Corpus-Based Anaphora Resolution Towards Antecedent Preference

Michael PAUL, Kazuhide YAMAMOTO and **Eiichiro SUMITA**
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{paul, yamamoto, sumita}@itl.atr.co.jp

Abstract

In this paper we propose a corpus-based approach to anaphora resolution combining a machine learning method and statistical information. First, a decision tree trained on an annotated corpus determines the coreference relation of a given anaphor and antecedent candidates and is utilized as a filter in order to reduce the number of potential candidates. In the second step, preference selection is achieved by taking into account the frequency information of coreferential and non-referential pairs tagged in the training corpus as well as distance features within the current discourse. Preliminary experiments concerning the resolution of Japanese pronouns in spoken-language dialogs result in a success rate of 80.6%.

1 Introduction

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

So far various approaches[1] to anaphora resolution have been proposed. In this paper a *machine learning approach* (decision tree) is combined with a preference selection method based on the *frequency* information of non-/coreferential 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.

Machine learning approaches using decision trees proposed so far have focused on preference selection criteria directly derived from the decision tree results. The work described in (Conolly et al., 1994) utilized a decision tree capable of judging which one of two given anaphor-antecedent pairs is "better". Due to the lack of a strong assumption on "transitivity", however, this sorting algorithm is more like a greedy heuristic search as it may be unable to find the "best" solution.

The preference selection for a single antecedent in (Aone and Bennett, 1995) is based on the maximization of confidence values returned from a pruned decision tree for given anaphor-candidate pairs. However, decision trees are characterized by an independent learning of specific features, i.e., relations between single attributes cannot be obtained automatically. Accordingly, the use of dependency factors for preference selection during decision tree training requires that the artificially created attributes expressing these dependencies be defined. However, this not only extends human intervention into the automatic learning procedure (i.e., which dependencies are important?), but can also result in some drawbacks on the contextual adaptation of preference selection methods.

The preference selection in our approach is based on the combination of statistical frequency information and distance features in the discourse. Therefore, our decision tree is not applied directly to the task of preference selection, but aims at the elimination of irrelevant candidates based on the knowledge obtained from the training data.

[1] See section 4 for a more detailed comparison with related research.

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

2 Corpus-Based Anaphora Resolution

In this section we introduce a new approach to anaphora resolution based on coreferential properties automatically extracted from a training corpus.

In the first step, the decison tree filter is trained on the linguistic, discourse and coreference information annotated in the training corpus which is described in section 2.1.

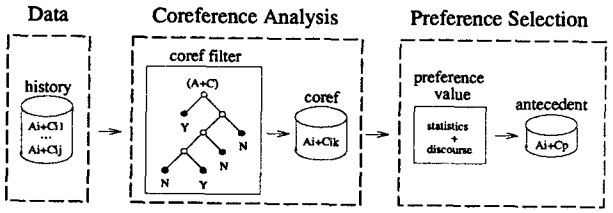

Figure 1: System outline

The resolution system in Figure 1 applies the coreference filter (cf. section 2.2) to all anaphor-candidate pairs ($A_i + C_{ij}$) found in the discourse history. The detection of anaphoric expressions is out of the scope of this paper and just reduced to tags in our annotated corpus. Antecedent candidates are identified according to noun phrase part-of-speech tags. The reduced set ($A_i + C_{ik}$) forms the input of the preference algorithm which selects the most salient candidate C_p as described 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-ITL Speech and Language Database* (Takezawa et al., 1998) consisting of 500 Japanese spoken-language dialogs annotated with coreferential tags. It includes nominal, pronominal, and ellipsis annotations, whereby the anaphoric expressions used in our experiments are limited to those referring to nominal antecedents (nominal: 2160, pronominal: 526, ellipsis: 3843).

