Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7014–7021
Marseille, 20-25 June 2022

A Large-Scale Japanese Dataset for Aspect-based Sentiment Analysis

Yuki Nakayama, Koji Murakami, Gautam Kumar, Sudha Bhingardive, Ikuko Hardaway
Rakuten Institute of Technology, Rakuten Group Inc.
Tokyo, Japan
{yuki.b.nakayama, koji.murakami, gautam.kumar, sudha.bhingardive, ikuko.hardaway}@rakuten.com

Abstract

There has been significant progress in the field of sentiment analysis. However, aspect-based sentiment analysis (ABSA) has not been explored in the Japanese language even though it has a huge scope in many natural language processing applications such as 1) tracking sentiment towards products, movies, politicians *etc*; 2) improving customer relation models. The main reason behind this is that there is no standard Japanese dataset available for ABSA task. In this paper, we present the first standard Japanese dataset for the hotel reviews domain. The proposed dataset contains 53,192 review sentences with seven aspect categories and two polarity labels. We perform experiments on this dataset using popular ABSA approaches and report error analysis. Our experiments show that contextual models such as BERT works very well for the ABSA task in the Japanese language and also show the need to focus on other NLP tasks for better performance through our error analysis.

Keywords: Aspect-based Sentiment Analysis, Dataset Construction, Less-Resourced

1. Introduction

Sentiment analysis is a very well-known and essential field in NLP. It is used to identify the sentiment of a text or text span and classify it into pre-defined categories like *positive*, *negative*, *neutral*, *etc*. Aspect-based sentiment analysis is a fine-grained task in sentiment analysis. In this task, it is necessary to recognize words representing "aspect" in addition to identifying the polarity. For example, hotel reviews not only express the overall sentiment of a meal, but also sentiments regarding its service, atmosphere, location, amenity, and so on. Let us consider a hotel review sentence in the Japanese language with the following example:

"朝食は美味しかったですが、スタッフのサービスはいまいちでした。" (*Breakfast was delicious, but the service of the staff was not so good.*)

In this sentence, there are two aspects: 朝食 (*breakfast*) and サービス (*service*). It has positive polarity with respect to *breakfast* but negative polarity regarding *service*. In many businesses, it is indispensable to analyze such fine-grained polarities concerning different aspects to improve the quality of services. There are many versions of aspect-based sentiment analysis (ABSA) tasks which are aspect-category sentiment analysis (ACSA), aspect-term sentiment analysis (ATSA), and targeted-aspect sentiment analysis (TABSA). In sentiment analysis literature, there has been various approaches proposed so far on various languages including Japanese.

For research on ABSA, several datasets have been constructed, including SemEval-2014 Restaurant Review dataset, Laptop Review dataset (Pontiki et al., 2014), Twitter (Dong et al., 2014), SentiHood (Saeidi et al., 2016), SemEval-2015 (Pontiki et al., 2015) and another large-scale Multi-Aspect Multi-Sentiment (MAMS) dataset (Jiang et al., 2019). These datasets are based on GGM (Consumer Generated Media) and have become the benchmark datasets for the ABSA task in English. Table 1 shows the statistics of these datasets.

However, for the Japanese language, there are very few datasets publicly available for even general sentiment analysis, such as Rakuten Travel reviews[1] and NTCIR dataset (Seki et al., 2010) for sentiment analysis. Only chABSA dataset [2] is available for ABSA task. This dataset consists of $3,215$ sentences with one or more aspects and is based on not CGM but an overview of business on financial reports published by Japanese public companies. Financial report has a formal writing style than other types of documents, such as news article, Weblog, or review.

Intending to advance and facilitate research in the field of aspect-based sentiment analysis in Japanese, in this paper, we present a large-scale new user review-based dataset. The dataset is based on Rakuten Travel Review publicly available. The dataset includes 12,476 hotel reviews including 72,624 sentences. In this dataset, sentences are annotated with more than one aspect category and polarity by both crowdsourcing and internal human annotators. We empirically evaluate two BERT-based approaches and the attention-based LSTM approach on our dataset. Experimental results demonstrate that our dataset consists of a wide variety of samples ranging from easy to hard. We also conduct an error analysis and it indicates the difficulty to handle review sentences through simple features.

