Multi-modal Reference Resolution in Situated Dialogue by Integrating Linguistic and Extra-Linguistic Clues

Ryu Iida, Masaaki Yasuhara, Takenobu Tokunaga
Department of Computer Science,
Tokyo Institute of Technology
W8-73, 2-12-1 Ohokayama Meguro Tokyo, 152-8552 Japan
{ryu-i,yasuhara,take}@cl.cs.titech.ac.jp

Abstract

This paper focuses on examining the effect of extra-linguistic information, such as eye gaze, integrated with linguistic information on multi-modal reference resolution. In our evaluation, we employ eye gaze information together with other linguistic factors in machine learning, while in prior work such as Kelleher (2006) and Prasov and Chai (2008) the incorporation of eye gaze and linguistic clues was heuristically realised. Conducting our empirical evaluation using a data set extended the REX-J corpus (Spanger et al., 2010) including eye gaze information, we examine which types of clues are useful on these three data sets, which consist largely of pronouns, non-pronouns and both respectively. Our results demonstrate that a dynamically moving visible indicator within the computer display (e.g. a mouse cursor) contributes to reference resolution for pronouns, while eye gaze information is more useful for the resolution of non-pronouns.

1 Introduction

The task of reference resolution has received much attention because it is important for applications that require interpreting text. In recent work on reference resolution within a text, several machine learning-based approaches have been proposed (McCarthy and Lehnert, 1995; Ge et al., 1998; Soon et al., 2001; Ng and Cardie, 2002; Iida et al., 2003; Yang et al., 2003; Denis and Baldridge, 2008), each of which mainly exploits linguistic clues motivated by the Centering Theory (Grosz et al., 1995) to model the discourse salience of all candidate antecedents. For instance, Yang et al. (2003) and Iida et al. (2003) presented machine learning-based reference resolution models where a pairwise comparison of candidate antecedents, in line with the basic idea of the Centering Theory, leads to the selection of the candidate with the highest salience for a given context. Denis and Baldridge (2008) extended the model by integrating the set of pairwise comparisons into ranking candidates to directly learn which clues of antecedents are useful.

Through the empirical evaluations using the data sets provided by the Message Understanding Conference (MUC)[1] and the Automatic Content Extraction (ACE)[2], which consist of newspaper articles and transcripts of broadcasts, linguistically motivated approaches have achieved better performance than state-of-the-art rule-based reference resolution systems (e.g. Soon et al. (2001) and Ng and Cardie (2002)).

In contrast to this research paradigm (i.e. research focusing on only the linguistic aspect of reference), research in the area of multi-modal interfaces has focused on referring expressions used in multi-modal conversations, in other words, identifying referents of referring expressions in a static scene or a situated world (e.g. objects depicted in a computer display), taking extra-linguistic clues into account (Byron, 2005; Prasov and Chai, 2008; Prasov and Chai, 2010; Schütte et al., 2010, etc.). For instance, Kelleher and van Genabith (2004) used the centrality and size of a object in the display to determine its visual salience. Prasov and Chai (2008) and Prasov and Chai (2010) exploited eye fixations to detect users' focus of attention in terms of visual prominence; their research has been motivated by work in the cognitive sciences (Tanenhaus et al., 1995; Tanenhaus et al., 2000; Hanna et al., 2003; Hanna and Tanenhaus, 2004; Hanna and Brennan, 2007; Metzing and Brennan, 2003; Ferreira and Tanenhaus, 2007; Brown-Schmidt et al., 2002).

[1] www-nlpir.nist.gov/related projects/muc/

[2] www.itl.nist.gov/iad/mig/tests/ace/

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 84–92,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

These previous studies have shown how promising using eye gaze information for multi-modal reference resolution can be. However, they rely on heuristic techniques for determining visual salience. Hence, there is still room for improvement by introducing eye gaze information in a more systematic and principled manner[3]. This paper, therefore, focuses on a multi-modal reference resolution model that integrates eye gaze and linguistic information by using a machine learning technique. Adapting a ranking-based anaphora resolution model, such as was proposed by Denis and Baldridge (2008), we integrate extra-linguistic information with other linguistic factors for more accurate reference resolution. With the above as a suitable background, this paper focuses on the issue of how to effectively combine linguistic and extra-linguistic factors for multi-modal reference resolution, taking collaborative task dialogues in Japanese as our target data set.

