The Role of Interactivity in Human-Machine Conversation for Automatic Word Acquisition

Shaolin Qu **Joyce Y. Chai**
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{qushaoli,jchai}@cse.msu.edu

Abstract

Motivated by the psycholinguistic finding that human eye gaze is tightly linked to speech production, previous work has applied naturally occurring eye gaze for automatic vocabulary acquisition. However, unlike in the typical settings for psycholinguistic studies, eye gaze can serve different functions in human-machine conversation. Some gaze streams do not link to the content of the spoken utterances and thus can be potentially detrimental to word acquisition. To address this problem, this paper investigates the incorporation of interactivity in identifying the close coupling of speech and gaze streams for word acquisition. Our empirical results indicate that automatic identification of closely coupled gaze-speech streams leads to significantly better word acquisition performance.

1 Introduction

Spoken conversational interfaces have become increasingly important in many applications such as remote interaction with robots (Lemon et al., 2002), intelligent space station control (Aist et al., 2003), and automated training and education (Razzaq and Heffernan, 2004). As in any conversational system, one major bottleneck in conversational interfaces is robust language interpretation. To address this problem, previous multimodal conversational systems have utilized pen-based or deictic gestures (Bangalore and Johnston, 2004; Qu and Chai, 2006) to improve interpretation. Besides gestures, eye movements that naturally occur during interaction provide another important channel for language understanding, for example, reference resolution (Byron et al., 2005; Prasov and Chai, 2008). Recent work has also shown that what users look at on the interface (e.g., natural scenes or generated graphic displays) during speech production provides unique opportunities for word acquisition, namely automatically acquiring semantic meanings of spoken words by grounding them to visual entities (Liu et al., 2007) or domain concepts (Qu and Chai, 2008).

Psycholinguistic studies have shown that eye gaze indicates a person's attention (Just and Carpenter, 1976), and eye movement can facilitate spoken language comprehension (Tanenhaus et al., 1995; Eberhard et al., 1995). It has been found that users' eyes move to the mentioned object directly before speaking a word (Meyer et al., 1998; Rayner, 1998; Griffin and Bock, 2000). This parallel behavior of eye gaze and speech production motivates our previous work on word acquisition (Liu et al., 2007; Qu and Chai, 2008). However, in interactive conversation, human gaze behavior is much more complex than in the typical controlled settings used in psycholinguistic studies. There are different types of eye movements (Kahneman, 1973). The naturally occurring eye gaze during speech production may serve different functions, for example, to engage in the conversation or to manage turn taking (Nakano et al., 2003). Furthermore, while interacting with a graphic display, a user could be talking about objects that were previously seen on the display or something completely unrelated to any object the user is looking at. Therefore using every speech-gaze pair for word acquisition can be detrimental. The type of gaze that is mostly useful for word acquisition is the kind that reflects the underlying attention and tightly links to the content of the co-occurring speech. Thus, one important question is how to identify the closely coupled speech and gaze streams to improve word acquisition.

To address this question, we develop an approach that incorporates interactivity (e.g., speech,

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 188–195,
Queen Mary University of London, September 2009. ©2009 Association for Computational Linguistics

user activity, conversation context) with eye gaze to identify closely coupled speech and gaze streams. We further use the identified speech and gaze streams to acquire words with a translation model. Our empirical evaluation demonstrates that automatic identification of closely coupled gaze-speech streams can lead to significantly better word acquisition performance.

2 Related Work

Previous work has explored word acquisition by grounding words to visual entities. In (Roy and Pentland, 2002), given speech paired with video images of objects, mutual information between auditory and visual signals was used to acquire words by associating acoustic phone sequences with the visual prototypes (e.g., color, size, shape) of objects. Given parallel pictures and description texts, generative models were used to acquire words by associating words with image regions in (Barnard et al., 2003). Different from this previous work, in our work, the visual attention foci accompanying speech are indicated by eye gaze. As an implicit and subconscious input, eye gaze brings additional challenges in word acquisition.

