Extraction of Relations among Characters in Literary Texts

Hiromasa Nishihara (1310054)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 12, 2015

Keywords: Information Extraction, Relation Extraction, Literary Text, Acquisition of Pattern.

Many people have a lot of fun for reading novels. Their reading is often suspended if they have a time only when they take a train to work or have lunch break in an office. When they resume reading, they sometimes forget the story of the novel and have to read it again a little. Showing a summary of information about characters in the novel would help readers to remind the story and restart their reading smoothly. To support the readers to understand the contents of the novel, this research proposes a method to automatically extract relations among characters in literary texts. The examples of character relations are family such as "John is a brother of James." and friendship such as "Michael and Robert are friends." To implement and evaluate our proposed method, we use a collection of novels excerpted from the website 'Aozora Bunko' which collects a large amount of copyright expired Japanese novels.

The task of this research can be regarded as a kind of relation extraction task. Many previous studies aimed at extraction of various relations; not only relations between people but also general ones such as causal relation and 'is-a' relation. Some researches focused on extracting social relations from the literary text. They tried to just quantify a degree of relationships between the characters or extract only limited types of the relations. On the other hand, the goal of this research is to extract many kinds of relations such as family and friends from the novel. Although a few methods to automatically obtain patterns for extraction of the general relations have been already proposed, they cannot be simply applied for our task. Therefore, we proposed an original method to semi-automatically obtain the patterns to extract the character relations.

The outline of the proposed method is as follows. First, we manually construct a dictionary called 'Relation Word Dictionary' that compiles relation words. In this paper, 'relation word' is defined as a word that expresses a relation between people. We also construct a set of patterns that can extract the relations between the characters. After the above preparation, preprocessing such as morphological analysis is performed to the text in the novel. A list of the characters is also built by identifying the characters in the novel in this preprocessing. In addition, zero anaphora resolution is performed to compensate ellipses in Japanese sentences. By referring the character list and Relation Word Dictionary, sentences that may express the relations between the characters (called 'character relation sentence' hereafter) are extracted. Then, the character relations are extracted by pattern matching. In this study, the character relation to be extracted is either pair relation such as 'John - father' or triplet relation such as 'John - sister - Mary'. Finally, a diagram of the relations among the characters is constructed by concatenating the extracted relations.

In the preprocessing, the public tool MeCab and CaboCha are used for morphological analysis and dependency parsing, respectively. The characters in the novel are identified based on a named entity extraction module of CaboCha, semantic classes in a thesaurus and selectional restriction of case frames. Both semantic classes and case frames are provided by the lexical resource 'Nihongo Goi Taikei' (NGT hereafter). In addition, 'character stop words' are used for preventing from identifying wrong words as the characters. They are words that do not express a meaning of a person, and manually collected by checking the frequently appearing characters in 500 novels. Since ellipses are often found in Japanese texts, we implement zero anaphora resolution based on Nariyama's work with some modifications. In our zero anaphora resolution method, omitted cases in the sentences are detected by case frame analysis, and ellipses are filled by several rules. Relation Word Dictionary is manually developed by referring the two thesauri, ‘Kadokawa Ruigo Shin-Jiten’ and NGT. Two dictionaries are built: Relation Word Dictionary A that compiles only the unambiguous relation words, and Relation Word Dictionary B that compiles both the ambiguous and unambiguous relation words. Here, ‘ambiguous relation word’ stands for a word that has meanings of a relation and not a relation. We propose two approaches to construct a set of patterns for extracting the character relations: one is manual development and the other is semi-automatic development. In the former approach, 8 patterns are made by examination of 10 novels. Each is a string-based pattern that includes characters, relation word, postpositions and so on. In the latter approach, the patterns are designed to extract the characters and relation word from ‘bunsetsu’, a chunk of a basic phrase in Japanese, by checking the postposition appearing after the character or relation word as well as dependencies between two characters or one character and one relation word. In addition, each pattern is associated with one relation word. At first, 2 templates of patterns for extraction of the pair relation and 12 templates for the triplet relation are prepared. They are applied to the 500 novels to acquire candidates of patterns. Furthermore, more candidates of patterns are produced by generalization of the obtained patterns. Then reliability of each candidate of the pattern is estimated. It is defined as the ratio of the number of the sentences where the correct relation is extracted by the pattern to the total number of the sentences matched by the pattern. The reliability is manually calculated by using the 500 novels. Finally, only candidates whose reliability is higher than a threshold are selected as the patterns. After extracting the character relations by pattern matching, the diagram of the relations among the characters is constructed. A node represents the character and an edge represents the relation. The edge of the triplet relation is labeled with the relation word.

The proposed methods were evaluated by the experiment. Ten novels were used as a test data. Precision, recall and F-measure of extraction of the character relations from the test data were measured. The same relations in different notations (e.g. ‘David - Father’ and ‘David - Dad’) were manually merged into a single relation in both the gold standard and the output of the proposed system. Comparing the system with and without zero anaphora resolution, the performance of the latter was slightly better. This might be because the performance of zero anaphora resolution was poor. When the manually constructed patterns were used, the system using Relation Word Dictionary A was better than that using B. The best F-measure was 0.30. On the other hand, when the semi-automatically constructed patterns were used, the system using Relation Word Dictionary B was better. Comparing the various thresholds for the reliability of the patterns, it was found that the best threshold was 0.6. F-measure of the system with the best threshold was 0.34, which was slightly better than the method using manually constructed patterns. Both proposed methods outperformed the baseline where all possible relations between the characters from 'character relation sentences' were extracted without pattern matching. We conducted an error analysis on the proposed methods. The major causes of false-positive errors were failure of identification of the characters in the novel and the use of the inappropriate patterns. On the other hand, the major causes of false-negative errors were failure to retrieve the characters and the lack of the extraction patterns. Although the performance of the proposed method is not high enough and there is much room to improve it, one of the contribution of the paper is to show several possible solutions via our error analysis. Finally, a diagram of the character relations was constructed and evaluated. It was far from the ideal one.

In future, the performance of the zero anaphora resolution should be improved. The method to fully automatically construct the character relation extraction patterns should be investigated, too.