This paper is organised as follows. We first explain related work and our stance on multi-modal reference resolution in Section 2; we then present which multi-modal task we chose and how we merge eye gaze information into the predefined multi-modal task in Section 3. Section 4 introduces what types of information are used in the experiments shown in Section 5. We finally conclude this paper and discuss future directions in Section 6.

2 Related work

Within the field of computational linguistics, researchers have focused on developing computational models of reference resolution, taking into account various linguistic factors, such as grammatical, semantic and discourse clues mainly acquired from the relationship between an anaphor and any candidate antecedents (Mitkov, 2002; Lappin and Leass, 1994; Brennan et al., 1987; Strube and Hahn, 1996, etc.). Research trends for reference resolution have shifted from handcrafted rule-based approaches to corpus-based approaches due to the growing success of machine learning algorithms (e.g. Support Vector Machines (Vapnik, 1998)). For instance, an approach to coreference resolution proposed by Soon et al. (2001), in which the problem of reference resolution is decomposed into a set of binary classification problems of whether a pair of markables (e.g. NP) are anaphoric or not, achieved performance comparable to the state-of-the-art rule-based system, even though they used only a limited number of simple features. Researchers' concerns in this area cover a broad range of research topics from modeling the coreferential transitivity of a set of markables, to integrating discourse salience motivated by the Centering Theory (Grosz et al., 1995). This research area has continued to produce novel reference resolution models over the years, but the target of reference resolution is limited to only written texts or transcripts of speech.

In contrast to the above research area, researchers in the multi-modal community also have paid attention to reference resolution because it is also a crucial task for realising interaction between humans and computers. In this area, the evaluation is typically conducted in the situation where a set of objects (i.e. candidate referents) are depicted within a computer display. For instance, Stoia et al. (2008) designed an experiment where two participants controlled an avatar in a virtual world for exploring hidden treasures. In this case, the task of reference resolution is to identify an object shown on the computer display as referred to by a referring expression used by the participants during dialogue. The task becomes more complicated than typical coreference resolution for written texts because a referent is considered as either *anaphoric* (i.e. it has already appeared in the previous discourse history) or *exophoric*, (i.e. the reference resolution system needs to search for the referent from the set of objects shown in a computer display).

In order to capture the characteristics of exophoric cases, extra-linguistic information acquired from participants' eye gaze data and the visual prominence of each object are also exploited together with linguistic information. A series of research by Kelleher and his colleagues (Kelleher and van Genabith, 2004; Kelleher et al., 2005; Kelleher, 2006; Schütte et al., 2010) tackled the problem of modeling visual salience of objects in situated dialogue. In their algorithm, the visual salience of each object is estimated based on its centrality within the scene and its size; their hy-

[3] Frampton et al. (2009) employed the incorporation of linguistic and visual features on reference resolution of multi-party dialogues. However, their target was limited to only the expression *you* in dialogues, while our focus is to investigate the use of the expressions bridging between a dialogue and the real world (e.g. expressions referring to puzzle pieces on a computer display).

pothesis was that the salience is higher if a object is larger and is placed nearer the centre of the computer display. In Kelleher (2006)'s approach to reference resolution, linguistic clues such as ranking rules of candidate referents based on the Centering Theory (Grosz et al., 1995) were introduced in addition to using visual salience, but the integration of both clues was done in a heuristic way.

In addition to the visual salience assessed from the characteristics of objects in the world, eye gaze has received much attention as a clue for reference resolution. Prasov and Chai (2008), for example, employed eye gaze on the task of identifying a referent in the situation where objects are placed in a static scene. The time span after a speaker most recently fixates on an object is incorporated into their reference resolution model as well as the information of how recently the object was referred to by a referring expression. Although the results of their evaluation demonstrated that eye gaze significantly contributes to increasing performance, there is still room for improvement by adapting machine learning techniques, because in their work the linguistic and visual attention information was heuristically integrated.

In contrast, our previous work (Iida et al., 2010) employed a machine learning technique to identify the most likely candidate referent, taking into account linguistic features together with cues capturing visual salience found within the situated dialogues contained in the REX-J corpus (Spanger et al., 2010). We reported that extra-linguistic information contributes to improving performance (especially, in pronominal reference). However, in Iida et al. (2010) eye gaze information was not considered, even though in the area of cognitive science researchers have demonstrated that a speaker's eye fixations are strong clues for identifying a referent of a referring expression (Tanenhaus et al., 1995; Tanenhaus et al., 2000; Hanna et al., 2003; Hanna and Tanenhaus, 2004; Hanna and Brennan, 2007; Metzing and Brennan, 2003; Ferreira and Tanenhaus, 2007; Brown-Schmidt et al., 2002). Against this background, we investigate the effect of linguistic and extra-linguistic information including eye gaze on multi-modal reference resolution, extending Iida et al. (2010)'s reference resolution model.

