Automated Vocabulary Acquisition and Interpretation in Multimodal Conversational Systems

Yi Liu **Joyce Y. Chai** **Rong Jin**
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
{liuyi3, jchai, rongjin}@cse.msu.edu

Abstract

Motivated by psycholinguistic findings that eye gaze is tightly linked to human language production, we developed an unsupervised approach based on translation models to automatically learn the mappings between words and objects on a graphic display during human machine conversation. The experimental results indicate that user eye gaze can provide useful information to establish such mappings, which have important implications in automatically acquiring and interpreting user vocabularies for conversational systems.

1 Introduction

To facilitate effective human machine conversation, it is important for a conversational system to have knowledge about user vocabularies and understand how these vocabularies are mapped to the internal entities for which the system has representations. For example, in a multimodal conversational system that allows users to converse with a graphic interface, the system needs to know what vocabularies users tend to use to describe objects on the graphic display and what (type of) object(s) a user is attending to when a particular word is expressed. Here, we use *acquisition* to refer to the process of acquiring relevant vocabularies describing internal entities, and *interpretation* to refer to the process of automatically identifying internal entities given a particular word. Both acquisition and interpretation have been traditionally approached by either knowledge engineering (e.g., manually created lexicons) or supervised learning from annotated data. In this paper, we describe an unsupervised approach that relies on naturally co-occurred eye gaze and spoken utterances during human machine conversation to automatically acquire and interpret vocabularies.

Motivated by psycholinguistic studies (Just and Carpenter, 1976; Griffin and Bock, 2000; Tenenhaus et al., 1995) and recent investigations on computational models for language acquisition and grounding (Siskind, 1995; Roy and Pentland, 2002; Yu and Ballard, 2004), we are particularly interested in two unique questions related to multimodal conversational systems: (1) In a multimodal conversation that involves more complex tasks (e.g., both user initiated tasks and system initiated tasks), is there a reliable temporal alignment between eye gaze and spoken references so that the coupled inputs can be used for automated vocabulary acquisition and interpretation? (2) If such an alignment exists, how can we model this alignment and automatically acquire and interpret the vocabularies?

To address the first question, we conducted an empirical study to examine the temporal relationships between eye fixations and their corresponding spoken references. As shown later in section 4, although a larger variance (compared to the findings from psycholinguistic studies) exists in terms of how eye gaze is linked to speech production during human machine conversation, eye fixations and the corresponding spoken references still occur in a very close vicinity to each other. This natural coupling between eye gaze and speech provides an opportunity to automatically learn the mappings between

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 368–375,
Prague, Czech Republic, June 2007.

words and objects without any human supervision.

Because of the larger variance, it is difficult to apply rule-based approaches to quantify this alignment. Therefore, to address the second question, we developed an approach based on statistical translation models to explore the co-occurrence patterns between eye fixated objects and spoken references. Our preliminary experiment results indicate that the translation model can reliably capture the mappings between the eye fixated objects and the corresponding spoken references. Given an object, this model can provide possible words describing this object, which represents the acquisition process; given a word, this model can also provide possible objects that are likely to be described, which represents the interpretation process.

In the following sections, we first review some related work and introduce the procedures used to collect eye gaze and speech data during human machine conversation. We then describe our empirical study and the unsupervised approach based on translation models. Finally, we present experiment results and discuss their implications in natural language processing applications.

2 Related Work

Our work is motivated by previous work in the following three areas: psycholinguistics studies, multimodal interactive systems, and computational modeling of language acquisition and grounding.

Previous psycholinguistics studies have shown that the direction of gaze carries information about the focus of the user's attention (Just and Carpenter, 1976). Specifically, in human language processing tasks, eye gaze is tightly linked to language production. The perceived visual context influences spoken word recognition and mediates syntactic processing (Tenenhaus et al., 1995). Additionally, before speaking a word, the eyes usually move to the objects to be mentioned (Griffin and Bock, 2000). These psycholinguistics findings have provided a foundation for our investigation.

In research on multimodal interactive systems, recent work indicates that the speech and gaze integration patterns can be modeled reliably for individual users and therefore be used to improve multimodal system performances (Kaur et al., 2003). Studies have also shown that eye gaze has a potential to improve resolution of underspecified referring expressions in spoken dialog systems (Campana et al., 2001) and to disambiguate speech input (Tanaka, 1999). In contrast to these earlier studies, our work focuses on a different goal of using eye gaze for automated vocabulary acquisition and interpretation.

