Automatic Acquisition of Script Knowledge from a Text Collection

Toshiaki Fujiki
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuta-cho, Midori-ku, Yokohama, JAPAN
fujiki@lr.pi.titech.ac.jp

Hidetsugu Nanba
Graduate School of Information Sciences
Hiroshima City University
3-4-1 Otsukahigashi, Asaminami-ku, Hiroshima, JAPAN
nanba@its.hiroshima-cu.ac.jp

Manabu Okumura
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuta-cho, Midori-ku, Yokohama, JAPAN
oku@pi.titech.ac.jp

Abstract

In this paper, we describe a method for automatic acquisition of script knowledge from a Japanese text collection. Script knowledge represents a typical sequence of actions that occur in a particular situation. We extracted sequences (pairs) of actions occurring in time order from a Japanese text collection and then chose those that were typical of certain situations by ranking these sequences (pairs) in terms of the frequency of their occurrence. To extract sequences of actions occurring in time order, we constructed a text collection in which texts describing facts relating to a similar situation were clustered together and arranged in time order.

We also describe a preliminary experiment with our acquisition system and discuss the results.

1 Introduction

Script is a term proposed by Schank, and it refers to a form of knowledge representation. Script knowledge is a body of knowledge that describes a typical sequence of actions people do in a particular situation(Schank and Abelson, 1977). For example, when we go to a restaurant, we usually 'enter the restaurant', 'wait', 'sit down', 'get the menu and decide what to eat', 'order the dish', 'wait until the dish has come', and so on. This sequence can be said to be script knowledge in the situation of 'eating at a restaurant'.

Script knowledge has been used in natural language processing, especially for word sense disambiguation, text generation, and automatic text summarization(Dejong, 1982). However, most studies have used only small portions of script knowledge manually generated by the authors. We need a large-scale knowledge database; however, manually producing such a database would cost too much.

In this paper, we propose a method for automatic acquisition of script knowledge from a Japanese text collection. Because script knowledge represents a typical sequence of actions formed in a particular situation, we extracted sequences (pairs) of actions that occur in time order. We then chose among these actions the ones that are typical by ranking them in terms of the frequency of their occurrence. To extract sequences of actions that occur in time order, we constructed a text collection in which texts describing facts relating to a similar situation were clustered together and arranged in time order.

In Section 2, we describe our proposed method and show how we constructed the text collection. In Section 3, we describe a preliminary experiment with our acquisition system and discuss the results.

2 Proposed Method

Our method consists of the following three steps:

1. Constructing a text collection.

2. Extracting sequences (pairs) of actions from the text collection.

3. Selecting typical sequences.

We show the outline of our method in Figure 1, where the process of automatic acquisition of script knowledge related to 'murder case' is described. In the following subsections we explain these steps in greater detail.

2.1 Constructing a Text Collection

To use our method, we need to construct a text collection that has the following three features: The texts in the collection describe only facts. The texts are arranged in time order. The texts are on a similar topic.

To construct a text collection that satisfies the above conditions, we used a corpus of newspaper articles and performed the following steps: First, we retrieved news reports on a topic, and clustered them into subtopics. In this step, we used an automatic text clustering system(IPA, 2002). Next, we extracted only the first paragraph from each report, and arranged the paragraphs in clusters based on the date of issue of the report. We used only the first paragraphs of the news reports because they tend to describe facts in time order.

After that, the sentences in the text collection were syntactically analyzed by using a Japanese syntactic analyzer, KNP(Kurohashi and Nagao, 1994).

2.2 Extracting Pairs of Actions

In this section, we describe three cases where two actions occur one after the other and can be extracted as a pair of actions. Let us first explain what we mean by 'action', 'pair of actions', and 'sequence of actions' in this paper. In this work, an action is defined as a tuple of a transitive verb, its subject, and its object. We use the Japanese postpositional particles 'が' and 'は' to detect subjects, and 'を' to detect objects. A 'pair of actions' consists of two actions that occur in time order. A 'sequence of actions' can be defined as a transitive closure of all the pairs of actions.

1. Cases where verbs in different sentences have the same subject and object

   When two verbs in different sentences in a cluster have the same subject and object, a pair of actions can be extracted. For example, consider the following two sentences.

   > 警察は容疑者を発見した。
   > (The police found the suspect.)
   > 警察は容疑者を逮捕した。
   > (The police arrested the suspect.)