3 Collecting eye gaze data in situated dialogues

In our evaluation of automatic reference resolution, we focus on investigating the interaction between linguistic and extra-linguistic clues including eye fixations on multi-modal reference resolution. Therefore, corpora where participants frequently utter both anaphoric and exophoric referring expressions are preferable for our evaluation.

In recent multi-modal problem settings for data collection, researchers have been concerned with more realistic situations, such as dynamically changing scenes rendered in a 3D virtual world (e.g. (Byron, 2005)). However, if we use data collected from such a scenario, referring expressions will be relatively skewed to exophoric cases because of frequently occurring scene updates. On the other hand, if we adopt the data collected using a static scene, we will have a disadvantage in that the change of visual salience of objects is not observed because the centrality and size of each object is fixed through dialogues.

For these reasons, we adopt the same task setting as introduced in the REX-J corpus (Spanger et al., 2010), which consists of collaborative work (solving Tangram puzzles) by two participants; the setting of this corpus is more suitable for our purposes because of the frequent occurrence of both anaphoric and exophoric referring expressions.

For collecting data, we recruited 18 Japanese graduate students, and split them into 9 pairs[4]. All pairs knew each other previously and were of the same gender and approximately the same age. Each pair was instructed to solve four different Tangram puzzles. The goal of the puzzle is to construct a given shape by arranging seven pieces (of different simple shapes) as shown in Figure 1. The precise positions 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 1/65 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

[4] Note that the first pair was used to adjust the settings of our data collection, so 4 dialogues collected from that pair were not included in the evaluation data set used in Section 5.

	time	OP-UT	SV-UT	OP-REX	SV-REX	ERR-OP	ERR-SV
total	4:22:20	2,382	4,613	239 / 270	434 / 1,192	–	–
average	9:43	88.2	170.9	8.85 / 10.0	16.1 / 44.1	14.0%	13.9%
SD	3:32	69.8	86.8	10.2 / 11.3	15.9 / 24.4	9.9	10.4

OP-UT (SV-UT) stands for the number of utterances of operators (solvers). The right side of OP-REX (SV-REX) is the frequency of referring expressions uttered by the operators (solvers), whereas the left side stands for the frequency of pronominal expressions uttered by the operators (solvers). ERR-OP (ERR-SV) is the error rate of measuring the operators' (solvers') eye gaze. SD means the standard derivation.

Table 1: Referring expressions in the extended REX-J corpus

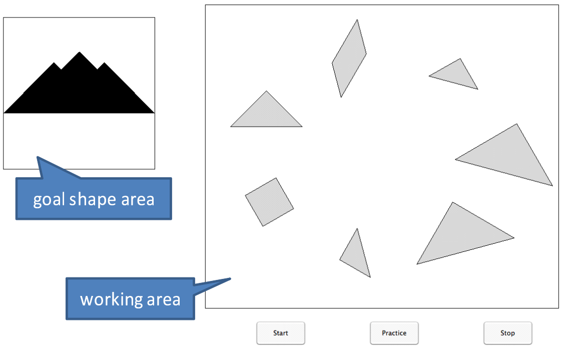

Figure 1: Screenshot of the Tangram simulator

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 is needed to distinguish between the different puzzle pieces. This collaboration is achieved by placing a set of participants side by side, each with their own display showing the work area and the mouse cursor begin manipulated by the operator in real time, 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. We put no constraint on the contents of their dialogues.

In addition to the attributes considered in the original REX-J corpus, we also collected eye gaze data synchronized with speech by using the Tobii T60 Eye Tracker, sampling at 60 Hz for recording users' eye gaze with 0.5 degrees in accuracy. Because the tracking results acquired from Tobii contain tracking errors, 5 dialogues in which the tracking results contain more than 40% errors were removed from the data set used in our evaluation.