The third area of research that influenced our work is computational modeling of language acquisition and grounding. Recent studies have shown that multisensory information (e.g., through vision and language processing) can be combined to effectively acquire words to their perceptually grounded objects in the environment (Siskind, 1995; Roy and Pentland, 2002; Yu and Ballard, 2004). Especially in (Yu and Ballard, 2004), an unsupervised approach based on a generative correspondence model was developed to capture the mapping between spoken words and the occurring perceptual features of objects. This approach is most similar to the translation model used in our work. However, compared to this work where multisensory information comes from vision and language processing, our work focuses on a different aspect. Here, instead of applying vision processing on objects, we are interested in eye gaze behavior when users interact with a graphic display. Eye gaze is an implicit and subconscious input modality during human machine interaction. Eye gaze data inevitably contain a significant amount of noise. Therefore, it is the goal of this paper to examine whether this modality can be utilized for vocabulary acquisition for conversational systems.

3 Data Collection

We used a *simplified* multimodal conversational system to collect synchronized speech and eye gaze data. A room interior scene was displayed on a computer screen, as shown in Figure 1. While watching the graphical display, users were asked to communicate with the system on topics about the room decorations. A total of 28 objects (e.g., multiple lamps and picture frames, a bed, two chairs, a candle, a dresser, etc., as marked in Figure 1) are explicitly modeled in this scene. The system is *simplified* in the sense that it only supports 14 tasks during human machine interaction. These tasks are designed to cover both open-ended utterances (e.g., the system

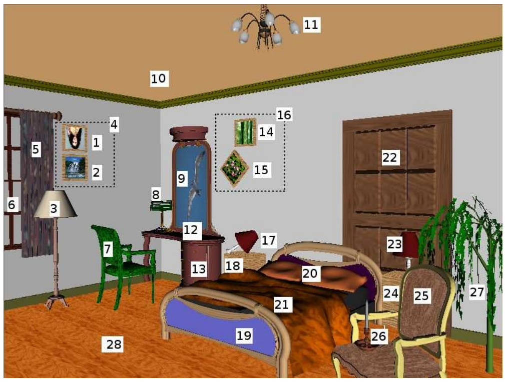

Figure 1: The room interior scene for user studies. For easy reference, we give each object an ID. These IDs are hidden from the system users.

asks users to describe the room) and more restricted utterances (e.g., the system asks the user whether he/she likes the bed) that are commonly supported in conversational systems. Seven human subjects participated in our study.

User speech inputs were recorded using the Audacity software[1], with each utterance time-stamped. Eye movements were recorded using an EyeLink II eye tracker sampled at 250Hz. The eye tracker automatically saved two-dimensional coordinates of a user's eye fixations as well as the time-stamps when the fixations occurred.

The collected raw gaze data is extremely noisy. To refine the gaze data, we further eliminated invalid and saccadic gaze points (known as "saccadic suppression" in vision studies). Since eyes do not stay still but rather make small, frequent jerky movements, we also smoothed the data by averaging nearby gaze locations to identify fixations.

4 Empirical Study on Speech-Gaze Alignment

Based on the data collected, we investigated the temporal alignment between co-occurred eye gaze and spoken utterances. In particular, we examined the temporal alignment between eye gaze fixations and the corresponding spoken references (i.e., the spoken words that are used to refer to the objects on the graphic display).

According to the time-stamp information, we can measure the length of time gap between a user's eye fixation falling on an object and the corresponding spoken reference being uttered (which we refer to as "length of time gap" for brevity). Also, we can count the number of times that user fixations happen to change their target objects during this time gap (which we refer to as "number of fixated object changes" for brevity). The nine most frequently occurred spoken references in utterances from all users (as shown in Table 1) are chosen for this empirical study. For each of those spoken references, we use human judgment to decide which object is referred to. Then, from both before and after the onset of the spoken reference, we find the closest occurrence of the fixation falling on that particular object. Altogether we have 96 such speech-gaze pairs. In 54 pairs, the eye gaze fixation occurred before the corresponding speech reference was uttered; and in the other 42 pairs, the eye fixation occurred after the corresponding speech reference was uttered. This observation suggests that in human machine conversation, eye fixation on an object does not necessarily always proceed the utterance of the corresponding speech reference.

