Extracting Pairs of Aspects and Opinion Words for Hotel Reviews in Myanmar Language

Cho Cho Hnin[#], Naw Naw[*]
[#,*]*Department of Information Science, University of Technnology (Yatanarpon Cyber City), Myanmar*

***Abstract*— This paper presents a rule based approach for extracting pairs of aspects and opinion words from informal user reviews written in Myanmar language. The system analyses hotel user reviews collected from hotel sites. The task of aspect level opinion mining mainly depends on identifying the relatedness between the aspects and opinion words in the reviews. Thus, it is an important fundamental task of aspect level opinion mining. It is one of the challenges because of informal writing styles of reviews. Especially it is a difficult task of aspects and opinion words extraction for Myanmar reviews due to the nature of Myanmar language. Firstly, the frequent nouns and noun phrases are identified as the aspects. Then, our focus is for extracting the relevant pairs of aspects and opinion words by using the linguistic rules.**

***Keywords*— aspect extraction, Myanmar language, opinion mining, opinion word extraction, syntactic patterns**

I. INTRODUCTION

Some users describe their opinions about aspects in details instead of the entire document. In such situation, it is necessary to analyze their comments for identifying these opinions. Some users want to know what aspects are liked or dislike by other users. Thus, the aspect level analysis is preferred for this work.

The extraction of aspects and opinion words is an essential task of feature level opinion mining to determine aspects and opinion words from the opinionated sentences of user reviews. Moreover, it is important to extract pairs of aspects and opinion words correctly, especially when the review sentences contain two or more aspects and opinion words. The main tasks of this level are identification of aspects, determining the opinion word and the calculation of opinion words for each aspect.

Most of the tasks in aspect level opinion mining are carried out in English language. Furthermore, some works perform the opinion mining in Chinese language, Hindi language and Korean language. It is relevant to perform the aspect and opinion words extraction using our natural languages, Myanmar language, prior performing sentiment classification task. Thus, this paper addresses how to approach extracting pairs of aspects and opinion words for Myanmar informal reviews.

II. RELATED WORK

There are various methods to perform feature extraction task in the aspect level opinion mining. We describe other related methods for features extraction and opinion words extraction.

One of the extraction works is in [6]. They proposed an approach for identifying aspects on which reviewers have expressed their opinions and classifying which aspects are positive or negative. They performed the feature extraction task by applying association rule mining. The frequent nouns or noun phrases in reviews are assumed to be aspects. They achieved the average accuracy of 84% over five products.

The methods proposed in [6], have been extended to use it to the field hotel reviews by the authors in [4]. They determined the sentiment scores for each aspect is calculated by using special linguistic rules. Their approach determined the sentiment orientation with the precision and recall of 90%.

In [9], the authors proposed a frequent pattern mining algorithm called H-mine for extracting features from the customer reviews. The system mainly focuses on those aspects that have been reviewed mostly by the customers.

The authors in [11], a two-fold rule-based model (TF-RBM) have described by using rules based on the sequential patterns extracted from the user reviews. Firstly, they extracted aspects which are correlated to domain independent opinions. Then, they extracted aspects related with domain dependent opinions. Their model improves the accuracy for extraction of aspects by applying frequency based and similarity based approach.

In [1], the authors compared Apriori and Generalized Sequential Pattern (GSP) algorithms in identifying frequent aspects and opinion words. This paper described that GSP is more substantial for mining data than Apriori.

A method which based on a graph for creating a subjective lexicon for Hindi language was proposed in [7]. They built the subjective lexicon with the help of the lexical resource, Word Net. Some opinion words are initialized as a seed list by using Word Net. It also contains the synonyms and antonyms of the opinion words. They considered every word in the Word net as

a node by traversing Word Net as a graph and by connecting to the synonyms and antonyms. The system obtained the accuracy of 74% for the classification phase. The 69% accuracy is also obtained in matching the Hindi words annotated by the humans.

Lizean Liu et al. [5] have performed a technique for feature extraction of opinion mining in Chinese language. They proposed an algorithm for clustering product features based on the structure of the Chinese reviews. They extract features by using opinion words as feature indicators. They also identify implicit features and cluster the features depending on the context-dependent information.

Q. Liau et al. [8] presented an approach to select rules automatically for identification of aspects. They performed rule set selection algorithm with three steps: the evaluation for rule, ranking of this rule and final selection phase.

