Title	教師なし学習で推定した確率分布によるゼロ代名詞の解析
Author(s)	杉原，大悟
Citation	
Issue Date	2003-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1654
Rights	
Description	Supervisor:鳥澤 健太郎，情報科学研究科，修士

An analyzing Japanese zero anaphora based on the probability distribution estimated by an unsupervised learning method

Daigo Sugihara (110065)

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

February 14, 2003

Keywords: analyzing Japanese anaphora, zero pronoun, unsupervised Learning, EM algorithm.

1 Objective

We propose a method for analyzing Japanese zero anaphora by using the probability distribution estimated only by an unsupervised learning method. In the process of communicating information through natural language, the expression whose referent can easily be identified by a receiver of information is often replaced by a pronoun. When a linguistic expression refers to one object and another expression occurring after the first also refers to the same object, these expressions are said to have an anaphoric relation, and the former expression is called a antecedent, and the later is a anaphor. Zero pronoun is the anaphor omitted in text or conversation, and zero anaphora is the anaphoric relation by zero pronoun. Zero Anaphora is a phenomenon widely observed in Japanese text, and its proper analysis are crucial for application of natural language processing automatic systems, such as machine translation, making summary of text, and so on.

In prior researches on the analyses of zero anaphora, various methods have been proposed. Seki[2001] and Kawahara[2001] automatically acquire the linguistic resources, which is required for large scale unannotated corpora. Kawahara automatically constructed the case frame dictionaries from unannotated corpora and applied it to analysis for Japanese zero anaphora. Seki analyzed Japanese zero anaphora by using the probability distribution estimated from corpora. But, Kawahara used the Japanese hand-crafted thesaurus in their construction of case frame dictionaries and analyzing for zero anaphora. Seki used not only hand-crafted thesaurus but also the corpora annotated with anaphoric relation to estimate their probability distribution. In short, they used hand-crafted resources in analysis of Japanese zero anaphor. On the other hand, our method acquires the resources for analysis of Japanese zero anaphora only from unannotated corpora by unsupervised learning method.

2 Methodology

The first resource we use in analyzing for Japanese zero anaphora is the probability distribution indicating the likelihood of co-occurrences between nouns and verbs, which is estimated by the unsupervised methods by Rooth[1999]. Actually, we use the probability distribution estimated by the method by Torisawa[2001]. Torisawa extended the unsupervised method by Rooth to Japanese. Torisawa augmented the training data with the co-occurrences likelihoods obtained by a statistical parser. This statistical parser was trained by using EDR corpus. But, the hand-crafted resources in EDR corpus cannot be the correct answer on estimation of the probability distribution by the method of Torisawa, so we recognize the method by Torisawa as unsupervised learning method. And the second resource we use in analyzing for Japanese zero anaphora is the probability distribution indicating the likelihood that a noun is the antecedent of zero pronoun when the noun takes one positional relation between the verb where the zero pronoun arises. The second probability distribution is estimated by using the our unsupervised learning method based on EM algorithm. The likelihood that one noun is the antecedent of zero pronoun is influenced not only by the semantic relation between the noun and the verb but also various syntactic elements, for example, the distance between the noun and the verb where the zero pronoun arises, post-positional particles of the nouns, and so on. We call these various elements "the positional relation between a noun and a verb", and regard R as the vector represent it. We formalize R as follows,

$$R = \langle d, c1, rentai, head, eachid \rangle$$

d: The distance between a noun and a verb. In the case where they occur in the same sentence, its value takes 0. In the case where a noun occurs in n sentences previous to the sentence including the verb, its value is n. $d \in \{0, 1, 2, ...\}$

c1: Post-positional particle of noun. Such as we distinguish the particles by integer numbers. $c1 \in \{0, 1, 2, 3, ...\}$

rentai: This feature denotes whether a noun is included in a relative clause or not. $rentai \in \{0, 1\}$

head: This feature denotes whether a noun is included in the first paragraph of article or not. $head \in \{0, 1\}$

eachid: This feature distinguishes the same patterns of $\langle d, c1, rentai, head \rangle$ occurring previous to one verb. $eachid \in \{1, 2, ...\}$

We distinguish "the positional relation between a noun and a verb" in text with the pattern of values of R. We constructed the probabilistic model expressing the situation one noun in text is the antecedent of the zero pronoun. Then, from this model, we derived the equation to estimate the probability distribution that the noun with the positional relation R is the antecedent of the zero pronoun by applying standard derivation steps in the EM method. This probability distribution on R can be estimated by using an iterative procedure from unannotated corpora and the probability distribution of co-occurrences between nouns and verbs obtained by the unsupervised methods by Rooth. We analyze the zero anaphora by evaluating likelihood that given nouns become the antecedent of zero pronouns in text by using these two probability distributions.

3 Experiments

We estimated the probability distribution on R according to our estimation equation from 3,120 articles in *Yomiuri Shimbun* newspaper articles. We conducted experiment about identification of antecedent of zero pronoun by using the estimated probability distribution. For this experiment, we randomly selected 10 articles from what weren't contained in the training data, and manually annotated those articles with anaphoric relation for zero pronouns. We treated these articles as the correct answer corpus for our experiment. We conducted our experiment under the assumption that all the zero pronouns were correctly detected. For each zero pronoun we extracted antecedent candidates from the preceding contexts, which are ordered according to the extent to which they can be the antecedent for the target zero pronoun, and evaluated the likelihood each candidate become the antecedent by using the probability distributions. We compared the two cases. The first case is where we identified antecedents using only the probability distribution of co-occurrences between verbs and nouns obtained by the methods proposed by Rooth. The second case is where we identified antecedents using the combination of the first probability distribution and the probability distribution on R. The second case is our method for analyzing Japanese zero anaphora. We calculated the accuracy by the ratio between the number of zero pronouns whose antecedents were correctly identified and the number of all the zero pronouns. In the situation one verb had k zero pronouns of one case, if the correct answer was included in the k-best candidates, we judged the antecedent for the zero pronoun identified correctly. The accuracy of our method resulted in 41.07%. Actually this figure is inferior to the precedence researches. But, when we compared the accuracy of the first case with that of the second case, our accuracy was 11% higher rather than that of the first case by using only the probability distribution of co-occurrences between verbs and nouns obtained by the methods proposed by Rooth. From this result, we concluded that we could attain the improvement in accuracy of analysis for Japanese zero anaphora by our unsupervised learning method.