Our main contributions are summarized as follows: (1) We present the first standard large-scale dataset for the ABSA task in the Japanese language, which was released for public use. We manually annotate hotel re-

[1] https://www.nii.ac.jp/dsc/idr/rakuten/

[2] https://github.com/chakki-works/chABSA-dataset

Dataset	Language	Type	Size
Restaurant	En	ATSA	4,827
Restaurant	En	ACSA	4,738
Laptop	En	ATSA	3,012
Twitter	En	ATSA	6,940
SentiHood	En	ATSA	5,215
MAMS	En	ATSA	13,854
MAMS	En	ACSA	8,879
chABSA	Ja	ATSA	3,215

Table 1: Statistics of existing datasets for ABSA

views with seven aspects with two sentiments. (2) We perform experiments of ABSA task on this dataset and experimental results prove that contextual models such as BERT works very well in the Japanese language. (3) We report the challenges associated with the ABSA task in the Japanese language with detailed error analysis on this dataset.

2. Related Work

2.1. Aspect-based Sentiment Analysis

Traditional ABSA approaches utilize hand-crafted features. For English language, with the abundance of sentiment lexicons (Rao and Ravichandran, 2009; Kaji and Kitsuregawa, 2007) NRC-Canada-2014 (Kiritchenko et al., 2014) built lexicon-based features for sentiment analysis. Traditional ML approaches focus on building classifiers such as SVM (Mullen and Collier, 2004) after extracting features. However, those sentiment lexicons-based approaches suffer from the following limitations, (1) The lexicons can not apply to other languages; (2) It is very time-consuming to keep dictionaries with better quality.

In the last couple of years, deep learning has been quite popular for ABSA as well as many other NLP tasks. Transfer learning has made even deep learning models less data hungry. Wang et al. (2016) and Xue and Li (2018) proposed models based on attention-based LSTM, CNN with gating mechanisms. Some of the transfer learning techniques include Sun et al. (2019) and Xu et al. (2019) which utilize pre-trained BERT (Devlin et al., 2019). There are some other approaches such as Song et al. (2019) and Zeng et al. (2019) based on attention encoder network and local context focus mechanisms respectively. Also, Yang et al. (2019) proposed a multi-task learning approach based on BERT and local context focus mechanism. Wang et al. (2020) proposed a relational graph attention network (R-GAT) to encode the new tree structure for sentiment prediction.

Nakayama and Fujii (2015) proposed a CRF (Lafferty et al., 2001) based method while utilizing features associated with lexical and syntactic information to extract condition-opinion relations from Rakuten Travel reviews. Nio and Murakami (2018) proposed a BiLSTM network and tried to use PoS tag, Japanese SentiWordnet [3] [4] feature and Japanese polar dictionary (Kobayashi et al., 2004) to improve the performance. Also, another technique based on stacked denoising auto-encoders was proposed by Zhang and Komachi (2015) that uses distributed word representations. Bataa and Wu (2019) showed the better performance of transfer learning techniques including BERT (Devlin et al., 2019), ULMFiT (Howard and Ruder, 2018), ELMo (Peters et al., 2018) on Japanese e-commerce reviews.

2.2. Dataset for ABSA

Sentiment analysis has been focusing on CGM, such as Twitter, and reviews in the last couple of decades. At SemEval-2014, ABSA benchmark datasets including laptop and restaurant reviews were released (Pontiki et al., 2014). As another good language resource, many researchers have been focusing on Twitter and Dong et al. (2014) have constructed Twitter dataset. SentiHood (Saeidi et al., 2016) focuses the texts on the QA platform of Yahoo! in the domain of neighborhoods of London. MAMS dataset (Jiang et al., 2019) consists of restaurant reviews. For the Korean language, a dataset in an automotive domain has been released (Hyun et al., 2020) There is an interesting ABSA Japanese dataset namely chABSA, which is based on financial reports published by public companies. An example of the dataset is shown below:

"国内のきのこ事業の売上高は422億96100万円（同3.1%増）となりました。" (*Sales of domestic mushroom business was 42.961 billion yen (3.1% increased).*) → { "きのこ事業 (*domestic mushroom business*)" : business#sales#positive }

In this case, the polarity "Positive" is identified by "増 (*increased*)" in parenthesis. It is necessary to have different techniques to work on this dataset.