Eye gaze has been explored for word acquisition in previous work. In (Yu and Ballard, 2004), given speech paired with eye gaze and video images, a translation model was used to acquire words by associating acoustic phone sequences with visual representations of objects and actions. Word acquisition from transcribed speech and eye gaze during human-machine conversation has been investigated recently. In (Liu et al., 2007), a translation model was developed to associate words with visual objects on a graphical display. In our previous work (Qu and Chai, 2008), enhanced translation models incorporating speech-gaze temporal information and domain knowledge were developed to improve word acquisition. However, none of these previous works has investigated the role of interactivity in word acquisition, which is the focus of this paper.

3 Data Collection

We collected speech and eye gaze data through user studies. This data set is different from the data set used in our previous work (Qu and Chai, 2008). The difference lies in two aspects: 1) the data for this investigation was collected during mixed initiative human-machine conversation whereas the data in (Qu and Chai, 2008) was based only on question and answering; 2) user studies were conducted in a more complex domain for this investigation, which resulted in a richer data set that contains a larger vocabulary.

3.1 Domain

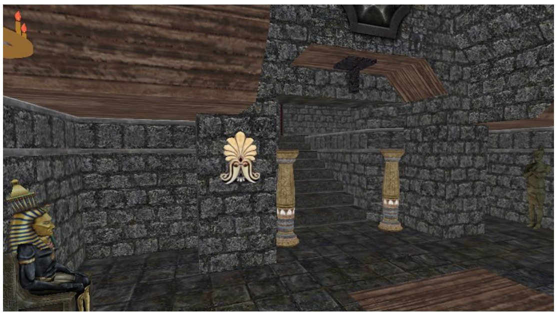


Figure 1: Treasure hunting domain

Figure 1 shows the 3D treasure hunting domain used in our work. In this application, the user needs to consult with a remote "expert" (i.e., an artificial system) to find hidden treasures in a castle with 115 3D objects. The expert has some knowledge about the treasures but can not see the castle. The user has to talk to the expert for advice regarding finding the treasures. The application is developed based on a game engine and provides an immersive environment for the user to navigate in the 3D space. During the experiment, each user's speech was recorded, and the user's eye gaze was captured by a Tobii eye tracker.

3.2 Data Preprocessing

From 20 users' experiments, we collected 3709 utterances with accompanying gaze fixations. We transcribed the collected speech. The vocabulary size of the speech transcript is 1082, among which 227 are either nouns or adjectives. The user's speech was also automatically recognized online by the Microsoft speech recognizer with a word error rate (WER) of 48.1% for the 1-best recognition. The vocabulary size of the 1-best speech recognition is 3041, among which 1643 are either nouns or adjectives.

The collected speech and gaze streams were automatically paired together by the system. Each time the system detected a sentence boundary (indicated by a long pause of 500 milliseconds) of the user's speech, it paired the recognized speech with the gaze fixations that the system had been accumulating since the previously detected sentence

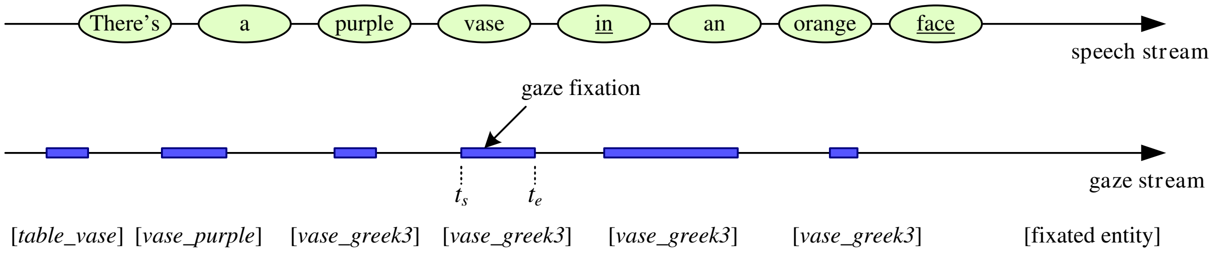


Figure 2: Accompanying gaze fixations and the 1-best recognition of a user's utterance "There's a purple vase and an orange vase." (There are two incorrectly recognized words "in" and "face" in the 1-best recognition)

boundary. Figure 2 shows a pair of user speech and accompanying stream of gaze fixations. In the speech stream, each spoken word was timestamped by the speech recognizer. In the gaze stream, each gaze fixation has a starting timestamp t_s and an ending timestamp t_e provided by the eye tracker. Each gaze fixation results in a fixated entity (3D object). When multiple entities are fixated by one gaze fixation due to the overlapping of entities, the one in the forefront is chosen.

