Using extra-linguistic information for generating demonstrative pronouns in a situated collaboration task

Philipp Spanger (philipp@cl.cs.titech.ac.jp)
Yasuhara Masaaki (yasuhara@cl.cs.titech.ac.jp)
Iida Ryu (ryu-i@cl.cs.titech.ac.jp)
Tokunaga Takenobu (take@cl.cs.titech.ac.jp)
Department of Computer Science, Tokyo Institute of Technology
Tokyo Meguro Ôokayama 2-12-1, 152-8552 Japan

Abstract

This paper addresses the generation of demonstrative pronouns in dialogues of a collaborative situated task. We built a Japanese corpus of dialogues in which two participants collaboratively solve the Tangram puzzle. The corpus records every action by participants and the arrangement of the puzzle pieces in synchronisation with the course of dialogues. This extra-linguistic information as well as transcribed dialogues were used to build a generation model of demonstrative pronouns. We adopted a machine learning technique (Support Vector Machine) to constructed a classifier deciding if a demonstrative pronoun is appropriate for referring to a given target puzzle piece in a given situation. Through a series of experiments, we found that extra-linguistic information, particularly those concerning current operations, plays a key role in learning usage of demonstrative pronouns.

Keywords: natural language generation, demonstrative pronouns, collaborative situate dialogue

Introduction

There has been increasing interest in empirical methods for Generating Referring Expressions (GRE). This includes the organisation of the GRE challenges (Belz & Gatt, 2007) which is designed to reproduce human reference behaviour as appears in the TUNA corpus (van Deemter, 2007). Thanks to the TUNA corpus, many participants of the challenges adopted a machine learning approach. (Dale & Viethen, 2009) also took a machine learning approach using a different corpus. They focused on relational referring expressions and tried to learn "description patterns" (sets of attributes). However, their corpus (GRE3D3) has a limitation similar to the TUNA corpus; i.e. they deal with a static situation, allowing for only "one-shot" referring expressions given a situation. This type of setting is quite far from actual human reference behaviour, where the collaborative aspect plays a central role (see for example (Clark & Wilkes-Gibbs, 1986; Heeman & Hirst, 1995)).

To cope with such limitations, several researchers have constructed and worked on corpora concerning collaboration tasks. They exploited machine learning approaches for GRE in dynamic interaction. (Jordan & Walker, 2005) utilised the COCONUT corpus (Di Eugenio, Jordan, Thomason, & Moore, 2000) and tried several feature sets based on different models for content selection of referring expressions. (Stoia, Shockley, Byron, & Fosler-Lussier, 2006) modeled various context variables as features in a learning algorithm to generate referring expressions given NP frame slots as input. They particularly note the importance of integrating extra-linguistic information as part of the context and introduced "spatial/visual features" in addition to dialogue history. (Jordan & Walker, 2005) provide a whole series of theoretically motivated features like "intentional influences" (modelling the task situation and agreement by the participants).

However, the extra-linguistic information these works dealt with, neither included features on the current operations nor on the actions that have been performed by participants during the course of the collaboration. It has been noted that in collaborative tasks (e.g. constructing an object), participants' actions on objects influence their reference behaviour (Foster et al., 2008). Thus, we propose to integrate features concerning the current operation as well as the action history as extra-linguistic information.

There has been some recent work in psycholinguistics, investigating the connection between language production and action (Roy, 2005), which in a broad sense is related to the present work. In this paper we focus on generating demonstrative pronouns, since they are commonly used in situated tasks (Piwek, 2007).

In the following, we start with a brief description of our Japanese corpus which was collected from dialogues in which two participants collaboratively solve the Tangram puzzle. Then we explain the experimental setting and the results followed by an error analysis. Finally we summarise our conclusions and give an outlook on our future work.

The Corpus

In order to utilise machine learning methods in the field of GRE, creating appropriate corpora is a critical question. Over the recent period, a number of situated corpora have been created.

The COCONUT corpus is collected from keyboard-input dialogues between two participants who are collaborating on a simple 2-D design task (Di Eugenio et al., 2000). The recorded object information including object location is limited to symbolic information. In contrast, both the QUAKE (Byron, 2005) and SCARE corpus (Stoia, Shockley, Byron, & Fosler-Lussier, 2008) are based on an interaction captured in a 3-D virtual reality. While those corpora deal with a relatively more complex domain (3-D virtual world), the subjects are only able to carry out limited kinds of actions (pushing buttons, picking up or dropping objects) as compared with the complexity of the three-dimensional target domain.