3. Dataset Construction

3.1. Data Collection and Target Aspects

Our research goal is to automatically produce a radar chart based on users' reviews on the travel domain. We use Rakuten Travel Review as our data to construct the ABSA dataset because of the following reasons. First, the data provides over 100k review sentences so that we are allowed to construct a large-scale dataset and we do not need to find other data resources. Second, 6 aspect categories are already pre-defined in the original data, and we are able to follow their business direction. The most important reason is that the data include their radar chart through (1-5) stars as in Figure 1 which users gave so that it is easily possible to compare a given chart and an automatically generated chart for system evaluation.

[3] http://sentiwordnet.isti.cnr.it
[4] http://compling.hss.ntu.edu.sg/wnja/

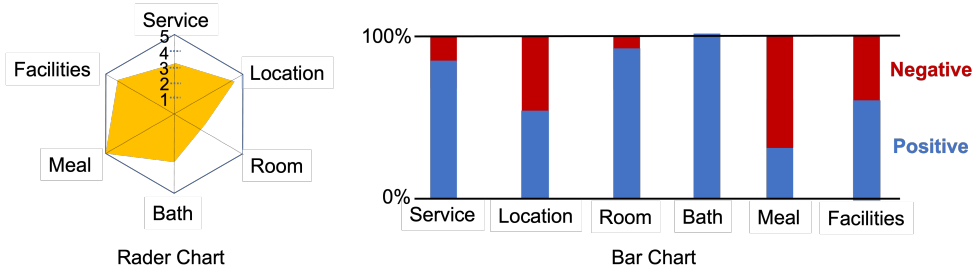

Figure 1: Examples of visualizing evaluations on an aspect by aspect basis in ABSA

Total number of reviews	12,476
Total number of sentences	76,624
Total number of unique sentences	72,525
Average of sentences per review	6.14
Average of words per review	65.75
Total number of unique words	19,735
Total number of reviewed hotel	1,329

Table 2: Statistics of our dataset perspective

In order to construct our dataset, we first collected hotel reviews through Rakuten Data Release[5]. All the reviews have been written from 2017 to 2019. We removed reviews that consist of only meaningless sentences, such as a greeting of a few words.

Table 2 shows statistics of our target dataset. It is very interesting that reviewers have written over 6 sentences in a review with a variety of words and it means the quality of the contents in the dataset is reliable. In other ABSA datasets introduced in Table 1, each sentence is a unit for annotation. However, we give annotators not sentences, but a complete review. Our data includes over 6 sentences in a review so it is likely that an annotator sometimes faces co-reference issues.

Review sentences mention a lot of aspects with different granularity. For example, a reviewer writes about meals, facilities, parking lots, or concierge, but another reviewer mentions the freshness of vegetables in a salad or the color of towels in their room. In our dataset construction, we follow the original 6 aspect categories. However, we made a decision to divide 食事(*Meal*) to 夕食(*Dinner*) and 朝食(*Breakfast*), because in a lot of cases, the meal type is different between dinner and breakfast even in a traditional hotel in Japan so they should be handled separately. Table 3 shows the definition of all 7 aspect categories, examples, and their polarity.

For designing ATSA (Aspect Term Sentiment Analysis) task, annotating aspect term (or OTE (Pontiki et al., 2016)) is crucial. However, our current goal is to create a visualization of the evaluation on an aspect category, so that we annotate only seven aspect categories to review sentences in this paper. Expanding the dataset to ATST is our future work.