Given the paired speech and gaze streams, we build a set of parallel word sequence and gaze fixated entity sequence $\{(\mathbf{w}, \mathbf{e})\}$ for the task of word acquisition. In section 6, we will evaluate word acquisition in two settings: 1) word sequence $\mathbf{w}$ contains all of the nouns/adjectives in the speech transcript, and 2) $\mathbf{w}$ contains all of the recognized nouns/adjectives in the 1-best speech recognition.

4 Word Acquisition With Eye Gaze

The task of word acquisition in our application is to ground words to the visual entities. Specifically, given the parallel word and entity sequences $\{(\mathbf{w}, \mathbf{e})\}$, we want to find the best match between the words and the entities. Following our previous work (Qu and Chai, 2008), we formulate word acquisition as a translation problem and use translation models for word acquisition. For each entity e, we first estimate the word-entity association probability $p(w|e)$ with a translation model, then choose the words with the highest probabilities as acquired words for e.

Inspired by the psycholinguistic findings that users' eyes move to the mentioned object before speaking a word (Meyer et al., 1998; Rayner, 1998; Griffin and Bock, 2000), in our previous work (Qu and Chai, 2008), we have incorporated the gaze-speech temporal information in the translation model as follows (referred as Model-2t through the rest of this paper):

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p_t(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i)$$

where l and m are the lengths of entity and word sequences respectively. In this equation, $p_t(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the temporal alignment probability representing the probability that w_j is aligned with e_i, which is further defined by:

$$p_t(a_j = i|j, \mathbf{e}, \mathbf{w}) = \begin{cases} 0 & d(e_i, w_j) > 0 \\ \frac{\exp[\alpha \cdot d(e_i, w_j)]}{\sum_i \exp[\alpha \cdot d(e_i, w_j)]} & d(e_i, w_j) \leq 0 \end{cases}$$

where α is a scaling factor, and $d(e_i, w_j)$ is the temporal distance between e_i and w_j. Based on the psycholinguistic finding that eye gaze happens before a spoken word, w_j is not allowed to be aligned with e_i when w_j happens earlier than e_i (i.e., $d(e_i, w_j) > 0$). When w_j happens no earlier than e_i (i.e., $d(e_i, w_j) \leq 0$), the closer they are, the more likely they are aligned. An EM algorithm is used to estimate $p(w|e)$ and α in the model.

Our evaluation in (Qu and Chai, 2008) has shown that Model-2t that incorporates temporal alignment between speech and eye gaze achieves significantly better word acquisition performance compared to the model where no temporal alignment is introduced. Therefore, this model is used for the investigation in this paper.

5 Identification of Closely Coupled Gaze-Speech Pairs

Successful word acquisition with the translation models relies on the tight coupling between the gaze fixations and the speech content. As mentioned earlier, not all gaze-speech pairs have this tight coupling. In a gaze-speech pair, if the speech does not have any word that relates to any of the gaze fixated entities, this instance only adds noise to word acquisition. Therefore, we should identify the closely coupled gaze-speech pairs and only use them for word acquisition.

In this section, we first describe the feature extraction, then evaluate the application of a logistic regression classifier to predict whether a gaze-speech pair is a **closely coupled gaze-speech instance** – an instance where at least one noun or adjective in the speech stream describes some entity fixated by the gaze stream. For the training of the classifier, we manually labeled each instance as either a coupled instance or not based on the speech transcript and the gaze fixations.

5.1 Feature Extraction

For a gaze-speech instance, the following sets of features are automatically extracted.

5.1.1 Speech Features (S)

The following features are extracted from speech:

- c_w – count of nouns and adjectives.
  More nouns and adjectives are expected in the user's utterance describing entities.
- c_w/l_s – normalized noun/adjective count.
  The effect of speech length l_s on c_w is considered.