Further, we computed the average *absolute* length of the time gap and the average number of fixated object changes, as well as their variances for each of 5 selected users[2] as shown in Table 1. From Table 1, it is easy to observe that: **(I)** A spoken reference always appears within a short period of time (usually 1-2 seconds) *before or after* the corresponding eye gaze fixation. But, the exact length of the period is far from constant. **(II)** It is not necessary for a user to utter the corresponding spoken reference *immediately* before or after the eye gaze fixation falls on that particular object. Eye gaze fixations may move back and forth. Between the time an object is fixated and the corresponding spoken reference is uttered, a user's eye gaze may fixate on a few other objects (reflected by the average number of eye fixated object changes shown in the table). **(III)** There is a large variance in both the length of time gap and the number of fixated object changes in terms of 1) the same user and the same spoken reference at different time-stamps, 2) the same user but different spo-

[1]http://audacity.sourceforge.net/

[2]The other two users are not selected because the nine selected words do not appear frequently in their utterances.

Spoken Reference	Average Absolute Length of Time Gap (in seconds)					Average Number of Eye Fixated Object Changes				
	User 1	User 2	User 3	User 4	User 5	User 1	User 2	User 3	User 4	User 5
bed	1.27 ± 1.40	1.02 ± 0.65	0.32 ± 0.21	0.59 ± 0.77	2.57 ± 3.25	2.1 ± 3.2	2.1 ± 2.2	0.4 ± 0.5	1.4 ± 2.2	5.3 ± 7.9
tree	-	0.24 ± 0.24	-	-	-	-	0.0 ± 0.0	-	-	-
window	-	0.67 ± 0.74	-	-	1.95 ± 3.20	-	0.0 ± 0.0	-	-	3.3 ± 5.9
mirror	-	1.04 ± 1.36	-	-	-	-	1.0 ± 1.4	-	-	-
candle	-	-	3.64 ± 0.59	-	-	-	-	8.5 ± 2.1	-	-
waterfall	1.80 ± 1.12	-	-	-	-	5.5 ± 4.9	-	-	-	-
painting	0.10 ± 0.10	-	-	-	-	0.2 ± 0.4	-	-	-	-
lamp	0.74 ± 0.54	1.70 ± 0.99	0.26 ± 0.35	1.98 ± 1.72	2.84 ± 2.42	1.3 ± 1.3	1.8 ± 1.5	0.3 ± 0.6	4.8 ± 4.3	2.7 ± 2.2
door	2.47 ± 0.84	-	-	2.49 ± 1.90	6.36 ± 2.29	5.0 ± 2.6	-	-	6.7 ± 5.5	13.3 ± 6.7

Table 1: The average absolute length of time and the number of eye fixated object changes within the time gap of eye gaze and corresponding spoken references. Variances are also listed. Some of the entries are not available because the spoken references were never or rarely used by the corresponding users.

ken references, and 3) the same spoken reference but different users. We believe this is due to the different dialog scenarios and user language habits.

To summarize our empirical study, we find that in human machine conversation, there still exists a natural temporal coupling between user speech and eye gaze, i.e. the spoken reference and the corresponding eye fixation happen within a close vicinity of each other. However, a large variance is also observed in terms of these temporal vicinities, which indicates an intrinsically more complex gaze-speech pattern. Therefore, it is hard to directly quantify the temporal or ordering relationship between spoken references and corresponding eye fixated objects (for example, through rules).

To better handle the complexity in the gaze-speech pattern, we propose to use statistical translation models. Given a time window of enough length, a speech input that contains a list of spoken references (e.g., definite noun phrases) is always accompanied by a list of naturally occurred eye fixations and therefore a list of objects receiving those fixations. All those pairs of speech references and corresponding fixated objects could be viewed as *parallel*, i.e. they *co-occur within the time window*. This situation is very similar to the training process of translation models in statistical machine translation (Brown et al., 1993), where parallel corpus is used to find the mappings between words from different languages by exploiting their co-occurrence patterns. The same idea can be borrowed here: by exploring the co-occurrence statistics, we hope to uncover the exact mapping between those eye fixated objects and spoken references. The intuition is that, the more often a fixation is found to exclusively co-occur with a spoken reference, the more likely a mapping should be established between them.