Annotating referring expressions and their referents were conducted in the same manner as Spanger et al. (2010), i.e. annotation was conducted using a multimedia annotation tool, ELAN[5]; an annotator manually detects a referring expression and then selects its referent out of the possible puzzle pieces shown on the computer display. Note that only Tangram pieces were tagged as referents of referring expressions, therefore the expressions referring to abstract entities such as an action and event were not annotated. In the corpus multiple pieces were annotated as a single referent, but such referents were excluded in our evaluation because of their infrequent occurrence. Table 1 summarises the statistics of our new version of the REX-J corpus, consisting of 27 dialogues.

4 Multi-modal reference resolution

4.1 Base models

To investigate the impact of extra-linguistic information on reference resolution, we conducted 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 on the computer display.

As a basis of 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.) focus 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 discussed mainly 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. More precisely, we em-

[5] www.lat-mpi.eu/tools/elan/

eye gaze features		
GZ1:	[0,1]	the frequency of fixating P in the time period $[t-T,t]$, normalised by the frequency of the total fixations during the period.
GZ2:	[0,1]	the length of a fixation on P in the time period $[t-T,t]$, nomalised by T.
GZ3:	[0,1]	the length of a fixation on P in the time period $[t-T,t]$, nomalised by the total length of fixation.
GZ4:	[0,1]	the frequency of fixating P in the time period uttering a referring expression, normalised by the frequency of the total fixations during the period.
GZ5:	[0,1]	the length of a fixation on P in the time period uttering a referring expression, nominalised by T.
GZ6:	[0,1]	the length of a fixation on P in the time period uttering a referring expression, nominalised by the total length of fixation.
GZ7:	yes,no	whether the frequency of fixating P in the time period $[t-T,t]$ is most frequent.
GZ8:	yes,no	whether the frequency of fixating P in the time period $[t-T,t]$ is more than 1.
GZ9:	yes,no	whether the fixation time of P in the time period $[t-T,t]$ is longest out of all pieces.
GZ10:	yes,no	whether there exists the fixation time of P in the time period $[t-T,t]$.
GZ11:	yes,no	whether the frequency of fixating P in the time period uttering a referring expression is most frequent.
GZ12:	yes,no	whether the frequency of fixating P in the time period uttering a referring expression is more than 1.
GZ13:	yes,no	whether the fixation time of P in the time period uttering a referring expression is longest out of all pieces.
GZ14:	yes,no	whether there exists the fixation time of P in the time period uttering a referring expression.

t is the onset time of a referring expression. P denotes a piece, T is a fixed time window (1500ms).

Table 2: Eye gaze features

ploy Denis and Baldridge (2008)'s ranking-based model because they demonstrated their model outperformed the model based on simple pairwise ranking (e.g. Yang et al. (2003)).

In Denis and Baldridge (2008)'s ranking-based model, the most likely candidate antecedent is decided by simultaneously ranking all candidate antecedents. To induce a ranker used in the ranking process, we adopt the Ranking SVM algorithm (Joachims, 2002)[6], which learns a weight vector to rank candidates for a given partial ranking of each referent, while the original work by Denis and Baldridge (2008) uses Maximum Entropy to create their ranking-based model. 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 of a given referring expression as first place and any other referents as second place.

4.2 Eye gaze features

As we mentioned in Section 2, a speaker's eye gaze contributes to disambiguating referents appearing in the speaker's utterances because the speaker tends to see the target object before it is referred to by a referring expression (Spivey et al., 2002). Several aspects must be considered in order to integrate a speaker's eye gaze data. First, because the eye gaze data includes saccades, the inhibition factor of perceptual sensitivity, we extract only eye fixations as discussed in Richardson et al. (2007). For separating saccades and eye fixations, we employ Dispersion-threshold identification (Salvucci and Anderson, 2001), detecting fixations by using the concentration of eye gaze based on the fact the fixations are relatively slower than saccades. Second, because of the errors in measuring eye gaze by the eye tracker, the fixation data needs to be interpolated by the surrounding data. More specifically, if the error interval is less than 100 msec and the difference of the centers of two fixations is smaller then 16 pixels, these fixations are concatenated according to the work by Richardson et al. (2007).

The clues exploited in this paper are based on the fact that the direction of eye gaze directly reflects the focus of attention (Richardson et al., 2007; Just and Carpenter, 1976) , i.e. when one utters a referring expression, he potentially focuses on the object involved by fixating his eyes on it. Therefore, we use the eye fixations as clues for identifying the pieces focused on using the following criteria: the nearest piece to the eye fixation point is more likely a target of focus over all other pieces. To reflect this, we introduce the feature set shown in Table 2. We henceforth call these features the *eye gaze features*. Note that the parameter T is set to 1,500 ms based on the previous work done by Prasov and Chai (2010).