5.1.2 Gaze Features (G)

For each fixated entity e_i, let l_e^i be its temporal fixation length. Note that several gaze fixations may have the same fixated entity, l_e^i is the total length of all the gaze fixations that fixate on entity e_i. We extract the following features from gaze stream:

- c_e – count of different gaze fixated entities.
  Fewer fixated entities are expected when the user is describing entities while looking at them.
- c_e/l_s – normalized entity count.
  The effect of temporal spoken utterance length l_s on c_e is considered.
- $\max_i(l_e^i)$ – maximal fixation length.
  At least one fixated entity's fixation is expected to be long enough when the user is describing entities while looking at them.
- $mean(l_e^i)$ – average fixation length.
  The average gaze fixation length is expected to be longer when the user is describing entities while looking at them.
- $var(l_e^i)$ – variance of fixation lengths.
  The variance of the fixation lengths is expected to be smaller when the user is describing entities while looking at them.

The number of gaze fixated entities is not only determined by the user's eye gaze, but also affected by the visual scene. Let c_e^s be the count of all the entities that have been visible during the time period concurrent with the gaze stream. We also extract the following scene related feature:

- c_e/c_e^s – scene-normalized fixated entity count.
  The effect of the visual scene on c_e is considered.

5.1.3 User Activity Features (UA)

While interacting with the system, the user's activity can also be helpful in determining whether the user's eye gaze is tightly linked to the content of the speech. The following features are extracted from the user's activities:

- *maximal distance of the user's movements* – the maximal change of user position (3D coordinates) during speech.
  The user is expected to move within a smaller range while looking at entities and describing them.
- *variance of the user's positions*
  The user is expected to move less frequently while looking at entities and describing them.

5.1.4 Conversation Context Features (CC)

While talking to the system (i.e., the "expert"), the user's language and gaze behavior are influenced by the state of the conversation. For each gaze-speech instance, we use the previous system response type as a nominal feature to predict whether this is a closely coupled gaze-speech instance.

In our treasure hunting domain, there are 8 types of system responses in 2 categories:

System Initiative Responses:

- *specific-see* – the system asks whether the user sees a certain entity, e.g., "Do you see another couch?".
- *nonspecific-see* – the system asks whether the user sees anything, e.g., "Do you see anything else?", "Tell me what you see".

- *previous-see* – the system asks whether the user has previously seen something, e.g., “Have you previously seen a similar object?”.
- *describe* – the system asks the user to describe in detail what the user sees, e.g., “Describe it”, “Tell me more about it”.
- *compare* – the system asks the user to compare what the user sees, e.g., “Compare these objects”.
- *repair-request* – the system asks the user to make clarification, e.g., “I did not understand that”, “Please repeat that”.
- *action-request* – the system asks the user to take action, e.g., “Go back”, “Try moving it”.

User Initiative Responses:

- *misc* – the system hands the initiative back to the user without specifying further requirements, e.g., “I don’t know”, “Yes”.

5.2 Evaluation of Gaze-Speech Identification

Given the extracted features and the “closely coupled” label of each instance in the training set, we train a logistic regression classifier (le Cessie and van Houwelingen, 1992) to predict whether an instance is a closely coupled gaze-speech instance.

Since the goal of identifying closely coupled gaze-speech instances is to improve word acquisition and we are only interested in acquiring nouns and adjectives, only the instances with recognized nouns/adjectives are used for training the logistic regression classifier. Among the 2969 instances with recognized nouns/adjectives and gaze fixations, 2002 (67.4%) instances are labeled as “closely coupled”. The prediction is evaluated by a 10-fold cross validation.

Feature sets	Precision	Recall
Null (*baseline*)	0.674	1
S	0.686	0.995
G	0.707	0.958
UA	0.704	0.942
CC	0.688	0.936
G + UA	0.719	0.948
G + UA + S	0.741	0.908
G + UA + CC	0.731	0.918
G + UA + CC + S	**0.748**	0.899

Table 1: Gaze-speech prediction performance for the instances with 1-best speech recognition

Table 1 shows the prediction precision and recall when different sets of features are used. As seen in the table, as more features are used, the prediction precision goes up and the recall goes down. It is important to note that prediction precision is more critical than recall for word acquisition when sufficient amount data is available. *Noisy* instances where the gaze is not coupled with the speech content will only hurt word acquisition since they will guide the translation models to ground words to the wrong entities. Although higher recall can be helpful, its effect is expected to be reduced when more data becomes available.

