and labeled keywords. Once the document is classified, unknown words in that document can served as new keywords to classify other documents [15][16]. However, such a method is difficult to be applied to aspect-level classification without aspect-level labeled data and an appropriate method of separating multiple topics in one review.

2.2 Guest Survey

The Survey Concerning Customer Satisfaction was implemented by the Hokkaido Government during 6/1/2016 to 2/28/2017 [8]. The participants needed to answer how satisfied they were towards 11 aspects (Table.1).

Table 1 Aspects in the Hokkaido Survey.

#	Aspects
1	for the entire trip and sightseeing
2	meals at each tourist destination
3	souvenirs
4	accommodations
5	tourist attractions
6	wifi accessibility
7	multilingual informational signs
8	local staff's linguistic abilities
9	transportation system
10	customer service
11	scenery

3. Methodology

3.1 Collecting reviews

As potential equivalent data to the ones used in the guest survey, every review posted during the survey period in the survey area on TripAdvisor is collected. Altogether 60,125 reviews concerning hotels, restaurants and attractions were collected using a data crawling tool named Octopus Data Collector. Among those reviews, 18,338 are non-Japanese review. Furthermore, 5,673 are English reviews written by tourists with location information which suggests they come from Singapore, Australian, America, Hong Kong and Great Britain; these 5,673 reviews will be used for keywords extraction.

3.2 Data cleaning and morphological parsing

The following basic text cleanings are applied to the reviews to improve the performance of morphological parsing: (1) replace abbreviation with original words, e.g. *not* instead of *n't*, (2) replace emoji with words, e.g. *(smile face)* instead of ^_^, (3) remove accent, e.g. *e* instead of *é*, (4) replace numbers with 0, and (5) replace characters with space except for -.,!?:; and alphabets.

A tool named Tree-tagger [17] is used to split the reviews into separated words and annotate each word with its lemma and part-of-speech information. Only nouns and proper nouns will be used for now.

3.3 Keywords extraction and exclusive classification

This session explains the extraction of keywords for a set of given aspects. The general idea is to define several keywords for each aspect manually, and then use them as clues to find new keywords based on co-occurrence. The following three steps will be repeated until there are no calculable words left.

3.3.1 Temporarily label the unknown words

Starting from the first word i in one review, if i is not a keyword, go to the next word; if i is a keyword, do the following steps before go to the next word.

- If i is the first keyword found, or it belongs to the same aspect as the previously found keyword, label all unlabeled words before i with a tag named after i 's aspect.
- If i belongs to a different aspect, go find the split point between i and the previously found keyword. A split point can be the nearest period mark before i, or i itself if no period mark exists. And then, label all unlabeled words before the split point with the previously found keyword's aspect, and unlabeled words after the split point and before i with i 's aspect.

3.3.2 Find potential keywords

We want to find potential keywords that have high occurrence frequencies in one aspect but appear in as less aspects as possible, which can be achieved using the calculation of tf-idf (Formula 1). For word i labeled with aspect j, $t_{i,j}$ is the value of tf-idf for i in j, $n_{i,j}$ is the occurrence frequency of i in j, $|A_i|$ is the amount of aspects that have i, and $|A|$ is the amount of all aspects which equals to 11 in this case.

$$t_{i,j} = n_{i,j} \times (\log \frac{|A|}{|A_i|} + 1) \quad (1)$$

Then, labeled words will be order by the sum of tf-idf (i.e. t_i in Formula 2) in descending order and alphabetically if they have the same t_i value.

$$t_i = \sum_{k=1}^{|A|} t_{i,k} \quad (2)$$

Finally, the top M words (M is initially set to the value of 20) will be considered as potential keywords for the current iteration.

3.3.3 Decide the aspect for top potential keywords

Potential keyword i will be added to aspect j as a keyword if it meets all three conditions as follows:

(1) $n_{i,j} = \max\{n_{i,k}, k \in A\} > 2\text{nd}\max\{n_{i,k}, k \in A\}$

(2) $\frac{n_{i,j}}{n_i} > \alpha$, where n_i the total occurrence frequency of i in all aspects, and parameter α is initially set to 0.5

(3) i is not a stopword.