5 Empirical Evaluation

In order to investigate the effect of extra-linguistic information with or without linguistic factors, we conducted empirical evaluations using the updated version of the REX-J corpus explained in

[6] www.cs.cornell.edu/People/tj/svm_light/svm_rank.html

(a) **Linguistic features**		
L1 :	yes, no	whether P is referred to by the most recent referring expression.
L2 :	yes, no	whether the time distance to the last mention of P is less than or equal to 10 sec.
L3 :	yes, no	whether the time distance to the last mention of P is more than 10 sec and less than or equal to 20 sec.
L4 :	yes, no	whether the time distance to the last mention of P is more than 20 sec.
L5 :	yes, no	whether P has never been referred to by any mentions in the preceding utterances.
L6 :	yes, no, N/A	whether the attributes of P are compatible with the attributes of R.
L7 :	yes, no	whether R is followed by the case marker '*o* (accusative)'.
L8 :	yes, no	whether R is followed by the case marker '*ni* (dative)'.
L9 :	yes, no	whether R is a pronoun and the most recent reference to P is not a pronoun.
L10 :	yes, no	whether R is not a pronoun and was most recently referred to by a pronoun.
(b) **Task specific features**		
T1 :	yes, no	whether the mouse cursor was over P at the beginning of uttering R.
T2 :	yes, no	whether P is the last piece that the mouse cursor was over when feature T1 is 'no'.
T3 :	yes, no	whether the time distance is less than or equal to 10 sec after the mouse cursor was over P.
T4 :	yes, no	whether the time distance is more than 10 sec and less than or equal to 20 sec after the mouse cursor was over P.
T5 :	yes, no	whether the time distance is more than 20 sec after the mouse cursor was over P.
T6 :	yes, no	whether the mouse cursor was never over P in the preceding utterances.
T7 :	yes, no	whether P is being manipulated at the beginning of uttering R.
T8 :	yes, no	whether P is the most recently manipulated piece when feature T7 is 'no'.
T9 :	yes, no	whether the time distance is less than or equal to 10 sec after P was most recently manipulated.
T10 :	yes, no	whether the time distance is more than 10 sec and less than or equal to 20 sec after P was most recently manipulated.
T11 :	yes, no	whether the time distance is more than 20 sec after P was most recently manipulated.
T12 :	yes, no	whether 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.

Table 3: Feature set

Section 3.

5.1 Experimental settings

We employed two models as baselines: a model using only discourse history features, and one using only eye gaze features.

Because the task setting is the same as the evaluation conducted in Iida et al. (2010), we employ the same feature set, consisting of linguistically motivated features, and also features which capture the task specific extra-linguistic information of each object. We call these two kinds of features the *linguistic features* and *task specific features*, respectively. The details of these features are summarised in Table 3.

As reported in Iida et al. (2010), the referential behaviour of pronouns is completely different from non-pronouns. For this reason, we separately create two reference resolution models; one called the *pronoun model*, which identifies a referent of a given pronoun, and another called the *non-pronoun model*, which is for all other expressions. During the training phase, we use only training instances whose referring expressions are pronouns for creating the pronoun model, and all other training instances for the non-pronoun model. We group these two models together, selecting which one to use based on the referring expression. In other words, the pronoun model is selected if a referring expression is a pronoun, and the non-pronoun model otherwise. We will hereafter refer to the selectional model which alternatively picks between the pronoun and non-pronoun models as the *separated model*.

model	pronoun	non-pronoun
Ling	56.0	65.4
Gaze	56.7	48.0
TaskSp	**79.2**	21.1
Ling+Gaze	66.5	75.7
Ling+TaskSp	79.0	67.1
TaskSp+Gaze	78.0	48.4
Ling+TaskSp+Gaze	78.7	**76.0**

Ling, TaskSp and Gaze stand for the models using the linguistic, task specific and eye gaze features respectively.

Table 4: results in the separated model (accuracy)

We also train a third model using all training instances without distinguishing between pronouns and non-pronouns. This model we will refer to as the *combined model*.