Table 1: Syntactic and semantic elements of referring expressions

Feature	types	tokens	Example
demonstrative	118	745	
adjective	100	196	"*ano migigawa no sankakkei* (that triangle at the right side)"
pronoun	19	547	"*kore* (this)"
attribute	303	641	
size	165	267	"*tittyai sankakkei* (the small triangle)"
shape	271	605	"*ôkii sankakkei* (the large triangle)"
direction	6	6	"*ano sita muiteru dekai sankakkei* (that large triangle facing to the bottom)"
spatial relations	129	148	
projective	125	144	"*hidari no okkii sankakkei* (the small triangle on the left)"
topological	2	2	"*ôkii hanareteiru yatu* (the big distant one)"
overlapping	2	2	"*sono sita ni aru sankakkei* (the triangle underneath it)"
action-mentioning	78	85	"*migi ue ni doketa sankakkei* (the triangle you put away to the top right)"
others	29	30	
remaining	15	15	"*nokotteiru ôkii sankakkei* (the remaining large triangle)"
similarity	14	15	"*sore to onazi katati no* (the one of the same shape as that one)"

In contrast to those existing corpora, we created a corpus recording a whole range of information potentially relevant in the collaborative human reference process in a situated setting. While our domain is simple, we allowed comparatively large flexibility in the actions recorded. Providing this larger freedom of actions to the participants led us to capture aspects of referring expressions in collaboration which previous corpora failed to record.

We recruited 12 Japanese graduate students of the Cognitive Science department, 4 females and 8 males, and split them into 6 pairs. Each pair was instructed to solve the Tangram puzzle cooperatively; constructing a given shape by arranging seven pieces of simple figures (two large triangles, a medium-size triangle, two small triangles, a parallelogram and a square).

We implemented a Tangram simulator (Figure 1), in order to record the precise position of every piece and every action the participants made during the solving process. Within the simulator, pieces can be moved, rotated and flipped on the computer display with simple mouse operations. The simulator displays two areas: a goal shape area (the left area in Figure 1) and a working area (the right area in Figure 1) where pieces are shown and can be manipulated.

We assigned a different role to each participant of a pair: a *solver* and an *operator*. Given a goal shape, the solver thinks of the necessary arrangement of the pieces and gives instructions to the operator how to move them. The operator manipulates the pieces with the mouse according to the solver's instructions. The participants of a pair sit side by side. A shield screen was set between the solver and operator to prevent the operator from seeing the goal shape on the solver's screen, and to restrict their interaction to speech only.

In summary, the solver can see the goal but can not manipulate the pieces; by contrast, the operator can manipulate the pieces but can not see the goal. This asymmetric setting is similar to the experiment by (Piwek, 2007) except that pointing is not allowed for the solver in our case.

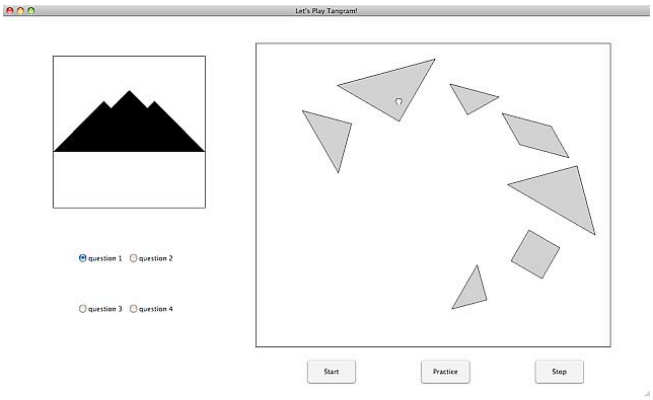

Figure 1: Screenshot of the Tangram simulator

Each participant pair is assigned 4 exercises. The participants exchange their roles after two exercises. We set a time limit of 15 minutes for an exercise. In order to prevent the solver from getting into deep thought and keeping silent, the simulator is designed to give a hint every five minutes by showing a correct piece position in the goal shape area. A dialogue ends when the goal shape is constructed or the time is up. Utterances by the participants are recorded separately in stereo through headset microphones in synchronisation with the position of the pieces and mouse operations. We collected 24 dialogues (4 exercises by 6 pairs) of about four hours in total. The average length of a dialogue was 10 minutes 43 seconds.

Recorded dialogues were transcribed with a time code attached to each utterance. In the transcribed text, referring expressions were annotated together with their referents by using the multiple-purpose annotation tool SLAT (Noguchi et al., 2008). Two annotators (two of the authors) separately annotated transcribed texts and corrected discrepancies by discussion between them. Finally, time codes were manually assigned to the starting and ending point of every annotated

referring expression.

We collected a total of 1,510 tokens and 450 types of Japanese referring expressions. Because of the asymmetric role assignment of the task, most of the referring expressions were used by solvers. Among these, we used 1,245 tokens referring to a single referent in the following experiments.