Stopwords are words that appear in multiple aspects. And they are automatically extracted (1) if a word appear in all aspects, (2) if a word contains any numbers, (3) if a word is the name of cities in Japan in urls extracted from TripAdvisor, (4) if a word belong to words from news; for now that is, nouns appears in 5 or more groups in the 20 Newsgroups data set (see http://qwone.com/~jason/20Newsgroups).

3.4 Review classification

This session explains two methods for text classification: aspect-level and sentence-level classification.

3.4.1 Aspect-level text classification

When given one review and keywords (including pre-defined ones and extracted ones), first label each word in that review using the same method shown in session 3.3.1. Then, for each aspect, extract all words with corresponding label as its related text.

For example, assuming we have three aspects = {food, service, view} and each aspect contains the following keywords: food = {food, beef}, service = {service} and view = {view}. The following review can be classified as shown in Fig.1.

This restaurant is great. The food is good, and you get great (food) //view outside the window. (view) //Their service is very attentive. (service) //My favorite is the beef. A little expensive, but worth it. (food)

Figure 1 Example of Aspect-level classification.

3.4.2 Sentence-level text classification

When given one review and keywords, first separate the review into sentences by period, exclamation or question mark. Then for each sentence, count the amount of keywords in each aspect. One sentence will be labeled after the aspect with the only and highest keywords amount. If the highest amount is a tie, this sentence will be labeled as mixed aspects. If no keyword exists, this sentence will be omitted. Fig.2 is an example of sentence-level classification using the same aspects and keywords setting explained in session 3.4.1.

This restaurant is great. // The food is good, and you get great view outside the window. (mixed) // Their service is very attentive. (service) // My favorite is the beef. (food) // A little expensive, but worth it.

Figure 2 Example of Sentence-level classification.

Table 2 Pre-defined Keywords.

#	Keywords
2	food, breakfast, restaurant, drink
3	buy, shop, souvenir, shopping, purchase, duty-free, drug, outlet, gift
4	hotel, room, bed
5	facility, bench, museum, golf, park, skiing, observation, wonderland, zoo, garden, ticket
6	wifi, wi-fi
7	sign, map, translation, explanation, pamphlet, signage
8	language, english, speak, communication, communicate
9	locate, bus, jr, highway, passenger, tram, bike, cable, walkway, pathway, road, flight, shuttle, ropeway, carriage, train, boat, path, traffic, tunnel, climb, fly
10	service, waiter, server, waitress, management, valet, manager, guide, staff
11	view, crater, snow, flower, tree, lavender, sakura, see, firework, landscape, show, mountain, performance

Table 3 Examples of Extracted Keywords.

#	Amount	Examples
2	354	dinner, buffet, meal, dish, pizza ... ... grease, seicomoart, gelato, apia, koisk
3	7	mall, arcade, tanukikoji, donut, shopper, , daiso, handicraft
4	137	lobby, luggage, amenity, shower, tatami ... ... cloth, stroller, air_con, modern, compare
5	10	odori, penguin, deck, enclosure, monkey, … illumination, asahiyama, lakes, northern
6	0	-
7	1	cemetery
8	1	mandarin
9	12	nous, teine, cape, shin, cts ... ..., jujigai, convent, terminus, foliage
10	1	speaking
11	30	festival, sunset, yotei, sculpture, tomita ... ... performer, snowfall, dusk, riding, dolphin

4. Results

4.1 Extracted Keywords

We used 84 pre-defined keywords for 10 aspects. Question 1 was ruled out from this experiment because the average answer rate for this question is only 0.4% in reviews. The pre-defined keywords are listed in Table.2.

We experimented on words that appear more than 5 times in

all reviews (i.e. $n_i > 5$), and 553 keywords were extracted. The examples of the first and last 4-5 keywords in each aspect are listed in Table.3 in descending n_i order. We can see that co-occurrence based method functioned well when extracting nouns about food and hotel. However, it is less successful for aspects containing more alternative nouns (e.g. People who went to the *zoo* won't write about the *museum*). There are also a few wrongly extracted keywords (e.g. *cemetery* in aspect 7), so we may need to re-train words with low occurrence frequency or low $n_{i,j}$ / n_i value.

4.2 Review classification

Aspect-level and sentence-level classification are applied to the 1,158 reviews used in the previous manual analysis.

