Automatic Acquisition of Noun and Classifier Agreement from a Corpus.

Shuhei Yano (0710072)

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

February 5, 2009

Keywords: Noun, Classifier, Pattern mining, Knowledge acquisition, Japanese.

In Japanese, classifier is generally used to count things. Numbers of kinds of classifiers are rich. Moreover, there exists agreement between nouns and classifiers that is, a noun is used with only specific classifiers. In generation or analysis of Japanese, knowledge about agreement between nouns and classifiers is important in order to use correct classifier. For example, this knowledge can be used to improve accuracy of word sense disambiguation. In this research, we collect a large amount of pairs of nouns and their corresponding classifiers from a corpus. It enables us to construct a lexicon of nouns including their agreeing classifiers.

There are some related works about this problem. Instead of assigning agreeing classifier for each noun, Bond et al. proposed a method to assign classifiers to semantic classes of nouns. Sornlertlamvanich's research is similar to ours but for Thai. He proposed a method to acquire agreement pairs of noun and classifier from a corpus. In that research, he used patterns which are created manually. Our approach is different from these previous work in the two aspects. The first one is that we collect comprehensively concord relationships between noun and classifier from a corpus, because nouns in the same semantic class does not always agree with the same classifier. The second one is that we learn patterns to extract pairs of nouns and classifiers based on pattern mining methods. Nouns and their

corresponding classifieres occur in various patterns, so it is very difficult to design patterns manually.

For preliminary investigation, at first, we use a simple pattern matching to try extracting pairs (n, c) from the corpus. Here n and c is a noun and classifier respectively, while (n, c) is a pairs of noun n and classifier c under agreement. We created patterns to extract nouns and classifiers by matching the following word sequence: $[noun + number + classifier]$, $[number + classifier + (no) + noun]$, $[noun + (ga) + number + classifier]$. It is assumed that these three patterns are typical to discover nouns and classifiers. However, there are many errors in (n, c) acquired by these patterns. Even when applying the best pattern, accurracy was just 54%.

Based on the above investigation, we proposed a method to extract automatically patterns (n, c) by pattern mining technique. Our method is divided into three steps: search sentences, extracting patterns, acquiring (n, c). Repeating these steps, (n, c) are acquired iteratively and then are added to the NC-DB database. For creating initial NC-DB, we utilize seeds which are small numbers of correct pairs of (n, c) used to learn patterns to extract (n, c). There are some approaches to create seeds but at present they are created manually. In order to extract precisely (n, c), we perform the following pre-processing to a corpus: (1) consider a sequence of numbers as a single word, (2) exclude classifiers which is used as unit, (3) remove uncountable nouns like abstract nouns and consider compound nouns as one word.

After the above pre-processing, we search all sentences which contain both noun n and classifier c in NC-DB. Next, we search word sequence which often occur in extracted sentences and then acquire patterns for extracting (n, c). When creating patterns, we investigate two methods, (1) a sequence of word in written form is extracted as a pattern, (2) a sequence of word in base form is extracted. Each candidate pattern created by these two methods is evaluated with three criteria to choose good patterns. The first one is number of sample sentences matched with patterns. The second one is the rate of correct (n, c) to the all extracted (n, c) pairs. In this case, there are two methods to define "correct (n, c)": (A) only (n, c) in seed are correct, (B) all (n, c) pairs extracted before are correct. The third one is the proportion of most frequent (n, c). Finally, using acquired patterns,

we extract new (n, c) from the corpus and add them to NC-DB.

The proposed method is empirically evaluated, Nikkei Newspaper for 1996 is used as a corpus. I proposed the method (1) and (2) for creating patterns, and the method (A) and (B) for selection of patterns. Therefore, I tried acquisition of (n, c) with four methods. Because a new extraction pattern was not acquired when I repeated three times of operation of the "example sentence search", "pattern acquisition", "extraction of (n, c)" in method (1)(A), I finished processing. There were 1,845 pairs of the (n, c), and the estimated number of correct (n, c) was 1,482, the accuracy was about 80%. I repeated two iteration steps in the method(1)(B). There were 1,721 pairs of (n, c), and the estimated number of correct (n, c) was 1,454, the accuracy was about 84%. The method (2)(A) was finished after three iteration steps. There were 1,817 pairs of (n, c), and the estimated number of correct (n, c) was 1,412, the accuracy was about 78%. The method (2)(B) was finished after two iteration steps. There were 1,704 pairs of (n, c), and the estimated number of correct (n, c) was 1,437, the accuracy was about 84%. When I checked (n, c) acquired by the proposed method, many pairs of nouns and classifiers have been acquired for nouns which occur in newspaper articles. Therefore, it would be possible to I acquire the agreement between nouns and classifiers for the technical terms from domain corpora. Furthermore, when I checked the acquired pattern, I found that precise patterns that might be difficult to develop manually are obtained.