The results show that speech features (S) and conversation context features (CC), when used alone, do not improve prediction precision much compared to the baseline of predicting all instances as closely coupled (with a precision of 67.4%). When used alone, gaze features (G) and user activity features (UA) are the two most useful feature sets for increasing prediction precision. When they are used together, the prediction precision is further increased. Adding either speech features or conversation context features to gaze and user activity features (G + UA + S/CC) further increases the prediction precision. Using all features (G + UA + CC + S) achieves the highest prediction precision, which is significantly better than the baseline: $z = 5.93, p < 0.001$. Therefore, we choose to use all feature sets to identify the closely coupled gaze-speech instances for word acquisition.

To compare the effects of the automatic gaze-speech identification on word acquisition from various speech input (1-best speech recognition, speech transcript), we also use the logistic regression classifier with all feature sets to identify the closely coupled gaze-speech instances for the instances with speech transcript. For the instances with speech transcript, there are 2948 instances with nouns/adjectives and gaze fixations, 2128 (72.2%) of them being labeled as “closely coupled”. The prediction precision is 77.9% and the recall is 93.8%. The prediction precision is significantly better than the baseline of predicting all instances as coupled: $z = 4.92, p < 0.001$.

6 Evaluation of Word Acquisition

Every conversational system has an initial vocabulary where words are associated with domain concepts of entities. In our evaluation, we assume that the system's vocabulary has one default word for each entity that indicates the semantic type of the entity. For example, the word "barrel" is the default word for the entity *barrel*. For each entity, we only evaluate those new words that are not in the system's vocabulary.

The acquired words are evaluated against the "gold standard" words that were manually compiled for each entity and its properties based on all users' speech transcripts. For the 115 entities in our domain, each entity has 1 to 20 "gold standard" words. The average number of "gold standard" words for an entity is 6.7.

6.1 Evaluation Metrics

We evaluate the n-best acquired words (words grounded to domain concepts of entities) using precision, recall, and F-measure. When a different n is chosen, we will have different precision, recall, and F-measure.

We also evaluate the whole ranked candidate word list on Mean Reciprocal Rank Rate (MRRR) as in (Qu and Chai, 2008):

$$\text{MRRR} = \frac{\sum_e \frac{\sum_{i=1}^{N_e} 1/index(w_e^i)}{\sum_{i=1}^{N_e} 1/i}}{\#e}$$

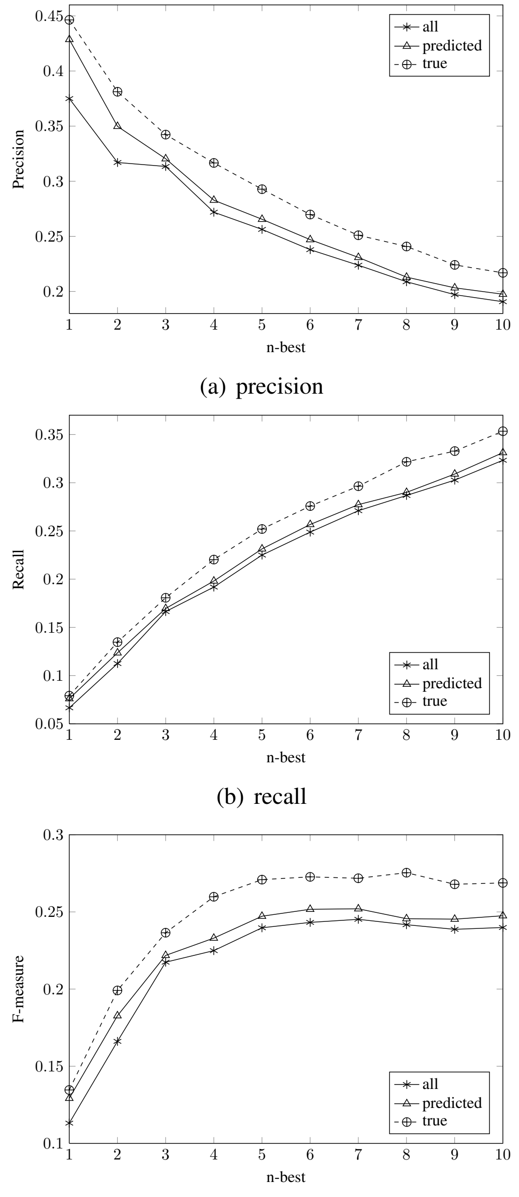
where N_e is the number of all "gold standard" words $\{w_e^i\}$ for entity e, $index(w_e^i)$ is the index of word w_e^i in the ranked list of candidate words for entity e.

