Corpus-based Resolution of Pronominal References

Michael Paul Kazuhide Yamamoto Eiichiro Sumita

ATR Interpreting Telecommunications Research Laboratories
e-mail: {paul,yamamoto,sumita}@itl.atr.co.jp

Abstract

In this paper we propose a corpus-based approach to pronominal anaphora resolution combining a machine learning method and statistical as well as discourse information. First, a decision tree trained on an annotated corpus determines the co-reference relation of a given anaphora and antecedent candidates and is utilized as a filter in order to reduce the amount of potential candidates. In a second step, preference selection is achieved by taking into account frequency information of co-/non-referential pairs tagged in the training corpus and distance features within the current discourse.

1 Introduction

Co-reference information is relevant for numerous NLP systems. Our interest in anaphora resolution is based on the demand for practical machine translation systems to be able to translate anaphoric expressions in agreement with the morphosyntactic characteristics of the referred object in order to prevent contextual misinterpretations.

So far various approaches to anaphora resolution have been proposed. In this paper a *machine learning approach* is combined with a preference selection method based on the *frequency* information of co-/non-referential pairs tagged in the corpus as well as *distance* features within the current discourse.

The advantage of machine learning approaches is that they result in modular anaphora resolution systems automatically trainable from a corpus with no or only a minimal amount of human intervention. In the case of decision trees we do have to provide information about possible antecedent indicators (syntactic, semantic, and pragmatic features) contained in the corpus, but the relevance of features for the resolution task is extracted automatically from the training data.

The distinction of our approach to related research reported in [1], [2] is the usage of the decision tree towards the selection of the most salient canddiate. They focus on preference selection criteria adopted from the decision tree itself. However, decision trees are characterized by an independent learning of specific attributes, i.e., relations between single attributes cannot be obtained automatically. Accordingly, the usage of dependency factors for preference selection during decision tree training requires that the artificially created attributes expressing these dependencies be defined. However, this would extent the human intervention into the automatic learning procedure (which dependencies are important?) and thus should be avoided.

The preference selection in our approach is based on statistical frequency information **and** discourse features. Therefore, our decision tree is not applied directly to the task of preference selection, but used as a preprocessing filter aiming at the elimination of irrelevant candidates.

The decision tree is trained on syntactic (lexical word attributes), semantic, and primitive discourse (distance, frequency) information and determines the co-referential relation between an anaphora and antecedent candidate in the given context. Irrelevant antecedent candidates are filtered out, achieving a noise reduction for the preference selection algorithm. A saliency factor is assigned to each potential anaphora-candidate pair depending on the proportion of non-/co-referential occurrences of the pair in the training corpus (*frequency ratio*) and the relative position of both elements in the discourse (*distance*). The most salient candidate is resolved as the antecedent of the anaphoric expression.

2 Corpus-Based Anaphora Resolution

After a short overview over the data corpus in section 2.1, section 2.2 focuses on the analysis of co-referential relationships by means of machine learning. Details of the preference selection algorithm are given in section 2.3. Preliminary experiments are conducted for the task of pronominal anaphora resolution and the performance of our system is evaluated in section 3.

2.1 Data Corpus

For our experiments we use the *ATR Speech and Language Database* [4] consisting of 500 Japanese spoken-language dialogs annotated with co-referential tags. Anaphoric expressions used in our experiments are limited to those referring to nominal antecedents (nominal: 2160, pronominal: 526, ellipsis: 3843).

Besides the anaphora type, we also include morphosyntactic information for each surface word as well as semantic codes for content words in this corpus. According to the tagging criteria used for our corpus an anaphoric tag refers to the most recent antecedent found in the dialog. The analysis of (possible) references from this antecedent to previous ones (*anaphoric chaining*) allows us to identify all correct antecedents for the given anaphoric expression.

Based on the corpus annotations we extract the frequency information of co-referential anaphora-antecedent pairs and non-referential pairs, whereby the latter pairs consist of the anaphoric expression and nominal candidates in the discourse history that are not tagged co-referentially.