5 Translation Models for Vocabulary Acquisition and Interpretation

Formally, we denote the set of observations by $\mathbf{D} = \{\mathbf{w}_i, \mathbf{o}_i\}_{i=1}^{N}$ where $\mathbf{w}_i$ and $\mathbf{o}_i$ refers to the i-th speech utterance (i.e., a list of words of spoken references) and the i-th corresponding eye gaze pattern (i.e., a list of eye fixated objects) respectively. When we study the problem of mapping given objects to words (for vocabulary acquisition), the parameter space $\mathbf{\Theta} = \{\Pr(w_j|o_k), 1 \leq j \leq m^w, 1 \leq k \leq m^o\}$ consists of the mapping probabilities of an arbitrary word w_j to an arbitrary object o_k, where m^w and m^o represent the total number of unique words and objects respectively. Those mapping probabilities are subject to constraints $\sum_{j=1}^{m^w} \Pr(w_j|o_k) = 1$. Note that $\Pr(w_j|o_k) = 0$ if the corresponding word w_j and o_k never co-occur in any observed list pair $(\mathbf{w}_i, \mathbf{o}_i)$.

Let l_i^w and l_i^o denote the length of lists $\mathbf{w}_i$ and $\mathbf{o}_i$ respectively. To distinguish with the notations w_j and o_k whose subscripts are indices for *unique* words and objects respectively, we use $\tilde{w}_{i,j}$ to denote the word in the j-th position of the list $\mathbf{w}_i$ and $\tilde{o}_{i,k}$ to denote the object in the k-th position of the list $\mathbf{o}_i$. In translation models, we assume that any word in the list $\mathbf{w}_i$ is mapped to an object in the corresponding list $\mathbf{o}_i$ or a *null object* (we reserve the position 0 for it in every object list). To denote all the word-object mappings in the i-th list pair, we introduce an alignment vector $\mathbf{a}_i$, whose element $a_{i,j}$ takes the value k if the word $\tilde{w}_{i,j}$ is mapped to $\tilde{o}_{i,k}$.

Then, the likelihood of the observations given the parameters can be computed as follows

$$\begin{aligned}\Pr(\mathbf{D};\mathbf{\Theta}) &= \prod_{i=1}^{N}\Pr(\mathbf{w}_i|\mathbf{o}_i) = \prod_{i=1}^{N}\sum_{\mathbf{a}_i}\Pr(\mathbf{w}_i,\mathbf{a}_i|\mathbf{o}_i)\\ &= \prod_{i=1}^{N}\sum_{\mathbf{a}_i}\frac{\Pr(l_i^w|\mathbf{o}_i)}{(l_i^o+1)^{l_i^w}}\prod_{j=1}^{l_i^w}\Pr(\tilde{w}_{i,j}|\tilde{o}_{a_{i,j}})\\ &= \prod_{i=1}^{N}\frac{\Pr(l_i^w|\mathbf{o}_i)}{(l_i^o+1)^{l_i^w}}\sum_{\mathbf{a}_i}\prod_{j=1}^{l_i^w}\Pr(\tilde{w}_{i,j}|\tilde{o}_{a_{i,j}})\end{aligned}$$

Note that the following equation holds:

$$\prod_{j=1}^{l_i^w}\sum_{k=0}^{l_i^o}\Pr(\tilde{w}_{i,j}|\tilde{o}_{i,k}) = \sum_{a_{i,1}=1}^{l_i^o}\cdots\sum_{a_{i,l_i^w}=1}^{l_i^o}\prod_{j=1}^{l_i^w}\Pr(\tilde{w}_{i,j}|\tilde{o}_{a_{i,j}})$$

where the right-hand side is actually the expansion of $\sum_{\mathbf{a}_i}\prod_j^{l_i^w}\Pr(\tilde{w}_{i,j}|\tilde{o}_{a_{i,j}})$. Therefore, the likelihood can be simplified as

$$\Pr(\mathbf{D};\mathbf{\Theta}) = \prod_{i=1}^{N}\frac{\Pr(l_i^w|\mathbf{o}_i)}{(l_i^o+1)^{l_i^w}}\prod_{j=1}^{l_i^w}\sum_{k=0}^{l_i^o}\Pr(\tilde{w}_{i,j}|\tilde{o}_{i,k})$$

