Integration of Referential Scope Limitations into Japanese Pronoun Resolution

Michael Paul and **Eiichiro Sumita**

ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun,
619-0288 Kyoto, Japan
paul@slt.atr.co.jp, sumita@slt.atr.co.jp

Abstract

We propose a practical approach to the anaphora resolution of Japanese pronouns incorporating knowledge about referential scope limitations extracted from an annotated corpus. A machine learning approach (decision tree) is utilized for the classification of the coreference relation of a given anaphor and antecedent candidates. The resolution scope of each pronoun is limited according to the relative distance distribution of the training data, resulting in increases in the classification accuracy and analysis speed by causing only a minor decrease in the recall performance.

1 Introduction

Various approaches have been proposed for anaphora resolution, like rule-based (Murata and Nagao, 1997) and machine learning (Aone and Bennett, 1995) approaches. These approaches select the most salient candidate from among previously mentioned noun phrases (*history*). A problem with these systems is that the resolution costs increase in proportion to the history size. This problem becomes especially serious in the analysis of long conversations like in dialog understanding or spoken language translation systems.

Most of the candidates in the history, however, are non-referential. This means that it might not be necessary to analyze the complete history. This paper focuses on the question of how far do we have to look back in history to find the antecedent of a specific referential expression? We propose a classification scheme that limits the scope of the resolution analysis according to the distribution of the relative distances between anaphora and antecedents tagged in the training corpus.

Our anaphora resolution[1] is carried out using a machine learning approach. A decision tree classifier is trained on the annotated corpus described in Section 2 and determines the coreferential relationship of the given anaphor-candidate pairs (cf. Section 3). In the general framework, the decision tree is applied to all of the noun phrases preceding the anaphoric expression in the history. Its system performance is utilized for a baseline comparison to the practical resolution scheme proposed in this paper (cf. Section 5).

An investigation into the statistics of the training corpus (cf. Section 4) reveals quite different characteristics concerning the referential scope of specific anaphoric expressions capable of being exploited not only to decrease the costs of the resolution process, but also to increase the accuracy of the decision tree classifier.

The proposed approach carries out the classification of coreferential relationships and does not select a *single* candidate as the antecedent of a given anaphor. However, the decision tree classifier can be seen as a filter that reduces noise, i.e., the elimination of non-referential candidates, for a successive preference selection scheme, e.g., that in (Kameyama, 1997).

[1] For the analysis of coreferential relationships we utilize the framework introduced in (Paul et al., 1999).

2 Tagged 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. The anaphoric expressions used in our experiments (described in Section 5) are limited to pronominal ones referring to nominal antecedents (637 pronouns). We also include morphosyntactic information like stem forms as well as semantic codes (Ohno and Hamanishi, 1981) for content words in this corpus.

In the example dialog between a clerk (r) and a customer (c) listed in Figure 1, noun phrases (*candidates*) are underlined and pronouns (*anaphora*) are marked with a box.

r1: はいサーカスサーカスでございます。
[Circus Circus] [be]
"Thank you for calling Circus Circus."

c1: こちらロサンジエルス滞在中の鈴木和夫と申します。
[here] [Los Angeles] [stay] [Kazuo Suzuki][be called]
"Hello, my name is Kazuo Suzuki. I'm staying in Los Angeles right now."

c2: 九月二十日に家族でラスベガスへの旅行を計画しているんですが。
[September,20th][family][LasVegas] [travel] [plan]
"We are planning to visit Las Vegas on the 20th of September."

c3: そちらにはカジノのほかに家族が楽しめる所があると聞いたんです。
[there] [casino] [others][family][enjoy] [place] [hear]
"I heard there are other family attractions besides the casinos."

r2: はいわたくしどもには無料のサーカスがございまして"
[yes][we] [free] [circus] [have]
"Yes, the circus is free of charge."
お客様と御家族皆さんで楽しいんでいただきます。
[guest] [family][all] [enjoy]
"You can enjoy it together with your family."

c4: それはおもしろそうですね子どもが喜びます。
[that] [interesting] [children][be glad]
"That sounds interesting. The kids will be delighted."