Table 1 shows the syntactic and semantic elements of the referring expressions we found as well as their respective frequency. Note that multiple features can be used in a single expression. The right-most column shows an example with its English translation. The identified element in the referring expression is underlined.

We note a strong tendency to employ object attributes, particularly the attribute "shape" as well as heavy use of demonstratives. While both of these elements are quite general and also appear in a variety of other non-situated settings, the heavy use of demonstratives (pronouns and adjectives) in particular is a characteristic of a collaborative setting.

Learning usage of demonstrative pronouns

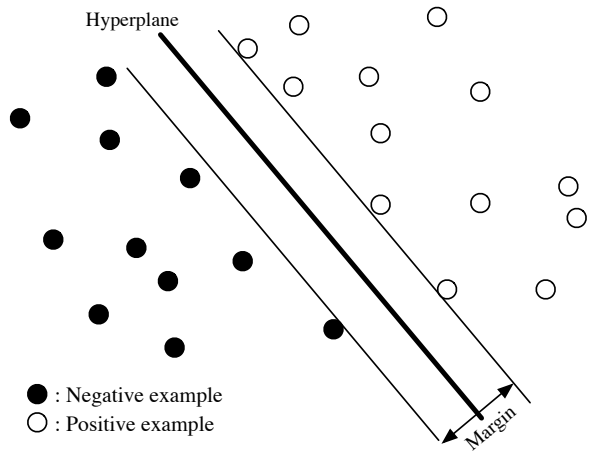

Figure 2: Example of SVM hyperplane

Using the collected corpus as training data, we employed a machine learning approach to replicate human referring expressions by introducing extra-linguistic information. In this paper, we focus on the generation of demonstrative pronouns since they are one of the most common referring expressions and will be influenced by extra-linguistic information. In our corpus it is the 2nd most frequent expression with 547 instances (see Table 1).

The Method

For determining whether to use a demonstrative pronoun or not, we employed Support Vector Machines (SVMs), which are supervised learning methods for binary classification (Vapnik, 1998). SVMs are used widely for various natural language processing tasks, such as text classification (Joachims, 1998) and multi-document summarisation (Fuentes, Alfonseca, & Rodríguez, 2007), showing a high generalisation performance and robustness against overfitting.

Given a set of examples of a positive and a negative class, the SVM seeks a separating hyperplane so as to maximise the *margin* between these classes as depicted in Figure 2. The training data of the SVM is represented as follows

$$\{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathcal{R}^n, y_i \in \{+1, -1\}\}_{i=1}^m,$$

where $\mathbf{x}_i$ is a n-dimensional feature vector of the i-th example and y_i is the corresponding class label ($+1$: positive, -1: negative). A hyperplane is represented as

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0, \quad \mathbf{w}, \mathbf{x} \in \mathcal{R}^n, b \in \mathcal{R},$$

where $\mathbf{w}$ is a normal vector and b is a parameter.

Possible hyperplanes for a specific dataset differ in the distance of samples to it. This margin represents exactly the boundaries within which a certain hyperplane can be moved without misclassifying any samples[1]. The SVM looks for a hyperplane that maximises this margin (i.e. minimising $||w||$), at the same time as avoiding misclassification (i.e. $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$).

In our experiments, we utilised the SVM-light software[2] (Joachims, 1999) with 1,245 instances of referring expressions extracted from our corpus as the training data. Since the size of our data is small, we conducted a 10-fold cross validation.

The features

As input data to the SVM, we need to define the feature-vectors representing a situation when a target is mentioned. We chose the 12 features shown in Table 2. Each feature was split up again according to its values, which are also shown in the table. For example, the time distance features (A1, A4, D1 and D4) were quantised into three intervals: $(0, 10]$ sec, $(10, 20]$ sec and $(20, \infty]$ sec and denoted as "< 10", "< 20" and "$< \infty$" respectively. Considering all these combinations results in 27 features.

The features are categorised into three categories: discourse history features (D1∼D5), action history features (A1∼A5) and the current operation features (O1 and O2). A1∼A5 correspond to D1∼D5 respectively, i.e. the feature Ai models the respective effect in the action domain, while the feature Di does so in the dialogue history.

The dialogue features (Dx) largely correspond to a model of the dialogue history. The operation (Ox) and action features (Ax) capture essential aspects of the collaboration that might have an impact on the accessibility of the target and thus the usage of demonstratives (for a theoretical discussion on what information needs to be captured by a dialogue history, see (Landragin & Romary, 2004)). We introduced the features D3 and A3 in order to represent distractor effect on recency.

[1] For a non-separable case, the SVM allows for some instances to be on the wrong side of the margin by introducing the slack variables ξ to the optimisation problem.

