An Improvement of Coordinate Structure Analysis for Legal Texts

Hiroki Matuyama (1010062)

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

February 6, 2012

Keywords: Legal engineering, Coordinate structure, Natural language analysis.

The coordinate structure analysis in natural language processing is one of difficult problems. The reason is that language features are different among document domains. Kurohashi and Nagao proposed a method to analyze coordinate structures using Dynamic Programming matching method. In their research, as they assumed that phrases in a coordinate structure are similar each other, they calculated the similarity between a forward phrase and backward phrase in a coordinate structure for analysis of coordinate structures. In fact, this method is implemented as a parsing tool called "KNP". However, when we analyze legal text with KNP, we find many errors for identification of coordinate structures. That is because legal texts have a different language features from others. In this thesis, we aim at improving performance of detection and identification of ranges of coordinate structures for legal texts. That is, we propose a new method of analyzing coordinate structures specialized for legal domain.

In this research, we define the coordinate structure as a sequence of one or more forward phrases, one coordinate key and one backward phrase. The coordinate key means a word connecting phrases in the coordinate structure. The forward phrase stands for a phrase which appears before the coordinate key, while the backward phrase stands for a phrase which appears after the key. They are under coordination in a coordinate structure. The

flow of procedures of the proposed coordination structure analysis is as follows. First, coordinate keys are identified in a sentence. In this research, we define eight coordinate keys;"mataha","oyobi","moshikuha","narabini", "to","ya","katu" and "sonota". Next, a head word of the forward phrase is identified. The head word is basically a word just before the key. Next, a set of possible candidates of the backward phrase is extracted. The start position of a backward phrase is basically a position just after the key. While the end position of the backward phrase is searched backward from the coordinate key. If a part of speech of the head of the forward phrase is a particle, particles are picked up as end positions of the backward phrase. If it is a verb, verbs are picked up as end positions. In the case of a noun, end positions are positions of words that are the three most similar words with the head of the forward phrase among all the last content words in a bunsetu phrase. However, the search of the end positions will be stopped when a comma, a period or other coordinate key is encountered. Next, a set of possible candidates of the forward phrase, also represented as pairs of a start and end position, is extracted. The end position of the forward phrase is basically a position of the head. To detect start positions, initial words in a bunsetu phrase are searched forward from the coordinate key. However, the search will be stopped when a comma, the beginning of a sentence or other coordinate key is encountered. Next, similarities between obtained candidates of the backward and forward phrases are calculated. Assuming that the phrases in the coordination are similar each other, a pair of phrases whose similarity is the highest among all combinations is chosen as the backward and forward phrase. The phrase similarity is defined based on alignment of words in two phrases. More similar aligned words are, more greater the phrase similarity becomes. And, when there is a word which is not aligned to any words, the similarity becomes low as a penalty. Furthermore, unaligned word appearing near the head of the phrase is more penalized. In addition, "dai","jô","kô" and "gô" are special words to denote identification numbers of statements in legal texts. So, these words are only allowed to be aligned to the same words. Finally, it is checked if another forward phrases before the forward phrase exist. If there is, a set of possible candidates of the next forward phrases is extracted in the same way. Then, the similarity between each candidate and

the already identified forward and backward phrases is calculated, and one with the highest score is chosen as the next forward phrase. This procedure is repeated until no more forward phrase is found.

Next, a method of analyzing hierarchical coordinate structures will be described. In this research, it is necessity to analyze inner coordinate structure first. That is, the system detects from inner coordination structure to outer. In other words, analyzing coordinate structure in a bottom up manner. First, the coordinate structures connected by the coordinate key "mataha" and "oyobi" are detected, second "moshikuha" and "narabini", and finally "to", "ya", "katu" and "sonota". This order of the coordinate keys follows convention of legal texts. It is said that "mataha" and "oyobi" should be used for inner coordinate structure, while "moshikuha" and "narabini" should be used for outer ones. When analyzing hierarchical coordinate structures, if inner coordinate structure is included in a forward or backward phrase of outer coordinate structure, there is a problem that the similarity between the forward and backward phrases is not able to be estimated appropriately, because lengths of two phrases tend to be different greatly. Therefore, when the forward or backward phrase or both in outer coordinate structure includes inner coordinate structures, we replaced them with the inner backward phrase when we calculate similarity between phrases. This is because we consider the balance between forward and backward phrases in outer coordination structure.

Based on the proposed method, a system to analyze the coordinate structure, which can handle three conjuncts and hierarchical coordinate structure, is implemented. 300 sentences in legal texts, 200 sentences as a development data and 100 sentences as an evaluation data, are analyzed using this system and the detected coordinate structures are evaluated. As a result, the F-measure of the detected coordinate structures is 50% on the evaluation data, the F-measure of the detected coordinate keys is 93%, the F-measure of the detected forward phrase is 65%, and the F-measure of the detected backward phrase is 64%. In addition, the proposed method outperformed KNP. F-measure of the detected coordinate structure by KNP is 26%. Thus proposed method was better 24% against KNP.