Figure 1: Example dialog

According to the tagging guidelines used for our corpus, an anaphoric tag refers to the most recent antecedent found in the dialog, but this antecedent might also refer to a previous one. Therefore, the *transitive closure* between the anaphora and the first mention of the antecedent in the history defines the set of positive examples, whereas the nominal candidates outside the transitive closure are considered negative examples for coreferential relationships. In our example, the proper noun (c2)" ラスベガス [Las Vegas]" is tagged as the antecedent of the pronoun (c3)"そちら [there]". On the other hand, the anaphor-candidate pair {(c3)"そちら [there]", (r1)"サーカスサーカス [Circus Circus]"} is not corefential, and therefore forms a negative example.

The difficulty of our task can be verified according to the average number of antecedent candidates, i.e., the sum of positive and negative examples, for a given pronoun. In our corpus, the average number is 36.7.

3 Coreference analysis

For the experiments described in Section 5, we utilize a trainable resolution approach using shallow information, i.e., syntactic and semantic word attributes as well as primitive discourse information extracted from a morphological analysis of the input.

To learn the coreferential relationships from our corpus, we use the C5.0 machine learning algorithm (Quinlan, 2000). The set of attributes employed for the decision tree learning consists of discrete and continuous values extracted from the training corpus. Two decision tree classes are used to determine whether there is a coreferential relationship (class: *coref*) or not (class: *no-rel*).

3.1 Training attributes

For the learning of the decision tree we distinguish attributes by the stem forms of content words, their semantic classifications, and their parts-of-speech as illustrated in Table 1.

Table 1: Training attributes

category		sample
content word:		"そちら", "行く"
	semantic code	{name}, {shop}
	part-of-speech	代名詞 [pronoun]
		普通名詞 [common noun]
		本動詞 [verb]
functional word:	particle	"は", "を", "と", "や"
	conjunction	"ので", "たら"
	conjugation	"ない", "れる", "た"
discourse:	distance	(*continuous values*)
	count	(*continuous values*)

Moreover, we use information about syntactical markers like particles or sentence conjunctions as well as primitive discourse information about distances and numbers of occurrences for the determination of coreferential relationships.

content word

For the resolution of pronouns, we check not only which anaphoric expressions are involved, but also the existence of other content words, like sentence predicates, for the respective input sentences.

semantic code
For the semantic classification of content words, we use the *Ruigo-Shin-Jiten*, a three-layered semantic hierarchy (Ohno and Hamanishi, 1981). The top two layers are utilized; they distinguish 100 classes.

part-of-speech
We distinguish 33 parts-of-speech for verbs (e.g., 本動詞, 助動詞, 判定詞), nominal expressions (e.g., 普通名詞, 代名詞), adjectives (e.g., 形容詞, 数詞), and functional words (e.g., 格助詞, 接続詞).

functional word
In Japanese, the grammatical role of specific content words is marked according to particles succeeding the expression. We distinguish *case particles* (e.g., は, が, を, に), conjunction particles (e.g., と, や), and adverbial particles (e.g., とか, など). Moreover, the existence of specific conjunctions (e.g., ながら, ので) and the conjugation form of the sentence predicate, are verified for the determination of coreferential relationships.

discourse
We use information concerned with the number of occurrences of specific content words and their distances in the discourse.

For the training of the decision tree, we provide the complete set of attributes described above. No other coreference indicators are used in our approach, such as the analysis of discourse marker or topic and focus information. This is because these indicators, which were proposed for previous resolution systems, require a more sophisticated linguistic analysis of the input data.

3.2 Learning phase

During the iterative analysis of each dialog, anaphoric expressions are identified according to the assigned coreference tags. Previously mentioned nouns are considered as possible antecedent candidates.

Questions are applied to each anaphor-candidate pair either by matching specified expressions in the respective utterances (discrete values) or by calculating attribute values in the given context (continuous values).

The application of these questions yields a single attribute vector classifying the characteristics of the given reference. In the case of antecedents, this vector is assigned to the coreference class *coref*, whereas a separate class *no-rel* is used for the vectors of non-referential candidates.

