Incorporating Temporal and Semantic Information with Eye Gaze for Automatic Word Acquisition in Multimodal Conversational Systems

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

Abstract

One major bottleneck in conversational systems is their incapability in interpreting unexpected user language inputs such as out-of-vocabulary words. To overcome this problem, conversational systems must be able to learn new words automatically during human machine conversation. Motivated by psycholinguistic findings on eye gaze and human language processing, we are developing techniques to incorporate human eye gaze for automatic word acquisition in multimodal conversational systems. This paper investigates the use of temporal alignment between speech and eye gaze and the use of domain knowledge in word acquisition. Our experiment results indicate that eye gaze provides a potential channel for automatically acquiring new words. The use of extra temporal and domain knowledge can significantly improve acquisition performance.

1 Introduction

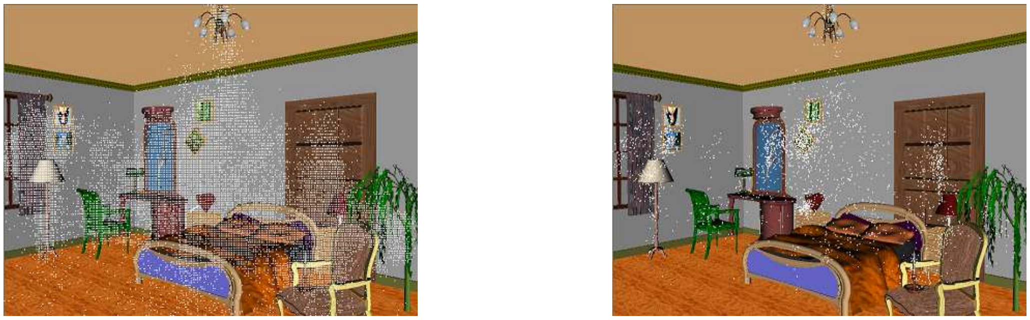
Interpreting human language is a challenging problem in human machine conversational systems due to the flexibility of human language behavior. When the encountered vocabulary is outside of the system's knowledge, conversational systems tend to fail. It is desirable that conversational systems can learn new words automatically during human machine conversation. While automatic word acquisition in general is quite challenging, multimodal conversational systems offer an unique opportunity to explore word acquisition. In a multimodal conversational system where users can talk and interact with a graphical display, users' eye gaze, which occurs naturally with speech production, provides a potential channel for the system to learn new words automatically during human machine conversation.

Psycholinguistic studies have shown that eye gaze is tightly linked to human language processing. Eye gaze is one of the reliable indicators of what a person is "thinking about" (Henderson and Ferreira, 2004). The direction of eye gaze carries information about the focus of the user's attention (Just and Carpenter, 1976). The perceived visual context influences spoken word recognition and mediates syntactic processing of spoken sentences (Tanenhaus et al., 1995). In addition, directly before speaking a word, the eyes move to the mentioned object (Griffin and Bock, 2000).

Motivated by these psycholinguistic findings, we are investigating the use of eye gaze for automatic word acquisition in multimodal conversation. Particulary, this paper investigates the use of temporal information about speech and eye gaze and domain semantic relatedness for automatic word acquisition. The domain semantic and temporal information are incorporated in statistical translation models for word acquisition. Our experiments show that the use of domain semantic and temporal information significantly improves word acquisition performance.

In the following sections, we first describe the basic translation models for word acquisition. Then, we describe the enhanced models that incorporate temporal and semantic information about speech and eye gaze for word acquisition. Finally, we present the results of empirical evaluation.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 244–253,
Honolulu, October 2008. ©2008 Association for Computational Linguistics



(a) Raw gaze points

(b) Processed gaze fixations

Figure 1: Domain scene with a user's gaze fixations

2 Related Work

Word acquisition by grounding words to visual entities has been studied in many language grounding systems. For example, given speech paired with video images of single objects, mutual information between audio and visual signals was used to acquire words by associating acoustic phone sequences with the visual prototypes (e.g., color, size, shape) of objects (Roy and Pentland, 2002). Generative models were used to acquire words by associating words with image regions given parallel data of pictures and description text (Barnard et al., 2003). Different from these works, in our work, the visual attention foci accompanying speech are indicated by eye gaze. Eye gaze is an implicit and subconscious input, which 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 information and video images, a translation model was used to acquire words by associating acoustic phone sequences with visual representations of objects and actions. A recent investigation on word acquisition from transcribed speech and eye gaze in human machine conversation was reported in (Liu et al., 2007). In this work, a translation model was developed to associate words with visual objects on a graphical display. Different from these previous works, here we investigate the incorporation of extra knowledge, specifically speech-gaze temporal information and domain knowledge, with eye gaze to facilitate word acquisition.