Besides the anaphor type, we also include morphosyntactic information like stem form and inflection attributes for each surface word as well as semantic codes for content words (Ohno and Hamanishi, 1981) in this corpus.

r1: ありがとうございます。シティホテルでございます。
[thank you very much] [City Hotel]
"Thank you for calling City Hotel."
c1: もしもし、私田中弘子と言いますが、
[hello] [I][Hiroko Tanaka][the name is]
"Hello, my name is Hiroko Tanaka."
そちらのホテルの予約したいんですが。
[there] [hotel] [reservation][would like to have]
"I would like to make a reservation at your hotel."
r2: お客様のお名前のスペルを頂けますでしょうか。
[your] [name] [spelling] [can I have]
"Can you spell your name for me, please?
c2: はい。ティーエーエヌエーケイエーです。
[yes] [T] [A] [N] [A] [K] [A] [be]
"It's T A N A K A."
r3: はい。十日にこちらにご到着ということでございますね。
[yes] [tenth] [here] [arrival] [be]
"Okay, you will arrive here on the tenth, right?"

Figure 2: Example dialog

In the example dialog between the hotel reception (r) and a customer (c) listed in Figure 2 the proper noun (r1)"シティホテル [City Hotel]" is tagged as the antecedent of the pronoun (c1)"そちら [there]" as well as the noun (c1)"ホテル [hotel]". An example for ellipsis is the ommitted subject (c2)"∅[it]" referring to (r2)"スペル [spelling]".

According to the tagging guidelines used for our corpus an anaphoric tag refers to the most recent antecedent found in the dialog. However, this antecedent might also refer to a previous one, e.g. (r3)"こちら [here]"→(c1)" そちら [there]"→(r1) " シティホテル [City Hotel]". Thus, the *transitive closure* between the anaphora and the first mention of the antecedent in the discourse history defines the set of positive examples, e.g. (そちら, シティホテル), whereas the nominal candidates outside the transitive closure are considered negative examples, e.g. (そちら, 田中), for coreferential relationships.

Based on the corpus annotations we extract the frequency information of coreferential anaphor-antecedent pairs and non-referential pairs from the training data. For each non-/coreferential pair the occurrences of surface and stem form as well as semantic code combinations are counted.

Table 1: Frequency data

type	anaphor	candidate	$freq^+$	$freq^-$	*ratio*
word-word	そちら	シティホテル	6	0	1
	そちら	田中	0	11	-1
	こちら	十日	0	0	-0.1
word-sem	こちら	{shop}	33	33	0
sem-sem	{demonstratives}	{shop}	51	18	0.48

In Table 1 some examples are given for pronoun anaphora, whereas the expressions "{...}" denote semantic classes assigned to the respective words.

The values $freq^+$, $freq^-$ and *ratio* and their usage are described in more detailed in section 2.3.

Moreover, each dialog is subdivided into *utterances* consisting of one or more *clauses*. Therefore, distance features are available on the utterance, clause, candidate, and morpheme levels. For example, the distance values of the pronoun (r3)"こちら [here]" and the antecedent (r1)" シティホテル [City Hotel]" in our sample dialog in Figure 2 are d_{utter}=4, d_{clause}=7, d_{cand}=14, d_{morph}=40.

2.2 Coreference Analysis

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

During anaphora resolution the decision tree is used as a module determining the coreferential property of each anaphor-candidate pair. For each detected anaphoric expression a candidate list[3] is created. The decision tree filter is then successively applied to all anaphor-candidate pairs.

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

2.3 Preference Selection

The primary order of candidates is given by their word distance from the anaphoric expression. A straightforward preference strategy we could choose is the selection of the most recent candidate (MRC) as the antecedent, i.e., the first element of the candidate list. The success rate of this baseline test, however, is quite low as shown in section 3.

But, this result does not mean that the *recency* factor is not important at all for the determination of saliency in this task. One reason for the bad performance is the application of the baseline test to the unfiltered set of candidates resulting in the frequent selection of non-referential antecedents. Additionally, long-range references to candidates introduced first in the dialog are quite frequent in our data.

[2]cf. (Quinlan, 1993)

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

An examination of our corpus gives rise to suspicion that similarities to references in our training data might be useful for the identification of those antecedents. Therefore, we propose a preference selection scheme based on the combination of *distance* and *frequency* information.