3.2. Annotation Strategy

In other datasets described earlier, "Positive", "Negative" and "Neutral" are commonly used as polarity labels. Identifying "Neutral" is important to understand the reviews. However, unlike the other two polarities labels, this polarity does not play an important role to overview the distribution of evaluation on an aspect-by-aspect basis with visualization such as radar chart and bar chart as shown in Figure 1. Therefore, we focus on only "Positive" and "Negativ" as polarity labels in our dataset. When we had preliminary annotation to 200 reviews including around 1,100 sentences, 80% of sentences were annotated for either "Positive" or "Negative" and half of unannotated sentences are greetings or fact descriptions. This means annotating only two polarity labels still covers enough sentences and we are able to reduce annotation labor because annotators just need to consider 14 candidates (7 aspect categories $*$ 2 polarities) rather than 21 candidates ($7 * 3$). However, it is still very hard even for experienced annotators to work against thousands of sentences from scratch. From our preliminary annotation, we estimated that 400 reviews are able to be annotated by an annotator in a week so that accomplishing data annotation needs over 30 weeks. In order to reduce annotators' burden and to complete the work in a shorter period, we tried two-stage annotation. The first stage is trial annotation, which is performed using crowdsourcing, and at the second stage, experienced annotators correct the information annotated at the first stage.

We employed 43 Japanese native people via a Japanese agent company[6] for 2 months to work on annotation according to our specification. This stage aimed to make overlap annotation, one review can be annotated by multiple annotators. We assume that different annotators should work the same if a sentence can be annotated easily while annotation results can be unsynchronized when they face the difficult examples. we assigned only 1 annotator to the short reviews, which consist of less than 4 sentences while more than 2 annotators were asked to work on the longer reviews. Fi-

[5] https://www.nii.ac.jp/dsc/idr/rakuten/

[6] http://www.abejainc.com

Aspect	Definition of expressions	Examples and their polarity ((P)ositive or (N)egative)
Service	Actions of helping for clients' satisfaction	(P) 名古屋駅から少し歩きますが、朝は送ってもらえたので助かりました。(It takes a little walk from Nagoya station, but a ride to the station in the morning was nice.)
Dinner	Evening meal, including taste, plate or ingredients	(P) 季節に合わせた懐石料理は見た目も美しく美味しかったです。 (Seasonal Kaiseki cuisine was beautiful and delicious.)
Breakfast	Morning meal, including taste, plate or ingredients	(P) 朝食は、種類も多く、満足でした。 (The breakfast has lots of different dishes and was satisfactory.)
Location	Access or landscape	(P) 隣には郵便局もあり買ったお土産もすぐ送れる。 (You can send souvenirs you brought, from the post office next door.)
Facilities	Resources in the hotel except room and bath	(N) ただ唯一、部屋の金庫が小さいのがマイナスポイント。 (The only downside was the safebox in the room was small.)
Room	Attribute of the room	(P) 広いお部屋でゆっくりとした時間を過ごす事が出来ました。 (I had a relaxing time in the spacious room.)
Bath	Public bath or bath in the room	(N) 風呂は混んでいて入浴出来ませんでした。 (I couldn't take a bath at the common bath, because it was too crowded.)

Table 3: Definition of aspects and examples in the dataset

nally, each of longer reviews (7,335 reviews in total) has been annotated by 2.61 annotators on average. We asked 2 our annotators to correct the annotated information. These annotators are native Japanese speakers and have over 8 years annotation experience.

Our annotation guideline was based on the experience of preliminary annotation described above. When we annotated 200 reviews in the preliminary annotation, a lot of controversial cases were found and considered. We developed a annotation guideline based on our specification and distributed to all annotators. The following cases are some of interesting difficulties in our review data.

3.2.1. Fact and Opinion

Lots of sentence in a review can include not only some aspect or expressions with polarity, but just facts. It is necessary to focus on only the contents of the sentences and any aspect or polarity are not annotated when the sentences do not include any opinion. For example, the sentence

"夕食は舟盛りと蟹でした。" (*Sashimi boat and crabs were served as main dishes at dinner.*)

shows what the main dishes were at the dinner as a fact. Even if a sentence mentions a dish for celebration, like turkey at Thanksgiving, we do not annotate because the sentence does not include any opinion.