3 Data Collection

We recruited users to interact with a simplified multimodal conversational system to collect speech and eye gaze data.

3.1 Domain

We are working on a 3D room decoration domain. Figure 1 shows the 3D room scene that was shown to the user in the experiments. There are 28 3D objects (bed, chairs, paintings, lamp, etc.) in the room scene. During the human machine conversation, the system verbally asked the user a question (e.g., "*what do you dislike about the arrangement of the room*?") or issued a request (e.g., "*describe the left wall*") about the room. The user provided responses by speaking to the system.

During the experiments, users' speech was recorded through an open microphone and users' eye gaze was captured by an Eye Link II eye tracker. Eye gaze data consists of the screen coordinates of each gaze point that was captured by the eye tracker at a sampling rate of 250hz.

3.2 Data Preprocessing

As for speech data, we collected 357 spoken utterances from 7 users' experiments. The vocabulary size is 480, among which 227 words are nouns and adjectives. We manually transcribed the collected speech.

As for gaze data, the first step is to identify gaze fixation from raw gaze points. As shown in Figure 1(a), the collected raw gaze points are very noisy. They can not be used directly for identifying gaze fixated entities in the scene. We processed the raw gaze data to eliminate invalid and saccadic gaze points. Invalid gaze points occur when users look off the screen. Saccadic gaze points occur during ballistic eye movements between gaze fixations. Vision studies have shown that no visual processing occurs in the human mind during saccades (i.e., saccadic suppression) (Matin, 1974). Since eyes do not stay still but rather make small, frequent jerky movements, we average nearby gaze points to better identify gaze fixations. The processed eye gaze fixations are shown in Figure 1(b).

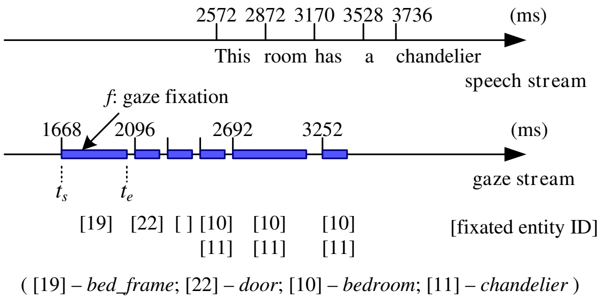


Figure 2: Parallel speech and gaze streams

Figure 2 shows an excerpt of the collected speech and gaze fixation in one experiment. In the speech stream, each word starts at a particular timestamp. In the gaze stream, each gaze fixation has a starting timestamp t_s and an ending timestamp t_e. Each gaze fixation also has a list of fixated entities (3D objects). An entity e on the graphical display is fixated by gaze fixation f if the area of e contains fixation point of f.

Given the collected speech and gaze fixations, we build parallel speech-gaze data set as follows. For each spoken utterance and its accompanying gaze fixations, we construct a pair of word sequence and entity sequence $(\mathbf{w}, \mathbf{e})$. The word sequence $\mathbf{w}$ consists of only nouns and adjectives in the utterance. Each gaze fixation results in a fixated entity in the entity sequence $\mathbf{e}$. When multiple entities are fixated by one gaze fixation due to the overlapping of the entities, the forefront one is chosen. Also, we merge the neighboring gaze fixations that contain the same fixated entities. For the parallel speech and gaze streams shown in Figure 2, the resulting word sequence is $\mathbf{w}$ = [room chandelier] and the entity sequence is $\mathbf{e}$ = [*bed_frame door chandelier*].

