# Incorporating Extra-linguistic Information into Reference Resolution in Collaborative Task Dialogue

**Ryu Iida** **Shumpei Kobayashi** **Takenobu Tokunaga**
Tokyo Institute of Technology
2-12-1, Ôokayama, Meguro, Tokyo 152-8552, Japan
{ryu-i,skobayashi,take}@cl.cs.titech.ac.jp

## Abstract

This paper proposes an approach to reference resolution in situated dialogues by exploiting extra-linguistic information. Recently, investigations of referential behaviours involved in situations in the real world have received increasing attention by researchers (Di Eugenio et al., 2000; Byron, 2005; van Deemter, 2007; Spanger et al., 2009). In order to create an accurate reference resolution model, we need to handle extra-linguistic information as well as textual information examined by existing approaches (Soon et al., 2001; Ng and Cardie, 2002, etc.). In this paper, we incorporate extra-linguistic information into an existing corpus-based reference resolution model, and investigate its effects on reference resolution problems within a corpus of Japanese dialogues. The results demonstrate that our proposed model achieves an accuracy of 79.0% for this task.

## 1 Introduction

The task of identifying reference relations including anaphora and coreferences within texts has received a great deal of attention in natural language processing, from both theoretical and empirical perspectives. Recently, research trends for reference resolution have drastically shifted from handcrafted rule-based approaches to corpus-based approaches, due predominately to the growing success of machine learning algorithms (such as Support Vector Machines (Vapnik, 1998)); many researchers have examined ways for introducing various linguistic clues into machine learning-based models (Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002; Yang et al., 2003; Iida et al., 2005; Yang et al., 2005; Yang et al., 2008; Poon and Domingos, 2008, etc.). Research has continued to progress each year, focusing on tackling the problem as it is represented in the annotated data sets provided by the Message Understanding Conference (MUC)[1] and the Automatic Content Extraction (ACE)[2]. In these data sets, coreference relations are defined as a limited version of a typical coreference; this generally means that only the relations where expressions refer to the same named entities are addressed, because it makes the coreference resolution task more information extraction-oriented. In other words, the coreference task as defined by MUC and ACE is geared toward only identifying coreference relations anchored to an entity within the text.

In contrast to this research trend, investigations of referential behaviour in real world situations have continued to gain interest in the language generation community (Di Eugenio et al., 2000; Byron, 2005; van Deemter, 2007; Foster et al., 2008; Spanger et al., 2009), aiming at applications such as human-robot interaction. Spanger et al. (2009) for example constructed a corpus by recording dialogues of two participants collaboratively solving the Tangram puzzle. The corpus includes extra-lingustic information synchronised with utterances (such as operations on the puzzle pieces). They analysed the relations between referring expressions and the extra-linguistic information, and reported that the pronominal usage of referring expressions is predominant. They also revealed that the multi-modal perspective of reference should be dealt with for more realistic reference understanding. Thus, a challenging issue in reference resolution is to create a model bridging a referring expression in the text and its object in the real world. As a first step, this paper focuses on incorporating extra-linguistic information into an existing corpus-based approach, taking Spanger et al. (2009)'s REX-J corpus[3] as the data set. In our

[1] www-nlpir.nist.gov/related_projects/muc/
[2] www.itl.nist.gov/iad/mig//tests/ace/
[3] The corpus was named REX-J after their publication of



*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1259–1267,
Uppsala, Sweden, 11-16 July 2010. 

problem setting, a referent needs to be identified by taking into account extra-linguistic information, such as the spatiala relations of puzzle pieces and the participants' operations on them, as well as any preceding utterances in the dialogue. We particularly focus on the participants' operation of pieces and so introduce it as several features in a machine learning-based approach.

This paper is organised as follows. We first explain the corpus of collaborative work dialogues in Section 2, and then present our approach for identifying a referent given a referring expression in situated dialogues in Section 3. Section 4 shows the results of our empirical evaluation. In Section 5 we compare our work with existing work on reference resolution, and then conclude this paper and discuss future directions in Section 6.