MRRR measures how close the ranks of the "gold standard" words in the candidate word lists are to the best-case scenario where the top N_e words are the "gold standard" words for e. The higher the MRRR, the better is the acquisition performance.

6.2 Evaluation Results

We evaluate the effect of the closely coupled gaze-speech instances on word acquisition from the 1-best speech recognition and speech transcript. The predicted closely coupled gaze-speech instances are generated by a 10-fold cross validation with the logistic regression classifier.

Figure 3 shows the precision, recall, and F-measure of the n-best words acquired from 1-best speech recognition by Model-2t using all instances (*all*), predicted coupled instances (*predicted*), and true (manually labeled) coupled instances (*true*). As shown in the figure, using predicted coupled instances achieves consistently better performance than using all instances. These results show that the identification of coupled gaze-speech prediction helps word acquisition. When the true coupled instances are used, the performance is further improved. This means that reliable identification of coupled gaze-speech instances can lead to better word acquisition.



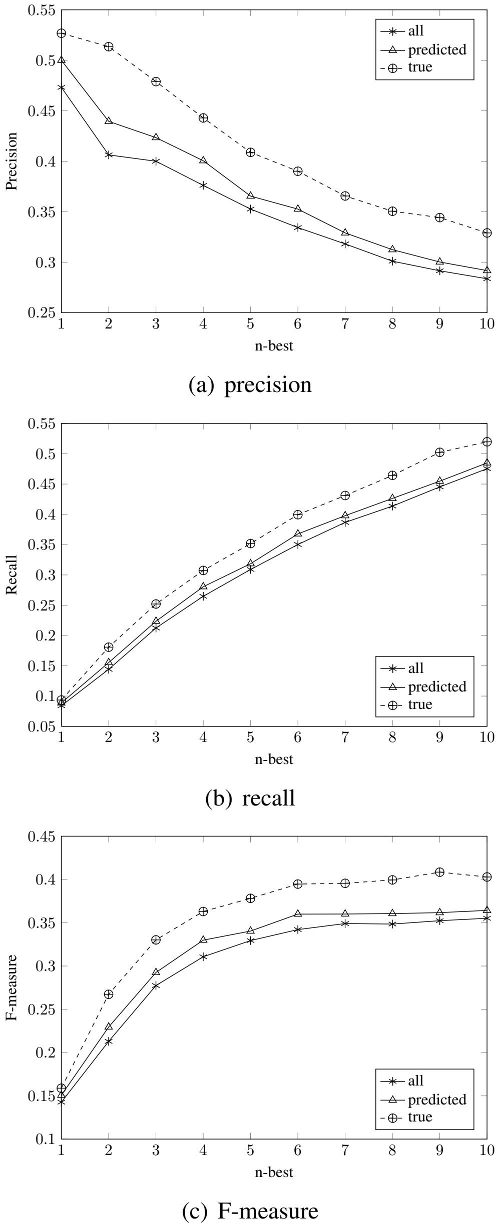
(a) precision

(b) recall

(c) F-measure

Figure 3: Performance of word acquisition on 1-best speech recognition

Figure 4 shows the precision, recall, and F-measure of the n-best words acquired from speech transcript by Model-2t using all instances, predicted coupled instances, and true coupled instances. Consistent with the performance based on the 1-best speech recognition, we can observe



(a) precision

(b) recall

(c) F-measure

Figure 4: Performance of word acquisition on speech transcript

that automatic identification of coupled instances results in better word acquisition performance and using the true coupled instances results in even better performance.

Table 2 presents the MRRRs achieved by Model-2t when words are acquired from different speech input (speech transcript, 1-best recognition) with different set of instances (all instances, predicted coupled instances, true coupled instances). These results also show the consistent behavior. Using predicted coupled instances achieves significantly better MRRR than using all instances no matter the words are acquired from 1-best speech recognition ($t = 2.59, p < 0.006$) or speech transcript($t = 3.15, p < 0.002$). When the true coupled instances are used, the performances are further improved for both 1-best recognition ($t = 2.29, p < 0.013$) and speech transcript ($t = 5.21, p < 0.001$) compared to using predicted coupled instances.