4 Translation Models for Automatic Word Acquisition

Since we are working on conversational systems where users interact with a visual scene, we consider the task of word acquisition as associating words with visual entities in the domain. Given the parallel speech and gaze fixated entities $\{(\mathbf{w}, \mathbf{e})\}$, we formulate word acquisition as a translation problem and use translation models to estimate word-entity association probabilities $p(w|e)$. The words with the highest association probabilities are chosen as acquired words for entity e.

4.1 Base Model I

Using the translation model I (Brown et al., 1993), where each word is equally likely to be aligned with each entity, we have

$$p(\mathbf{w}|\mathbf{e}) = \frac{1}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} p(w_j|e_i) \qquad (1)$$

where l and m are the lengths of entity and word sequences respectively. This is the model used in (Liu et al., 2007) and (Yu and Ballard, 2004). We refer to this model as **Model-1** throughout the rest of this paper.

4.2 Base Model II

Using the translation model II (Brown et al., 1993), where alignments are dependent on word/entity positions and word/entity sequence lengths, we have

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p(a_j = i|j, m, l) p(w_j|e_i) \qquad (2)$$

where $a_j = i$ means that w_j is aligned with e_i. When $a_j = 0$, w_j is not aligned with any entity (e_0 represents a *null* entity). We refer to this model as **Model-2**.

Compared to Model-1, Model-2 considers the ordering of words and entities in word acquisition. EM algorithms are used to estimate the probabilities $p(w|e)$ in the translation models.

5 Using Speech-Gaze Temporal Information for Word Acquisition

In Model-2, word-entity alignments are estimated from co-occurring word and entity sequences in an unsupervised way. The estimated alignments are dependent on where the words/entities appear in the word/entity sequences, not on when those words and gaze fixated entities actually occur. Motivated by the finding that users move their eyes to the mentioned object directly before speaking a word (Griffin and Bock, 2000), we make the word-entity alignments dependent on their temporal relation in a new model (referred as **Model-2t**):

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p_t(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i) \quad (3)$$

where $p_t(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the temporal alignment probability computed based on the temporal distance between entity e_i and word w_j.

We define the temporal distance between e_i and w_j as

$$d(e_i, w_j) = \begin{cases} 0 & t_s(e_i) \leq t_s(w_j) \leq t_e(e_i) \\ t_e(e_i) - t_s(w_j) & t_s(w_j) > t_e(e_i) \\ t_s(e_i) - t_s(w_j) & t_s(w_j) < t_s(e_i) \end{cases} \quad (4)$$

where $t_s(w_j)$ is the starting timestamp (ms) of word w_j, $t_s(e_i)$ and $t_e(e_i)$ are the starting and ending timestamps (ms) of gaze fixation on entity e_i.

The alignment of word w_j and entity e_i is decided by their temporal distance $d(e_i, w_j)$. 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. Specifically, the temporal alignment probability of w_j and e_i in each co-occurring instance $(\mathbf{w}, \mathbf{e})$ is computed as

$$p_t(a_j = i|j, \mathbf{e}, \mathbf{w}) = \begin{cases} 0 & d(e_i, w_j) > 0 \\ \frac{\exp[\alpha \cdot d(e_i, w_j)]}{\Sigma_i \exp[\alpha \cdot d(e_i, w_j)]} & d(e_i, w_j) \leq 0 \end{cases} \quad (5)$$

where α is a constant for scaling $d(e_i, w_j)$. In our experiments, α is set to 0.005.

An EM algorithm is used to estimate probabilities $p(w|e)$ in Model-2t.

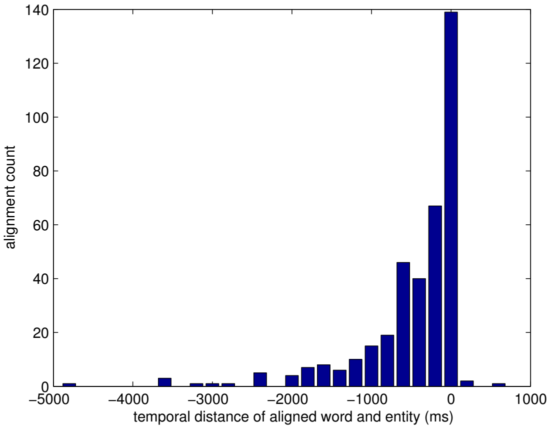


Figure 3: Histogram of truly aligned word and entity pairs over temporal distance (bin width = 200ms)

For the purpose of evaluation, we manually annotated the truly aligned word and entity pairs. Figure 3 shows the histogram of those truly aligned word and entity pairs over the temporal distance of aligned word and entity. We can observe in the figure that 1) almost no eye gaze happens after a spoken word, and 2) the number of word-entity pairs with closer temporal distance is generally larger than the number of those with farther temporal distance. This is consistent with our modeling of the temporal alignment probability of word and entity (Equation (5)).

