TR-IT-0327

Resolution of Referential Expressions within TDMT

Michael PAUL Kazuhide YAMAMOTO

2000.01

Abstract

This report deals with the processing of contextual phenomena within the framework of the spoken-language translation system TDMT (Transfer-Driven Machine Translation). We apply a corpus-based approach to resolve referential expressions in Japanese utterances. In our approach a machine-learning algorithm (decision tree) is utilized to select automatically the attributes from a tagged training set necessary for the resolution task. The task-specific decision tree is applied to the input data and the knowledge about the obtained reference objects is used for a context-adopted translation of the source utterances into English and German.

ATR Interpreting Telecommunications Research Laboratories

Contents

1 Introduction

Contextual processing is not peculiar to spoken-language, but demands for contextual processing are usually higher because the dialogue attendants tend to use many anaphoric or elliptical expressions for information that is mutually understood. In the case of ellipsis the source language doesn't express the subject or other grammatical cases and the target must express it. Moreover, the correct analysis of coreferences is essential to avoid misinterpretations and to allow context-sensitive translations.

In our spoken-language machine translation system [Sumita et al. 99], we apply a corpus-based approach to resolve referential expressions in order to achieve context-adopted translations of the source utterances.

2 Data Corpus

We use the *ATR-ITL Speech and Language Database* [Takezawa et al. 98] containing 500 annotated Japanese spoken-language dialogs. It includes tags for subject ellipsis (8291 samples) as well as nominal (2160 samples), pronominal (526 samples), and ellipsis (3843 samples) coreferential annotations, whereby the anaphoric expressions are limited to those referring to nominal antecedents.

Besides the referential 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 & Hamanishi 81] 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 1: Example dialog

In the example dialog between the hotel reception (r) and a customer (c) listed in Figure 1 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)"Ø[it]" referring to (r2)"スペル [spelling]".

2.1 Ellipsis Tagging

In order to train and evaluate our ellipsis resolver, we tagged some ellipsis types to a dialogue corpus. The ellipsis types used to tag the corpus are shown in Table 1.

Table 1: Tagged Ellipsis Types

Tag	Meaning
<1sg>	first person, singular
<1pl>	first person, plural
<2sg>	second person, singular
<2pl>	second person, plural
<g>	person(s) in general
<a>	anaphoric

Each ellipsis marker is tagged at the predicate. We made a distinction between first or second person and person(s) in general. Note that *'person(s) in general'* refers to either an unidentified or an unspecified person or persons. In Far-Eastern languages such as Japanese, Korean, and Chinese, there is no grammatically obligatory case such as the subject in English. It is thus necessary to distinguish such ellipses.

2.2 Anaphora Tagging

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 2: 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 2 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 5.2.

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 1 are d_{utter}=4, d_{clause}=7, d_{cand}=14, d_{morph}=40.

3 Decision Tree

For the resolution of referential expressions in our system we utilize a decision-tree learning approach. To learn the referential relations from our corpus we have chosen a C4.5[1]-like machine learning algorithm without pruning.

The input of the learning algorithm is task-specific and consists of feature vectors representing the attribute values of the morphologically analyzed entities of the training data whereby each utterance is subdivided into clauses.

In the ellipsis task a feature vector is extracted for each single clause and the assigned class characterizes the respective type of the subject filler.

Anaphora resolution, however, requires an iterative analysis of each dialog. Each clause is checked successively for anaphoric expressions. Questions are applied by either matching specified expressions in the respective clause (discrete values) or calculating attribute values in the current context (continuous values). The application of question sets to anaphor, candidate and clause constituents yields in a single vector classifying the relevant features for the given reference. In the case of antecedents this vector is assigned to the coreference class, whereas vectors of all non-tagged candidates form a separate class classifying non-referential relations.

The number of feature vectors for all training samples forms the input of the learning method. By optimizing the entropy value for each subset, the automatic classifier algorithm produces a binary decision tree ranking important features higher in the tree in order to achieve an early decision about the classification of the specified input.