In [2], the authors proposed a technique for feature segmentation and feature categorization. First, their system segments the review sentences which consist of multiple features, into the individual features. Then, it identifies the irrelevant feature with the help of speech dictionary and context information. It also determines the scores of the feature using opinion words. Finally, the product features are categorized based on clustering.

In [10], the authors focus on to extract the patterns of opinion words or phrase for the aspects written by the users. These patterns consist of noun, adjective, adverb and verb. These aspects and opinion words are needed in the opinion summarization phase for the users to know which aspects are suitable and which are not. They performed their experiments on five products.

III. Proposed System and Methodology

The proposed system mainly performs the extraction for pairs of aspects and opinion words in the detailed aspect level for Myanmar language.

Firstly, the input user reviews are split into individual sentences. Then, word segmentation process is performed for individual sentences. After the pre-processing phase, the frequent nouns are extracted that are related with the opinion words in the opinion lexicon. The extract aspects are also matched with the opinion words lexicon to form the relevant pairs. The overall system design for extraction process is shown in Fig. 1.

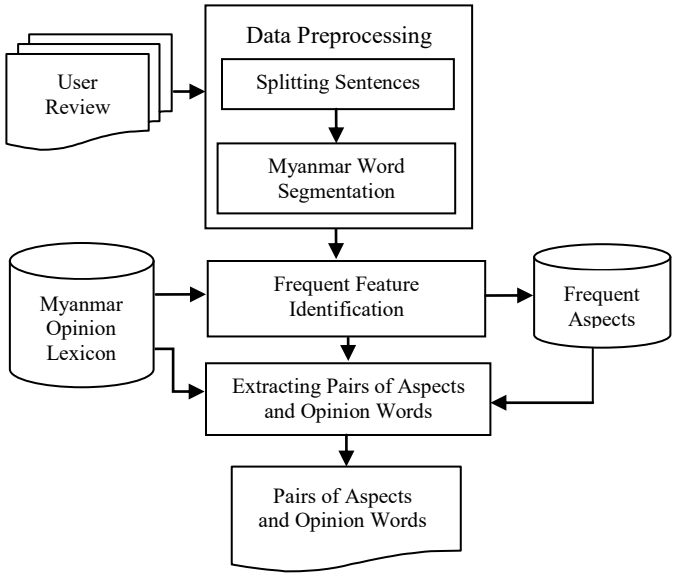

Fig. 1. Overview design for the extraction system

A. Lexicon Creation

Myanmar sentiment lexicon for hotel domain is also compiled in order to use for both the extraction phase and the sentiment classification phase. This lexicon cosists of the lists of positive, negative and neutral opinion words. Morever, it contains some Myanmar idioms that are recently talked by users about Hotel reviews and the intensifiers that modify the opinion words.

Furthermore, we also consturct Myanmar dictionary which consists of the synonyms and antonyms in order to user in feature categorization.

B. Data Preprocessing

A major preprocessing task is the word segmentation because Myanmar language does not have delimiter or white space between words. Myanmar word segmentation includes two subprocess called syllable segmentation and syllable merging. Our system performs the segmentation task with the help of the syllable segmentation method proposed in [3].

C. Frequent Aspects Identification

After the Myanmar word segmentation phase, it is necessary to identity the aspects on which many users have expressed their opinions. The important task is to find what users like and dislike about a particular feature. Aspects are the important features commented by the users (e.g. ဝန်ဆောင်မှု (service), အစားအစာ (food)). Aspects may be the nouns and noun phrases which is appeared in review sentences. If the nouns have the their synonyms in matching with the dictionary, these nouns are grouped. For example, the words အစားအစာ (food) and အစားအသောက် (food) have the same meaning. Thus, such words are categorized into the one.

Frequent nouns can most likely be considered as aspect words. So, the frequent pattern mining method is applied to extract the frequent aspects. The frequent nouns which are related with the opinion words are considered as the possible aspects.

D. Extraction of Aspects and Opinion Words

The essential task in aspect level sentiment analysis is to extract pairs of aspects and opinion words correctly. In the task of opinion mining at this level, the aspect and its corresponding opinion words must be determined. Opinion words are the indicators that describe opinions about aspects (e.g. ကောင်းမွန် (good), ညစ်ပတ် (dirty)). The adjectives, adverbs and verbs contained in the lexicon are considered as opinion words.