Instances	All	Predicted	True
Transcript	0.462	0.480	0.526
1-best reco	0.343	0.369	0.390

Table 2: MRRRs based on different data set

The quality of speech recognition is critical to word acquisition performance. Comparing word acquisition based on speech transcript and 1-best speech recognition, as expected, word acquisition performance on speech transcript is much better than on recognized speech. However, the acquisition performance based on speech transcript is still comparably low. For example, the recall of acquired words is still below 55% even when the 10 best word candidates are acquired for each entity. This is mainly due to the scarcity of words. Many words appear less than three times in the data, which makes them unlikely to be associated with any entity by the translation model. When more data is available, we expect to see better acquisition performance.

Note that our current evaluation is based on a two-stage approach, i.e., first identifying closely-coupled streams based on supervised classification and then automatically establishing mappings between words and entities in an unsupervised manner. There could be other approaches to address the word acquisition problem (e.g., supervised learning to directly identify whether a word is mapped to an object). Our two-stage approach has the advantage of requiring minimum supervision since the models learned from the first stage is application-independent and is potentially portable to different domains.

7 Conclusions

Unlike in the typical settings for psycholinguistic studies, human eye gaze can serve different functions during human machine conversation. Some gaze and speech streams may not be tightly coupled and thus can be detrimental to word acquisition. Therefore, this paper describes an approach that incorporates features from the interaction context to identify closely coupled gaze and speech streams. Our empirical results indicate that the word acquisition based on these automatically identified gaze-speech streams achieves significantly better performance than the word acquisition based on all gaze-speech streams. Our future work will combine gaze-based word acquisition with multiple speech recognition hypotheses (e.g., word lattices) to further improve word acquisition and language interpretation performance.

Acknowledgments

This work was supported by grants IIS-0347548 and IIS-0535112 from the National Science Foundation. We thank anonymous reviewers for their valuable comments and suggestions.

References

G. Aist, J. Dowding, B. A. Hockey, M. Rayner, J. Hieronymus, D. Bohus, B. Boven, N. Blaylock, E. Campana, S. Early, G. Gorrell, and S. Phan. 2003. Talking through procedures: An intelligent space station procedure assistant. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

S. Bangalore and M. Johnston. 2004. Robust multimodal understanding. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

D. Byron, T. Mampilly, V. Sharma, and T. Xu. 2005. Utilizing visual attention for cross-modal coreference interpretation. In *Proceedings of the Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, pages 83–96.

K. Eberhard, M. Spivey-Knowiton, J. Sedivy, and M. Tanenhaus. 1995. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24:409–436.

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

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

D. Kahneman. 1973. *Attention and Effort*. Prentice-Hall, Inc., Englewood Cliffs.

S. le Cessie and J. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

O. Lemon, A. Gruenstein, and S. Peters. 2002. Collaborative activities and multitasking in dialogue systems. *Traitement Automatique des Langues*, 43(2):131–154.

Y. Liu, J. Chai, and R. Jin. 2007. Automated vocabulary acquisition and interpretation in multimodal conversational systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*.

A. Meyer, A. Sleiderink, and W. Levelt. 1998. Viewing and naming objects: eye movements during noun phrase production. *Cognition*, 66(22):25–33.

Y. Nakano, G. Reinstein, T. Stocky, and J. Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Z. Prasov and J. Chai. 2008. What's in a gaze? the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of ACM 12th International Conference on Intelligent User interfaces (IUI)*.

S. Qu and J. Chai. 2006. Salience modeling based on non-verbal modalities for spoken language understanding. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 193–200.

S. Qu and J. Chai. 2008. Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 244–253.

K. Rayner. 1998. Eye movements in reading and information processing - 20 years of research. *Psychological Bulletin*, 124(3):372–422.

L. Razzaq and N. Heffernan. 2004. Tutorial dialog in an equation solving intelligent tutoring system. In *Proceedings of the Workshop on Dialog-based Intelligent Tutoring Systems: State of the art and new research directions*.

D. Roy and A. Pentland. 2002. Learning words from sights and sounds, a computational model. *Cognitive Science*, 26(1):113–146.

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

C. Yu and D. Ballard. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions*, 1(1):57–80.