5.2 Results

Table 4 shows the accuracy results of our empirical evaluation separately evaluating pronouns and non-pronouns. In reference resolution of pronouns

model	combined	separated
Ling	62.7	61.8
Gaze	51.1	51.2
TaskSp	43.7	42.8
Ling+Gaze	69.9	72.3
Ling+TaskSp	69.9	71.5
TaskSp+Gaze	55.2	59.5
Ling+TaskSp+Gaze	**72.5**	**77.0**

Table 5: Overall results (accuracy)

the results show that the model using only the linguistic features (Ling) achieved performance comparable to the one using only the eye gaze features (Gaze). Moreover, the model using only the task specific features (TaskSp) obtained performance significantly better than the others. This is because a mouse cursor is the only shared visual stimulus between the operator and solver. Therefore, it becomes the most important clue for pronouns, while the eye fixations of a speaker are not necessarily shared between them.

In contrast to pronouns, the non-pronoun model using only the linguistic features (Ling) outperforms the one using either eye gaze features or the task specific features (Gaze and TaskSp). This may be because one linguistic feature (L6) works more effectively than the other features. As shown later (see Table 6), in non-pronoun cases, the feature L6, which is the binary value indicating the compatibility of the attributes between two referring expressions, has the highest feature weight, leading to the best performance out of all three models (Ling, Gaze and TaskSp).

In addition, combining the linguistic and eye gaze features (Ling+Gaze) on non-pronoun reference resolution contributes to increasing performance. This means that these two features work in a complementary manner when a referring expression cannot be judged on a superficial level whether it refers to a discourse referent or a visually focused referent. From these results, we can see that the clues from utterances of participants are also essential for precise reference resolution, while the previous work focusing on eye fixations tends to concentrate on modeling only eye gaze information.

The accuracy results in Table 5 show the performance of the combined and separated models for different settings of feature selection. Table 5 shows that the two models achieved almost the same performance when the linguistic, eye gaze and task specific features are individually used. However, it also shows that the separated model outperforms the combined model when more than two feature types are utilised. This indicates that separating the models with regard to the type of referring expression does make sense even when we employ eye fixations as a clue for recognising referent objects. It also shows that both the combined and separated models obtained the best performance for each model using all the features. In other words, the three types of features work in a complementary manner on multi-modal reference resolution.

	pronoun model		non-pronoun model	
rank	feature	weight	feature	weight
1	T1	0.4744	L6	0.6149
2	T3	0.2684	GZ10	0.1566
3	L1	0.2298	GZ9	0.1566
4	T7	0.1929	GZ7	0.1255
5	T9	0.1605	GZ11	0.1225
6	GZ10	0.1547	GZ14	0.1134
7	GZ9	0.1547	GZ13	0.1134
8	L6	0.1442	GZ12	0.1026
9	GZ7	0.1267	L2	0.1014
10	L2	0.1164	GZ1	0.0750

Table 6: 10 highest weights of the features in each model

We next investigated the significance of each feature for the pronoun and non-pronoun models. We calculate the weight of a feature f shown in Table 6 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 the Ranking SVM algorithm, 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.

Table 6 shows the top 10 features with the highest weights of each model. It demonstrates that in the pronoun model the task specific features have the highest weight, while in the non-pronoun model these features are less significant. As shown in Table 4, pronouns are strongly related to the situation where the mouse cursor is over a piece, which is consistent with the results reported in Iida et al. (2010).

In contrast, the highest features in the non-pronoun model are occupied by the eye gaze features, except for L6. This indicates that in the situation where a speaker mentions pieces realised as non-pronouns, the eye fixations become a good clue for identifying the current focus of the speaker, while the task specific features such as the location of the mouse cursor are less significant. In addition, Table 6 also shows that the discourse feature L6 obtains the highest significance. This means that exploiting the linguistic factors together with eye fixations is essential for more accurate reference resolution.

6 Conclusion

In this paper we focused on investigating the impact of eye fixations on reference resolution compared to using other extra-linguistic information. We conducted an empirical evaluation using referring expressions appearing in collaborative work dialogues from the extended REX-J corpus, synchronised with eye gaze information. We demonstrated that the referents of pronouns are relatively easily identified, as they rely on the visual salience such as is indicated by moving the mouse cursor, and that non-pronouns are strongly related to eye fixations on its referent. In addition, our results also show that combining linguistic, eye gaze and other extra-linguistic factors contribute to increasing the overall performance of identifying all referring expressions.

There are several future directions for making the multi-modal reference resolution more accurate and robust. First, we need to introduce more task dependent information reflecting the characteristics of each multi-modal task. In the Tangram puzzle task, for example, once a piece becomes part of a partially constructed shape, the piece tends to be less salient because a solver typically gives an instruction to move a scattered piece to a partially constructed shape. We expect that introducing such task specific clues into the reference resolution model as features will contribute to improving performance.