First, utilizing statistical information about the frequency of coreferential anaphor-antecedent pairs ($freq^+$) and non-referential pairs ($freq^-$) extracted from the training data, we define the *ratio* of a given reference pair as follows[4]:

$$ratio = \begin{cases} -\delta & : (freq^+ = freq^- = 0) \\ \frac{freq^+ - freq^-}{freq^+ + freq^-} & : otherwise \end{cases}$$

The value of *ratio* is in the range of $[-1, +1]$, whereby $ratio = -1$ in the case of exclusive non-referential relations and $ratio = +1$ in the case of exclusive coreferential relationships. In order for referential pairs occurring in the training corpus with $ratio = 0$ to be preferred to those without frequency information, we slightly decrease the *ratio* value of the latter ones by a factor δ.

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* given by the primary order of the candidates in the discourse.

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

The $pref$ value is calculated for each candidate and the precedence ordered list of candidates is resorted towards the maximization of the preference factor. Similarly to the baseline test, 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 thus 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. For our experiments reported in section 3 we adopted the selection of the dialog-initial candidate as the backup strategy.

3 Evaluation

For the evaluation of the experimental results described in this section we use *F-measure* metrics calculated by the *recall* and *precision* of the system performance. Let $\sum_t$ denote the total number of tagged

[4]In order to keep the formula simple the frequency types are omitted (cf. Table 1)

anaphor-antecedent pairs contained in the test data, $\sum_f$ the number of these pairs passing the decision tree filter, and $\sum_c$ the number of correctly selected antecedents.

During evaluation we distinguish three classes: whether the correct antecedent is the first element of the candidate list (f), is in the candidate list (i), or is filtered out by the decision tree (o). The metrics *F*, *recall* (R) and *precision* (P) are defined as follows:

$$F = \frac{2 \times P \times R}{P + R} \qquad R = \frac{\sum_c}{\sum_t} \qquad P = \frac{\sum_c}{\sum_f}$$

$$\sum_c = |f| \qquad \sum_f = |f| + |i| \qquad \sum_t = |f| + |i| + |o|$$

In order to prove the feasibility of our approach we compare the four preference selection methods listed in Figure 3.

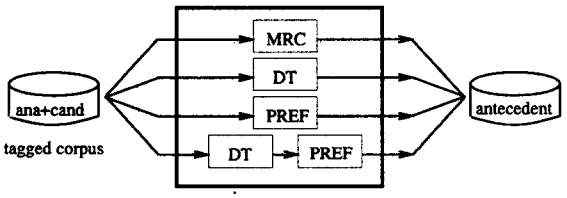

Figure 3: Preference selection experiments

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 decision tree 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).

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 and semantic attributes.

3.1 Training Size

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 Figure 4.

The baseline test *MRC* succeeds in resolving only 43.9% of the most recent candidates correctly as the antecedent. The best *F-measure* rate for *DT* is 65.0% and for *PREF* the best rate is 78.1% whereas

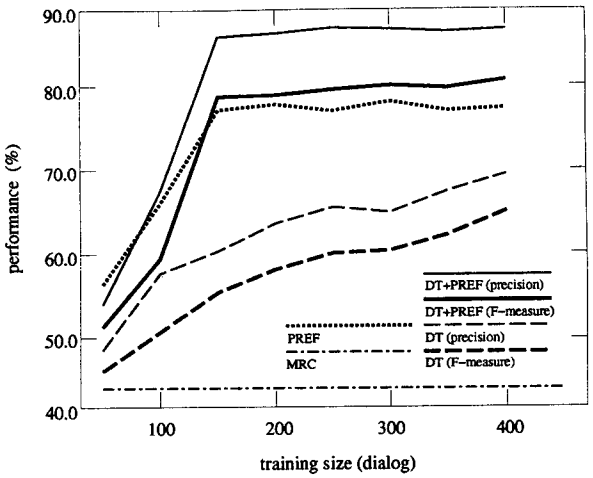

Figure 4: Training size versus performance

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 borne 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 of 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. The average number of all candidates (*history*) for a given anaphor in our open data is 39 candidates which is reduced to 11 potential candidates by the decision tree filter resulting in a reduction rate of 71.8% (closed test: 81%). The number of trivial selection cases (only one candidate) increases from 2.7% (history) to 11.4% (filter; closed test: 21%). On average, two candidates are skipped in the history to select the correct antecedent.

