Mining Advices from Weblogs

Alfan Farizki Wicaksono, Sung-Hyon Myaeng
Department of Computer Science
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
{alfan.farizki, myaeng}@kaist.ac.kr

ABSTRACT

Weblog, one of the fastest growing user generated contents, often contains key learnings gleaned from people's past experiences which are really worthy to be well presented to other people. One of the key learnings contained in weblogs is often vented in the form of advice. In this paper, we aim to provide a methodology to extract sentences that reveal advices on weblogs. We observed our data to discover the characteristics of advices contained in weblogs. Based on this observation, we define our task as a classification problem using various linguistic features. We show that our proposed method significantly outperforms the baseline. The presence or absence of imperative mood expression appears to be the most important feature in this task. It is also worth noting that the work presented in this paper is the first attempt on mining advices from English data.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Performance, Experimentation

Keywords

Text Mining, Advice Mining

1. INTRODUCTION

Previous survey has shown that the largest percentage of bloggers (37%) cited "*my life and personal experiences*" as the main topic [4]. As the manifestation of personal experiences, weblogs often contain explicit key learnings gleaned from people's past experiences which are really worthy to be well presented to other people. One can surely learn from past experiences that others have discovered so that one can get new information or perspective that one might not have discovered before. Advice is one of the key learnings that can be found in weblogs.

Specifically, in travel weblogs, key learnings from one's travel experiences are often vented in the form of travel advices. This kind of advice can range from guiding people in performing action at a particular tourism place, to specific recommendation concerning what people have to do. Moreover, advice can be also recommendation against a particular course of action [1]. Table 1 shows some examples of travel advices extracted from well-known travel weblogs. Storing those advice-revealing sentences is obviously very useful for future use.

Such travel advices are very important information among online travelers (i.e., those travelers who have internet connection). This information gives them perspective on where they should travel, what they should do, and what they should be aware of. According to the Travel Industry Association of America [8], 67% of online travelers in the United States search for information on destination via the Internet. Unfortunately, current search techniques are not designed for in-depth advice seeking. Top ranking pages from search engine may not contain the desired advices. Even if they do, people are still needed to read through the retrieved pages and capture manually which part of the pages contains desired advices. Even these difficulties also exist when people go directly to well-known travel websites (or weblogs) for seeking advices since advices are mostly mingled with other unrelated information in the text content. It would be very nice if these travel advices were all captured and subsequently indexed by well-organized user-friendly indices enabling people to find easily what kind of travel advice they would like to seek. Therefore, the need of system that can capture and collect advices automatically on the Web is very urgent.

To address those aforementioned problems, we introduce a new task so called advice mining. The goal of this task is to capture automatically advices on the Web and subsequently store them in the depository called *advice depository*. The stored advices are in well-organized user-friendly structured information comprising 6 slots as listed below.

- **Advice mention**: This slot refers to a sentence that reveals the advice.
- **Advice context**: In what situation, where, or when a person should follow the advice.
- **Advice type**: Type of advice, e.g., *advice for* and *advice against.*

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

- **Reason**: The reason why the author issues the advice. Knowing the reason is important since other people may have their own way to handle the issue addressed in the advice.
- **Associated experience mention**: List of experience-revealing sentences that give explanation about the extracted advices.
- **Link to resource**: Link to corresponding weblog entry (URL).

A full solution to the task of mining aforementioned complete advice representations is beyond the scope of this paper. We leave the full solution of advice mining task for future work. But, this paper addresses an important subtask so called *advice-revealing sentence extraction.* The examples of extracted sentences are shown in table 1. The results from this task are used to fill slot number 1 in the 6-slot structure mentioned before, i.e., advice mentions. In our research, we extract sentences which reveal advices from travel weblogs due to its usefulness and importance in travel domain. Although our data belongs to travel domain, our proposed method is completely domain-independent. We define the problem as a classification task using some linguistic features and opinionated lexical resource. A full solution obviously needs this task as the first step because advice context, advice type, reason, associated experience mention, and link to resource could only be extracted after knowing the advice mention.