Second, in our evaluation we adopted collaborative work dialogues where two participants solve Tangram puzzles. Since all objects (i.e. puzzle pieces) have nearly the same size, this results in explicitly rejecting the factor that a relatively larger object occupying the computer display has higher prominence over smaller objects, which has been considered by Byron (2005). In order to take such a factor into account, we need further data collection and then to incorporate additional factors into the current reference resolution model.

A third possible direction for future work is to examine the relation between linguistic and intentional structures, which are discussed in Grosz and Sidner (1986). In our problem setting, when a solver instructs an operator how to construct a goal shape, a series of utterances by the solver reflects the solver's intentions. As we already mentioned above, objects which a solver wants an operator to manipulate tend to draw a solver's attention, while the other objects (especially, the objects representing the partially constructed shape) are considered less salient. Exploiting the importance of the speaker's intentions also needs to be considered in future work.

References

S. E. Brennan, M. W. Friedman, and C. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 155–162.

S. Brown-Schmidt, E. Campana, and M. K. Tanenhaus. 2002. Reference resolution in the wild: On-line circumscription of referential domains in a natural, interactive, problem-solving task. In *Proceedings of the 24th annual meeting of the Cognitive Science Society*, pages 148–153.

D. K. Byron. 2005. Utilizing visual attention for crossmodal coreference interpretation. In *In Proceedings of Fifth International and Interdisciplinary Conference on Modeling and Using Context*, pages 83–96.

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.

F. Ferreira and M. K. Tanenhaus. 2007. Introduction to the special issue on language–vision interactions. *Journal of Memory and Language*, 57:455–459.

M. Frampton, R. Fernández, P. Ehlen, M. Christoudias, T. Darrell, and S. Peters. 2009. Who is "you"? combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of the 12th Conference of the European Chapter of the ACL* (*EACL 2009*), pages 273–281.

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.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

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.

J. E. Hanna and S. E. Brennan. 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57.

J. E. Hanna and M. K. Tanenhaus. 2004. Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, 28:105–115.

J. E. Hanna, M. K. Tanenhaus, and J. C. Trueswell. 2003. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1):43–61.

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, S. Kobayashi, and T. Tokunaga. 2010. Incorporating extra-linguistic information into reference resolution in collaborative task dialogue. In *Proceeding of the 48st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 1259–1267.

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.

M. Just and P. A. Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480.

J. Kelleher and J. van Genabith. 2004. Visual salience and reference resolution in simulated 3-d environments. *Artificial Intelligence Review*, 21(3):253–267.

J. Kelleher, F. Costello, and J. van Genabith. 2005. Dynamically structuring updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence*, 167:62–102.

J. D. Kelleher. 2006. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25:21–35.

S. Lappin and H. J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

J. F. McCarthy and W. G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1050–1055.

C. Metzing and S. E. Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49:201–213.

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

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.

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

Z. Prasov and J. Y. Chai. 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 471–481.

D. C. Richardson, R. Dale, and M. J. Spivey. 2007. Eye movements in language and cognition: A brief introduction, methods in cognitive linguistics. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, and M. J. Spivey, editors, *Methods in Cognitive Linguistics*, pages 323–344. John Benjamins.

D. D. Salvucci and J. R. Anderson. 2001. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16:39–86.

N. Schütte, J. D. Kelleher, and B. Mac Namee. 2010. Visual salience and reference resolution in situated dialogues: A corpus-based evaluation. In *In Proceedings of the AAAI Symposium on Dialog with Robots, Arlington, Virginia, USA. 11th - 13th Nov 2010.*

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, M. Yasuhara, R. Iida, T. Tokunaga, A. Terai, and N. Kuriyama. 2010. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources & Evaluation.*

M. J. Spivey, M. K. Tanenhaus, K. M. Eberhard, and J. C. Sedivy. 2002. Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4):447–481.

L. Stoia, D. M. Shockley, D. K. Byron, and E. Fosler-Lussier. 2008. Scare: A situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC 2008*).

M. Strube and U. Hahn. 1996. Functional centering. In *Proceeding of the 34st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pages 270–277.

M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

M. K. Tanenhaus, J. S. Magnuson, D. Dahan, and C. Chambers. 2000. Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, 29(6):557–580.

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.