3.2.2. Indirect Opinions

Sentences sometimes do not include explicit opinion or fact, but implicit opinion. It is necessary to identify if the reviewer is satisfied or frustrated with the aspect. For example, the sentence

"和洋室のお部屋は所々リフォームされていて苦になりません。" (*Japanese-Western style room was comfortable as it's renovated well.*)

represents the reviewer does not feel uncomfortable in the room so that "Room" and "Positive" can be annotated to the sentence.

3.2.3. Multiple Aspects and Opinions

Longer sentences tend to include multiple aspects and opinions. It is easy to recognize opinions when a sentence is criticizing multiple aspects. When a sentence includes opposed opinions, both "Positive" and "Negative" can be annotated even if both opinions are regarding the same aspect. For example, the sentence

"中身は素晴らしいので、外観が勿体無い、という感じでした。" (*Even though interior design is great, exterior should be improved.*)

includes two opposed opinions to the same aspect category, so that both "Positive" and "Negative" to "Facility" are annotated.

3.2.4. Wish and Suggestion

Review sentences represent not only reviewers' opinion, but their suggestions. For example, the sentence

"シャワーは温度調節機能付なら、100点とおもいます。" (*If the shower could have temperature control, it would be perfect.*)

includes a wish in the sentence. It is necessary to identify if the reviewer is satisfied or frustrated with the aspect, and "Bath(Pos)" can be annotated. In this case, the reviewer is already satisfied and proposes a better way.

3.3. Analyzing the Dataset

We provided $12,476$ reviews, including $76,624$ sentences to our annotators, and they annotated at least one aspect category to $53,192$ sentences. In these annotated sentences, the average category per sentence is calculated as 1.73. Furthermore, $19,700$ out of $53,192$ sentences have more than two aspect categories. This number is comparable in scale to MAMS (Jiang et al., 2019), which has 9,000 sentences in the ACSA setting. The number of un-annotated sentences is not small. However, 61.8% of those sentences are the first or the last sentence in a review. It is reasonable that both the

first and last sentence in review tend to not have neither aspect category nor polarity because a lot of reviewers start writing their sentence with some fact such as

"夫婦で‘6月15日に利用しました。" (*My wife and I stayed at the hotel on June 15th.*) or

"奈良での常宿です。" (*This hotel is my favorite in Nara city.*).

They can finish their review with greetings, like

"お世話になりました、ありがとうございました！" (*Great stay, thank you very much!*) or

"ぜひまた利用したいと思います。" (*I believe I will stay again.*).

Next, to evaluate inter-annotator agreement, 375 randomly selected sentences were annotated by two human annotators. The inter-annotator agreement was $\kappa = 0.78$. The agreement between them was 80.4% and 79.8% measured as micro and macro average F_1 scores, respectively, when the annotation results are evaluated in which one annotation result is treated as a gold standard and the others as the output of the system. The aspect category, which includes disagreed examples the most was *bath* (micro F1 = 72.80). 56.25% of disagreement was caused by missing additional aspect category annotation.
Table 4 shows the statistics of the number of *positive*, *negative* and *conflict* labeled review sentences with respect to each aspect category. There are total $66,405$ instances with *positive* label while $21,452$ instances with *negative* label and $2,208$ with *conflict* label. The *conflict* label applies when both polarities, *positive* and *negative* are expressed at the same aspect category.

Aspect	Positive	Negative	Conflict	Total
Breakfast	12,357	2,625	350	15,332
Dinner	9,299	1,893	257	11,449
Bath	7,642	2,269	353	10,264
Service	13,916	5,692	305	19,913
Location	6,144	931	187	7,262
Facilities	8,784	5,507	413	14,704
Room	8,263	2,535	343	11,141
Total	66,405	21,452	2,208	90,065

Table 4: Statistics of Aspect Category and Polarity

4. Evaluation

In order to evaluate the quality of the proposed dataset, we perform experiments on following two sub-tasks.