Some review sentence contains an aspect and its related opinion words. However, some sentence consists of two or more aspects and opinion words. In such case, it is a big challenge to determine the relevant pairs of aspects and opinion words correctly. In other research works, the nearby opinion word is extracted as its qualified word of the aspect. But, it is unable to extract the relevant opinion words correctly for the related aspects.

By the arrangement of the Myanmar words in the informal reviews, the opinion words may be both on the left of the aspect but also on the right of it. The possible patterns of aspects and corresponding opinion words are analyzed in order to extract the correct pairs of aspects and opinion words with the help of some syntactic rules.

In the extraction phase, it is also important to handle negation words and intensifier. These words can affect the task of sentiment classification. Myanmar negation words can be in the front of the opinion words. These words are words such as " မ " (not/no), "လုံးဝ" (never). Thus, the polarity score of the next closest opinion word will be reversed in the aspect scoring phase. The intensifiers are words such as အလွန် (very), အရမ်း (too) and အနည်းငယ် (a few). These words can change the orientation of the opinion words.

TABLE I. RULES FOR ASPECTS AND OPINION WORDS EXTRACTION

Type	Observation	Aspect & Opinion Word Pairs
1	A , OW	{A+OW}
2	A , Mod , OW	{A+Mod+OW}
3	A , OW_1 , Conj , OW_2	{A+OW_1} , {A+OW_2}
4	A_1 , Conj , A_2 , OW	{A_1+OW} , {A_2+OW}
5	A_1 , Conj , A_2 , Mod , OW	{A_1+Mod+OW} , {A_2+Mod+OW}
6	A_1 , Conj , A_2 , OW_1 , Conj , OW_2	{A_1+OW_1} , {A_1+OW_2}, {A_2+OW_1} , {A_2+OW_2}
7	A_1, OW_1 , Conj , A_2 , OW_2	{A_1+OW_1} , {A_2+OW_2}
8	A_1 , Mod , OW_1 , Conj , A_2 , OW_2	{A_1+Mod+OW_1} , {A_2+OW_2}
9	A_1 ,OW_1 , Conj ,A_2 ,Mod, OW_2	{A_1 +OW_1} , {A_2+Mod+OW_2}
10	A_1 , OW_1 , Conj , A_2 , NW OW_2	{A_1+OW_1}, {A_2+NW+OW_2}

