Anaphora Resolution for Named Entity Extraction in Japanese Newspaper Articles

Masaomi Satake (910052)

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

February 15, 2002

Keywords: Named Entity,Coreference,Anaphora Resolution,Reference Expression,Transcription Class.

Named entity recognition ("NE" hereafter) is a technique to extract proper nouns from a text and assign semantic tags ("NE tags" hereafter) such as organization, person etc. It is an important step for various applications of natural language processing or text processing. Many researchers have been devoted in research on NE. Especially, newspaper articles are used for the text from which named entities will be extracted, because there are many expressions which stands for time or money in them. In many previous works, rules to recognize named entities were automatically learned, based on the words or their POS tags surrounding target proper nouns. Furthermore, in the most of these works, each proper noun in a document was recognized independently with the others. This causes the following two problems:

- The named entities may not be extracted.

  For example, suppose that there are two proper nouns, "Kousei Torihiki Iinkai"(Fair Trade Commission) and "Koutorii"(an abbreviation for "Kousei Torihiki Iinkai") in the same document, and they refer the same entity. The NE tag for the first proper noun may be "organization" tag and no NE tag may be assigned for the second one. In other words, the second expression may not be extracted as a named

entity. However, it should be assigned "organization" tag, because both expression refer the same entity.

- The same NE tag is not assigned to the same entity.

  For example, suppose that there are the two proper nouns, "Yamagishi Akira" and "Yamagishi" in the same document, and they refer the same person. The NE tag for the first proper noun may be "person", while the one for the second may be "organization". However, the same NE tag should be assigned to the expressions.

In order to solve these problems, this paper proposes the method to improve the accuracy of named entity recognition from newspaper articles using the result of anaphora resolution for proper nouns. Furthermore, this paper also proposes the new method of anaphora resolution for a proper nouns.

The overview of the proposed named entity recognition system is as follows: First, initial NE tags are assigned to the proper nouns in newspaper articles. This is done by an existing NE recognition system. Next, expressions which refers the same entity are identified by anaphora resolution. Finally, initial NE tags are modified so that the same NE tags are assigned to all the proper nouns which refers the same entity.

The overview of the proposed anaphora resolution algorithm is as follows: First, the expressions whose antecedents should be identified are extracted from a document. In this work, target expressions for anaphora resolution are "proper noun" or "expression which refers other proper noun", called "reference expression" hereafter. There are three kinds of reference expression as follows:

- **abbreviation**

  ex. "Matsushita", when it is an abbreviation of "Matsushita Denki Sangyou" (Matsushita Electric Industrial)

- **an common noun which refers proper noun**

  ex. "Daigaku"(a university), when it refers a proper noun "Tokyo Daigaku"(Tokyo University)

- **an expression containing the morpheme "Dou"**

  ex. "Dou-sha"(the same company), "Dou-ken"(the same prefecture)

  "Dou" is the morpheme which means the "same", and it always refers an entity preceding in a document.

When identifying the antecedents of reference expressions, LCS (Longest Common Subsequences) between a candidate of an antecedent and an anaphora (reference expression) is calculated as a score. Then, the candidate which has high score is identified as the antecedent. Furthermore, the following three features are also used in anaphora resolution, "grammatical features", "distance" and "transcription class". "Grammatical features" are attributes of words about topicalization and surface cases, and they are used to rank the candidates of antecedents in the Centering Theory. "Distance" is defined as the distance between antecedent and anaphora, and it is regarded that an expression which located near an anaphora is preferred as an antecedent of it. "Transcription class" is the feature newly introduced in this research. By analyzing reference expression in newspaper articles, the following tendency was found: when a transcription form of a reference expression is different from that of its antecedent, and another expression which refers the same entity appears next in a document, its transcription form also tend to be different from the previous ones. That is, it is considered that entities which appear in various transcription forms in the same document are preferred as antecedents, especially when an anaphora is an expression containing the morpheme "dou". For such reasons, the following three class are defined: the class of entities which appear in various transcription forms as "different transcription class", the class of entities which appear first in a document as "not-appeared transcription class", the class of entities which appear in the same transcription forms as "same transcription class". According to the transcription class, the entities are preferred as antecedents in the order of "different", "not-appeared", "same". The anaphora resolution algorithm are newly developed based on the above three features.

Finally, an experiment are conducted for the evaluation. The recall and precision of anaphora resolution using proposed algorithm was 34.78% and 54.42%, respectively. One of the reasons why recall/precision was not so

good is errors of the morphological analysis. The segmentation of the words given by morphological analyzer ALTJAWS sometimes didn't coincident with the segmentation of reference expressions, and some reference expressions were failed to extract.

On the other hand, comparing with the existing named entity recognition system, which does not use the result of anaphora resolution, F-measure of the proposed system was slightly gained, by 1%. In spite of the low performance of anaphora resolution, F-measure of named entity recognition was improved. It indicates that using the result of anaphora resolution is effective to improve the performance of named entity recognition.