## 2 REX-J corpus: a corpus of collaborative work dialogue

For investigating dialogue from the multi-modal perspective, researchers have developed data sets including extra-linguistic information, bridging objects in the world and their referring expressions. The COCONUT corpus (Di Eugenio et al., 2000) is collected from keyboard-dialogues between two participants, who are collaborating on a simple 2D design task. The setting tends to encourage simple types of expressions by the participants. The COCONUT corpus is also limited to annotations with symbolic information about objects, such as object attributes and location in discrete coordinates. Thus, in addition to the artificial nature of interaction, such as using keyboard input, this corpus only records restricted types of data.

On the other hand, though the annotated corpus by Spanger et al. (2009) focuses on a limited domain (i.e. collaborative work dialogues for solving the Tangram puzzle using a puzzle simulator on the computer), the required operations to solve the puzzle, and the situation as it is updated by a series of operations on the pieces are both recorded by the simulator. The relationship between a referring expression in a dialogue and its referent on a computer display is also annotated. For this reason, we selected the REX-J corpus for use in our empirical evaluations on reference resolution. Before explaining the details of our evaluation, we sketch out the REX-J corpus and some of its prominent statistics.

Spanger et al. (2009), which describes its construction.

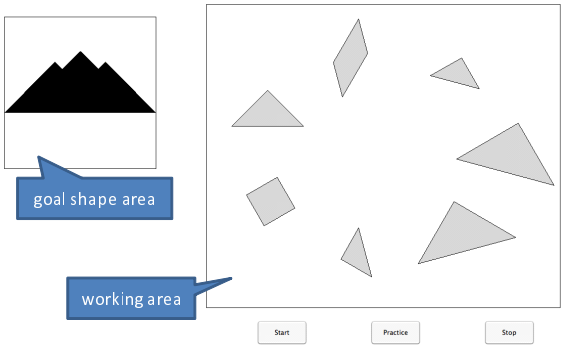


Figure 1: Screenshot of the Tangram simulator

### 2.1 The REX-J corpus

In the process of building the REX-J corpus, Spanger et al. (2009) recruited 12 Japanese graduate students (4 females and 8 males), and split them into 6 pairs. All pairs knew each other previously and were of the same sex and approximately the same age. Each pair was instructed to solve the Tangram puzzle. The goal of the puzzle is to construct a given shape by arranging seven pieces of simple figures as shown in Figure 1. The precise position of every piece and every action that the participants make are recorded by the Tangram simulator in which the pieces on the computer display can be moved, rotated and flipped with simple mouse operations. The piece position and the mouse actions were recorded at intervals of 10 msec. The simulator displays two areas: a goal shape area (the left side of Figure 1) and a working area (the right side of Figure 1) where pieces are shown and can be manipulated.

A different role was assigned to each participant of a pair: a *solver* and an *operator*. Given a certain goal shape, the solver thinks of the necessary arrangement of the pieces and gives instructions to the operator for how to move them. The operator manipulates the pieces with the mouse according to the solver's instructions. During this interaction, frequent uttering of referring expressions are needed to distinguish the pieces of the puzzle. This collaboration is achieved by placing a set of participants side by side, each with their own display showing the work area, and a shield screen set between them to prevent the operator from seeing the goal shape, which is visible only on the solver's screen, and to further restrict their

interaction to only speech.

### 2.2 Statistics

Table 1 lists the syntactic and semantic features of the referring expressions in the corpus with their respective frequencies. Note that multiple features can be used in a single expression. This list demonstrates that 'pronoun' and 'shape' features are frequently uttered in the corpus. This is because pronominal expressions are often used for pointing to a piece on a computer display. Expressions representing 'shape' frequently appear in dialogues even though they may be relatively redundant in the current utterance. From these statistics, capturing these two features can be judged as crucial as a first step toward accurate reference resolution.