These 1,158 reviews originally include reviews written in three languages (English, Simplified Chinese and Traditional Chinese) and posted by tourists from seven regions (America, Singapore, Australia, mainland China, Taiwan, Hong Kong and Britain). They were randomly selected from the 18,338 non-Japanese reviews. In this experiment, Chinese reviews are translated into English in advance using Google Translation. Our previous manual analysis confirmed that Google Translation can provide about 97% consistency in results between pre / post translation Chinese reviews [10].

Table 4 Results of Aspect-level Classification.

#	relevant reviews	extracted reviews	true positives
2	698	779	668
3	105	227	83
4	345	459	325
5	292	259	129
6	23	19	19
7	40	38	12
8	84	67	48
9	441	295	219
10	380	345	293
11	330	329	212

Table 5 Performance of Aspect-level Classification.

#	precision	recall	F1
2	0.858	0.957	0.905
3	0.366	0.790	0.500
4	0.708	0.942	0.808
5	0.498	0.442	0.468
6	1.000	0.826	0.905
7	0.316	0.300	0.308
8	0.716	0.571	0.636
9	0.742	0.497	0.595
10	0.849	0.771	0.808
11	0.644	0.642	0.643

Table 6 Results of Sentence-level Classification.

#	relevant reviews	extracted reviews	true positives
2	698	606	560
3	105	101	50
4	345	345	272
5	292	141	89
6	23	8	8
7	40	16	4
8	84	20	12
9	441	174	138
10	380	181	165
11	330	213	150

Table 7 Performance of Sentence-level Classification.

#	precision	recall	F1
2	0.924	0.802	0.859
3	0.495	0.476	0.485
4	0.788	0.788	0.788
5	0.631	0.305	0.411
6	1.000	0.348	0.516
7	0.250	0.100	0.143
8	0.600	0.143	0.231
9	0.793	0.313	0.449
10	0.912	0.434	0.588
11	0.704	0.455	0.552

Relevant reviews in Table.4 and Table.6 are results from document-level manual analysis performed by the first author [7]. Those results show whether the review contains the text related to a certain aspect, rather than pointing out the exact phrase or sentence related to that aspect. Therefore, the precision, recall and F1 value in Table.5 and Table.7 are based on document-level comparison, which only show a brief image of the performance of the classification.

Compare to sentence-level text classification, aspect-level classification reached higher recall and F1 values, but lower precision values in general. One possible cause can be that the object of the classification is word instead of phrase. The meaning of a single word may change when compounded with another word. Moreover, Aspect 7 - *multilingual informational signs*, has the lowest precision and recall value, combined with the result of wrongly extracted keywords. It is possible that co-occurrence based keyword extraction and dictionary-based text classification are not the appropriate method to be applied to certain aspects. On the other hand, Aspect 6 - *wifi accessibility* reached unexpectedly high F1 value with only two pre-defined keywords, which means some aspect may only have few keywords after all. Thus, the creation of pre-defined keywords should be done with cautious. Furthermore, Aspect 5 - *tourist attractions* and Aspect 9 - *transportation system* have

low recall values, which is possibly caused by the insufficient number of keywords. For those aspects containing more alternative nouns, learning materials for keywords should not be limited to the reviews themselves; the use of other resource should also be considered, such as extracted attractions names from TripAdvisor.

5. Discussion

Results of this research suggest that dictionary-based aspect-level classification is a partial solution to the extraction of text related to the questions used in traditional travel survey from travel reviews when labeled data are absent. The performance of the presented method is promising in the aspects of *food*, *accommodation*, *wifi* and *service* (F1>0.8); however, it is unsuitable for extracting text about aspects such as *multilingual informational signs*. Also, it should be noticed that the results depend on the pre-defined keywords, stopwords and the value of M and α; therefore, it may take several tunings to get the best results.

It is expected that the performance can be improved if a single word can be automatically recognized as either an independent keyword or part of a compound word. This can be achieved by considering the probability distribution of each n-gram words combination [18] or by cross-checking the hit results of each combination from search engines. For compounded proper nouns, we can also consider the uses of other resources such as extracted attractions names from TripAdvisor. However, it is observed that reviewers tend to use abbreviations or only part of the full spelling when writing proper nouns. Therefore, search engine techniques including spelling error detecting may be necessary. In addition, only nouns are used in this paper, but other words such as adjectives and verbs can also be useful.