[2] `http://svmlight.joachims.org/`

Table 2: Features representing a situation

Dialogue history features	
D1 : $<10, <20, <\infty$	the time distance to the last mention of the target
D2 : yes, no	a binary value indicating whether the last expression referring to the target used a demonstrative pronoun or not
D3 : integer	the number of other pieces mentioned during the time period of D1
D4 : $<10, <20, <\infty$	the time distance to the last mention of another piece
D5 : yes, no	a binary value indicating if the target is the latest mentioned piece
Action history features	
A1 : $<10, <20, <\infty$	the time distance to the last action on the target
A2 : flip, move, rotate	the last operation type on the target
A3 : integer	the number of other pieces that were operated during the time period of A1
A4 : $<10, <20, <\infty$	the time distance to the last operation on another piece
A5 : yes, no	a binary value indicating if the target is the latest operated piece
Current Operation features	
O1 : yes, no	a binary value indicating if the target is under operation at the beginning of a referring expression
O2 : yes, no	a binary value indicating if the target is under the mouse cursor at the beginning of a referring expression

The Results

Given these features, our task is to decide whether to use a demonstrative pronoun for mentioning a target. We constructed an SVM classifier which classifies a pair comprised of a target piece and a situation represented by the above features into two classes: "demonstrative pronoun" and "other".

Table 3: Results of classification

Features	Recall	Precision	F-measure
All	0.698 (382/547)	0.674 (382/567)	0.686
w/o Dx	0.735 (402/547)	0.661 (402/608)	0.696
w/o Ax	0.698 (382/547)	0.673 (382/568)	0.685
w/o Ox	0.587 (321/547)	0.572 (321/561)	0.579

Table 3 shows the results of the classification. We first investigated the overall effects of features of the different categories (dialogue history, action history and current operation). The rows show the combinations of the feature categories; "All" means using all features in Table 2, "w/o Dx" means using all features except for discourse history (D1$\sim$ D5), "w/o Ax" and "w/o Ox" mean removing those respective features. We can observe significant performance degradation when removing the current operation features (O1 and O2). This indicates that information of the ongoing action has a strong impact on the usage of demonstrative pronouns. Since in our setting most of the referring expressions are used by solvers (out of all demonstrative pronouns, 400 are by the solver, 147 by the operator), who are not allowed to point at pieces, a situation where the mouse cursor is on the target (O2) cannot be regarded as a "pointing" action in the ordinary sense. In a broader sense, however, it could be considered as a joint action where a solver uses a linguistic expression while an operator points to a piece in order to identify it as target. We might be able to call this phenomenon "collaborative pointing".

Table 4: Learnt weight of features

rank	feature	weight	rank	feature	weight
1.	O2=yes	1.6174	16.	A1=<20	0.0032
2.	O1=yes	0.3587	17.	A2=rotate	0.0001
3.	D5=yes	0.2232	18.	D4=$<\infty$	-0.0206
4.	D2=yes	0.1685	19.	A1=$<\infty$	-0.0261
5.	A1=<10	0.1587	20.	A2=move	-0.0339
6.	D4=<10	0.1504	21.	D5=no	-0.0467
7.	D1=<10	0.1008	22.	D1=<20	-0.0735
8.	D2=no	0.0996	23.	A4=<10	-0.1249
9.	A5=no	0.0735	24.	D1=$<\infty$	-0.1260
10.	A4=<20	0.0551	25.	O1=no	-0.1625
11.	A4=$<\infty$	0.0551	26.	O2=no	-0.1625
12.	A2=flip	0.0405	27.	D4=<20	-0.1765
13.	D3	0.0270			
14.	A5=yes	0.0147			
15.	A3	0.0092			

Piwek observed a tendency for speakers to use shorter linguistic expressions when using pointing actions in a similar setting (Piwek, 2007). Unlike our setting, however, a solver ("instructor" in their terminology) can point to pieces as well as an operator ("builder"). Although we consider only demonstrative pronouns in this paper, our observation supports their claim (in spite of the language difference of Japanese vs. English), that pointing and operation on the target encourage use of pronouns. In addition, this observation suggests that deictic usage of pronouns is dominant in our corpus. Actually, 402 demonstrative pronouns out of 547 were used with the mouse cursor being on the target (O2=yes).

Compared with the current operation features (Ox), there is no significant difference between discourse history features and action history features. The recall slightly improves when removing discourse history features at the cost of precision.

In order to investigate effective features in detail, we constructed a SVM classifier using all features and all 1,245 instances as training data, and we calculated the learnt weight of each feature. The weight of a specific feature here can be thought of as representing its "importance" for the classifier in determining its prediction result. Table 4 shows the ranked features according to their weights. The feature O2=yes (the target is under the mouse) has the highest weight followed by O1=yes (the target is under operation), confirming the importance of information of the current operation as discussed earlier.