## 3 Reference Resolution using Extra-linguistic Information

Before explaining the treatment of extra-linguistic information, let us first describe the task definition, taking the REX-J corpus as target data. In the task of reference resolution, the reference resolution model has to identify a referent (i.e. a piece on a computer display)[4]. In comparison to conventional problem settings for anaphora resolution, where the model searches for an antecedent out of a set of candidate antecedents from preceding utterances, expressions corresponding to antecedents are sometimes omitted because referring expressions are used as deixis (i.e. physically pointing to a piece on a computer display); they may also refer to a piece that has just been manipulated by an operator due to the temporal salience in a series of operations. For these reasons, even though the model checks all candidates in the preceding utterances, it may not find the antecedent of a given referring expression. However, we do know that each referent exists as a piece on the display. We can therefore establish that when a referring expression is uttered by either a solver or an operator, the model can choose one of seven pieces as a referent of the current referring expression.

### 3.1 Ranking model to identify referents

To investigate the impact of extra-linguistic information on reference resolution, we conduct an empirical evaluation in which a reference resolution model chooses a referent (i.e. a piece) for a given referring expression from the set of pieces illustrated on the computer display.

As a basis for our reference resolution model, we adopt an existing model for reference resolution. Recently, machine learning-based approaches to reference resolution (Soon et al., 2001; Ng and Cardie, 2002, etc.) have been developed, particularly focussing on identifying anaphoric relations in texts, and have achieved better performance than hand-crafted rule-based approaches. These models for reference resolution take into account linguistic factors, such as relative salience of candidate antecedents, which have been modeled in Centering Theory (Grosz et al., 1995) by ranking candidate antecedents appearing in the preceding discourse (Iida et al., 2003; Yang et al., 2003; Denis and Baldridge, 2008). In order to take advantage of existing models, we adopt the ranking-based approach as a basis for our reference resolution model.

In conventional ranking-based models, Yang et al. (2003) and Iida et al. (2003) decompose the ranking process into a set of pairwise comparisons of two candidate antecedents. However, recent work by Denis and Baldridge (2008) reports that appropriately constructing a model for ranking all candidates yields improved performance over those utilising pairwise ranking.

Similarly we adopt a *ranking-based* model, in which all candidate antecedents compete with one another to decide the most likely candidate antecedent. Although the work by Denis and Baldridge (2008) uses Maximum Entropy to create their ranking-based model, we adopt the Ranking SVM algorithm (Joachims, 2002), which learns a weight vector to rank candidates for a given partial ranking of each referent. Each training instance is created from the set of all referents for each referring expression. To define the partial ranking of referents, we simply rank referents referred to by a given referring expression as first place and other referents as second place.

### 3.2 Use of extra-linguistic information

Recent work on multi-modal reference resolution or referring expression generation (Prasov and Chai, 2008; Foster et al., 2008; Carletta et al., 2010) indicates that extra-linguistic information, such as eye-gaze and manipulation of objects, is

[4] In the current task on reference resolution, we deal only with referring expressions referring to a single piece to minimise complexity.

Table 1: Referring expressions in REX-J corpus

| feature | tokens | example |
|---|---|---|
| demonstratives | 742 | |
| adjective | 194 | "*ano migigawa no sankakkei* (that triangle at the right side)" |
| pronoun | 548 | "*kore* (this)" |
| attribute | 795 | |
| size | 223 | "*tittyai sankakkei* (the small triangle)" |
| shape | 566 | "*ôkii sankakkei* (the large triangle)" |
| direction | 6 | "*ano sita muiteru dekai sankakkei* (that large triangle facing to the bottom)" |
| spatial relations | 147 | |
| projective | 143 | "*hidari no okkii sankakkei* (the small triangle on the left)" |
| topological | 2 | "*ôkii hanareteiru yatu* (the big distant one)" |
| overlapping | 2 | " *sono sita ni aru sankakkei* (the triangle underneath it)" |
| action-mentioning | 85 | "*migi ue ni doketa sankakkei* (the triangle you put away to the top right)" |

one of essential clues for distinguishing deictic reference from endophoric reference.