The amount of attribute vectors for all of the training samples forms the input of the learning method. By optimizing the entropy value for each subset, the automatic classifier algorithm produces a decision tree that ranks important attributes higher in the tree in order to achieve an early decision about the classification of the input (Quinlan, 1993).

3.3 Application phase

For each anaphoric expression of the test data, a candidate list, i.e., a list of the nominal candidates preceding the anaphoric element in the current discourse, is created. The decision tree classifier is then successively applied to all of the anaphor-candidate pairs.

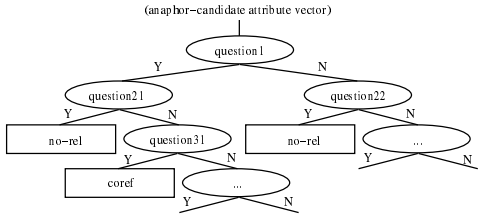

Figure 2: Decision tree classifier

Starting with the top node of the decision tree, the question assigned to this node is tested against the input, i.e., the respective anaphor-candidate attribute vector. Depending on the truth value of the question, the procedure descends to the respective sub-branch. The verification procedure is continued until a leaf containing the classification result (*coref* vs. *no-rel*) is reached (cf. Figure 2).

4 Referential scope

An investigation into the distribution of the relative distances of annotated anaphor-antecedent pairs in the training corpus shows quite different characteristics concerning the referential scope of the respective anaphoric expressions. Each relative distance is mea-

sured as the number of single nouns (candidates) between the anaphor and the antecedent. Figure 3 illustrates the relative distance distributions of the anaphoric expressions (こちら [here], そちら [there], これ [this one], それ [that one], この [this], その [that]) utilized in the experiments described in Section 5. The graphs describe the coverage of the training samples according to the referential scope limitations, i.e., the X-axis shows the percentage of training samples whose relative distance is less than the referential scope value plotted on the Y-axis.

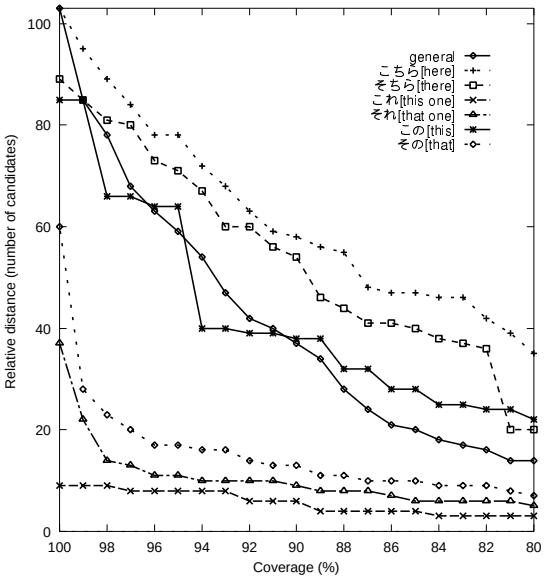

Figure 3: Distribution of referential scope

The distribution of all anaphoric expressions (*general*) shows that we have to verify the last 103 candidates of the history in order to cover all of the coreferential relationships of the training data. However, if we decide to cover only 80% of the data, the analysis scope can be reduced to 13 candidates.

Looking at the relative distance distribution of single anaphora, we can see a similar behaviour for the pronouns こちら [here] (giving a direction or referring to oneself) and そちら [there] (giving a direction or referring to a third person) which is characterized by an even distribution and a large referential scope. On the other hand, the demonstratives これ [this one] and それ [that one] as well as the determiner その [that][2] form a group of anaphoric expressions whose referential scope is quite limited besides some rare exceptions. The determiner この [this] shows a distribution similar to *general* with large differences in the referential scope for a coverage variation between 94% and 99%.

One of the benefits we can reap from this investigation is to know the upper boundary of the analysis scope for each anaphoric expression. The incorporation of this scope limitation into the analysis of the respective anaphor can be seen as a practical extension of the general framework. Here, the decision tree classifier is applied only to those candidates whose relative distance to the anaphoric expression lies within the scope. All candidates beyond this limit are ignored.

The advantage of this approach is a reduction in the size of antecedent candidates, which in turn decreases the *costs* of the analysis process. The scope limitation also prevents the misclassification of non-referential candidates beyond the limit, which in turn increasing the *accuracy* of the decision tree classifier. However, we do have to expect, that at least some of the correct antecedents of the open test data will come to lie outside the analysis scope, i.e., they will be ignored by the classifier, leading to a decrease in the system *recall*, i.e., the proportion of correct antecedents identified correctly.

In Section 5 we try to give an answer to the question of how close can we come to a coreferential classifier with a high accuracy based on referential scope limitations extracted from the training corpus that does not affect the system recall?

5 Evaluation

Five-way cross-validation experiments are conducted for the resolution of 637 input samples consisting of the pronouns most frequently occurring in the corpus (こちら [here]: 142, そちら [there]: 168, これ [this one]: 40, それ [that one]: 136, この [this]: 49, その [that]: 102).

[2]Strictly speaking, the determiners この [this] and その [that] do not belong to the group of Japanese pronouns, but they are anaphoric and therefore included in our investigation.

In order to prove the feasibility of our approach we compare the following classification systems:

- *general*: a single decision tree classifier trained on the input samples of all of the pronouns
- *specific*: decision tree classifiers (one for each pronoun) trained on the input samples of their respective pronoun

Concerning the analysis scope of the above systems we distinguish:

- *history*: all of the candidates preceding the anaphoric expression
- *scope*: the candidates within a relative distance defined as the coverage (in %) of the distance distribution of the training samples

The performance of the baseline system *general+history* and each specific classification system (*specific+history*) are reported in Section 5.2 and utilized for a comparison to those systems with scope limitations, i.e., *general+scope* and *specific+scope*, described in Section 5.3.

5.1 Criteria

For the evaluation of the system performance we calculate the *resolution costs* (i.e., the number of anaphor-candidate attribute vectors (*cases*) to which the decision tree is applied), the *accuracy* of the decision tree classifier (i.e., the proportion of correct classified objects), and the *recall* of the classification algorithm (i.e., the proportion of annotated antecedents (*target cases*) that the system identifies correctly).

Let a denote the number of target cases classified correctly, b the number of non-referential cases classified coreferentially, c the number of target cases classified non-referentially, and d the number of non-referential cases classified correctly as illustrated below.

classification

coref	no-rel	← classified as
a	c	coref
b	d	no-rel

annotation

The *costs*, *accuracy*, and *recall* of the system are defined as follows:

$$costs = a + b + c + d$$

$$accuracy = \frac{a+c}{a+b+c+d}$$

$$recall = \frac{a}{a+c}$$

In the case of a scope limitation all antecedent candidates beyond the limit are not classified by the decision tree. However, for evaluation purposes, we assign the default class *no-rel* to all out-of-scope candidates and modify the evaluation criteria as given below.

out-of-scope *classification*

no-rel	coref	no-rel	← classified as
e	a	c	coref
f	b	d	no-rel

annotation

Here, e denotes the number of correct antecedents dropped due to the scope limitation and f is the number of out-of-scope candidates classified correctly by the default class *no-rel*. In the case of a scope limitation, the evaluation measures of the system are defined as follows:

$$costs_{scope} = a + b + c + d$$

$$accuracy_{scope} = \frac{a+c+f}{a+b+c+d+e+f}$$

$$recall_{scope} = \frac{a}{a+c+e}$$

5.2 General framework

In order to be able to judge the performance of the proposed approach, we utilize the general framework, i.e., the validation of all candidates in the history, for the baseline evaluation. In Table 2 we summarize the accuracy and recall for the open test evaluation of the baseline system (*general+history*) and the anaphor-specific classification system (*specific+history*) trained only on samples of the respective anaphoric expressions.

Table 2: Baseline performance

classification	accuracy	recall
general+history	62.7%	82.1%
specific+history	−7.5%	−5.7%

The baseline accuracy is 62.7% and its recall is 82.1%. However, the application of the specific classification schemes to all candidates in

history results in a performance drop of 7.5% in accuracy and 5.7% in recall.

5.3 Scope limitation

According to the limitation of the analysis scope introduced in Section 4, all candidates beyond the scope limitation are ignored. Figure 4 describes the cost reduction dependent on the coverage of the relative distance distribution of the training data.

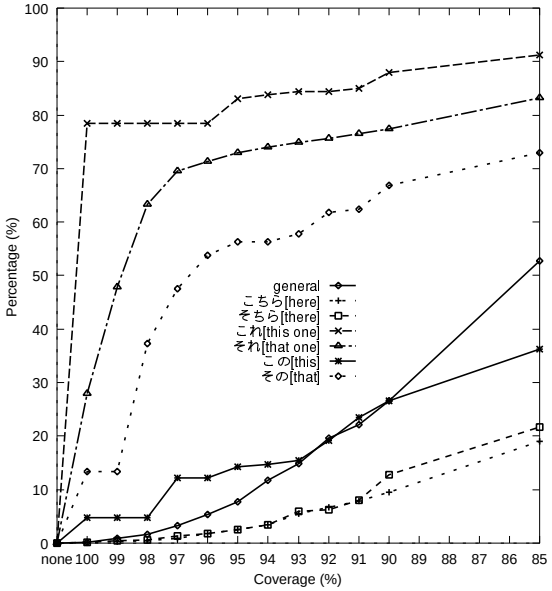

Figure 4: Cost reduction

A sharp increase in the number of out-of-scope candidates can be seen for これ [this one] (∼90%), それ [that one] (∼80%), and その [that] (∼70%). この [this] increases up to 40%. In contrast, almost no reduction in the costs can be achieved for small decreases of the coverage rate for こちら [here] and そちら [there].

Based on these numbers, we can expect a large improvement in accuracy for those classification systems with a high cost reduction rate which Figure 5 verifies.

The largest gain of around 65% in the accuracy rate is achieved for これ [this one] and それ [that one] followed by その [that] with an improvement of 35.7%. The accuracy of この [this] increases by 13.8%. However, the accuracy rates of こちら [here] and そちら [there] do not improve at all.

On the other hand, a large increase in the resolution costs shows that a system comes to be prone to a decrease in the recall performance. Despite large differences in the resolution costs and system accuracy, however, the recall values of all classification system decrease monotonically in a similar way. For the coverage rate of 85% the recall drops by 10-15% (cf. Figure 6).

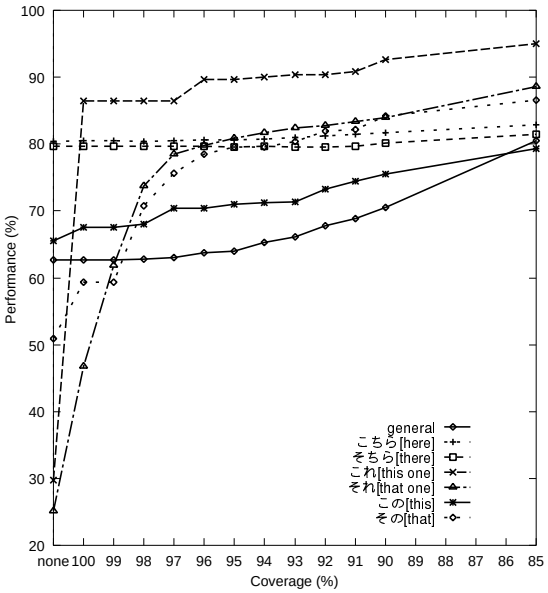

Figure 5: Accuracy

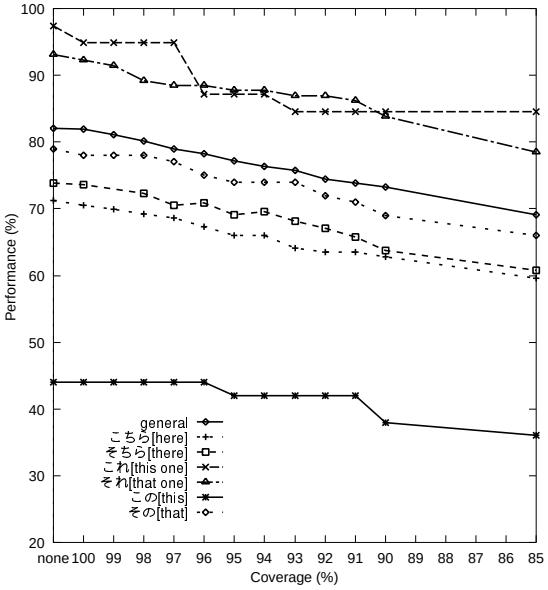

Figure 6: Recall

These results show that our approach is effective for the anaphora これ [this one], それ [that one], この [this], and その [that], but not for the pronouns こちら [here] and そちら [there].

However, it is still an open question as to what are the best scope limit values for achieving our goal of a high accuracy with a minor decrease in the system recall.

The misclassification of non-referential candidates is less harmful than the omission of correct antecedents, because there is no recovery from the latter case; non-referential candidates can still be separated from coreferential ones later on using saliency-based selection or similar schemes.

Therefore, we focus on the regression of the system recall for the selection of the optimal system parameter. In Table 3, we use a threshold (5%) for the maximal recall decrease of each classification system towards its *history* results that we are willing to accept.

Table 3: Effect of scope limitation

classification	(coverage)	costs	accuracy	recall
general+scope	(95%)	−7.7%	+1.3%	−4.9%
こちら [here]	(95%)	−1.7%	+0.2%	−3.9%
そちら [there]	(94%)	−3.4%	+0.0%	−4.2%
これ [this one]	(97%)	−78.5%	+56.8%	−2.5%
それ [that one]	(96%)	−71.4%	+54.7%	−4.6%
この [this]	(91%)	−23.5%	+9.0%	−2.0%
その [that]	(93%)	−57.9%	+29.5%	−2.5%

A threshold larger than 5% causes an increase in the cost reduction, but only a small improvement in the system accuracy that does not warrant a drop in the recall performance anymore. If we do not apply any scope limitations to the resolution of the anaphora こちら [here] and そちら [there], since no gain in accuracy can be achieved, there is no cost reduction, but we can reduce the recall regression of the overall system performance. Table 4 shows the selected coverage rates for the limitation of the analysis scope of the specific decision tree classifiers and its performance.

Table 4: Scope limitation

こちら [here]:	none	そちら [there]:	none
これ [this one]:	97%	それ [that one]:	96%
この [this]:	91%	その [that]:	93%

classification	costs	accuracy	recall
specific+scope	−33.2%	+17.4%	−7.1%

The overall system performance of the classification scheme *specific+scope* is then a cost reduction of 33.2%, an increase of 17.4% in accuracy, and a drop of 7.1% in recall.

6 Related Research

Most of the resolution systems described in literature, focus on the selection of a single history candidate, whereby the *recency* of candidates is frequently utilized as a saliency measure. However, only a few systems try to limit the scope of their resolution modules according to the referential characteristics of the respective anaphoric expressions.

(Kameyama, 1997) introduces a *locality* assumption, which restricts the analysis scope according to the anaphor type.[3] However, these limits are selected arbitrarily by the author. Moreover, the pronominal anaphora contained in the evaluation of thirty newspaper articles (MUC-6 coreference task) consist mainly of 3rd person pronouns with intrasentential references.

(Ide and Cristea, 2000) analyzes the discourse structure of text taken from the MUC corpus in order to determine domains of referential *accessibility* for each referential expression. The search space is reduced by skipping subordinated discourse segments. However, this approach requires an enhanced structural analysis and does not exploit any upper boundary for the maximal referential scope of the respective anaphoric expressions.

7 Conclusion

This paper focuses on the incorporation of referential scope characteristics of anaphora into a corpus-based classification scheme for the resolution of Japanese pronouns. The result of this incorporation is an increase in the classification accuracy and a decrease in the analysis costs as shown in Table 5.