   Two verbs('発見 (find)' and '逮捕 (arrest)') have the same subject ('警察 (police)') and object ('容疑者 (suspect)'). In this case, the latter action (sentence) occurs after the former. Therefore, a pair of actions (('警察が容疑者を発見する (police finds suspect)'), ('警察が容疑者を逮捕する (police arrests suspect)')) can be extracted.

   Note that in this case we take advantage of the fact that the sentences in the cluster are arranged in time order.

2. Cases where two verbs describe a continuous modification of the object in a sentence

   > 警察は容疑者を発見し、逮捕した。
   > (The police found the suspect, and arrested him.)

   In the above example, the verb '発見する (find)' and the verb '逮捕する (arrest)' describe a continuous modification relation. When two verbs have this relationship, they tend to be in time order. Therefore, in this case a pair of actions can be extracted.

3. Cases where the main verb and the verb of the relative clause have the same noun as the object
   In these cases, the verb in the relative clause should be in the past tense (auxiliary verb 'た' should be attached to the verb).

   > 警察は発見した容疑者を逮捕した。
   > (The police arrested the suspect whom they had found.)

   In the above example, the verb '発見する (find)' modifies the noun '容疑者 (suspect)', and '容疑者 (suspect)' is the object of '逮捕

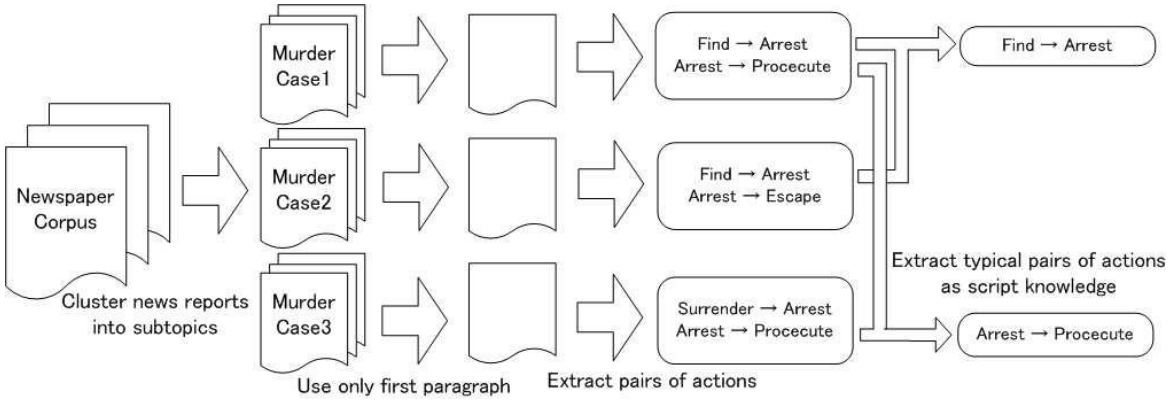

Figure 1: Outline of Our Method

する (arrest)'. In such a case, two actions can be thought to occur in time order. Therefore, a pair of actions can be extracted.

The verb in the relative clause must be in the past tense, because the action in the relative clause does not necessarily occur before the action in the main clause when the verb in the relative clause is in the present tense. Consider, for example, the following two sentences:

彼は信頼する医者を訪ねた。
(He visited the doctor whom he trusted.)
彼は訪ねる医者を決めた。
(He decided on the doctor to visit.)

The verbs('信頼する (trust)' and '訪ねる (visit)') in the relative clauses of both the sentences are in the present tense, and the actions in the relative clauses do not occur before the actions in the main clauses. Therefore, we extract a pair of actions only when the verb in the relative clause is in the past tense.

In Japanese, the subject and object of a verb in a sentence may be omitted. In such cases, to extract a pair of actions, we naively supplemented them by using the noun phrase that appeared before the verb and satisfied the selectional restriction of the verb.

Passive sentences also cause a problem, since passivization often changes the case of the verb. Therefore, we generated patterns where the case of the verb changes to obtain passive cases from the active ones.

2.3 Selecting Typical Pairs

At this step, we selected typical pairs of actions from the extracted pairs. First, we generalized the extracted pairs by changing subjects and objects into semantic features and merging similar verbs into one by using Japanese thesaurus 'Bunrui Goi Hyo'(NLRI, 1964). As a result of this generalization, we could easily determine whether two pairs are same and count the frequency of occurrence.

Next, pairs of actions were assigned a score based on the frequency of occurrence. Pairs with a score exceeding predetermined threshold value were considered typical. Typical sequences of actions were then constructed as a transitive closure of all the selected typical pairs and, acquired as script knowledge.