D5 gained more weight than A5, meaning that the last men-

tioned piece tends to be referred to by pronoun (D5), but this is not the case as much for the last operated piece (A5). The high rank of D2 could be interpreted that a piece referred to by a pronoun last tends to be subsequently mentioned by pronoun. These observations are consistent with past research on anaphora resolution (Mitkov, 2002).

Another remarkable tendency is the rank of A1 and A4 features. Among A1 features, the most recent one (< 10) has the highest rank (5), while the two more distant cases (< 20, $< \infty$) have much lower ranks (16, 19). In contrast, the ranks of A4 features show the exact opposite tendency. That is, the most recent one (< 10) has the lowest rank (23) of all A4 features, while the two more distant cases (< 20, $< \infty$) have a much higher rank (10, 11). This indicates that in order to use pronouns, the target is better to have been operated recently (high rank of A1=< 10), in contrast the other pieces are better to have been operated a long time ago (higher rank of A4=$< 20, \infty$).

It is interesting that there is no such tendency for their counterparts (D1 and D4); both D1 and D4 reside close to one another. This suggests that the recency of operation is more salient than that of linguistic mention in this experimental setting. In addition, features A3 and D3 reside in close ranking; this means the number of other pieces operated/mentioned during the period from the last mention/operation of the target up to now has a similar effect.

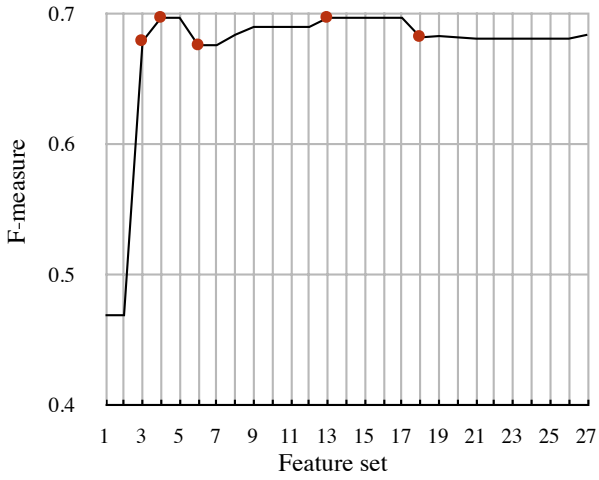

Figure 3: Results of step-wise expansion of feature set

Based on these results, we implemented follow-up experiments in order to investigate the impact of each feature by adding a feature at a time in ascending order of Table 4.

Figure 3 shows the graph of the development of the F-measure plotted over the feature combinations 1–27 (1: only the O2=yes feature, 27: all 27 features). There are two "peak-areas" around the feature combinations 4–5 and 13–17. We note that in all these cases the recall and precision values are equivalent (Recall: 0.735, Precision: 0.661). This is also the overall highest recall reached by any feature-combinations.

It is also notable, that in both combinations, where the F-measure decreases by over 0.1 (6 and 18), adding a D4 feature (D4=< 10, D4=$< \infty$) causes this decrease. This indicates that this information is rather counterproductive in this setting. Overall, we observe that the two current operation features (O1=yes, O2=yes) in combination with the highest-weight dialogue history features (D5=yes, D2=yes), i.e. combination 4, results in a local maximum of the F-measure.

Error analysis

In this section we analyse in detail the remaining errors produced by feature combination 4, which produced the best F-measure with the fewest number of features (Recall: 0.735, Precision: 0.661, F-measure: 0.696). Fundamentally, we need to distinguish two types of errors: *false positives* (FPs: when humans do not use demonstrative pronouns but the the classifier predicted a demonstrative pronoun) and the opposite case of *false negatives* (FNs: when humans used a demonstrative pronoun but the classifier wrongly predicted non-usage). We carried out a comparison of the set of cases where the classifier produced a correct answer to the set of respective error instances.

False Positives (FPs)

FPs are the type of error that negatively impact precision. There were 134 unique cases of FPs. Our main observation is that all false positive outputs by the the classifier have the feature O2=yes, i.e. we can not detect correctly the cases where the subject does not use a demonstrative pronoun even though the mouse cursor is on the target.

We note that the highest precision was achieved by feature combination 3, which includes the operation features and the dialogue feature D5=yes (i.e. the target piece is the last mentioned piece). This indicates that in the case where the target is under the mouse and being operated, the feature D5=yes is a relatively strong indicator of non-usage of demonstrative pronouns in this setting.