For the automation of review analysis, apart from dictionary-based approach, other potential solution can be the application of unsupervised machine learning or techniques of the reuse of available labeled data. For the second step of the automation, which is sentiment classification of extracted text, machines learning approaches such as the use of SVM are expected to be applicable.

6. Conclusions

In this paper, we present a method to automatically extract keywords for a set of given aspects and then classify text in travel review into each aspect. The performance of this method varies in different aspects. This method is expected to be further improved by introducing the process of compound words, the use of adjectives and verbs, other resources as learning materials and etc.

Reference

[1]Marrese-Taylor, E., Velásquez, J. D., Bravo-Marquez, F., & Matsuo, Y. (2013). Identifying customer preferences about tourism products using an aspect-based opinion mining approach. Procedia Computer Science, 22, 182-191.

[2]Liu, Y., Teichert, T., Hu, F., & Li, H. (2016). How do tourists evaluate Chinese hotels at different cities? Mining online tourist reviewers for new insights. In WHICEB (p. 67).

[3]Suzuki, S., & Kurata, Y. (2017). An Analysis of Characteristic of Tourism Destinations using User Profile of Twitter Society for Tourism Informatics, 13(1), 39-52. (in Japanese)

[4]Ferrer-Rosell, B., Coenders, G. & Marine-Roig, E. Is planning through the Internet (un)related to trip satisfaction?. Inf Technol Tourism (2017) 17: 229. https://doi.org/10.1007/s40558-017-0082-7

[5]Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management, 58, 51-65.

[6]Antonio, N., de Almeida, A., Nunes, L. et al. Hotel online reviews: different languages, different opinions. Inf Technol Tourism (2018) 18: 157. https://doi.org/10.1007/s40558-018-0107-x

[7]Song, S., Kawamura, H., Uchida, J. and Saito, H.: Towards a New Method for the Investigation of Tourists' Needs, The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 1B1-OS-11a-03 (2018)

[8]Hokkaido Government (2016). Survey Concerning Customer Satisfaction (in Japanese). http://www.pref.hokkaido.lg.jp/kz/kkd/H28doutai_home.htm [access: 8/4/2017]

[9]Schouten, K., Frasincar, F.: Survey on aspect-level sentiment analysis, IEEE Transactions on Knowledge and Data Engineering, Vol.28, No.3, pp.813-830 (2016)

[10]Song, S., Kawamura, H., Uchida, J. and Saito, H.: Towards a New Investigation Method for Tourists' Needs: from Travel Survey to the Analysis of Travel Reviews, Proceedings of The First International Conference on Digital Practice for Science, Technology, Education, and Management, Ebetsu, pp.11-16.(2018)

[11]Hai, Z., Chang, K., & Kim, J. J. (2011, February). Implicit feature identification via co-occurrence association rule mining. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 393-404). Springer, Berlin, Heidelberg.

[12]George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[13]Fukumoto, F. & Suzuki, Y. (2002). Using Synonyms and Their Hypernymy Relations of WordNet for Text Categorization. J.IPS Japan, 43(6), pp.1852-1865(in Japanese)

[14]Masaki Endo, Takuya Konda, Keisuke Saeki, Masaharu Hirota, Yohei Kurata, and Ohno, S. (2016) A Study of Classification and Extraction of Tourism Keyword Using Morpheme N-gram and RIDF for Tourism Information Retrieval in Visit Region. Tourism and Information, 12, 31-46 (in Japanese)

[15]Honda, T., Yamamoto, M. & Ohuchi, A. Automatic Website Classification for Constructing A Tourism-related Internet Directory. Tourism and Information, Vol.2(1), pp.49-57 (2006) (in Japanese)

[16]Yamamoto, J., Takuma, K., Fukuhara, K., Kamei, S., & Fujita, S. (2016). Generating a dictionay for hotel recommendation based on reviews. The 78th national convention of IPSJ, 2016(1), pp.519-520. (in Japanese)

[17]Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. Pro-ceedings of the ACL SIGDAT-Workshop. Dublin, Ire-land.

[18]Ji, L., Sum, M., Lu, Q., Li, W., & Chen, Y. (2007, February). Chinese terminology extraction using window-based contextual information. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 62-74). Springer, Berlin, Heidelberg.