In order to apply a decision tree to a given input in the resolution phase the question assigned to each node is tested against the feature vector of the input. Depending on the truth value of the question, the procedure descends to the determined sub-branch of the decision tree. The verification procedure is continued until a leaf containing the classification result is reached.

4 Ellipsis Resolution

Parts of utterances are often omitted in languages such as Japanese, Korean, and Chinese. In contrast, many Western languages such as English and German do not generally permit these omissions. Such ellipses must be resolved in order to translate the former languages into the latter.

Consider the Japanese utterance in sample (1):

customer: ∅ 奈良ホテル に 滞在 しています。
[I am staying at the Nara Hotel.] (1)

The subject is omitted in the above utterance, i.e., it is not explicitly expressed who stays at the Nara Hotel. However, native speakers understand that it is the speaker of the utterance who stays there.

In order to determine the subject of the utterance, it is necessary to consider various information surrounding the utterance, i.e.,

[1] cf. [Quinlan 93]

Table 3: Number of training attributes

Attributes	Num.
Content words (predicate)	100
Content words (case frame)	100
Func. words (case particle)	9
Func. words (conj. particle)	21
Func. words (auxiliary verb)	132
Func. words (other)	4
Exophoric information	1
Total	367

- the utterance has auxiliary verbs "- てい(る)-" and "- ます,"
- the utterance is declarative,
- the speaker of the utterance is a customer, and
- the agent of "滞在$_{[stay]}$" is a customer in most cases.

We have to determine the subject by considering the above elements in parallel. A manual rule construction of ellipsis resolution is a difficult and time-consuming task. With this in mind, a machine-learning approach has been utilized. Since various elements should be considered in resolving ellipses, it is difficult to exactly determine their relative degrees of influence. However, building a decision tree using a tagged training set automatically gives weight to every element through the criterion of entropy.

In our approach the machine-learning algorithm is used to select the attributes necessary for the ellipsis resolution. A decision tree is built, and used as the actual ellipsis resolver [Yamamoto & Sumita 98].

The training attributes that we prepared for Japanese ellipsis resolution are listed in Table 3. The training attributes in the table are classified into the following three groups:

- Exophoric information:
  Speaker's social role.
- Topic-dependent information:
  Predicates and their semantic categories.
- Topic-independent information:
  Functional words which express tense, modality, etc.

There is one approach that only uses topic-independent information to resolve ellipses that appear in dialogues. However, we took the position that both topic-dependent and -independent information should have different knowledge.

Thus, approaches utilizing only topic-independent knowledge must have a performance limit for developing an ellipsis resolution system. It is practical to seek an automatically trainable system that utilizes both types of knowledge.

The effective use of exophoric information, i.e., from the actual world, may perform well for resolving an ellipsis. Exophoric information consists of a lot of elements, such as the time, the place, the speaker, and the listener of the utterance. However, it is difficult to become aware of some of them, and some are rather difficult to prescribe. Thus we utilize one element, the speaker's social role, i.e., whether the speaker is the customer or the clerk. The reason for this is that it must be an influential attribute, and it is easy to detect in the actual world. Many of us would accept a real system such as a spoken-language translation system that detects speech with independent microphones.

It is generally agreed that attributes to resolve ellipses should be different in each case. Thus although we have to prepare them on a case by case basis, we trained a resolver with the same attributes.

Because we must deal with the noisy input that appears in real applications, the training attributes, other than the speaker's social role, are questioned on a morphological basis. We give each attribute its positional information, i.e., search space of morphemes from the target predicate. Positional information can be one of five kinds: before, at the latest, here, next, and afterward. For example, a case particle is given the position of '*before*', the search position of a prefix '*o-*' or '*go-*' is the '*latest*', and an auxiliary verb is '*after*' the predicate. The attributes of predicates, and their semantic categories are placed in '*here*'.