Switching to the notations w_j and o_k, we have

$$\Pr(\mathbf{D};\mathbf{\Theta}) = \prod_{i=1}^{N}\frac{\Pr(l_i^w|\mathbf{o}_i)}{(l_i^o+1)^{l_i^w}}\prod_{j=1}^{m^w}\left[\sum_{k=0}^{m^o}\Pr(w_j|o_k)\delta_{i,k}^o\right]^{\delta_{i,j}^w}$$

where $\delta_{i,j}^w = 1$ if $\tilde{w}_{i,j}\in\mathbf{w}_i$ and $\delta_{i,j}^w = 0$ otherwise, and $\delta_{i,k}^o = 1$ if $\tilde{o}_{i,k}\in\mathbf{o}_i$ and $\delta_{i,k}^o = 0$ otherwise.

Finally, the translation model can be formalized as the following optimization problem

$$\begin{aligned}\arg\max_{\mathbf{\Theta}} \quad & \log\Pr(\mathbf{D};\mathbf{\Theta})\\ s.t. \quad & \sum_{j=1}^{m^w}\Pr(w_j|o_k) = 1, \forall k\end{aligned}$$

This optimization problem can be solved by the EM algorithm (Brown et al., 1993).

The above model is developed in the context of mapping given objects to words, i.e., its solution yields a set of conditional probabilities $\{\Pr(w_j|o_k), \forall j\}$ for each object o_k, indicating how likely every word is mapped to it. Similarly, we can develop the model in the context of mapping given words to objects (for vocabulary interpretation), whose solution leads to another set of probabilities $\{\Pr(o_k|w_j), \forall k\}$ for each word w_j indicating how likely every object is mapped to it. In our experiments, both models are implemented and we will present the results later.

6 Experiments

We experimented our proposed statistical translation model on the collected data mentioned in Section 3.

6.1 Preprocessing

The main purpose of preprocessing is to create a "parallel corpus" for training a translation model. Here, the "parallel corpus" refers to a series of speech-gaze pairs, each of them consisting of a list of words from the spoken references in the user utterances and a list of objects that are fixated upon within the same time window.

Specifically, we first transcribed the user speech into scripts by automatic speech recognition software and then refined them manually. A time-stamp was associated with each word in the speech script. Further, we detected long pauses in the speech script as splitting points to create time windows, since a long pause usually marks the start of a sentence that indicates a user's attention shift. In our experiment, we set the threshold of judging a long pause to be 1 second. From all the data gathered from 7 users, we get 357 such time windows (which typically contain 10-20 spoken words and 5-10 fixated object changes).

Given a time window, we then found the objects being fixated upon by eye gaze (represented by their IDs as shown in Figure 1). Considering that eye gaze fixation could occur during the pauses in speech, we expanded each time window by a fixed length at both its start and end to find the fixations. In our experiments, the expansion length is set to 0.5 seconds.

Finally, we applied a part-of-speech tagger to each sentence in the user script and only singled out nouns as potential spoken references in the word list. The Porter stemming algorithm was also used to get the normalized forms of those nouns.

The translation model was trained based on this preprocessed parallel data.

6.2 Evaluation Metrics

As described in Section 5, by using a statistical translation model we can get a set of translation probabilities, either from any given spoken word to all the objects, or from any given object to all the spoken words. To evaluate the two sets of translation probabilities, we use *precision* and *recall* as

#Rank	Precision	Recall	#Rank	Precision	Recall
1	0.6667	0.2593	6	0.2302	0.5370
2	0.4524	0.3519	7	0.2041	0.5556
3	0.3810	0.4444	8	0.1905	0.5926
4	0.3095	0.4815	9	0.1799	0.6296
5	0.2667	0.5185	10	0.1619	0.6296

Table 2: Average precision/recall of mapping given objects to words (i.e., acquisition)

#Rank	Precision	Recall	#Rank	Precision	Recall
1	0.7826	0.3214	6	0.3043	0.7500
2	0.5870	0.4821	7	0.2671	0.7679
3	0.4638	0.5714	8	0.2446	0.8036
4	0.3804	0.6250	9	0.2293	0.8393
5	0.3478	0.7143	10	0.2124	0.8571

Table 3: Average precision/recall of mapping given words to objects.(i.e., interpretation)

evaluation metrics.