Once those advice-revealing sentences have been successfully extracted, there are two other points as additional benefits of this task besides providing easiness for people in accessing travel advices. **(1)** Extracted advices can be used as prepared information for context aware application [3]. The rapid development of mobile devices such as smart phones and tablet PCs triggers many researches in developing such information recommendation system considering users' contexts. After the system detects the users' contexts, e.g., place and time, the system could provide advices for the users in accordance with current place and time. For instance, user **A** is going to have lunch at restaurant **B**. Subsequently, the recommender system may present automatically some advices regarding what kind of menu user **A** should eat, what kind of menu user **A** should avoid, or maybe what kind of action user **A** should avoid. Those returned advices are retrieved from extracted advice-revealing sentences depository. Those advices were written by other people who have experiences in having lunch at restaurant **B** before. **(2)** Extracted advices provide a big boost to the tourist destination marketers. After people visited a particular tourist destination, they usually write down some advices in accordance with their experience on a travel weblog. These advices are manifestation of their experiences being communicated about the strengths and weakness of the destination. Understanding individual travel experiences from extracted advices is clearly a cost-effective method for destination marketers to assess their service quality and improve travelers' overall experiences [5].

This paper is organized as follows. In section 2, we review the related work. In section 3, we mention our observation to discover the characteristics of advice. Following that, section 4 gives explanation about our proposed method. Section 5 evaluates the proposed method. Finally, section 6 concludes the paper and mentions some future works.

Table 1: Example of extracted advices in travel weblogs

1	A request or advice, **make sure you** carry sufficient cash and filled up your car/bike tank, since you don't find anything in the hill except a SBI ATM.
2	**Be careful with** your money and count the change returned. When paying the national boarding tax the lady didn't returned me the change.
3	**Avoid to** pay in full in advance, especially if you are not coming back to Lhasa. Many agencies will ask you to pay everything in advance (like ours), but some others ask you to pay only 70% in advance.

2. RELATED WORKS

There have been some attempts to harvest people's experiences from weblogs [7]. Sentiment analysis is also another task which is related to our task. The goal of sentiment analysis is to find opinionated sentences and then extract objects contained in those sentences as well as their associated features and sentiment [6]. Based on our observation, some advice-revealing sentences contain opinionated words as shown in sentence "*it is **nice** idea to bring medicine wherever you go*". In this sentence, the opinionated word *nice* is a good indicator showing that the sentence reveals an advice. Later, we show that existing approach for sentiment analysis is used for determining the value of one of our features.

As far as we know, the only one closest work was done by Kozawa et al. [3]. They also addressed the problem of extracting advice-revealing sentences. However, the final goal of their research is different from ours. Moreover, they assumed that the target language is Japanese language which is different from our target language, i.e., English. They defined the problem as classification task using some features which are not suited to be applied directly on English data. Furthermore, we implemented their proposed method with some adaptations. Finally, we show that our proposed method is significantly better than their method.

3. CHARACTERISTICS OF ADVICE

In our research, we define advice as *a sentence made by person, usually as a suggestion or a guide to action and/or conduct relayed in a particular context.* Furthermore, we also observed each sentence in our data, especially advice-revealing sentence, to capture the characteristic of advices in English data. Firstly, we found that advice-revealing sentences tend to contain specific vocabulary aid such as *make sure you, be careful with, be sure to, be aware, you have to, i suggest, i strongly recommend, advice,* etc. Table 1 shows the examples of vocabulary aid in bolded words. Secondly, advice-revealing sentences often express modality such as obligation, suggestions, necessity, and command. These modalities are expressed using following modal verbs: *could, might, must, shall, should, will, and would.* Thirdly, we found that advice-revealing sentences often contain imperative mood expression that urge the readers to act a certain way. In our data, this imperative mood is often expressed by using a simple unconjugated form of the verb without any subject attached. The third sentence in table 1 shows the example. Lastly, advice-revealing sentences often contain opinionated words as shown in sentence "*it is **nice** idea to bring medicine wherever you go*". In this sentence, the opinionated word *nice* is a good indicator showing that the sentence reveals an advice. One more thing is that most of sentences described in past tense are not advices.

4. PROPOSED METHOD

As mentioned previously, we defined the problem as binary classification task. We extract a set of features that can help us label each sentence in the corpus whether or not it reveals advice mention. These features are selected mainly based on characteristics of advice addressed in the previous section.

Feature A (**FA**) is set of clue expressions defined through investigating our data. Suppose, there are m clue expressions in the set. FA is essentially a collection of binary feature functions $\{fa_i(x)\}_{i=1}^{m} \in \{1, 0\}$. If a clue expression corresponded with fa_i appears in an instance x, then $fa_i(x) = 1$; otherwise $fa_i(x) = 0$. There are currently 54 clue expressions in the set.

Feature B (**FB**) represents proper-noun and modal verb contained in a sentence. The proper-noun should be directly attached to the modal verb as a nominal subject. To determine this kind of proper noun, we use dependency parser[1] to find proper nouns that have "nominal subject" (*nsubj*) relation with modal verbs found in the sentence. The combination between identified proper-noun and modal verb is the value of this feature.

Feature C (**FC**) consists of three elements: set of clue verbs found in the sentence, proper-noun attached to the clue verbs as nominal subject, and POS tags of those clue verbs. Formally, we define $\{fcv_i(x)\}_{i=1}^{m}$ as binary feature functions corresponded with each clue verb in the set, where m is the number of clue verbs in the set. The value of $fcv_i(x)$ is determined based on the presence of clue verb in an instance x. Furthermore, we associate fcv_i with additional features (i.e., aforementioned proper-noun and POS tags), if only $fcv_i(x) = 1$. To construct the set of clue verbs, firstly, we used some seed clue verbs found directly by investigating our data (e.g., *suggest, recommend, advice,* etc.). Finally, we expanded the set of clue verbs by adding verbs from **Framenet**[2] using those seed clue verbs. Verbs that have the same frame with seed clue verbs were considered.

Feature D (**FD**) is binary feature indicating whether or not a sentence contains imperative mood expression. We defined 2 heuristic methods to determine the value of this feature. The first heuristic method consists of some rules leveraging only each word's POS tag information (Penn Treebank tagset). The rationale of this method is that if the sentence contains verb which is not preceded by subject, then the sentence most likely contains imperative mood expression. The detail rules are as follow.

1. If first word's POS is either **VB** (verb, base form) or **VBP** (verb, non-3rd person singular present), then the sentence contains imperative mood expression.

2. If first word's POS is **RB** (adverb) and second word's POS is either **VB** or **VBP**, then the sentence contains imperative mood expression.

3. Starting from any word position whose POS is either **SYM** (symbol) or **punctuation mark**, apply once again rule *number 1* and rule *number 2.*

4. Other conditions: no imperative mood expression found in the sentence.

The second heuristic method can be easily explained using an algorithm. If the sentence contains at least one verb without any subject attached, then it most likely contains imperative expression. Given initial set of detected verbs in the sentence, the algorithm eliminates any verb contained in the list which acts as governor in following dependency relation: "nominal subject" (*nsubj*), "clausal subject" (*csubj*), "clausal passive subject" (*csubjpass*), "passive nominal subject" (*nsubjpass*), "auxiliary" (*aux*), "controlling subject" (*xsubj*). To handle negative imperative expression such as "*do not bring smartphone here*", the algorithm pays attention on each auxiliary (aux) relation detected in the sentence. If governor of auxiliary relation is attached to any subject, then verb acting as dependent in the auxiliary relation is also eliminated from the list. In the end, if the list is not empty, then the sentence contains imperative expression; otherwise it does not.

Feature E (**FE**) is binary feature indicating whether or not a sentence contains opinionated copula. A copula is the relation between the complement of a copular verb (*is, am, are*) and the copular verb itself. For example, the sentence "*it is* ***better*** *to bring umbrella*" contains copula relation where word *better* acts as governor and word *is* acts as dependent. Opinionated copula means that the governor of the relation has subjectivity score above a given threshold. We found some evidences on our data showing that advice-revealing sentences contain opinionated copula. Suppose, w is a particular word. Subjectivity score of the word w is determined by leveraging **SentiWordNet** [2], i.e., a well-known existing resource for sentiment analysis. It is also worth noting here that $0 \leqslant Subjectivity(w) \leqslant 1$. If $Subjectivity(w)$ is above a given threshold[3] value, then the sentence has opinionated copula; otherwise it does not.

5. EXPERIMENTS AND RESULTS

We evaluated our proposed method on data crawled from travelblog.org[4]. Our data consists of weblog entries which tell about people's travel experiences around the world. We selected 650 weblog entries using some clue words which indicate the presence of advices, such as *suggest, advice, recommend, tips,* etc. Thus, we followed some pre-processing steps including sentence splitting. Finally, we judged manually whether or not a particular sentence reveals advice based on our definition of advice. Currently, we have already labeled 207 weblog entries consisting 8,109 sentences. Table 2 describes our data in detail.

We also implemented method proposed by Kozawa et al. [3] for our baseline since they also addressed the task of advice-revealing sentence extraction. We had to do some modifications on their features since some of their features cannot be directly applied for English sentences. Their preconstructed resources such as set of clue expressions and evaluative expressions are in Japanese language. Instead, we implemented their features using our own set of clue expressions and opinionated words obtained from SentiWordNet. Their features are summarized in table 3.

We used SVM as a machine learning tool since it is known to be the state-of-the-art algorithm for classification prob-

[1] We use Stanford Dependency Parser

[2] http://framenet.icsi.berkeley.edu

[3] We use threshold of 0.69

[4] http://www.travelblog.org

Table 2: Detail of our data

	# advices	# non-advices
Training data	670	5,239
Testing data	200	2,000
Total of sentences	870	7,239

Table 3: Features proposed by Kozawa et al. [3] for English data

Feature	Description
FKA	First modal verb found in the sentence is any of the following: *could, might, must, shall, should, will, and would.*
FKB	Set of clue expressions defined through observation.
FKC	Frequency of opinionated words.
FKD	Through FKA, FKB, and FKC for the previous and next two sentences of the target sentence.

lem. Table 4 shows the performance of our baseline method. In [3], FKD is shown to be a good feature. However, this is not true in our experiment. This is most probably because we used a completely different set of expression clues and opinionated words. The best performance for baseline is achieved when we use FKA, FKB, and FKC at the same time, i.e., 0.309 in terms of F-score.

Table 5 shows the performance of our proposed method. FD_1 is implementation of FD using first heuristic, i.e., set of rules using word's POS information. Likewise, FD_2 is implementation of FD using second heuristic. Based on the results, FD_2 does not seem to be a good feature. But, when we use FD_1 and FD_2 as features at the same time, the performance is higher than individual feature's performance. The best performance is achieved when all features are included. In this case, the F-score value reaches 0.586, which means that our proposed method is significantly better than the baseline on English data. Although it is not so significant, opinionated copula (FE) also improves the overall performance.

To see the contribution of each feature, we also measured the performance by excluding each feature. As the results[5], FD (combination between FD_1 and FD_2) is the most important feature in this task. Excluding FD_1 and FD_2 in the series of features decreases the performance as much as 21% from the best performance in terms of F-score. On the other hand, FB does not seem to be a good discriminator since excluding FB almost does not change the overall performance. The performance difference is just 0.3% from the best performance in terms of F-score.

6. CONCLUSIONS AND FUTURE WORK

We have shown first attempt to extract advice-revealing sentences from English weblogs. We also observed our data to discover the characteristics of those advice-revealing sentences. Based on this observation, we defined our task as classification problem using various linguistic features. Our experiments showed that our proposed method significantly outperforms the baseline. The presence or absence of imperative mood expression appears to be the most important features in this task.

Table 4: Performance of our baseline method

Features	Prec.	Rec.	F1
FKA+FKB+FKC	**0.452**	**0.235**	**0.309**
FKA+FKB+FKC+FKD	0.279	0.285	0.282

Table 5: Performance of our proposed method

Features	Prec.	Rec.	F1
FA+FB+FC	0.462	0.310	0.371
FA+FB+FC+FD_1	0.514	0.655	0.576
FA+FB+FC+FD_2	0.425	0.370	0.396
FA+FB+FC+FD_1+FD_2	0.519	0.660	0.581
FA+FB+FC+FD_1+FD_2+FE	**0.523**	**0.665**	**0.586**

Despite the significant improvements, the results are still far short of the ideal situation yet. There are two future directions that we would like to do. First direction for future work would be to define better features so that advice-revealing sentences can be more accurately extracted. Second direction would be to work on other subtask on advice mining such as finding advice context.

7. ACKNOWLEDGEMENT

This research was supported by a Korean Government Scholarship Grant from the Ministry of Education, Science, and Technology (MEST), Republic of Korea.

8. REFERENCES

[1] Dalal, R. S. and Bonaccio, S. What types of advice do decision-makers prefer ? *Organizational Behavior and Human Decision Processes, 112, 11-23.*, 2010.

[2] Esuli, A. and Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. *LREC*, 2006.

[3] Kozawa, S., Okamoto, M., Nagano, S., Cho, K., Matsubara, S. Advice extraction from Web for providing prior information concerning outdoor activities. *4th International Symposium on Intelligent Interactive Multimedia Systems and Services, pp. 251-260*, 2011.

[4] Lenhart, A. and Fox, S. Bloggers: A portrait of the internet's new storytellers. *Pew Interent and American Life Project. Available at http://www.pewinterent.org.*, 2006.

[5] Pan, B., MacLaurin, R., Crotts, J.C. Travel blogs and the implications for destination marketing. *Journal of Travel Research vol. 46 no. 1 35-45.*, 2007.

[6] Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2(1-2), 1-135.*, 2008.

[7] Park, K., Jeong, Y., Myaeng, S. Detecting experiences from weblogs. In *Proceedings of the 48th Annual Meeting of the ACL, pages 1464-1472*, 2010.

[8] Travel Industry Association. Executive summaries - Travelers' use of the internet, 2004 edition. *Washington, DC: Author*, 2005.

[5] We do not show the results here due to lack of space