Among the 134 FPs, there were 55 cases of demonstrative adjectives, corresponding to 44% of all FPs. While they are different from demonstrative pronouns, we can thus say that these situations show a tendency for use of demonstratives. Out of these 55 demonstrative adjectives, 18 cases were the expression: *"sono"* (its). "Sono" can be considered both as a demonstrative adjective and a contracted form of "*sore* (it) + *no* (of)" i.e. a demonstrative pronoun with a genitive particle. For instance, "*sono kado* (its corner)" can be rephrased as "*sore no kado*". An inspection of these 18 cases revealed that they are all the latter case, i.e. a contraction of a demonstrative pronoun combined with a genitive particle. We consider those cases obvious annotation errors. If we re-calculate the F-measure by counting those cases as correct outputs, we get an increase of 3% in precision (0.691) and an F-measure of 0.712.

We listed up all FPs and investigated the top 50 by distance to SVM hyperplane (i.e. the "worst" cases). In this

preliminary evaluation by one of the authors, it was notable that while the subject did not use a demonstrative pronoun, a demonstrative pronoun as predicted by the SVM classifier would certainly be acceptable and enable one to correctly select the target. Thus, the FN cases investigated in the following section, where actual human use of demonstrative pronouns was not predicted by the classifier, is more important.

False Negatives (FNs)

FNs are those cases when humans use a demonstrative pronoun but the constructed classifier predicts non-usage; it is the type of error that negatively impacts recall. There are 145 of these errors. We note that the recall achieved by our current optimal feature combination 4 (F-measure: 0.735) is not improved by any other combination (except combinations 1 and 2 which simply *always* predict a demonstrative pronoun and thus have a recall of 1.0 but very low precision (0.305)).

The clearest and most interesting difference in feature distribution between FNs and the correctly learned demonstrative pronouns appears in the current operation features (O1 and O2). The relative frequency of cases where the target is under operation (O1=yes) is over 15% less among the FN errors than among the correct answers. Furthermore, all FN errors have the feature O2=no, i.e. the mouse cursor is not over the target. In fact, all FN errors occurred in the situations where the mouse cursor is either not over any piece (101 cases) or over a different piece from the target (44 cases), but the subjects still used a demonstrative pronoun. This observation reveals the fundamental weakness in our current model; the current operation features are too dominant.

The obvious question then is what enables subjects to use demonstrative pronouns even though the mouse cursor is not currently over the target. In order to answer this question, we investigated the top 50 worst cases out of the 145 FN cases. Among those cases, demonstrative pronouns by the solver constitute a slight majority (27 cases) over those by the operator (23 cases). However, at the current state of our analysis, we have not found any clear tendencies among FN expressions by the solver that might provide some clues for further improvement.

One type of expression that we identified, is a clarification question by the solver such as "there is the small triangle, right?", and after confirmation by the operator, the solver says "take that" (4 cases). This kind of interaction explicitly establishes common ground. Hence, the use of a demonstrative pronoun in this case could be explained in that as the newest addition to the common ground, the target is very salient. Traditionally, in the area of pronoun resolution research has concentrated on integrating various kinds of factors to calculate saliency of antecedent candidates (Grosz, Joshi, & Weinstein, 1983; Passonneau, 1993). We tried to capture the effect of the discourse context in the traditional sense in terms of the discourse history features (D1∼D5). However, it is clear that at present the integration of the discourse structure information into our current feature set is deficient. This results in the current operation features "drowning out" the information of the dialogue history features.

In addition, there were three cases in which pieces recently integrated into the partly-constructed goal figure were referred to by demonstrative pronouns by the solver, even though they were not under operation nor under the mouse. This suggests that a further factor contributing to object salience is based on the object's role within the task goal. Hence, modelling the intentional structure of participants is also necessary (Grosz & Sidner, 1986).

In contrast to the large variety of cases among the expressions by the solver, there is a comparatively clear tendency in those used by the operator. In 19 of the investigated operator-expressions (23 cases), we found two major tendencies: a delay between the use of a demonstrative pronoun and mouse pointing, as well as a diversity of pointing modes.

In our current definition, the feature O2=yes is true exactly if the mouse cursor is over the target at the beginning of uttering a referring expression. In reality, however, there is some delay and the mouse cursor might not be over the target at the exact moment of the beginning of an expression. This phenomena has also been noted in recent research on pointing in dialogue (Kranstedt, Lücking, Pfeiffer, Rieser, & Wachsmuth, 2006). This observation indicates that our current definition of the O2=yes feature is probably too strict and that we need to introduce a certain time margin between pointing and utterance.