2.2 Decision Tree Filter

To learn the co-reference relations from our corpus we have chosen a C4.5-like machine learning algorithm without pruning [3]. The training attributes consist of *lexical word attributes* (surface word, regular expression, part-of-speech, semantic code, morphological attributes[1] like gender, person, number) applied to the anaphora, antecedent, and sentence predicate. In addition, binary features like *attribute agreement*, *distance* and *frequency ratio*[2] are checked for each anaphora-antecedent pair. The decision tree result consists of only two classes determining the co-reference relation between the given anaphora-candidate pair.

During anaphora resolution the decision tree is applied as a filter to reduce the amount of possible candidates. A candidate list[3] is created for each tagged anaphoric expression and the decision tree filter is then successively applied to all anaphora-candidate pairs.

If the decision tree results in the non-reference class, the respective candidate is judged as irrelevant and eliminated from the list of potential candidates forming the input of the preference selection algorithm.

2.3 Preference Selection

The primary order of candidates is given by their distance from the anaphoric expression. Therefore, a straightforward preference strategy is the selection of the most recent candidate (*MRC*), i.e., the first element of the candidate list. The success rate of this baseline test, however, is quite low as shown in Fig. 2. Moreover, an examination of our data corpus gives rise to suspicion that even if more recent candidates should be preferred in principle, dialog initial references to candidates introduced first in the dialog are quite frequent. Similarities to other references in our corpus, however, seem to be useful for the correct

[1] Due to the poor Japanese morphology we adopt the attributes of a corresponding German translation.

[2] This value is defined as the ratio of the co-referential and non-referential occurrences of the given anaphora-candidate pair in the training corpus.

[3] A list of noun phrase candidates preceding the anaphora element in the current discourse.

identification of the antecedent, too. Therefore, we propose a preference scheme based on the combination of these features.

In a first step, we define the *ratio* of a given reference pair utilizing statistical information about the frequency of co-referential ($freq^+$) anaphora-antecedent and non-referential ($freq^-$) pairs extracted from our corpus.

$$ratio = \frac{freq^+ - freq^-}{freq^+ + freq^-}$$

As mentioned above the distance plays a crucial role in our selection method, too. We define a preference value *pref* by normalizing the *ratio* value according to the distance *dist*.

$$pref = \frac{ratio}{dist}$$

The *pref* value is calculated for each candidate and the precedence ordered list of candidates is resorted towards the maximization of the preference factor. Again, the first element of the preferenced candidate list is chosen as the antecedent. The precedence order between candidates of the same confidence continues to remain so and therefore a final decision is made in the case of a draw.

The robustness of our approach is ensured by the definition of a *backup* strategy which ultimately selects one candidate occurring in the history in the case that all antecedent candidates are rejected by the decision tree filter.

3 Evaluation

For the evaluation of the experimental results described in this section we use a *F-measure* metrics calculated by the *recall R* and *precision P* of the system performance. Let $\sum_{total}$ denote the total number of tagged anaphora-antecedent pairs contained in the test data, $\sum_{filter}$ the amount of these pairs passing the decision tree filter, and $\sum_{correct}$ the number of correctly selected antecedents.

During evaluation we distinguish 3 classes: whether the correct antecedent is the first element of the candidate list (c_f), is in the candidate list (c_i), or is filtered by the decision tree (c_o). The metrics F, R and P are then defined as follows:

$$F = \frac{2 \times P \times R}{P+R} \qquad R = \frac{\sum_{correct}}{\sum_{total}} \qquad P = \frac{\sum_{correct}}{\sum_{filter}}$$

$$\sum_{correct} = | c_f | \qquad \sum_{filter} = | c_f | + | c_i | \qquad \sum_{total} = | c_f | + | c_i | + | c_o |$$

In order to prove the feasibility of our approach we compare the four preference selection methods listed in Fig. 1. First, the baseline test *MRC* selects the most recent candidate as the antecedent of an anaphoric expression. The necessity of the filter and preference selection components is shown by comparing the filter scheme *DT* (i.e., select the first element of the filtered candidate list) and preference scheme *PREF* (i.e., resort the complete candidate list) against our combined method *DT+PREF* (i.e., resort the filtered candidate list).