6 Using Domain Semantic Relatedness for Word Acquisition

Speech-gaze temporal alignment and occurrence statistics sometimes are not sufficient to associate words to an entity correctly. For example, suppose a user says "*there is a lamp on the dresser*" while looking at a lamp object on a table object. Due to their co-occurring with the lamp object, words *dresser* and *lamp* are both likely to be associated with the lamp object in the translation models. As a result, word *dresser* is likely to be incorrectly acquired for the lamp object. For the same reason, the word *lamp* could be acquired incorrectly for the table object. To solve this type of association problem, the semantic knowledge about the domain and words can be helpful. For example, the knowledge that the word *lamp* is more semantically related to the object lamp can help the system avoid associating the word *dresser* to the lamp object. Therefore, we are interested in investigating the use of semantic knowledge in word acquisition.

On one hand, each conversational system has a *domain model*, which is the knowledge representation about its domain such as the types of objects and their properties and relations. On the other hand, there are available resources about domain independent lexical knowledge (e.g., WordNet (Fellbaum, 1998)). The question is whether we can utilize the domain model and external lexical knowledge resource to improve word acquisition. To address this question, we link the domain concepts in the domain model with WordNet concepts, and define semantic relatedness of word and entity to help the system acquire domain semantically compatible words.

In the following sections, we first describe our domain modeling, then define the semantic relatedness of word and entity based on domain modeling and WordNet semantic lexicon, and finally describe different ways of using the semantic relatedness of word and entity to help word acquisition.

6.1 Domain Modeling

We model the 3D room decoration domain as shown in Figure 4. The domain model contains all domain related semantic concepts. These concepts are linked to the WordNet concepts (i.e., synsets in the format of "word#part-of-speech#sense-id"). Each of the entities in the domain has one or more properties (e.g., semantic type, color, size) that are denoted by domain concepts. For example, the entity *dresser_1* has domain concepts *SEM_DRESSER* and *COLOR*. These domain concepts are linked to "dresser#n#4" and "color#n#1" in WordNet.

Note that in the domain model, the domain concepts are not specific to a certain entity, they are general concepts for a certain type of entity. Multiple entities of the same type have the same properties and share the same set of domain concepts.

6.2 Semantic Relatedness of Word and Entity

We compute the semantic relatedness of a word w and an entity e based on the semantic similarity between w and the properties of e. Specifically, semantic relatedness $SR(e, w)$ is defined as

$$SR(e, w) = \max_{i,j} sim(s(c_e^i), s_j(w)) \qquad (6)$$

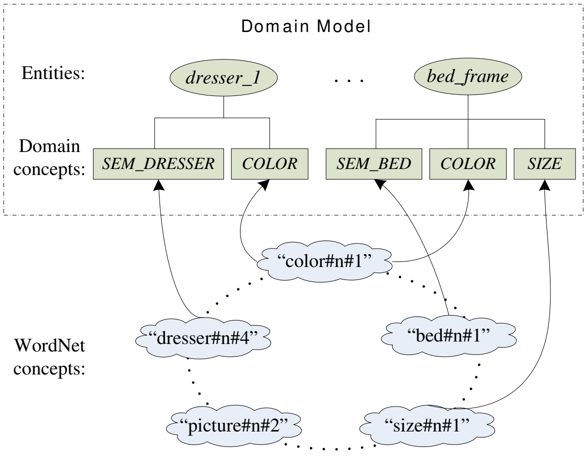


Figure 4: Domain model with domain concepts linked to WordNet synsets

where c_e^i is the i-th property of entity e, $s(c_e^i)$ is the synset of property c_e^i as designed in domain model, $s_j(w)$ is the j-th synset of word w as defined in WordNet, and $sim(\cdot, \cdot)$ is the similarity score of two synsets.