For predicate semantics, we utilized the top two layers of *Kadokawa Ruigo Shin-Jiten* [Ohno & Hamanishi 81], a three-layered hierarchical Japanese thesaurus.

5 Anaphora Resolution

For the resolution of coreferential relationship we proposed a corpus-based anaphora resolution method that combines a machine learning algorithm with a statistical preference scheme [Paul et al. 99].

Based on the corpus annotations, we extract the frequency information of coreferential anaphora-antecedent pairs and non-referential pairs as well as the relative distance between the anaphora and the candidates from the training data.

This knowledge is utilized to train a decision tree on the determination of coreferential relationship for a given anaphora and an antecedent candidate. Thus, the relevance of the respective features for the resolution task is automatically extracted from the training data.

In our resolution approach, we argue for a separation of the analysis of coreferential relationships and the determination of the most salient candidate as listed in Figure 2.

In the first step, we apply the decision tree as a coreference filter (cf. section 5.1) to all possible anaphora-candidate pairs $(A_i + C_{ij})$ in the discourse. In this step, irrelevant candidates are filtered out to reduce noise for the preference

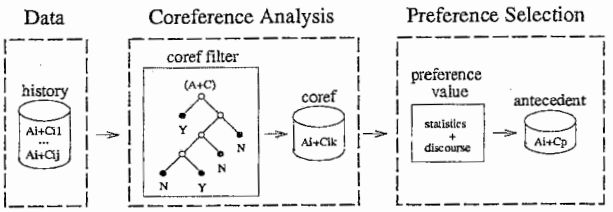

Figure 2: System outline

selection algorithm.

In the second step, the reduced set $(A_i + C_{ik})$ forms the input of the preference algorithm which selects the most salient candidate C_p by taking into account (I) the frequency information of the coreferential and non-referential pairs that were tagged in the training corpus and (II) the distance features within the current discourse (cf. section 5.2).

customer: (i) あした から 一週間 車 を借りたいんです。
[I would like to rent a car for one week from tomorrow on.]
(ii) 奈良ホテル に滞在しています。
[I am staying at the Nara Hotel.]
(iii) こちら の 部屋 の 番号 は 四〇七 電話番号 は 零七四二二二五五一五 です。
[The room number here is 407 and the telephone number is 0742-22-5515.]
clerk: (iv) そうですか 奈良ホテル でしたら ∅ お持ちできますよ。
[I see. We can bring it to the Nara Hotel.]

(2)

Sample (2) contains two anaphoric expressions, i.e. (I) the pronoun こちら$_{[here]}$ in utterance (iii), which refers to the proper noun 奈良ホテル$_{[Nara\ hotel]}$, and (II) the omitted direct object (ellipsis) ∅ of utterance (iv), which refers to 車$_{[car]}$ in utterance (i). The underlined nominal expressions preceding the respective anaphora in the discourse form the set of possible candidates.

In the case of the pronominal anaphora こちら$_{[here]}$ of this sample, it is sufficient to resolve the antecedent as the most recent candidate in the discourse. However, this straightforward resolution scheme has a low success rate due to its application to the unfiltered set of candidates resulting in the frequent selection of non-referential antecedents.

For example, the set of possible antecedents for the ellipsis anaphora in utterance (iv) consists of the ten underlined nominal expressions above. The most recent one is 奈良ホテル$_{[Nara\ hotel]}$, which should not be considered as the direct object of the transitive verb 持つ$_{[to\ bring]}$, because of its semantic attributes. In this example, the coreference filter successfully reduces the candidate set to two potential candidates, i.e. 番号$_{[number]}$ and 車$_{[car]}$.

Our preference selection scheme assigns a saliency value to the remaining candidates. This value is based on the occurrence of similar coreferences in the training data as well as the relative position of the respective candidate in the current discourse defining a balance between frequency statistics and recency constraints. Therefore, the candidate 車$_{[car]}$ is selected correctly in our

example as the antecedent of the omitted direct object instead of the more recent candidate 番号$_{[number]}$.

5.1 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.

5.2 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 6.

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.

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]:

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

[2]cf. [Quinlan 93]

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

[4]In order to keep the formula simple the frequency types are omitted (cf. Table 2)

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.

$$pref = \frac{ratio + 1}{dist}$$

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 6.2 we adopted the selection of the dialog-initial candidate as the backup strategy.

6 Evaluation

For the evaluation of the experimental results described in this section we use *F-measure* metrics calculated by the *recall* (R) and *precision* (P) of the system performance as listed in equation 3.

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

We conducted experiments on utterances that had not been subjected to decision-tree learning.

6.1 Ellipsis

In this section we discuss the feasibility of the ellipsis resolver via a decision tree in detail from three points of view: the amount of training data, the topic dependency, and the case difference. The first two are discussed against '*ga*(v.)' case (see subsection 6.1.3).

Recall and Precision for ellipsis of type i is defined as follows:

$$R_i = \frac{number\ of\ correct\ outputs\ for\ ellipsis\ of\ type\ i}{number\ of\ ellipsis\ of\ type\ i}$$

$$P_i = \frac{number\ of\ correct\ outputs\ for\ ellipsis\ of\ type\ i}{number\ of\ outputs\ of\ ellipsis\ of\ type\ i}$$

Table 4: Training size and performance

Dial.	Samp.	<1sg>	<2sg>
25	463	71.0	55.6
50	863	76.4	69.7
100	1710	82.1	76.4
200	3448	85.1	79.8
400	6906	84.7	81.1

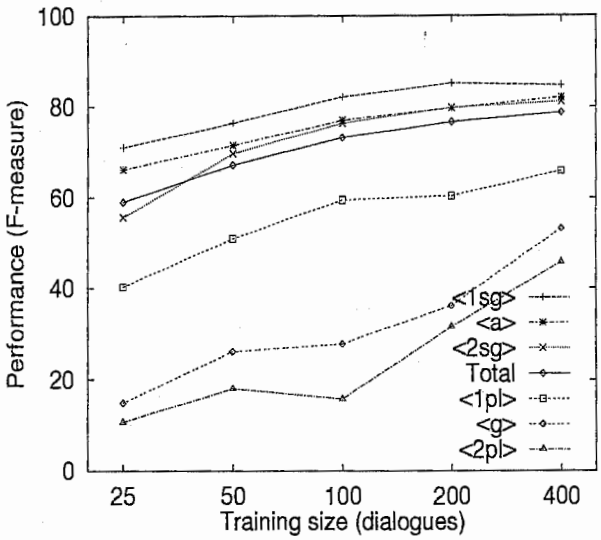

Figure 3: Training size and performance

6.1.1 Amount of Training Data

We trained decision trees with a varied number of training dialogues, namely 25, 50, 100, 200 and 400 dialogues, each of which included a smaller set of training dialogues. The experiment was done with 100 test dialogues (1685 subject ellipses), and none were included in the training dialogues.

Table 4 indicates the training size and performance calculated by F-measure. This illustrates that the performance improves as the training size increases in all types of ellipses. Although it is not shown in the table, we note that the results in both recall and precision improve continuously as well as those in F-measure.

The performance difference of all ellipsis types by training size is also plotted in Figure 3 on a semi-logarithmic scale. It is interesting to see from the figures that the rate of improvement gradually decelerates and that some of the ellipsis types seem to have practically stopped improving at around 400 training dialogues (6806 samples). [Aone & Bennett 95] claimed that the over-

all anaphora resolution performance seems to have reached a plateau at around 250 training examples. This result, however, indicates that $10^4 \sim 10^5$ training samples would be enough to train the trees in this task.

The chart gives us more information that performance limitation with our approach would be $80\% \sim 85\%$ because each ellipsis type seems to approach the similar value, in particular for those in large training samples <1sg> and <2sg>. Greater performance improvement is expected by conducting more training in <2pl> and <g>.

6.1.2 Topic Dependencies

It is completely satisfactory to build resolution knowledge only with topic-independent information. However, is it practical? We will discuss this question by conducting a few experiments.

The utilized ATR travel arrangement corpus contains dialogues exchanged between two people. Various topics of travel arrangements such as immigration, sightseeing, shopping, and ticket ordering are included in the corpus. A dialogue consists of 10 to 30 exchanges. We classified dialogues of the corpus into four topic categories:

H_1 Hotel room reservation, modification and cancellation

H_2 Hotel service inquiry and troubleshooting

H_R Other hotel arrangements, such as hotel selection and an explanation of hotel facilities

R Other travel arrangements

Fifty dialogues were chosen randomly from the corpus in the topic category H_1, H_2, R, and the overall topic $T(= H_1 + H_2 + H_R + R)$ as training dialogues. We used 100 unseen dialogues as test samples again, which were the same as the samples used in the training-size experiment.

Table 5: Topic dependency

Train/Test	$/H_1$	$/H_2$	$/H_R$	$/R$	Total
(%)	20.1	27.7	11.2	40.9	100.0
$H_1/$	78.1	55.9	65.3	61.6	63.7
$H_2/$	71.3	67.0	62.6	62.6	65.6
$R/$	75.1	61.7	61.1	75.4	69.9
$T/$	73.4	62.5	62.6	66.2	66.2
$T - H_R/$	73.7	61.9	59.5	63.9	64.8

Table 5 shows the topic-dependency of each topic category that we provide with the F-measure. For instance, the first figure in the '$T/$' row (73.4) denotes that the accuracy with the F-measure is 73.4% against topic H_1 test samples when training is conducted on T, i.e., all topics. Note that the second row of

the table indicates the ingredient of each topic in the test samples (and thus, the corpus).

The results illustrate that very high accuracy is obtained when a training topic and a test topic coincide. This implies the importance not to train dialogues of unnecessary topics if the resolution topic is imaginable or restricted, in order to obtain higher performance. Among four topic subcategories, topic R shows the highest accuracy (69.9%) in total performance. The reason is not that topic R has something important to train, but that topic R contains the most test dialogues chosen at random.

The table also illustrates that a resolver trained in various kinds of topics ('$T/$') demonstrates higher resolving accuracy against the testing data set. It performs with better than average accuracy in every topic compared to one which is trained in a biased topic. By looking at some examples it may be possible to build an all-around ellipsis resolver, but topic-dependent features are necessary for better performance. The '$T - H_R/$' resolver shows the lowest performance (59.5%) against '$/H_R$' test set. This result is more evidence supporting the importance of topic-dependent features.

6.1.3 Difference in Surface Case

We examined the feasibility of a machine-learned ellipsis resolver for three principal surface cases in Japanese, '*ga*', '*wo*', and '*ni*'[5]. Roughly speaking, they express the subject, the direct object, and the indirect object of a sentence respectively. We classified the '*ga*' case into two samples: a predicate of a sentence with a '*ga*' case ellipsis that is a verb or an adjective. In other words, this distinction corresponds to whether a sentence in English is a *be*-verb or a general-verb sentence. Henceforth, we call them '*ga*(v.)' and '*ga*(adj.)' respectively.

The training attributes provided are the same in all surface cases. They are listed in Table 3. In the experiment, 300 training dialogues and 100 unseen test dialogues were used. The following results are shown in Table 6[6].

Table 6: Performance of major types in case

Case	<1sg>	<2sg>
ga(adj.)	58.3	68.1
wo	66.7	—
ni	95.2	95.7
ga(v.)	84.7	81.1

The table illustrates that the *ga*(adj.) resolver has a similar performance to the *ga*(v.) resolver, whereas the former has a distinctive tendency toward the latter in each ellipsis type. The *ga*(adj.) case resolver produces unsatisfactory results in <1sg> and <2sg> ellipses, since insufficient samples appeared in the training set.

[5] We cannot investigate other optional cases due to a lack of samples.

[6] The result of the *ga*(v.) case is the same as '400' in Table 4.

In the '*wo*' case, more than 90% of the samples are tagged with <a>, thus they are easily recognized as anaphoric.

It is important to note that a satisfactory performance is presented for the '*ni*' case (mostly indirect object). One reason for this could be that many indirect objects refer to exophoric persons, and thus an approach utilizing a decision tree that makes a selection from fixed decision candidates is suitable for '*ni*' resolution.

6.2 Anaphora

The preliminary experiments reported in this section are conducted for pronominal anaphora, limited to the frequent ones (それ, これ, そちら, こちら, そこ) occurring in our training data.

Let $\sum_t$ denote the total number of tagged 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 *recall* (R_a) and *precision* (P_A) are defined as follows:

$$R_a = \frac{\sum_c}{\sum_t} \qquad \sum_c = |f|$$

$$P_a = \frac{\sum_c}{\sum_f} \qquad \sum_f = |f| + |i|$$

$$\sum_t = |f| + |i| + |o|$$

In order to prove the feasibility of our approach we compare the four preference selection methods listed in Figure 4.

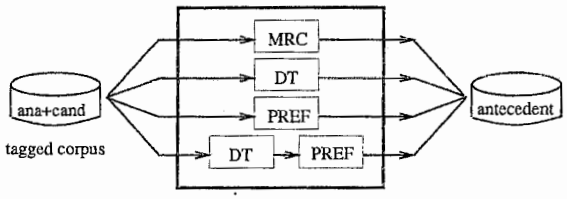

Figure 4: Preference selection experiments

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.

6.2.1 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 5.

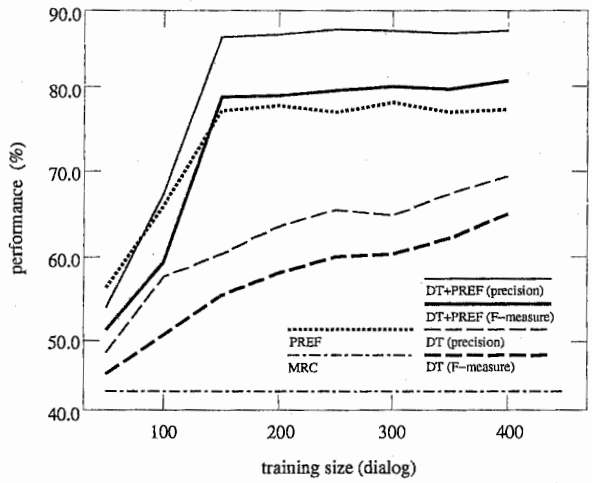

Figure 5: Training size versus performance

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 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.

6.2.2 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 7.

Table 7: 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

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. 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.

7 Incorporation into TDMT

TDMT achieves a multi-lingual spoken-language translation, which is based on a *constituent boundary parsing* method (CBP) in an example-based framework [Furuse & Iida 96]. The input sentence is incrementally parsed by matching meaningful units of linguistic structure (patterns) using a chart parsing algorithm. Given a set of translation examples TDMT tries to find the "closest" examples to the structured input using a *semantic distance calculation* (SDC) [Sumita & Iida 92]. By simulating the translation of the closest examples the

empirical transfer knowledge is applied to the analyzed source structure, resulting in a corresponding target structure, which is used to generate the translation (cf. Figure 6).

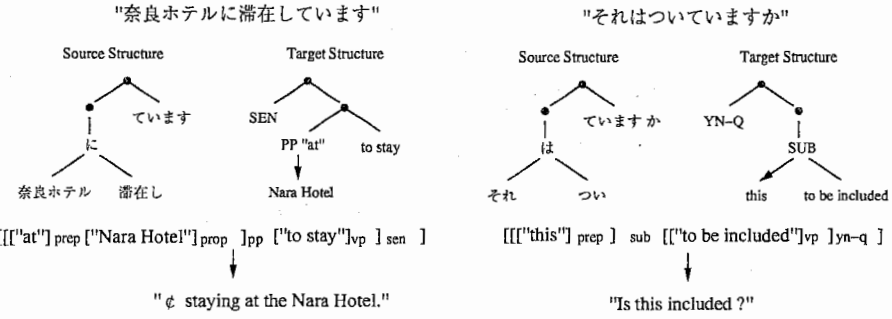

Figure 6: Translation examples

So far we used an *adhoc* solution for the handling of referential expressions, namely the introduction of default pronominal entities during the application of transfer knowledge based on syntactic restrictions of the sentence predicate, which are used during generation, whenever such an entity is missing in the transfer result. Thus the examples would be translated as "You are staying at the Nara Hotel." and "Is this included ?" , respectively.

The resolution modules described in this report can be incorporated into the TDMT framework as listed in Figure 7.

The input of the resolution modules depends on the characteristics of the respective tasks (cf. section 3). The output consists of a link to the reference object and is further utilized in the generation module.

In the case of the subject resolution, all features of the morphological analyzed utterance form the input of the ellipsis resolver. By parsing down the decision tree according to these attributes the resolution result is obtained as the ellipsis tag contained in the parse leaf. The omitted subject is than recovered by adding a corresponding pronominal expression to the target structure. In the case of the example "奈良ホテルに滞在しています" the analysis of the features described in section 4 leads to decision tree answer "<1sg>", which enables us to translate the utterance correctly as "I am staying at the Nara Hotel."

In the anaphora case, all possible candidates has to be memorized within in the dialog history. For a given anaphoric expressions the input of the resolution system consists of pairs of this expression with all elements in the candidate history. After filtering out non-referential anaphor-candidate pairs the most salient candidate is selected as the antecedent and used during the generation process to achieve a context-adopted translation[7]. Given the knowledge about the reference object we could replace the pronoun by a nominal expression in order to avoid human misinterpretations. But even if we still generate a pronominal expression, in languages with rich inflectional phenomena like German, the surface word has to agree with the inflectional characteristics of the

[7]Due to a heavy computational load caused by the required dialog management the anaphora resolution component is so far only incorporated into an experimental system