Moreover, the precision rates 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. Selecting the candidate first introduced in the dialog as a backup strategy shows the best performance due to the frequent dialog initial references contained in our data.

Table 2: Frequency and distance dependency

	DT	DT-no-dist	DT-no-freq	DT+PREF	DT+PREF-no-dist
recall	61.2	60.1	53.6	75.9	73.0
precision	69.4	68.7	64.5	86.0	82.8
F-measure	65.0	64.1	58.5	80.6	77.6
(filtered-out)	11.8	12.5	16.9	11.8	11.8

3.2 Feature Dependency

In our approach *frequency ratio* and *distance* information plays a crucial role not only for the identification of potential candidates during decision tree filtering, but also for the calculation of the preference value for each antecedent candidate.

In the first case these features are used independently to characterize the training samples whereas the preference selection method is based on the dependency between the frequency and distance values of the given anaphor-candidate pair in the context of the respective discourse. The relative importance of each factor is shown in Table 2.

First, we compare our decision tree filter *DT* to those methods that do not use either frequency (*DT-no-freq*) or distance (*DT-no-dist*) information. Frequency information does appear to be more relevant for the identification of potential candidates than distance features extracted from the training corpus. The recall performance of *DT-no-freq* decreases by 7.6% whereas *DT-no-dist* is only 1.1% below the result of the original *DT* filter[5]. Moreover, the number of correct antecedents not passing the filter increases by 5.1% (*DT-no-freq*) and 0.7% (*DT-no-dist*).

However, the distance factor proves to be quite important as a preference criterion. Relying only on the frequency ratio as the preference value, the recall performance of *DT+PREF-no-dist* is only 73.0%, down 2.9% of the original *DT+PREF* method.

The effectiveness of our approach is not only based on the usage of single antecedent indicators extracted from the corpus, but also on the combination of these features for the selection of the most preferable candidate in the context of the given discourse.

4 Related Research

Due to the characteristics of the underlying data used in these experiments a comparison involving absolute numbers to previous approaches gives us less evidence. However, the difficulty of our task can be verified according to the baseline experiment results reported in (Mitkov, 1998). Resolving pronouns in English technical manuals to the most recent candidate achieved a success rate of 62.5%, whereas in our experiments only 43.9% of the most recent candidates are resolved correctly as the antecedent (cf. section 3).

Whereas *knowledge-based* systems like (Carbonell and Brown, 1988) and (Rich and LuperFoy, 1988) combining multiple resolution strategies are expensive in the cost of human effort at development time and limited ability to scale to new domains, more recent *knowledge-poor* approaches like (Kennedy and Boguraev, 1996) and (Mitkov, 1998) address the problem without sophisticated linguistic knowledge. Similarly to them we do not use any sentence parsing or structural analysis, but just rely on morphosyntactic and semantic word information.

Moreover, clues are used about the *grammatical* and *pragmatic* functions of expressions as in (Grosz et al., 1995), (Strube, 1998), or (Azzam et al., 1998) as well as rule-based *empirical approaches* like (Nakaiwa and Shirai, 1996) or (Murata and Nagao, 1997), to determine the most salient referent. These kinds of manually defined scoring heuristics, however, involve quite an amount of human intervention which is avoided in machine learning approaches.

As briefly noted in section 1, the work described in (Conolly et al., 1994) and (Aone and Bennett, 1995) differs from our approach according to the usage of the decision tree in the resolution task. In (Conolly et al., 1994) a decision tree is trained on a small number of 15 features concerning anaphor type, grammatical function, recency, morphosyntactic agreement and subsuming concepts. Given two anaphor-candidate pairs the system judges which is "better". However, due to the lack of a strong assumption on "transitivity" this sorting algorithm may be unable to find the "best" solution.

Based on discourse markers extracted from lexical, syntactic, and semantic processing, the approach of (Aone and Bennett, 1995) uses 66 unary and binary attributes (lexical, syntactic, semantic, position, matching category, topic) during decision tree training. The confidence values returned from the pruned decision tree are utilized as a saliency measure for each anaphor-candidate pair in order to se-

[5]So far we have considered the decision tree filter just as a black-box tool. Further investigations on tree structures, however, should give us more evidence about the relative importance of the respective features.

lect a single antecedent. However, we use dependency factors for preference selection which cannot be learned automatically because of the independent learning of specific features during decision tree training. Therefore, our decision tree is not applied directly to the task of preference selection, but only used as a filter to reduce the number of potential candidates for preference selection.

In addition to salience preference, a statistically modeled *lexical preference* is exploited in (Dagan et al., 1995) by comparing the conditional probabilities of co-occurrence patterns given the occurrence of candidates. Experiments, however, are carried out on computer manual texts with mainly intra-sentential references. This kind of data is also characterized by the avoidance of disambiguities and only short discourse units, which prohibits almost any long-range references. In contrast to this research, our results show that the distance factor in addition to corpus-based frequency information is quite relevant for the selection of the most salient candidate in our task.

5 Conclusion

In this paper we proposed a corpus-based anaphora resolution method combining an automatic learning algorithm for coreferential relationships with statistical preference selection in the discourse context. We proved the applicability of our approach to pronoun resolution achieving a resolution accuracy of 86.0% (precision) and 75.9% (recall) for Japanese pronouns despite the limitation of sparse data. 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 (extra-sentential knowledge, etc.) which should lead 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 multilingual machine translation which enables us to handle a variety of referential relations in order to improve the translation quality.

Acknowledgement

We would like to thank Hitoshi Nishimura (ATR) for his programming support and Hideki Tanaka (ATR) for helpful personal communications.

References

C. Aone and S. Bennett. 1995. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. In *Proc. of the 33th ACL*, p. 122–129.

S. Azzam, K. Humphreys, and R. Gaizauskas. 1998. Evaluating a Focus-Based Approach to Anaphora Resolution. In *Proc. of the 17th COLING*, p. 74–78, Montreal, Canada.

J. Carbonell and R. Brown. 1988. Anaphora Resolution: A Multi-Strategy Approach. In *Proc. of the 12th COLING*, p. 96–101, Budapest, Hungary.

D. Conolly, J. Burger, and D. Day. 1994. A Machine Learning Approach to Anaphoric Reference. In *Proc. of NEMLAP'94*, p. 255–261, Manchester.

I. Dagan, J. Justeson, S. Lappin, H. Leass, and A. Ribak. 1995. Syntax and Lexical Statistics in Anaphora Resolution. *Applied Artificial Intelligence*, 9:633–644.

B. Grosz, A. Joshi, and S. Weinstein. 1995. A Framework for Modeling the Local Coherence of Discourse. *Comp. Linguistics*, 21(2):203–225.

C. Kennedy and B. Boguraev. 1996. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In *Proc. of the 16th COLING*, p. 113–118, Copenhagen, Denmark.

R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proc. of the 17th COLING*, p. 869–875, Montreal, Canada.

M. Murata and M. Nagao. 1997. An Estimate of Referents of Pronouns in Japanese Sentences using Examples and Surface Expressions. *Journal of Natural Language Processing*, 4(1):87–110.

H. Nakaiwa and S. Shirai. 1996. Anaphora Resolution of Japanese Zero Pronouns with Deictic Reference. In *Proc. of the 16th COLING*, p. 812–817, Copenhagen, Denmark.

S. Ohno and M. Hamanishi. 1981. *Ruigo-Shin-Jiten*. Kadokawa.

J. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

E. Rich and S. LuperFoy. 1988. An Architecture for Anaphora Resolution. In *Proc. of the 2nd Conference on Applied Natural Language Processing*, p. 18–23, Austin, TX.

M. Strube. 1998. Never Look Back: An Alternative to Centering. In *Proc. of the 17th COLING*, p. 1251–1257, Montreal, Canada.

T. Takezawa, T. Morimoto, and Y. Sagisaka. 1998. Speech and language database for speech translation research in ATR. In *Proc. of Oriental COCOSDA Workshop*, p. 148–155.