- Aspect Category Detection (ACD): Detection of aspect categories in the review
- Sentiment Classification (SC): Classification of sentiment of the review for each identified aspect

The dataset has been split into $80 : 10 : 10$ ratio for training, validation, and testing while maintaining the same distribution among all 14 labels (7 aspect categories $* 2$ sentiments (positive and negative)).

Metrics used for evaluation: Macro *F1-score* and *Accuracy* have been used for both tasks.

Approaches used: We perform experiments using BERT-based and Attention-based approaches, which are the most basic and popular in ABSA tasks.

4.1. BERT-based Approaches

We use pre-trained BERT model on Japanese Wikipedia provided by *Tohoku University* [7]. The number of Transformer blocks is 12, the hidden layer size is 768, the number of self-attention heads is 12, and the total number of parameters for the pre-trained model is $110M$. When fine-tuning, we keep the dropout probability at 0.1 and an optimum number of epochs was determined by the validation set. The initial learning rate is $1e^{-5}$, and the batch size is 32.

4.1.1. BERT Multi-label Classification (BERT-MLC)

In this setting, we formulate the ABSA task as multi-label classification with 14 labels which are a combination of 7 aspect categories and 2 sentiment labels. We use an output of hidden representation corresponding to the $[CLS]$ token for classification. We feed the output to linear layer with $sigmoid$ activation function. If a value of the sigmoid function is less than 0.5 for all output units, the system returns $None$. The label means that a sentence does not belong to any of the fourteen categories. We set the maximum sequence length to 275, which corresponds to the maximum number of tokens in the data, plus the number of special tokens, $[CLS]$ and $[SEP]$.

4.1.2. BERT Sentence Pair Classification (BERT-SPC)

Here, we convert ABSA task to Sentence Pair Classification similar to Song et al. (2019) approach which feeds sequence $[CLS] + sentence + [SEP] + aspect\ category + [SEP]$ into the basic $BERT$ model for sentence pair classification task. We define three labels *i.e.* {positive, negative, none} for each (sentence, aspect) pair. If sentence s does not mention anything about aspect category a, we assume the sentence pair (s, a) has class $none$. We set maximum sequence length to 278.

4.2. Attention-based LSTM with Aspect Embedding (ATAE-LSTM)

We use formulation proposed by Wang et al. (2016) and repository [8] for this experiment. In this setting,

[7] https://github.com/cl-tohoku/bert-japanese

[8] https://github.com/songyouwei/ABSA-PyTorch

we concatenate sentence token embeddings with the aspect embedding for each aspect category and feed it in Attention-based LSTM network. We also have three labels *i.e.* {positive, negative, none} for each (sentence,aspect) pair as well as BERT-SPC. We use $Mecab$ [9] tokenizer with $Unidic$ [10] dictionary (version: 2.3.0+2020-10-08) and pre-trained fasttext embeddings for Japanese [11]. The hidden layer size is 100, the dropout probability at 0.1, the number of epochs is 20 and maximum sequence length is 128. The initial learning rate is 0.0005, and the batch size is 64. Total number of trainable parameters is 922,703 and nontrainable parameters is 6,720,600. We pick the weights with best *F1-score* on validation data.

4.3. Results and Discussions

We compare the results for all approaches mentioned in Section 4.1 and Section 4.2. Table 5 presents experimental results for both ACD and SC tasks. The results demonstrate BERT-based approaches worked much better than Attention-based LSTM.

For ACD task, *Accuracy* score is much larger than *F1-score* over the three approaches. The gap is due to that *Accuracy* score covers an evaluation on unannotated cases where a sentence does not belong to any of the seven categories. BERT-SPC achieved the highest *F1-score* 0.771 and accuracy 92.3%. On the other hand, for the SC task, BERT-MLC achieved the highest *F1-score* 0.958 and accuracy 97.1%. This is because BERT-SPC was not able to deal with the *Conflict* cases described in Section 3.3. Also, BERT-SPC is found to be seven times more time-consuming than BERT-MLC as BERT-SPC need to make sentence pair for each aspect category.