TABLE II. EXAMPLE SENTENCES FOR ASPECTS AND OPINION WORDS EXTRACTION

Type	Examples
1	ဒီဟိုတယ်မှာ ဧည့်ခန်းက လှပ ပါတယ်။ ဧည့်ခန်း (A) , လှပ (OW) → { ဧည့်ခန်း (A) + လှပ (OW) }
2	ဒီဟိုတယ်မှာ ဧည့်ခန်းက အရမ်း လှပ ပါတယ်။ ဧည့်ခန်း (A) , အရမ်း (Mod) , လှပ (OW) → { ဧည့်ခန်း (A) + အရမ်း (Mod) + လှပ (OW) }
3	ဒီဟိုတယ်မှာ ဧည့်ခန်းက လှပ ပြီး ခုံညား ပါတယ်။ ဧည့်ခန်း (A) , လှပ (OW) , ပြီး(Conj) , ခုံညား (OW) → { ဧည့်ခန်း (A) + လှပ (OW) } , { ဧည့်ခန်း (A) + ခုံညား (OW) }
4	ဧည့်ခန်း နဲ့ အိပ်ခန်း တွေက လှပ ပါတယ်။ ဧည့်ခန်း (A) , နဲ့ (Conj) , အိပ်ခန်း (A) , လှပ (OW) → { ဧည့်ခန်း (A) + လှပ (OW) } , { အိပ်ခန်း (A) + လှပ (OW) }
5	ဧည့်ခန်း နဲ့ အိပ်ခန်း တွေက အရမ်း လှပ ပါတယ်။ ဧည့်ခန်း (A) , နဲ့ (Conj) , အိပ်ခန်း (A) , အရမ်း (Mod), လှပ(OW) → { ဧည့်ခန်း (A) + အရမ်း (Mod) + လှပ (OW) } , { အိပ်ခန်း (A) + အရမ်း (Mod) + လှပ (OW) }
6	ဧည့်ခန်း နဲ့ အိပ်ခန်း တွေက လှပ ပြီး ခုံညား ပါတယ်။ ဧည့်ခန်း(A), နဲ့(Conj), အိပ်ခန်း(A), လှပ(OW),ပြီး(Conj),ခုံညား (OW) → { ဧည့်ခန်း (A) + လှပ (OW) } , { ဧည့်ခန်း (A) + ခုံညား (OW) } , { အိပ်ခန်း (A) + လှပ (OW) } , { အိပ်ခန်း (A) + ခုံညား (OW) }
7	ဧည့်ခန်း က လှပ ပြီး ဝန်ဆောင်မှု က ကောင်း ပါတယ်။ ဧည့်ခန်း(A), လှပ(OW) , ပြီး(Conj), ဝန်ဆောင်မှု(A), ကောင်း(OW) → { ဧည့်ခန်း (A) + လှပ (OW) } , { ဝန်ဆောင်မှု (A) + ကောင်း (OW) }
8	ဧည့်ခန်း က အရမ်း လှပ ပြီး ဝန်ဆောင်မှု က ကောင်း ပါတယ်။ ဧည့်ခန်း (A) , အရမ်း (Mod), လှပ (OW), ပြီး (Conj), ဝန်ဆောင်မှု (A) , ကောင်း (OW) → { ဧည့်ခန်း (A) + အရမ်း (Mod) + လှပ (OW) } , { ဝန်ဆောင်မှု (A) + ကောင်း (OW) }
9	ဧည့်ခန်း က လှပ ပြီး ဝန်ဆောင်မှု က အရမ်း ကောင်း ပါတယ်။ ဧည့်ခန်း (A) , လှပ (OW) , ပြီး (Conj) , ဝန်ဆောင်မှု (A) , အရမ်း, ကောင်း (OW) → { ဧည့်ခန်း (A) + လှပ (OW) } , { ဝန်ဆောင်မှု (A) + အရမ်း (Mod) + ကောင်း (OW) }
10	ဧည့်ခန်း က လှပ ပြီး ဝန်ဆောင်မှု က မ ကောင်း ပါဘူး။ ဧည့်ခန်း (A) , လှပ (OW) , ပြီး (Conj) , ဝန်ဆောင်မှု (A) , မ (NW) , ကောင်း (OW) → { ဧည့်ခန်း (A) + လှပ (OW) } , { ဝန်ဆောင်မှု (A) + မ (NW) + ကောင်း (OW) }

Many different syntactic patterns are occurred in the Myanmar informal review texts. Among them, some syntactic patterns are shown in TABLE I. Here, A is for aspect, OW is for opinion word, Mod is for intensifier, NW is for negation word and Conj is for the conjunction.

For Type (1) sentence, there are an aspect (A) and an opinion word (OW). Thus, the opinion word modifies only one aspect. If the sentence contains one or two intensifiers in front of the opinion word as in pattern (2), it is needed to extract these intensifiers together with the opinion word. If the review sentence is in the form of pattern (3), only one aspect is qualified by two opinion words. In this case, two pairs of aspects and opinion words are extracted. In pattern (4), there are two aspects qualified by one opinion word. Therefore, each aspect is matching with this opinion word. Example sentences for each pattern described in TABLE I, are shown in TABLE II.

IV. CASE STUDY AND PERFORMANCE EVALUATION

There are 1000 hotel user reviews collected from three hotel sites. Firstly, the review documents are split into individual sentences. Then, these are segmented into words by Myanmar word segmentation algorithm. After the preprocessing task, the frequent aspects are identified using the co-occurrence relationship between the aspects and opinion words. Frequent aspects are extracted with the help of opinion words as the feature indicators. It can be seen from the TABLE III that there are about 25 aspects for Hotel domain in the aspect extraction phase.

And then, pairs of aspects and opinion words are extracted by special linguistic rules. This step is important because the correct calculation of the polarity scores for aspect level opinion mining relies on the identification of aspects and opinion words pairs. For the performance evaluation of the pairs of aspects and opinion words extraction process, the extracted pairs of aspects and opinion words are compared with previously tagged pairs of aspects and opinion words. The experimental results for this extraction phase are shown in TABLE IV.