Specifically, for a given object o_k the translation model will yield a set of probabilities $\{\Pr(w_j|o_k), \forall j\}$. We can sort the probabilities and get a ranked list. Let us assume that we have the ground truth about all the spoken words to which the given object should be mapped. Then, at a given number n of top ranked words, the *precision* of mapping the given object o_k to words is defined as

$$\frac{\text{\# words that } o_k \text{ is correctly mapped to}}{\text{\# words that } o_k \text{ is mapped to}}$$

and the *recall* is defined as

$$\frac{\text{\# words that } o_k \text{ is correctly mapped to}}{\text{\# words that } o_k \text{ should be mapped to}}$$

All the counting above is done within the top n rank. Therefore, we can get different precision/recall at different ranks. At each rank, the overall performance can be evaluated by averaging the precision/recall for all the given objects. Human judgment is used to decide whether an object-word mapping is correct or not, as ground truth for evaluation.

Similarly, based on the set of probabilities of mapping a given object with spoken words, we can find a ranked list of objects for a given word, i.e. $\{\Pr(o_k|w_j), \forall k\}$. Thus, at a given rank the *precision* and *recall* of mapping a given word w_j to objects can be measured.

6.3 Experiment Results

Vocabulary acquisition is the process of finding the appropriate word(s) for any given object. For the sake of statistical significance, our evaluation is done on 21 objects that were mentioned at least 3 times by the users.

Table 2 gives the average precision/recall evaluated at the top 10 ranks. As we can see, if we use the most probable word acquired for each object, about 66.67% of them are appropriate. With the rank increasing, more and more appropriate words can be acquired. About 62.96% of all the appropriate words are included within the top 10 probable words found. The results indicate that by using a translation model, we can obtain the words that are used by the users to describe the objects with reasonable accuracy.

Table 4 presents the top 3 most probable words found for each object. It shows that although there may be more than one word appropriate to describe a given object, those words with highest probabilities always suggest the most popular way of describing the corresponding object among the users. For example, for the object with ID 26, the word `candle` gets a higher probability than the word `candlestick`, which is in accordance with our observation that in our user study, on most occasions users tend to use the word `candle` rather than the word `candlestick`.

Vocabulary interpretation is the process of finding the appropriate object(s) for any given spoken word. Out of 176 nouns in the user vocabulary, we only evaluate those used at least three times for statistical significance concerns. Further, abstract words (such as `reason, position`) and general words (such as `room, furniture`) are not evaluated since they do not refer to any particular objects in the scene. Finally, 23 nouns remain for evaluation.

We manually enumerated all the object(s) that those 23 nouns refer to as the ground truth in our evaluation. Note that a given noun can possibly be used to refer to multiple objects, such as `lamp`, since we have several lamps (with object ID 3, 8, 17, and 23) in the experiment setting, and `bed`, since bed frame, bed spread, and pillows (with object ID 19, 21, and 20 respectively) are all part of a bed. Also, an object can be referred to by multiple nouns. For example, the words `painting, picture`, or `waterfall` can all be used to refer to the object with ID 15.

Object	Rank 1	Rank 2	Rank 3
1	paint (0.254) *	wall (0.191)	left (0.150)
2	pictur (0.305) *	girl (0.122)	niagara (0.095) *
3	wall (0.109)	lamp (0.093) *	floor (0.084)
4	upsid (0.174) *	left (0.151) *	paint (0.149) *
5	pictur (0.172)	window (0.157) *	wall (0.116)
6	window (0.287) *	curtain (0.115)	pictur (0.076)
7	chair (0.287) *	tabl (0.088)	bird (0.083)
9	mirror (0.161) *	dresser (0.137)	bird (0.098) *
12	room (0.131)	lamp (0.127)	left (0.069)
14	hang (0.104)	favourit (0.085)	natur (0.064)
15	thing (0.066)	size (0.059)	queen (0.057)
16	paint (0.211) *	pictur (0.116) *	forest (0.076) *
17	lamp (0.354) *	end (0.154)	tabl (0.097)
18	bedroom (0.158)	side (0.128)	bed (0.104)
19	bed (0.576) *	room (0.059)	candl (0.049)
20	bed (0.396) *	queen (0.211) *	size (0.176)
21	bed (0.180) *	chair (0.097)	orang (0.078)
22	bed (0.282)	door (0.235) *	chair (0.128)
25	chair (0.215) *	bed (0.162)	candlestick (0.124)
26	candl (0.145) *	chair (0.114)	candlestick (0.092) *
27	tree (0.246) *	chair (0.107)	floor (0.096)