Table 5: System performance

classification	cost reduction	accuracy	recall
general+history	0.0%	62.7%	82.1%
specific+history	0.0%	55.2%	76.4%
general+scope	7.7%	64.0%	77.2%
specific+scope	33.2%	80.1%	75.0%

The accuracy of the baseline system (*general+history*) is 62.7% and its recall is 82.1%. The usage of anaphor-specific classifiers (*specific+history*) results in a lower performance of 55.2% and 76.4%, respectively, because the learned referential characteristics of single anaphora leads to a performance drop when applied to all candidates in history.

[3] Unrestricted for proper nouns, 10 sentences for definite noun phrase references, three sentences for pronouns, and only the current sentence for reflexives.

With a scope limitation applied to the general framework (*general+scope*), we achieve an accuracy of 64.0%, a recall of 77.2%, and a cost reduction of 7.7%. However, the largest improvement in the overall system performance resulting in an accuracy of 80.1%, a recall of 75.0%, and a cost reduction of 33.2%, is achieved by the *specific+scope* approach, i.e., the utilization of anaphor-specific classification systems in combination with analysis scope limitation according to the coverage of the relative distance distribution of the training data.

Large differences in the feasibility of this approach can be seen for the various anaphoric expressions. An investigation into the relative distance distribution of annotated anaphor-antecedent pairs in the training corpus revealed an even distribution with a large referential scope for the pronouns こちら [here] and そちら [there]. Therefore, almost no effect could be achieved through the limitation of the analysis scope, i.e., the validation of the complete history is required in order to achieve a high system performance for the resolution of these anaphoric expressions.

On the other hand, a drastic increase of around 55% in accuracy in combination with a high system recall of 90% (and more) could be achieved for the demonstratives これ [this one] (accuracy: 86.5%, recall: 94.9%) and それ [that one] (accuracy: 89.8%, recall: 88.5%) due to a majority of short-ranged references.

The application of the scope limitation also resulted in a high accuracy of over 75% and a small decrease in the recall of 2% for the determiners この [this] (accuracy: 74.5%, recall: 42.0%) and その [that] (accuracy: 80.4%, recall: 74%).

The system proposed in this paper does not select a single candidate as the antecedent of the anaphoric expression to be resolved, but the high accuracy rates of the system enable a large restriction of the search space, i.e., an identification of around 80% of non-referential candidates, for selection schemes using some kinds of preference measures for the determination of the most salient candidate.

A problem with the current system is the large number (around 20%) of correct antecedents classified as non-referential. One reason for this misclassification is an insufficient amount of training data. We used different numbers of training dialogs (50-400 dialogs) for the training of the decision tree. The steadily increasing performance results implied a lack of training data for the identification of potential candidates. Currently, we are extending our corpus and we expect that a larger number of coreferential variants will lead to an improvement of the system recall.

Moreover, investigations into the feasibility of our approach for languages other than Japanese, e.g. the Englich MUC corpus, will enable us to compare this approach more precisely towards related research.

References

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

N. Ide and D. Cristea. 2000. A Hierarchical Account of Referential Accessibility. In *Proc. of the 38th ACL*, pages 416–424, Hong Kong.

M. Kameyama. 1997. Recognizing Referential Links: An Information Extraction Perspective. In *Proc. of the 36th ACL, Workshop* Operational factors in practical, robust, anaphora resolution for unrestricted texts, pages 46–53.

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

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

M. Paul, K. Yamamoto, and E. Sumita. 1999. Corpus-Based Anaphora Resolution Towards Antecedent Preference. In *Proc. of the 37th ACL, Workshop* Coreference and Its Applications, pages 47–52, Maryland, USA.

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

J. Quinlan. 2000. C50. http://rulequest.com/.

T. Takezawa, T. Morimoto, and Y. Sagisaka. 1998. Speech and language database for speech translation research in ATR. In *Proc. of Oriental COCOSDA Workshop'98*, pages 148–155.