For instance, Prasov and Chai (2008) demonstrated that integrating eye-gaze information (especially, relative fixation intensity, the amount of time spent fixating a candidate object) into the conventional dialogue history-based model improved the performance of reference resolution. Foster et al. (2008) investigated the relationship of referring expressions and the manupluation of objects on a collaborative construction task, which is similar to our Tangram task[5]. They reported about 36% of the initial mentioned referring expressions in their corpus were involved with participant's operations of objects, such as mouse manipulation.

From these background, in addition to the information about the history of the preceding discourse, which has been used in previous machine learning-based approaches, we integrate extra-linguistic information into the reference resolution model shown in Section 3.1. More precisely, we introduce the following extra-linguistic information: the information with regards to the history of a piece's movement and the mouse cursor positions, and the information of the piece currently manipulated by an operator. We next elaborate on these three kinds of features. All the features are summarised in Table 2.

#### 3.2.1 Discourse history features

First, 'type of' features are acquired from the expressions of a given referring expression and its antecedent in the preceding discourse if the antecedent explicitly appears. These features have been examined by approaches to anaphora or coreference resolution (Soon et al., 2001; Ng and Cardie, 2002, etc.) to capture the salience of a candidate antecedent. To capture the textual aspect of dialogues for solving Tangram puzzle, we exploit the features such as a binary value indicating whether a referring expression has no antecedent in the preceding discourse and case markers following a candidate antecedent.

#### 3.2.2 Action history features

The history of the operations may yield important clues that indicate the salience in terms of the temporal recency of a piece within a series of operations. To introduce this aspect as a set of features, we can use, for example, the time distance of a candidate referent (i.e. a piece in the Tangram puzzle) since the mouse cursor was moved over it. We call this type of feature the *action history feature*.

#### 3.2.3 Current operation features

The recency of operations of a piece is also an important factor on reference resolution because it is directly associated with the focus of attention in terms of the cognition in a series of operations. For example, since a piece which was most recently manipulated is most salient from cognitive perspectives, it might be expected that the piece tends to be referred to by unmarked referring expressions such as pronouns. To incorporate such clues into the reference resolution model, we can use, for example, the time distance of a candidate referent since it was last manipulated in the preceding utterances. We call this type of feature the *current operation feature*.

[5]Note that the task defined in Foster et al. (2008) makes no distinction between two roles; a operator and a solver. Thus, two partipants both can mamipulate pieces on a computer display, but need to jointly construct to create a predefined goal shape.