Table 4: Words found for given objects. Each row lists the top 3 most probable spoken words (being stemmed) for the corresponding given object, with the mapping probabilities in parentheses. Asterisks indicate correctly identified spoken words. Note that some objects are heavily overlapped, so the corresponding words are considered correct for all the overlapping objects, such as `bed` being considered correct for objects with ID 19, 20, and 21.

Word	Rank 1	Rank 2	Rank 3	Rank 4
curtain	6 (0.305) *	5 (0.305) *	7 (0.133)	1 (0.121)
candlestick	25 (0.147) *	28 (0.135)	24 (0.131)	22 (0.117)
lamp	22 (0.126)	12 (0.094)	17 (0.093) *	25 (0.093)
dresser	12 (0.298) *	9 (0.294) *	13 (0.173) *	7 (0.104)
queen	20 (0.187) *	21 (0.182) *	22 (0.136)	19 (0.136) *
door	22 (0.200) *	27 (0.124)	25 (0.108)	24 (0.106)
tabl	9 (0.152) *	12 (0.125) *	13 (0.112) *	22 (0.107)
mirror	9 (0.251) *	12 (0.238)	8 (0.109)	13 (0.081)
girl	2 (0.173)	22 (0.128)	16 (0.099)	10 (0.074)
chair	22 (0.132)	25 (0.099) *	28 (0.085)	24 (0.082)
waterfal	6 (0.226)	5 (0.215)	1 (0.118)	9 (0.083)
candl	19 (0.156)	22 (0.139)	28 (0.134)	24 (0.131)
niagara	4 (0.359) *	2 (0.262) *	1 (0.226)	7 (0.045)
plant	27 (0.230) *	22 (0.181)	23 (0.131)	28 (0.117)
tree	27 (0.352) *	22 (0.218)	26 (0.100)	13 (0.062)
upsid	4 (0.204) *	12 (0.188)	9 (0.153)	1 (0.104) *
bird	9 (0.142) *	10 (0.138)	12 (0.131)	7 (0.121)
desk	12 (0.170) *	9 (0.141) *	19 (0.118)	8 (0.118)
bed	19 (0.207) *	22 (0.141)	20 (0.111) *	28 (0.090)
upsidedown	4 (0.243) *	3 (0.219)	6 (0.203)	5 (0.188)
paint	4 (0.188) *	16 (0.148) *	1 (0.137) *	15 (0.118) *
window	6 (0.305) *	5 (0.290) *	3 (0.085)	22 (0.065)
lampshad	3 (0.223) *	7 (0.137)	11 (0.137)	10 (0.137)

Table 5: Objects found for given words. Each row lists the 4 most probable object IDs for the corresponding given words (being stemmed), with the mapping probabilities in parentheses. Asterisks indicate correctly identified objects. Note that some objects are heavily overlapped, such as the candle (with object ID 26) and the chair (with object ID 25), and both were considered correct for the respective spoken words.

Table 3 gives the average precision/recall evaluated at the top 10 ranks. As we can see, if we use the most probable object found for each speech word, about 78.26% of them are appropriate. With the rank increasing, more and more appropriate objects can be found. About 85.71% of all the appropriate objects are included within the top 10 probable objects found. The results indicate that by using a translation model, we can predict the objects from user spoken words with reasonable accuracy.

Table 5 lists the top 4 probable objects found for each spoken word being evaluated. A close look reveals that in general, the top ranked objects tend to gather around the correct object for a given spoken word. This is consistent with the fact that eye gaze tends to move back and forth. It also indicates that the mappings established by the translation model can effectively find the approximate area of the corresponding fixated object, even if it cannot find the object due to the noisy and jerky nature of eye gaze.