Model	ACD		SC	
	F1	Acc.	F1	Acc.
BERT-MLC	0.761	**0.924**	**0.958**	**0.971**
BERT-SPC	**0.771**	0.923	0.947	0.964
ATAE-LSTM	0.530	0.769	0.847	0.870

Table 5: Comparison of results on Aspect Category Detection (ACD) and Sentiment Classification (SC)

F1-score for all 14 labels combined are reported in Table 6. Overall *F1-score* is found to be best for *breakfast-positive* category using BERT-MLC. On the other hand, overall *F1-score* is found to be lowest for the *room-negative* category. This is because *room* aspect is sometimes implicit in the review sentence and has a huge variety of topics with a lot of ambiguities. From the table, one can see that BERT-MLC tends to have better performance than BERT-SPC on negative sentiments for most of the aspect categories.

Aspect-Sentiment	BERT SPC	BERT MLC
Breakfast (Pos)	0.806	**0.818**
Breakfast (Neg)	0.614	**0.680**
Dinner (Pos)	**0.788**	0.787
Dinner (Neg)	0.545	**0.658**
Bath (Pos)	**0.774**	0.771
Bath (Neg)	0.634	**0.689**
Service (Pos)	**0.810**	0.799
Service (Neg)	0.630	**0.675**
Location (Pos)	**0.731**	0.726
Location (Neg)	**0.638**	0.623
Facilities (Pos)	**0.728**	0.706
Facilities (Neg)	**0.656**	0.651
Room (Pos)	0.759	**0.765**
Room (Neg)	0.579	**0.609**
Macro Average	0.692	**0.711**

Table 6: Macro *F1-score* for all 14 labels combined

5. Error Analysis and Next Challenges

To grasp characteristics of our dataset, we conducted error analysis of BERT-MLC approach. We found that the causes of errors characteristic in our dataset are *Zero Anaphora*, *Fine-grained similar category* and *Implicit negative opinions attributed to Japanese culture*.

Zero Anaphora There are 37 cases related to this issue in the dataset. Zero anaphora means that an aspect term was not explicitly mentioned in the sentence. This issue is frequently raised as a linguistic phenomenon in the Japanese language. For example, the sentence

"思いがけないことだったので、とても驚き、嬉しかったです。" (*I was surprised and pleased because of unexpected things.*)

is related to *Service* aspect with *Positive* label but no keyword related to *Service* is explicitly written here. Even if we tried to avoid co-reference/zero anaphora issue through offering not each sentence, but review, this issue frequently happens in customer-generated text, especially.

Fine-grained similar category There are 73 cases where a sentence with only *Dinner* aspect is classified into both *Breakfast* and *Dinner* aspect categories. For example, for the following sentence,

"料理も半個室の食事処で大変美味しく頂きました。" (*The food was very delicious at the semi-private room.*)

the system was not able to distinguish *breakfast* and *dinner* categories because reviewer has mentioned only coarse-grained term such as *meal*. Such error can be solved considering the context. The following sentence appeared just after the above sentence.

[9] https://pypi.org/project/mecab-python3/

[10] https://unidic.ninjal.ac.jp/

[11] https://fasttext.cc/docs/en/pretrained-vectors.html

“朝食は、バイキングで昨夜と同じ食事処で頂きましたが、...” (*I had breakfast at the same restaurant as last night at the buffet,...*).

Thus, we need to devise methods to select a context appropriate for aspect categories and effectively encode the selected sentence.

Implicit negative opinions attributed to Japanese culture 15 errors related to that Japanese people tend to be humble and polite in giving negative feedback. For the example, for the following sentence,

“できれば、どんな種類の料理が食べたいか、アンケートをとったらいかがでしょうか。” (*If possible, why don't you take a questionnaire about what kind of foods you want to eat?*)

the reviewer did not express an explicit negative opinion such as “The assortment of food is bad.” because Japanese people sometimes hesitate to directly tell their feelings with strong opinion words.