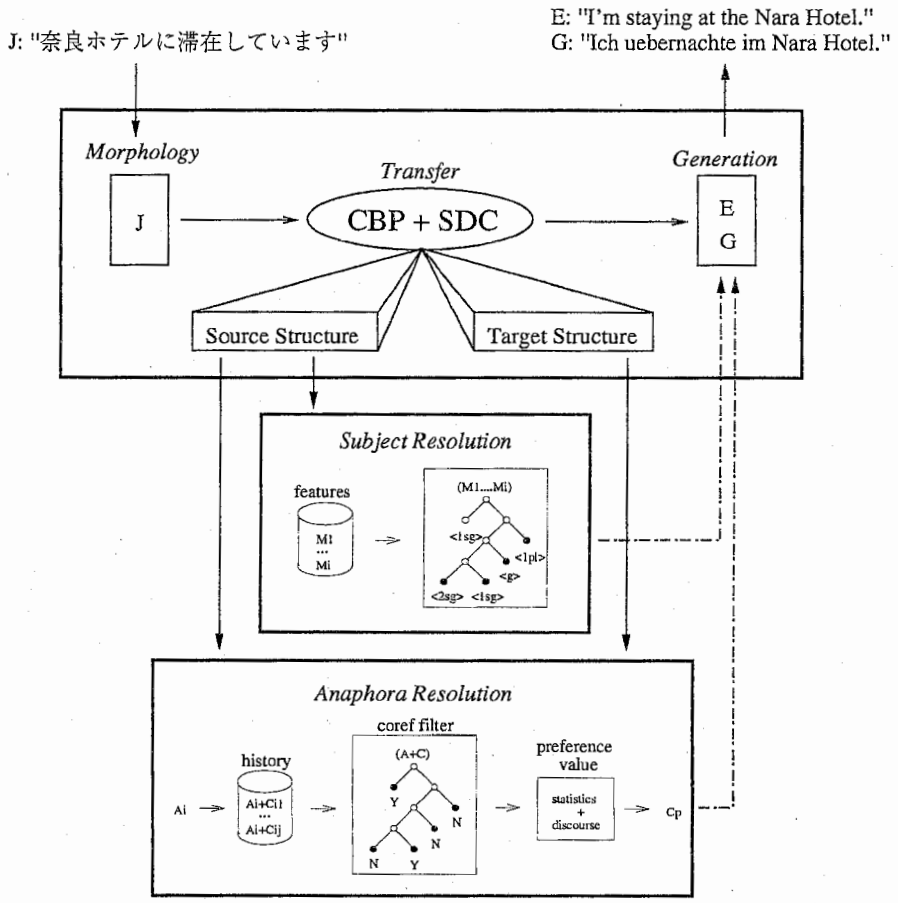

Figure 7: Resolution of referential expressions in TDMT

antecedent. For example, the German translation of the utterance "それはついていますか" would be translated as "Ist *es* enthalten", whereby "es" referes to an unidentified entity. However, in the context of "それ" $\overset{coref}{\rightarrow}$ "税金" (in German: "Steuer", female) we have to generate the female pronoun "sie", resulting in "Ist *sie* enthalten" ("is it included?"), in order to enable its correct interpretation by the hearer.

8 Conclusion

The ellipsis resolution module is incorporated into the TDMT system resolving omitted subjects for the translation of Japanese utterances into English and German. By investigating the decision tree of our experiments we found that topic-dependent attributes are necessary to obtain high performance resolution, and that indispensable attributes vary according to the grammatical case.

Even if the anaphora resolution component is not yet fully incorporated into the TDMT system, experimental results give rise to suspicion that the problem of corpus size is more severe for the anaphora task than for subject resolution. However, despite the lack of training data, 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.

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 performance.

Moreover, applying the proposed resolution approaches to other corpora, e.g. MUC tasks, should give us more insight in domain and language dependency of these modules.

Additionally, we are also developing TDMT for Japanese-to-Chinese, for which the resolution modules described in this report should also be incorporated in the near future.

References

[Aone & Bennett 95] C. Aone and S. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proc. of the 33th ACL*, p. 122–129.

[Furuse & Iida 96] O. Furuse and H. Iida. 1996. Incremental translation utilizing constituent boundary patterns. In *Proc. of the 16th COLING*, p. 412–417, Copenhagen, Denmark.

[Ohno & Hamanishi 81] S. Ohno and M. Hamanishi. 1981. *Ruigo-Shin-Jiten.* Kadokawa.

[Paul et al. 99] M. Paul, K. Yamamoto, and E. Sumita. 1999. Corpus-based anaphora resolution towards antecedent preference. In *Proc. of the 37th ACL, Workshop* Coreference and It's Applications, p. 47–52, Maryland.

[Quinlan 93] J. Quinlan. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

[Sumita et al. 99] E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai. 1999. Solutions to problems inherent in spoken-language translation: The atr-matrix approach. In *Proc. of the Machine Translation Summit VII*, p. 229–235, Singapore.

[Sumita & Iida 92] E. Sumita and H. Iida. 1992. Example-based transfer of japanese adnominal particles into english. *IEICE Trans*, E75-D, No.4:585–594.

[Takezawa et al. 98] 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*, p. 148–155.

[Yamamoto & Sumita 98] K. Yamamoto and E. Sumita. 1998. Feasibility study for ellipsis resolution in dialogues by machine-learning technique. In *Proc. of the 17th COLING*, p. 1428–1434, Montreal, Canada.