When a referring expression is ambiguous, the operator often tries to confirm the correct piece by consecutive pointing to several candidates together with using demonstrative pronouns. These actions tend to be so quick that the mouse cursor does not always pass over the candidate pieces. In addition, there is a variety of pointing modes: circling around the piece or a part of the piece, just indicating the direction of the piece, and moving along one side/edge of the piece etc. (Steininger, Schiel, & Louka, 2001). Some of these pointing modes do not necessarily require the mouse cursor to pass over the objects. It will be necessary to take into account this diversity of pointing modes and to integrate this information into the model.

Figure 4 shows an example of a situation where the operator clarifies an ambiguous expression during the following interaction.

Time [msec]	Speaker	Utterance
141020	solver	*sita ni aru tiisai sankakkei wo* (the small triangle at the bottom-ACC)
143950	operator	*kotti desuka?* (is this it?)
144390	operator	*kotti?* (this?)

This example is an actual FN example output by our SVM classifier; even though the subject uses two demonstrative pronouns, the classifier predicted non-usage in both cases. The figure denotes snapshots of the mouse cursor positions by circles with their respective time code.

When the solver started the utterance "*sita ni aru tiisai sankakkei wo* (the small triangle at the bottom-ACC)" to refer to piece (3), the mouse cursor was located at the right upper area (denoted by time code 141020). Just before the

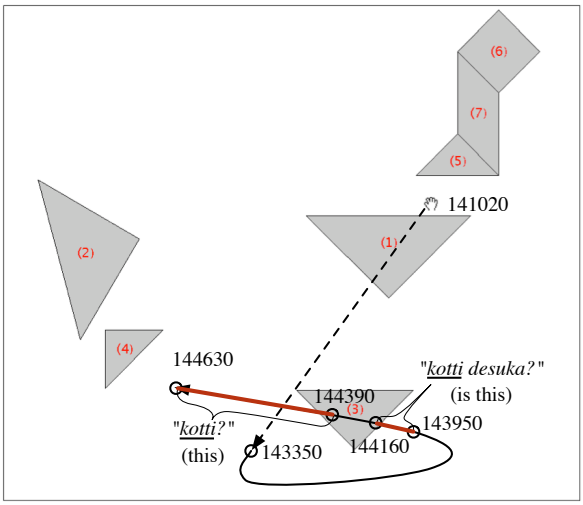

Figure 4: Example of an FN case with the trace of the mouse movement

solver finished his reference expression, which had a span of 141020–143390, the operator quickly moved the cursor to the center bottom position to the left of piece (3) at 143350. Following this, he moved the cursor smoothly to the right, then to the left of piece (3), up to the point marked 144630. During this smooth uninterrupted mouse movement (143350–144630), the operator used two referring expressions "*kotti* (this)" (one starting from 143950 and other from 144390) in order to confirm the correct referent. While at the beginning of the utterance, the mouse was not over the piece thus the feature O2 is "no" in this case, although the span of the first "*kotti* (this)" (143950–144160) partially overlapped with the duration when the mouse cursor was on the target (piece (3)) as shown in Figure 4. During the span of the second "*kotti* (this)" referring to piece (4) (144390–144630), the mouse cursor was not on the target piece (4) at all, thus again O2 became "no". The cursor gets, however, close enough to the target, thus the solver (hearer) could understand the correct referent (piece (4)).

On evaluation

Recently, there has been significant discussion in the NLG community on different evaluation measures and their usefulness (Gatt & Belz, 2008; Reiter & Belz, 2009; Khan, Deemter, Ritchie, Gatt, & Cleland, 2009). Within this context, two fundamental ways of evaluation have been stressed; namely *intrinsic* evaluation methods (evaluating system output relative to a corpus or an absolute evaluation metric) and *extrinsic* evaluation methods (assessing system output on something external, e.g. human performance on a task).

We note that the results reported in this paper (F-measure, etc.) were calculated with the collected human expressions as gold standard and thus limited to an intrinsic evaluation. However, as (Gatt & Belz, 2008) emphasize, intrinsic and extrinsic evaluation methods "yield results that are not significantly correlated". This in turn underlines the necessity of evaluating our system output on an extrinsic metric. In general, this is important in a situated collaboration domain such as is considered in this paper. The whole purpose of referring expressions in this context is the achievement of a task (here: to operate the correct piece).

Furthermore, given that in every scene there are in fact a whole number of acceptable descriptions enabling a subject to pick the correct piece, evaluating the algorithm output only against one of those possible descriptions is not an optimal metric. As indicated previously, in all 50 cases of FP errors that we reviewed, a demonstrative pronoun would actually allow task achievement. Thus, while the reported maximum F-measure of about 0.69 provides a certain (relatively strict) indication of performance, we will need to conduct a task-based (extrinsic) evaluation. In such an evaluation, our aim will be to *measure task performance*; e.g. whether a subject can correctly determine the target as well as the amount of time necessary for this.