6. Conclusion and Future Work

In this paper, we presented a large-scale ABSA dataset consisting of 76,624 review sentences for the Japanese language in the hotel reviews domain. This is the largest and first standard dataset that is publicly available for the Japanese language. We provide strong baselines for Japanese ABSA tasks using BERT and Attention-based LSTM networks on this dataset. Our results show that even Japanese language transfer learning techniques (in our case BERT) work better for Aspect Category Detection and Sentiment Classification. We also conducted error analysis to grasp the characteristics of our dataset and presented several important issues for further improvement.

Our dataset includes 7 aspect categories and 2 polarities, however, annotating both aspect terms (or OTE) and “Neutral” polarity to review sentences was out of our scope.

It is obvious that these kinds of information are indispensable for not only ATSA but also deeply understanding the review sentences. Our dataset, which is named “Rakuten Travel Review aspects and sentiment-tagged corpus”, is available at Rakuten Data Release[12]. We are currently working on additional annotation for the next version and the second release of our dataset will be made available in the future.

7. Bibliographical References

Bataa, E. and Wu, J. (2019). An investigation of transfer learning-based sentiment analysis in japanese. *CoRR*, abs/1905.09642.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 49––54.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Hyun, D., Cho, J., and Yu, H. (2020). Building large-scale English and Korean datasets for aspect-level sentiment analysis in automotive domain. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 961–966, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Jiang, Q., Chen, L., Xu, R., Ao, X., and Yang, M. (2019). A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China, November. Association for Computational Linguistics.

Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083, Prague, Czech Republic, June. Association for Computational Linguistics.

Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. (2014). NRC-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August. Association for Computational Linguistics.

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., and Fukushima, T. (2004). Collecting evaluative expressions for opinion extraction. In *International Conference on Natural Language Processing*, pages 596–605. Springer.

Lafferty, J., Mccallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 01.

Mullen, T. and Collier, N. (2004). Sentiment analysis

[12] https://rit.rakuten.com/data_release/

using support vector machines with diverse information sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 412–418, Barcelona, Spain, July. Association for Computational Linguistics.

Nakayama, Y. and Fujii, A. (2015). Extracting condition-opinion relations toward fine-grained opinion mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 622–631, Lisbon, Portugal, September. Association for Computational Linguistics.

Nio, L. and Murakami, K. (2018). Japanese sentiment classification using bidirectional long short-term memory recurrent neural network. In *"In Proceedings of the 24th Annual Meeting Association for Natural Language Processing (pp. 1119-1122)."*.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June. Association for Computational Linguistics.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June. Association for Computational Linguistics.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 675–682, Athens, Greece, March. Association for Computational Linguistics.

Saeidi, M., Bouchard, G., Liakata, M., and Riedel, S. (2016). SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Seki, Y., Ku, L.-W., Sun, L., Chen, H.-H., and Kando, N. (2010). Overview of multilingual opinion analysis task at NTCIR-8. In *Proceedings of the Seventh NTCIR Workshop*.

Song, Y., Wang, J., Jiang, T., Liu, Z., and Rao, Y. (2019). Attentional encoder network for targeted sentiment classification. *CoRR*, abs/1902.09314.

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *CoRR*, abs/1903.09588.

Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Wang, K., Shen, W., Yang, Y., Quan, X., and Wang, R. (2020). Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online, July. Association for Computational Linguistics.

Xu, H., Liu, B., Shu, L., and Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. *CoRR*, abs/1904.02232.

Xue, W. and Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia, July. Association for Computational Linguistics.

Yang, H., Zeng, B., Yang, J., Song, Y., and Xu, R. (2019). A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *arXiv preprint arXiv:1912.07976*.

Zeng, B., Yang, H., Xu, R., Zhou, W., and Han, X. (2019). Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9:3389, 08.

Zhang, P. and Komachi, M. (2015). Japanese sentiment classification with stacked denoising auto-encoder using distributed word representation. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 150–159, Shanghai, China, October.

8. Language Resource References