The precision/recall in vocabulary acquisition is not as high as that in vocabulary interpretation, partially due to the relatively small scale of our experiment data. For example, with only 7 users' speech data on 14 conversational tasks, some words were only spoken a few times to refer to an object, which prevented them from getting a significant portion of probability mass among all the words in the vocabulary. This degrades both precision and recall. We believe that in large scale experiments or real-world applications, the performance will be improved.

7 Discussion and Conclusion

Previous psycholinguistic findings have shown that eye gaze is tightly linked with human language production. During human machine conversation, our study shows that although a larger variance is observed on how eye fixations are exactly linked with corresponding spoken references (compared to the psycholinguistic findings), eye gaze in general is closely coupled with corresponding referring expressions in the utterances. This close coupling nature between eye gaze and speech utterances provides an opportunity for the system to automatically acquire different words related to different objects without any human supervision. To further explore this idea, we developed a novel unsupervised approach using statistical translation models.

Our experimental results have shown that this approach can reasonably uncover the mappings between words and objects on the graphical display. The main advantages of this approach include: 1) It is an unsupervised approach with minimum human inference; 2) It does not need any prior knowledge to train a statistical translation model; 3) It yields probabilities that indicate the reliability of the mappings.

Certainly, our current approach is built upon simplified assumptions. It is quite challenging to incorporate eye gaze information since it is extremely noisy with large variances. Recent work has shown that the effect of eye gaze in facilitating spoken language processing varies among different users (Qu and Chai, 2007). In addition, visual properties of the interface also affect user gaze behavior and thus influence the predication of attention (Prasov et al., 2007) based on eye gaze. Our future work will develop models to address these variations.

Nevertheless, the results from our current work have several important implications in building robust conversational interfaces. First of all, most conversational systems are built with static knowledge space (e.g., vocabularies) and can only be updated by the system developers. Our approach can potentially allow the system to automatically acquire knowledge and vocabularies based on the natural interactions with the users without human intervention. Furthermore, the automatically acquired mappings between words and objects can also help language interpretation tasks such as reference resolution. Given the recent advances in eye tracking technology (Duchowski, 2002), integrating non-intrusive and high performance eye trackers with conversational interfaces becomes feasible. The work reported here can potentially be integrated in practical systems to improve the overall robustness of human machine conversation.

Acknowledgment

This work was supported by funding from National Science Foundation (IIS-0347548, IIS-0535112, and IIS-0643494) and Disruptive Technology Office. The authors would like to thank Zahar Prasov for his contribution to data collection.

References

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

E. Campana, J. Baldridge, J. Dowding, B. A. Hockey, R. Remington, and L. S. Stone. 2001. Using eye movements to determine referents in a spoken dialog system. In *Proceedings of PUI'01*.

A. T. Duchowski. 2002. A breath-first survey of eye tracking applications. *Behavior Research methods, Instruments, and Computers*, 33(4).

Z. M. Griffin and K. Bock. 2000. What the eyes say about speaking. *Psychological Science*, 11:274–279.

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

M. Kaur, M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi. 2003. Where is "it"? Event synchronization in gaze-speech input systems. In *Proceedings of ICMI'03*, pages 151–157.

Z. Prasov, J. Y. Chai, and H. Jeong. 2007. Eye gaze for attention prediction in multimodal human-machine conversation. In *2007 Spring Symposium on Interaction Challenges for Artificial Assistants*, Palo Alto, California, March.

S. Qu and J. Y. Chai. 2007. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *NAACL'07*, pages 284–291, Rochester, New York, April.

D. Roy and A. Pentland. 2002. Learning words from sights and sounds, a computational model. *Cognitive Science*, 26(1):113–1146.

J. M. Siskind. 1995. Grounding language in perception. *Artificial Intelligence Review*, 8:371–391.

K. Tanaka. 1999. A robust selection system using real-time multi-modal user-agent interactions. In *Proceedings of IUI'99*, pages 105–108.

M. K. Tenenhaus, M. Sivey-Knowlton, E. Eberhard, and J. Sedivy. 1995. Integration of visual and linguistic information during spoken language comprehension. *Science*, 268:1632–1634.

C. Yu and D. H. Ballard. 2004. On the integration of grounding language and learning objects. *Proceedings of AAAI'04*.