Conclusion and Future Work

We built and analysed a Japanese corpus of referring expressions in a situated collaboration task. Using the corpus as training data, we constructed an SVM classifier to identify situations where demonstrative pronouns are suitable for referring to a target object. The experimental results show that the information on the current action, is important in deciding usage of pronouns.

Our error analysis focused on false negatives and some characteristics of scenes in such cases. We identified that additional analysis of this type of error as critical in further improving our approach. Among the types of information that we will need to integrate in our current feature set, we noted information on distractors, a more realistic model of pointing as well as possibly a model representing the role of the target piece in the task. It is also necessary to extend our current work to deal with other types of referring expressions and investigate the relationship.

References

Belz, A., & Gatt, A. (2007). The attribute selection for GRE challenge: Overview and evaluation results. In *Proceedings of the MT summit XI workshop using corpora for natural language generation: Language generation and machine translation (UCNLG+MT)* (p. 75-83).

Byron, D. K. (2005). *The OSU Quake 2004 corpus of two-party situated problem-solving dialogs* (Tech. Rep.). Department of Computer Science and Enginerring, The Ohio State University.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1-39.

Dale, R., & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th european workshop on natural language generation (ENLG 2009)* (pp. 58–65).

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

Foster, M. E., Bard, E. G., Guhe, M., Hill, R. L., Oberlander, J., & Knoll, A. (2008). The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of 3rd human-robot interaction* (p. 295-302).

Fuentes, M., Alfonseca, E., & Rodríguez, H. (2007). Support vector machines for query-focused summarization trained and evaluated on pyramid data. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 57–60). Prague, Czech Republic: Association for Computational Linguistics.

Gatt, A., & Belz, A. (2008). Attribute selection for referring expression generation: New algorithms and evaluation methods. In *Proceedings of the 5th international conference on natural language generation, INLG-08* (p. 50-58).

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting of the association for computational linguistics (ACL '83)* (p. 44-50).

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175-204.

Heeman, P. A., & Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics*, *21*, 351-382.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning (ECML)* (pp. 137–142). Berlin: Springer.

Joachims, T. (1999). *Making large-scale svm learning practical. advances in kernel methods - support vector learning, b. sch* SM*kopf and c. burges and a. smola (ed.)*. MIT-Press.

Jordan, P. W., & Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, *24*, 157-194.

Khan, I. H., Deemter, K. van, Ritchie, G., Gatt, A., & Cleland, A. A. (2009). A hearer-oriented evaluation of referring expression generation. In *Proceedings of the 12th european workshop on natural language generation (ENLG 2009)* (pp. 98–101). Athens, Greece: Association for Computational Linguistics.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006). Deixis: How to determine demonstrated objects using. In *In proceedings of the 6 th international workshop on gesture in human-computer interaction and simulation*.

Landragin, F., & Romary, L. (2004). Dialogue history modelling for multimodal human-computer interaction. In *Eighth workshop on the semantics and pragmatics of dialogue (Catalog '04)* (p. 41-48).

Mitkov, R. (2002). *Anaphora resolution*. Longman.

Noguchi, M., Miyoshi, K., Tokunaga, T., Iida, R., Komachi, M., & Inui, K. (2008). Multiple purpose annotation using SLAT – Segment and link-based annotation tool. In *Proceedings of 2nd linguistic annotation workshop* (p. 61-64).

Passonneau, R. J. (1993). Getting and keeping the center of attention. In M. Bates & R. M. Weischedel (Eds.), *Challenges in nnatural language processing* (p. 179-226). Cambridge University Press.

Piwek, P. L. (2007). Modality choise for generation of referring acts. In *Proceedings of the workshop on multimodal output generation (MOG 2007). CTIT workshop proceedings* (p. 129-139).

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *To appear in Computational Linguistics*.

Roy, D. (2005). Grounding words in perception and action: Computational insights. In *Trends in cognitive sciences* (p. 389-396).

Steininger, S., Schiel, F., & Louka, K. (2001). Gestures during overlapping speech in multimodal human-machine dialogues. In *International workshop on information presentation and natural multimodal dialogue*.

Stoia, L., Shockley, D. M., Byron, D. K., & Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the fourth international natural language generation conference* (pp. 81–88).

Stoia, L., Shockley, D. M., Byron, D. K., & Fosler-Lussier, E. (2008). SCARE: A situated corpus with annotated referring expressions. In *Proceedings of the sixth international conference on language resources and evaluation (LREC '08)* (p. 28-30).

van Deemter, K. (2007). *TUNA: Towards a unified algorithm for the generation of referring expressions* (Tech. Rep.). Aberdeen University. (`www.csd.abdn.ac.uk/research/tuna/pubs/TUNA-final-report.pdf`)

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons.