Event Mining from the Blogosphere Using Topic Words

Yoshihiko Suhara[†] suhara@ae.keio.ac.jp
Hiroyuki Toda[‡] toda.hiroyuki@lab.ntt.co.jp
Akito Sakurai[†,⋆] sakurai@ae.keio.ac.jp

[†]Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kouhoku, Yokohama, 223-8522, Japan
[‡]NTT Cyber Solutions Laboratories, NTT Corporation, 1-1 Hikarino-oka, Yokosuka, 239-0847, Japan
[⋆]CREST, Japan Science and Technology Agency, Kawaguchi-shi, 332-0012, Japan

Abstract

In this paper, we propose a method of extracting "action relations" between related topic words from Japanese weblogs (blogs). An action relation is a tuple of *agent*, *target* and *predicate*. Our method obtains blog articles that contain two keywords by AND search and outputs action relations constructed from the predicate and the two keywords with case particles that follow. However, since words are often omitted after they appear once, sentences with both of the keywords are scarce. Our method overcomes this problem by combining sentences that contain the common predicate and locate nearby.

Keywords

Information extraction, Relation extraction, Web mining

1. Introduction

The need for a blog search service has been growing with the recent popularization of blogging. Blogs contain a lot of useful information such as bloggersf tastes and interests. This is why there has been a variety of efforts to extract information from blog articles. For example, Technorati[1] lists most frequently searched keywords and tags. However, a single keyword could not provide sufficient information about a topical event - an occasion or an action occurring in the real world described in blogs. To let users presume a certain event, BLOGRANGER[2] [1] and kizasi.jp[3] extract topics words from blog articles and group them to display as related keywords. However it is difficult for users, unless they have prior knowledge about the event, to presume the event only from the presented keyword group. For instance, For instance, when the words "Israel" and "Lebanon" are presented as related topic words, users without prior knowledge about complicated political situations between Israel and Lebanon cannot realize what actually happened.

We have considered that an event could be presumed by users more easily and adequately by presenting the relations between related keywords. The relation between keywords may include action, affiliation, mission, location and social relation. From these relations, we focus on the action relation as we consider it describes an event best. The purpose of this research, therefore, is to extract the action relation between related topic words from blogs. In this paper, we define "event mining" as the extraction of event contents as simple expressions.

To extract action relations, we have focused attention on predicates and their case elements: nouns followed by a case particle. Since a case particle decides the case of the case element in Japanese, we can estimate the semantic role of the case element by its case particles. Although one case particle may have several meanings depending on its predicate and the context, we regard *ga* case as agent of the action, *ni* and *wo* case as *target* of the action, and the predicate as an action itself. Then we can extract the "action relation" between two keywords from the sentence in which these two keywords are co-occurred as case elements.

However, a word which has appeared in a preceding sentence is often omitted or replaced with a demonstrative word. Although it is necessary to complement the case element properly using anaphora resolution, anaphora resolution is not yet fully developed for practical use [2]. Therefore, we consider the simple way to solve this problem to extract action relations by using massive amount of blog articles.

2. Proposed method

We propose a method of extracting action relations by combining case elements obtained by analyzing sentences having at least one of the specified keywords as case elements. When one of the keywords appears as a case element in a sentence, the keyword, the case particle, and the corresponding predicate are extracted as a "predicate pattern". The proposed method forms action relations by combining predicate patters having identical predicates but possibly different keywords and different case particles. The pattern thus formed is referred to as an action pattern.

The semantic role in passive sentences is different from that in active sentences. Generally speaking, where a sentence is passive, the semantic role of the *ga* case and *ni* case are interchanged. Therefore, the *ga* case represents the target of the action and the *ni* case represents an agent. Where a juncture of a verb in imperfective form and the suffix *reru* is extracted, the algorithm decides the sentence as the passive voice. Then the ga case and ni case in the sentence are interchanged and the predicate becomes in active voice. In this paper, we prepare two types of relation extraction methods, which use / remove the sentences judged as passive voice by using this method. To distinguish these methods, we call the method that exploits passive voice "converted" and call the method that removes passive voice as "removed".

The proposed method consists of the following four steps.

[1] http://www.technorati.com/
[2] http://ranger.labs.goo.ne.jp/ (Japanese)
[3] http://kizasi.jp/ (Japanese)

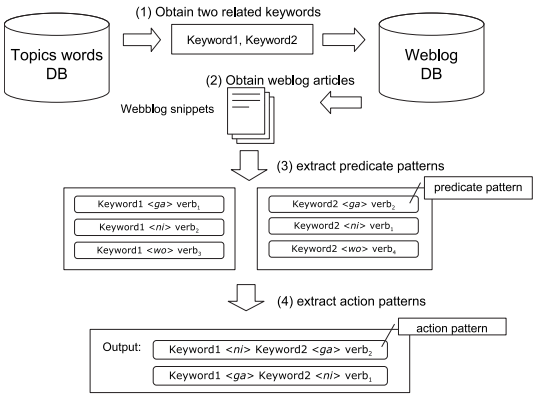

Fig. 1: *An overview of the proposed method*

1. Obtain two related keywords,
2. Obtain blog articles,
3. Extract predicate patterns, and
4. Extract action patterns.

Firstly, two associated topic words are obtained from the web page "The latest hot keywords" in BLOGRANGER[4]. The algorithm of keyword association and topic words extraction are adopted by BLOGRANGER [3][4]. Secondly, we exploit blog search engine to obtain blog articles for the relation extraction. An AND search is conducted to obtain a set of documents with a specific pair of keywords. In the third process, the blog articles obtained are divided into sentences with punctuation marks as separating characters. For each sentence, dependency structure analysis is conducted using the analyzer CaboCha[5]. Since Japanese is agglutinative, we have to first divide a sentence into words and then assign POS tags. CaboCha is a tool not only for morphological analysis but also for dependency analysis. In the last process, from the predicate patterns extracted with two keywords, the patterns with identical predicates are selected and combined to form action patterns. When the same case particle appears in the sentences, the ni and wo cases being considered the same, the patterns are not combined. The action patterns are sorted by their score and then are output. The sum of the predicate pattern frequencies is used as a score.

3. Evaluation

We evaluated the proposed method by comparing it with a baseline method, which selects sentences having two keywords o-occurred within one sentence and forms an action pattern with the keywords and their corresponding predicate. BLOGRANGER API[6] was used to search blog articles. An AND search was conducted for the 17 pairs of keywords, and 500 snippets having n words in length (n = 50 and 500) were obtained. The snippet with n words in length means a text string with n words that contains the relevant keywords. We use three measures: Correctness of the Highest Ranked pattern (CHR), Mean Reciprocal Rank (MRR) [5] and Discounted Cumulative Gain (DCG) [6].

The results are shown in Fig.2. Roughly speaking, the proposed method has attained significantly higher values than

[4] http://ranger.labs.goo.ne.jp/br1/
[5] http://chasen.org/ taku/software/cabocha/
[6] http://ranger.labs.goo.ne.jp/hacks/

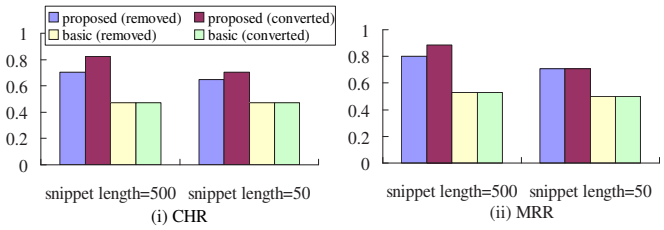

Fig. 2: *Evaluation results: CHR (left) and MRR (right)*

baseline method in the three measures. Specifically the proposed method (removed and converted) is statistically significantly better than the baseline method in cases when snippet length is 500 and 50 (t-test, $p < 0.05$) in CHR and MRR (Fig.2). Proposed method (converted) is statistically significantly better than proposed method (removed) (t-test, $p < 0.05$) in CHR (Fig.2 left). In other measures, there are no statistically significance difference between "converted" method and "removed" method.

A reason why the correctness for 500-word length snippets is better than that for 50-word cases might be:

1. The longer the snippets are, the more frequently correct keyword-predicate pairs appear.
2. The shorter the snippets are, the larger portion of the sentences is fragmented into non-sentences by the boundary caused by the length restriction.

4. Conclusion

From the results of the evaluation experiments, it has been confirmed that the proposed method extracts action relations with higher accuracy and in larger number than the baseline method which relies on co-occurrence of the two keywords in a single sentence.

The proposed method is still very primitive in that it does not utilize syntactic structure or dependency structure, which would help us to improve the method. Since blogosphere is still expanding, our method or improved one will be beneficial to many of bloggers.

References

[1] Fujimura, K., Toda, H., Inoue, T., Hiroshima, N., Kataoka, R. and Sugizaki, M. BLOGRANGER - A Multi-faceted Blog Search Engine. In *Proceedings of the WWW 2006 3nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (2006).

[2] Iida, R., Inui, K. and Matsumoto, Y. Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution. *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Lin-guistics*, pp. 625-632. (2006).

[3] Fujimura, K., Inoue, T. and Sugisaki, M. The EigenRumor Alogorithm for Ranking Blogs. In *Proceedings of the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (2005).

[4] Toda, H. and Kataoka, R. In *Proceedings of the 7th Annual ACM international Workshop on Web information and Data Management*, pp.81-86. (2005).

[5] Voorhees, E. The TREC-8 Question Answering Track Report. In *Proceedings of TREC-8, NIST Special Publication 500-246*, pp.77-82. (1999).

[6] Järvelin, K. and Kekäläinen, J. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Confer-ence*, pp.41-48. (2000).