Table 2: Feature set

| | | |
|---|---|---|
| (a) **Discourse history features** | | |
| DH1 : | yes, no | a binary value indicating that P is referred to by the most recent referring expression. |
| DH2 : | yes, no | a binary value indicating that the time distance to the last mention of P is less than or equal to 10 sec. |
| DH3 : | yes, no | a binary value indicating that the time distance to the last mention of P is more than 10 sec and less than or equal to 20 sec. |
| DH4 : | yes, no | a binary value indicating that the time distance to the last mention of P is more than 20 sec. |
| DH5 : | yes, no | a binary value indicating that P has never been referred to by any mentions in the preceding utterances. |
| DH6 : | yes, no, N/A | a binary value indicating that the attributes of P are compatible with the attributes of R. |
| DH7 : | yes, no | a binary value indicating that R is followed by the case marker '*o* (accusative)'. |
| DH8 : | yes, no | a binary value indicating that R is followed by the case marker '*ni* (dative)'. |
| DH9 : | yes, no | a binary value indicating that R is a pronoun and the most recent reference to P is not a pronoun. |
| DH10 : | yes, no | a binary value indicating that R is not a pronoun and was most recently referred to by a pronoun. |
| (b) **Action history features** | | |
| AH1 : | yes, no | a binary value indicating that the mouse cursor was over P at the beginning of uttering R. |
| AH2 : | yes, no | a binary value indicating that P is the last piece that the mouse cursor was over when feature AH1 is 'no'. |
| AH3 : | yes, no | a binary value indicating that the time distance is less than or equal to 10 sec after the mouse cursor was over P. |
| AH4 : | yes, no | a binary value indicating that the time distance is more than 10 sec and less than or equal to 20 sec after the mouse cursor was over P. |
| AH5 : | yes, no | a binary value indicating that the time distance is more than 20 sec after the mouse cursor was over P. |
| AH6 : | yes, no | a binary value indicating that the mouse cursor was never over P in the preceding utterances. |
| (c) **Current operation features** | | |
| CO1 : | yes, no | a binary value indicating that P is being manipulated at the beginning of uttering R. |
| CO2 : | yes, no | a binary value indicating that P is the most recently manipulated piece when feature CO1 is 'no'. |
| CO3 : | yes, no | a binary value indicating that the time distance is less than or equal to 10 sec after P was most recently manipulated. |
| CO4 : | yes, no | a binary value indicating that the time distance is more than 10 sec and less than or equal to 20 sec after P was most recently manipulated. |
| CO5 : | yes, no | a binary value indicating that the time distance is more than 20 sec after P was most recently manipulated. |
| CO6 : | yes, no | a binary value indicating that P has never been manipulated. |

P stands for a piece of the Tangram puzzle (i.e. a candidate referent of a referring expression) and R stands for the target referring expression.

## 4 Empirical Evaluation

In order to investigate the effect of the extra-linguistic information introduced in this paper, we conduct an empirical evaluation using the REX-J corpus.

### 4.1 Models

As we see in Section 2.2, the feature testing whether a referring expression is a pronoun or not is crucial because it is directly related to the 'deictic' usage of referring expressions, whereas other expressions tend to refer to an expression appearing in the preceding utterances. As described in Denis and Baldridge (2008), when the size of training instances is relatively small, the models induced by learning algorithms (e.g. SVM) should be separately created with regards to distinct features. Therefore, focusing on the difference of the pronominal usage of referring expressions, we separately create the reference resolution models; one is for identifying a referent of a given pronoun, and the other is for all other expressions. We henceforth call the former model the *pronoun model* and the latter one the *non-pronoun model* respectively. At the training phase, we use only training instances whose referring expressions are pronouns for creating the pronoun model, and all other training instances are used for the non-pronoun model. The model using one of these models depending on the referring expression to be solved is called the *separate model*.

To verify Denis and Baldridge (2008)'s premise mentioned above, we also create a model using all training instances without dividing pronouns and other. This model is called the *combined model* hereafter.

### 4.2 Experimental setting

We used 40 dialogues in the REX-J corpus[6], containing 2,048 referring expressions. To facilitate the experiments, we conduct 10-fold crossvalidation using 2,035 referring expressions, each of which refers to a single piece in a computer dis-

[6]Spanger et al. (2009)'s original corpus contains only 24 dialogues. In addition to this, we obtained anothor 16 dialogues by favour of the authors.

Table 3: Results on reference resolution: accuracy

| model | discourse history (baseline) | | +action history* | | +current operation | | +action history, +current operation* | |
|---|---|---|---|---|---|---|---|---|
| separated model (a+b) | 0.664 | (1352/2035) | 0.790 | (1608/2035) | 0.685 | (1394/2035) | 0.780 | (1587/2035) |
| a) pronoun model | 0.648 | (660/1018) | 0.886 | (902/1018) | 0.692 | (704/1018) | 0.875 | (891/1018) |
| b) non-pronoun model | 0.680 | (692/1017) | 0.694 | (706/1017) | 0.678 | (690/1017) | 0.684 | (696/1017) |
| combined model | 0.664 | (1352/2035) | 0.749 | (1524/2035) | 0.650 | (1322/2035) | 0.743 | (1513/2035) |

'*' means the extra-lingustic features (or the combinations of them) significantly contribute to improving performance. For the significant tests, we used McNemar test with Bonferroni's correction for multiple comparisons, i.e. $\alpha/K = 0.05/4 = 0.01$.

play[7].

As a baseline model, we adopted a model only using the discourse history features. We utilised *SVM*$^{rank}$[8] as an implementation of the Ranking SVM algorithm, in which the parameter *c* was set as 1.0 and the remaining parameters were set to their defaults.

### 4.3 Results

The results of each model are shown in Table 3. First of all, by comparing the models with and without extra-linguistic information (i.e. the model using all features shown in Table 2 and the baseline model), we can see the effectiveness of extra-linguistic information. The results typically show that the former achieved better performance than the latter. In particular, it indicates that exploiting the action history features are significantly useful for reference resolution in this data set.

Second, we can also see the impact of extra-linguistic information (especially, the action history features) with regards to the pronoun and non-pronoun models. In the former case, the model with extra-linguistic information improved by about 22% compared with the baseline model. On the other hand, in the latter case, the accuracy improved by only 7% over the baseline model. The difference may be caused by the fact that pronouns are more sensitive to the usage of the action history features because pronouns are often uttered as deixis (i.e. a pronoun tends to directly refer to a piece shown in a computer display).

The results also show that the model using the discourse history and action history features achieved better performance than the model using all the features. This may be due to the duplicated definitions between the action history and current operation features. As we can see in the feature definitions of CO1 and AH1, some current operation features partially overlap with the action history features, which is effectively used in the ranking process. However, the other current operation features may have bad effects for ranking referents due to their ill-formed definitions. To shed light on this problem, we need additional investigation of the usage of features, and to refine their definitions.

Finally, the results show that the performance of the separated model is significantly better than that of the combined model[9], which indicates that separately creating models to specialise in distinct factors (i.e. whether a referring expression is a pronoun or not) is important as suggested by Denis and Baldridge (2008).

We next investigated the significance of each

Table 4: Weights of the features in each model

| | pronoun model | | non-pronoun model | |
|---|---|---|---|---|
| rank | feature | weight | feature | weight |
| 1 | AH1 | 0.6371 | DH6 | 0.7060 |
| 2 | AH3 | 0.2721 | DH2 | 0.2271 |
| 3 | DH1 | 0.2239 | AH3 | 0.2035 |
| 4 | DH2 | 0.2191 | AH1 | 0.1839 |
| 5 | CO1 | 0.1911 | DH1 | 0.1573 |
| 6 | DH9 | 0.1055 | DH7 | 0.0669 |
| 7 | AH2 | 0.0988 | CO5 | 0.0433 |
| 8 | CO3 | 0.0852 | CO3 | 0.0393 |
| 9 | DH6 | 0.0314 | CO1 | 0.0324 |
| 10 | CO2 | 0.0249 | DH3 | 0.0177 |
| 11 | DH10 | 0 | AH4 | 0.0079 |
| 12 | DH7 | -0.0011 | AH2 | 0.0069 |
| 13 | DH3 | -0.0088 | CO4 | 0.0059 |
| 14 | CO6 | -0.0228 | DH10 | 0.0059 |
| 15 | CO4 | -0.0308 | DH9 | 0 |
| 16 | CO5 | -0.0317 | CO2 | -0.0167 |
| 17 | DH8 | -0.0371 | DH8 | -0.0728 |
| 18 | AH6 | -0.0600 | CO6 | -0.0885 |
| 19 | AH4 | -0.0761 | DH4 | -0.0924 |
| 20 | DH5 | -0.0910 | AH5 | -0.1042 |
| 21 | DH4 | -0.1193 | AH6 | -0.1072 |
| 22 | AH5 | -0.1361 | DH5 | -0.1524 |

[7]The remaining 13 instances referred to either more than one piece or a class of pieces, thus were excluded in this experiment.

[8]www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[9]For the significant tests, we used McNemar test ($\alpha = 0.05$).

Table 5: Frequencies of REs relating to on-mouse

| | pronouns | others | total |
|---|---|---|---|
| # all REs | 548 | 693 | 1,241 |
| # on-mouse | 452 | 155 | 607 |
| | (82.5%) | (22.4%) | (48.9%) |

'# all REs' stands for the frequency of referring expressions uttered in the corpus and '# on-mouse' is the frequency of referring expressions in the situation when a referring expression is uttered and a mouse cursor is over the piece referred to by the expression.

feature of the pronoun and non-pronoun models. We calculate the weight of feature $f$ shown in Table 2 according to the following formula.

$$\text{weight}(f) = \sum_{x \in SVs} w_x z_x(f) \qquad (1)$$

where *SVs* is a set of the support vectors in a ranker induced by *SVM*$^{rank}$, $w_x$ is the weight of the support vector $x$, $z_x(f)$ is the function that returns 1 if $f$ occurs in $x$, respectively.

The feature weights are shown in Table 4. This demonstrates that in the pronoun model the action history features have the highest weight, while with the non-pronoun model these features are less significant. As we can see in Table 5, pronouns are strongly related to the situation where a mouse cursor is over a piece, directly causing the weights of the features associated with the 'on-mouse' situation to become higher than other features.

On the other hand, in the non-pronoun model, the discourse history features, such as DH6 and DH2, are the most significant, indicating that the compatibility of the attributes of a piece and a referring expression is more crucial than other action history and current operation features. This is compatible with the previous research concerning textual reference resolution (Mitkov, 2002).

Table 4 shows that feature AH3 (aiming at capturing the recency in terms of a series of operations) is also significant. It empirically proves that the recent operation is strongly related to the salience of reference as a kind of 'focus' by humans.

## 5 Related Work

There have been increasing concerns about reference resolution in dialogue. Byron and Allen (1998) and Eckert and Strube (2000) reported about 50% of pronouns had no antecedent in TRAINS93 and Switchboard corpora respectively. Strube and Müller (2003) attempted to resolve pronominal anaphora in the Switchboard corpus by porting a corpus-based anaphora resolution model focusing on written texts (e.g. Soon et al. (2001) and Ng and Cardie (2002)). They used specialised features for spoken dialogues as well as conventional features. They reported relatively worse results than with written texts. The reason is that the features in their work capture only information derived from transcripts of dialogues, while it is also essential to bridge objects and concepts in the real (or virtual) world and their expressions (especially pronouns) for recognising referential relations intrinsically.

To improve performance on reference resolution in dialogue, researchers have focused on anaphoricity determination, which is the task of judging whether an expression explicitly has an antecedent in the text (i.e. in the preceding utterances) (Müller, 2006; Müller, 2007). Their work presented implementations of pronominal reference resolution in transcribed, multi-party dialogues. Müller (2006) focused on the determination of non-referential *it*, categorising instances of *it* in the ICSI Meeting Corpus (Janin et al., 2003) into six classes in terms of their grammatical categories. They also took into account each characteristic of these types by using a refined feature set. In the work by Müller (2007), they conducted an empirical evaluation including antecedent identification as well as anaphoricity determination. They used the relative frequencies of linguistic patterns as clues to introduce specific patterns for non-referentials. They reported that their performance for detecting non-referentials was relatively high (80.0% in precision and 60.9% in recall), while the overall performance was still low (18.2% in precision and 19.1% in recall). These results indicate the need for advancing research in reference resolution in dialogue.

In contrast to the above mentioned research, our task includes the treatment of entity disambiguation (i.e. selecting a referent out of a set of pieces on a computer display) as well as conventional anaphora resolution. Although our task setting is limited to the problem of solving the Tangram puzzle, we believe it is a good starting point for incorporating real (or virtual) world entities into coventional anaphora resolution.

## 6 Conclusion

This paper presented the task of reference resolution bridging pieces in the real world and their referents in dialogue. We presented an implementation of a reference resolution model exploiting extra-linguistic information, such as action history and current operation features, to capture the salience of operations by a participant and the arrangement of the pieces. Through our empirical evaluation, we demonstrated that the extra-linguistic information introduced in this paper contributed to improving performance. We also analysed the effect of each feature, showing that while action history features were useful for pronominal reference, discourse history features made sense for the other references.

In order to enhance this kind of reference resolution, there are several possible future directions. First, in the current problem setting, we exclude zero-anaphora (i.e. omitted expressions refer to either an expression in the previous utterances or an object on a display deictically). However, zero-anaphora is essential for precise modeling and recognition of reference because it is also directly related with the recency of referents, either textually or situationally. Second, representing distractors in a reference resolution model is also a key. Although, this paper presents an implementation of a reference model considering only the relationship between a referring expression and its candidate referents. However, there might be cases when the occurrence of expressions or manipulated pieces intervening between a referring expression and its referent need to be taken into account. Finally, more investigation is needed for considering other extra-linguistic information, such as eye-gaze, for exploring what kinds of information is critical to recognising reference in dialogue.

## References


D. K. Byron and J. F. Allen. 1998. Resolving demonstrative pronouns in the trains93 corpus. In *Proceedings of the 2nd Colloquium on Discourse Anaphora and Anaphor Resolution* (*DAARC2*), pages 68–81.

D. K. Byron. 2005. Utilizing visual attention for cross-model coreference interpretation. In *CONTEXT 2005*, pages 83–96.

J. Carletta, R. L. Hill, C. Nicol, T. Taylor, J. P. de Ruiter, and E. G. Bard. 2010. Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods*, 42:254–265.

P. Denis and J. Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 660–669.

B. P. W. Di Eugenio, R. H. Thomason, and J. D. Moore. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies*, 53(6):1017–1076.

M. Eckert and M. Strube. 2000. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

M. E. Foster, E. G. Bard, M. Guhe, R. L. Hill, J. Oberlander, and A. Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction* (*HRI '08*), pages 295–302.

N. Ge, J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the 6th Workshop on Very Large Corpora*, pages 161–170.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.

R. Iida, K. Inui, H. Takamura, and Y. Matsumoto. 2003. Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 10th EACL Workshop on The Computational Treatment of Anaphora*, pages 23–30.

R. Iida, K. Inui, and Y. Matsumoto. 2005. Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processing* (*TALIP*), 4(4):417–434.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 364–367.

T. Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142.

R. Mitkov. 2002. *Anaphora Resolution*. Studies in Language and Linguistics. Pearson Education.

C. Müller. 2006. Automatic detection of nonreferential *It* in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–56.

C. Müller. 2007. Resolving *It*, *This*, and *That* in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 816–823.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 104–111.

H. Poon and P. Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 650–659.

Z. Prasov and J. Y. Chai. 2008. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces* (*IUI '08*), pages 20–29.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

P. Spanger, Y. Masaaki, R. Iida, and T. Takenobu. 2009. Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proceedings of Workshop on Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*.

M. Strube and C. Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 168–175.

K. van Deemter. 2007. TUNA: Towards a unified algorithm for the generation of referring expressions. Technical report, Aberdeen University.

V. N. Vapnik. 1998. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons.

X. Yang, G. Zhou, J. Su, and C. L. Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 176–183.

X. Yang, J. Su, and C. L. Tan. 2005. Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 165–172.

X. Yang, J. Su, J. Lang, C. L. Tan, T. Liu, and S. Li. 2008. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of Annual Meeting of the Association for Computational Linguistics* (*ACL*)*: Human Language Technologies* (*HLT*), pages 843–851.