3 Preliminary Experiment

We conducted a preliminary experiment with our system for automatic acquisition of script knowledge and investigated the effectiveness of our method. We used issues of Nihon Keizai Shimbun for the past 11 years (1990-2000) as a newspaper corpus and GETA(IPA, 2002) for automatic text clustering. In the case of script knowledge related to 'murder case', using the keyword 'murder case', we collected 4489 news reports, and these were clustered into 617 clusters.

As a result, 41 pairs of actions were extracted (the threshold was set to 2). Figure 2 shows part of the acquired script knowledge. In the figure, the time order between the actions is indicated by the arrows. For example, '[organization] arrests [human]' follows '[organization] finds [human]'.

Unfortunately, the acquired script knowledge does contain some errors. For example, the pair of actions '[organization] arrests [human]' follows '[organization] re-arrests [human]' in Figure 2 (the arrow is marked with a cross) is not in the right time order. This error is considered to be caused by generalizing different suspects to the same semantic feature [human].

Some other errors included those made in the syntactic analysis, wrong results of text clustering(these are the errors of the tools used), errors in supplementing omitted subjects and objects, and errors resulting from incorrect interpretation of passive sentences (these are the errors originating in our naive methods of analyzing Japanese sentences).

4 Conclusion

In this paper, we proposed a method for automatic acquisition of script knowledge from a Japanese text collection. We described a preliminary experiment with our acquisition system and discussed the results.

In the future, we want to devise a better method for dealing with passive sentences and supplementing omitted subjects and objects. We also plan to objectively (extrinsically) evaluate our system for other tasks such as automatic text summarization.

We think our method can work with other languages, though there must be some modification on syntactic analysis and definition of 'action'. We think script knowledge and structure of newspaper articles are language independent.

Our method of script knowledge acquisition has a few limitations. First, the method can acquire only the script knowledge with common subjects and/or objects. This limitation comes from our restrictions in extracting pairs of actions. If we don't impose these restrictions, however, much erroneous script knowledge might be obtained.

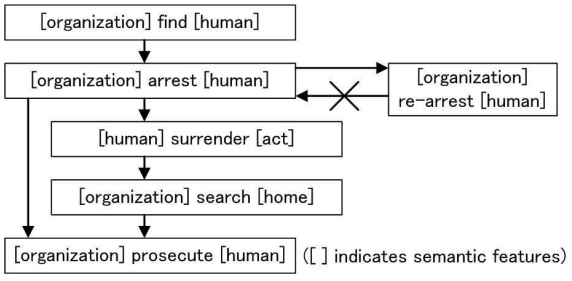

Figure 2: Result of the Experiment

Second, since our method is based on the characteristics of the text collection we construct (news reports in time order clustered into similar subtopics), it cannot correctly acquire script knowledge when the time order of the reports is not the same as the time order of the actions, as is the case, for example, with reports about kidnappings (news reports about kidnappings naturally come after the event). Third, because we used only the first paragraphs of the news reports, script knowledge could not be obtained from the background information in later paragraphs.

To cope with the third problem, we need to use whole news reports and analyze their rhetorical structure(Marcu, 2000) to clarify the time order among the sentences. We are now building such a rhetorical structure analyzer for Japanese news articles, and will report the results of our work in the future.

References

Gerald F. DeJong. 1982. *An overview of the FRUMP system.* Wendy G. Lehnert and Martin H. Ringle, editors, Strategies for Natural Language Processing, pages 149–176. Lawrence Erlbaum Associates.

Koichi Hori, Tadao Saito, and Hiroshi Inose. 1983. *Inductive Learning of a Script From the Text(in Japanese).* Information Processing Society of Japan, Special Interest Group on Natural Language Processing Research Report 37-5

IPA(Information-technology Promotion Agency, Japan). 2002. *Generic Engine for Transposable Association: GETA.* http://geta.ex.nii.ac.jp/

Sadao Kurohashi and Makoto Nagao. 1994. *KN Parser: Japanese Dependency/Case Structure Analyzer.* In Proceedings of The International Workshop on Sharable Natural Language Resources, Nara, Japan, pages 48–55.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* The MIT Press.

National Language Research Institute, editor. 1964. *Bunrui Goi Hyo(in Japanese).* Shuei Shuppan.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: an Inquiry into Human Knowledge Structures.* Lawrence Erlbaum Associates.