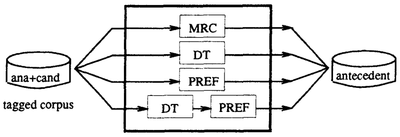

Figure 1: Outline of the experiments

5-way cross-validation experiments are conducted for pronominal anaphora resolution. The selected antecedents are checked against the annotated correct antecedents according to their morphosyntactic (gender, number, person) and semantic attributes[4].

We use varied numbers of training dialogs (50-400) for the training of the decision tree and the extraction of the frequency information from the corpus. Open tests are conducted on 100 non-training dialogs whereas closed tests use the training data for evaluation. The results of the different preference selection methods are shown in Fig. 2.

The baseline test *MRC* succeeds in resolving only 43.9% of the most recent candidates correctly as the antecedent. The best *F-measure* rates for *DT* and *PREF* are 65.0%

[4] Human misinterpretations are caused by an attributive disagreement. Therefore, these attributive criteria are considered sufficient for the evaluation task. However, *exact match* and *word match* are also used as identification criteria whereby the *F-measure* performance decreases by 4.1% (exact) and 1% (word).

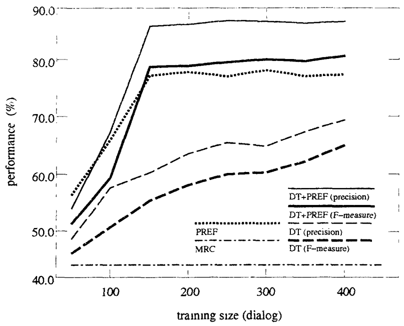

Figure 2: Training size versus performance

and 78.1% whereas the combination of both methods achieves a success rate of 80.6%.

The *PREF* method seems to reach a plateau at around 300 dialogs which is bared out by the closed test reaching a maximum of 81.1%. Comparing the *recall* rate of *DT* (61.2%) and *DT+PREF* (75.9%) with the *PREF* result, we might conclude that the decision tree is not much a help due to the side-effect of 11.8% of the correct antecedents being filtered out.

However, in contrast to the *PREF* algorithm, the *DT* method improves continuously according to the training size implying a lack of training data for the identification of potential candidates. Despite the sparse data the filtering method proves to be very effective. It achieves a reduction rate for the average number of all candidates of 71.8% (closed data: 81%). The amount of trivial selection cases (only one candidate) increases from 2.7% (history) to 11.4% (filter; closed data: 21%). On average, two candidates are skipped in the history to select the correct antecedent.

Moreover, the *precision* of *DT* (69.4%) and *DT+PREF* (86.0%) show that the utilization of the decision tree filter in combination with the statistical preference selection gains a relative improvement of 9% towards the preference and 16% towards the filter method.

Additionally, the system proves to be quite robust, because the decision tree filters out all candidates in only 1% of the open test samples whereby the selection of the last candidate of the history list as a backup strategy shows the best performance in our experiments.

4 Conclusion

In this paper we proposed a corpus-based anaphora resolution method combining an automatic learning algorithm for co-referential relationships with statistical preference selection in the discourse context. We proved the applicability of our approach to pronominal anaphora resolution despite the limitation of sparse data by achieving a resolution accuracy of 86.0% (precision) and 75.9% (recall) for Japanese pronouns. Improvements in these results can be expected by increasing the training data as well as utilizing more sophisticated linguistic knowledge (structural analysis of utterances, etc.) and discourse information due to a rise of the decision tree filter performance.

Preliminary experiments with nominal reference and ellipsis resolution showed promising results, too. We plan to incorporate this approach in multi-lingual machine translation which enables us to handle a variety of referential relations in order to improve the translation quality.

References

[1] C. Aone and S. Bennett, Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies, *Proc. of the 33th ACL*, pp. 122-129, 1995.

[2] D. Conolly, J. Burger and D. Day, A Machine Learning Approach to Anaphoric Reference, *Proc. of NEMLAP*, pp. 255-261, Manchester, UK, 1994.

[3] J. Quinlan, C4.5 Programs for Machine Learning, *Morgan Kaufmann*, 1993.

[4] T. Takezawa, T. Morimoto and Y. Sagisaka, Speech and language database for speech translation research in ATR, *Proc. of Oriental COCOSDA Workshop*, pp. 148-155, 1998.