TABLE III. EXTRACTED ASPECTS RESULTS

Aspects		
ဧည့်ခန်း (lobby)	ပန်းဥယျာဉ် (garden)	စားဖိုမှူး (chef)
အခန်း (room)	စားပွဲ (table)	ကော်ဖီ (coffee)
အိပ်ခန်း (bed room)	မြင်ဆင်ပုံ (decoration)	လက်ဖက်ရည် (tea)
ရေကူးကန် (swimming pool)	ဒီဇိုင်း (design)	စားပွဲထိုး (waiter)
စားသောက်ခန်း (dinning room)	ရှုခင်း (scene)	ဈေးနှုန်း (price)
အထောက်အပံ့ပစ္စည်း (facilities)	အစားအစာ (food)	ဝန်ဆောင်မှု (service)
နားနေခန်း (lounge)	မြင်ကွင်း (view)	အင်တာနက်လိုင်း (internet connection)
ရေချိုးခန်း (bath room)	ဝန်ထမ်း (staff)	ဧည့်ကြိုပြောန (reception)

TABLE IV. ACCURACY FOR EXTRACTING PAIRS OF ASPECTS AND OPINION WORDS

Aspect Group	Accuracy	Precision	Recall
Hotel Review 1	0.90	0.96	0.93
Hotel Review 2	0.89	0.95	0.91
Hotel Review 3	0.92	0.97	0.94

The overall accuracy over three hotel reviews is 93%, precision is 92%, and recall is 93% respectively. According to the experimental results, the precision is slightly higher then the recall due to the less extra pairs of aspects and opinion words are generated by our system.

V. CONCLUSIONS

The system proposes the important task in the detailed aspect level opinion mining for Myanmar language. Our proposed work mainly focuses on extracting pairs of aspects and opinion words using syntactic rules according to the structure of Myanmar language. Finally, our system can effectively identify the relevant aspects for Hotel dation with the help of opinion word lists in the Myanmar opinion lexicon.

In the future, user review data will be expanded to evaluate on large data set. Furthermore, we will perform the aspect level opinion mining task and will also construct opinion lexicon for other domains. We will also focus on the calculation of the opinion orientation whether which aspects are positive, which are negative or neutral.

REFERENCES

[1] A. Rashid, S. Asif, N. A. Butt and I. Ashraf, "Feature level opinion mining of Educational student feedback data using sequential pattern mining and association rule mining", International Journal of Computer Application, Vol. 81, No.10, 2013.

[2] B. Singh, S. Kushwah and S. Das, “Multi-Feature segmentation and cluster based approach for product feature categorization”, in International Journal of Information Technology and Computer Science, 2016, 3, 33-42.

[3] C. H. Cho and N. Naw, “A syllable segmentation algorithm for Myanmar language”, International Journal of Science and Research, Vol. 8, Issue 3, March, 2019.

[4] E. Marrese-Taylor, J. D. Velasquez, F. Bravo-Marquez, and Y. Matsuo, “Identifying customer preferences about tourism products using an aspect-based opinion mining approach” , in Procedia Computer Science, vol. 22, pp. 182-191, 2013.

[5] L. Liu, Z. Lr and H. Wang, "Extract Product Features in Chinese Web for Opinion Mining”, in Journal of Software, Vol. 8, No. 3, March, 2013.

[6] M. Hu and B. Liu, "Mining and summarizing customer reviews", in proceeding of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.

[7] Piyush Arora, Akshat Bakliwal and Vasudeva Varma, “Hindi Subjective Lexicon Generation using WordNet Graph Traversal”, in the proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics(CICLing), India.

[8] Q. Liu, Z. Gao, B. Liu and Y. Zhang, “Automated Rule Selection for Aspect Extraction in Opinion Mining”, in the Proceeding of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI, 2015.

[9] S. H. Ghorashi, T. Ibrahim, S. Noekha and N. S. Dastjerdi, “A frequent pattern mining algorithm for feature extraction of customer reviews”, International Journal of Computer Science, Vol. 9, Issue 4, No 1, July 2012.

[10] S. H. Su and T. L. Khin, “Extracting product features and opinion words using pattern knowledge in customer reviews”, in the Scientific World Journal, 2013.

[11] T. A. Tana, Y. N. Cheah, "A two-fold rule-based model for aspect extraction", in Expert Systems with Application, 2017, 273-285.