We computed the similarity score of two synsets based on the path length between them. The similarity score is inversely proportional to the number of nodes along the shortest path between the synsets as defined in WordNet. When the two synsets are the same, they have the maximal similarity score of 1. The WordNet-Similarity tool (Pedersen et al., 2004) was used for the synset similarity computation.

6.3 Word Acquisition with Word-Entity Semantic Relatedness

We can use the semantic relatedness of word and entity to help the system acquire semantically compatible words for each entity, and therefore improve word acquisition performance. The semantic relatedness can be applied for word acquisition in two ways: post process learned word-entity association probabilities by rescoring them with semantic relatedness, or directly affect the learning of word-entity associations by constraining the alignment of word and entity in the translation models.

6.3.1 Rescoring with semantic relatedness

In the acquired word list for an entity e_i, each word w_j has an association probability $p(w_j|e_i)$ that is learned from a translation model. We use the semantic relatedness $SR(e_i, w_j)$ to redistribute the probability mass for each w_j. The new association probability is given by:

$$p'(w_j|e_i) = \frac{p(w_j|e_i)SR(e_i, w_j)}{\sum_j p(w_j|e_i)SR(e_i, w_j)} \qquad (7)$$

6.3.2 Semantic alignment constraint in translation model

When used to constrain the word-entity alignment in the translation model, semantic relatedness can be used alone or used together with speech-gaze temporal information to decide the alignment probability of word and entity.

- Using only semantic relatedness to constrain word-entity alignments in **Model-2s**, we have

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i) \qquad (8)$$

  where $p_s(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the alignment probability based on semantic relatedness,

$$p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) = \frac{SR(e_i, w_j)}{\sum_i SR(e_i, w_j)} \qquad (9)$$

- Using semantic relatedness and temporal information to constrain word-entity alignments in **Model-2ts**, we have

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i) \qquad (10)$$

  where $p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the alignment probability that is decided by both temporal relation and semantic relatedness of e_i and w_j,

$$p_{ts}(a_j = i|j, \mathbf{e}, \mathbf{w}) = \frac{p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) p_t(a_j = i|j, \mathbf{e}, \mathbf{w})}{\sum_i p_s(a_j = i|j, \mathbf{e}, \mathbf{w}) p_t(a_j = i|j, \mathbf{e}, \mathbf{w})} \qquad (11)$$

  where $p_s(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the semantic alignment probability in Equation (9), and $p_t(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the temporal alignment probability given in Equation (5).

EM algorithms are used to estimate $p(w|e)$ in Model-2s and Model-2ts.

7 Grounding Words to Domain Concepts

As discussed above, based on translation models, we can incorporate temporal and domain semantic information to obtain $p(w|e)$. This probability only provides a means to ground words to entities. In conversational systems, the ultimate goal of word acquisition is to make the system understand the semantic meaning of new words. Word acquisition by grounding words to objects is not always sufficient for identifying their semantic meanings. Suppose the word *green* is grounded to a green chair object, so is the word *chair*. Although the system is aware that *green* is some word describing the green chair, it does not know that word *green* refers to the chair's color while the word *chair* refers to the chair's semantic type. Thus, after learning the word-entity associations $p(w|e)$ by the translation models, we need to further ground words to domain concepts of entity properties.

We further apply WordNet to ground words to domain concepts. For each entity e, based on association probabilities $p(w|e)$, we can choose the n-best words as acquired words for e. Those n-best words have the n highest association probabilities. For each word w acquired for e, the grounded concept c_e^* for w is chosen as the one that has the highest semantic relatedness with w:

$$c_e^* = \arg\max_i \left[\max_j sim(s(c_e^i), s_j(w))\right] \qquad (12)$$

where $sim(s(c_e^i), s_j(w))$ is the semantic similarity score defined in Equation (6).

8 Evaluation

We evaluate word acquisition performance of different models on the data collected from our user studies (see Section 3).

8.1 Evaluation Metrics

The following metrics are used to evaluate the words acquired for domain concepts (i.e., entity properties) $\{c_e^i\}$.

- Precision

$$\frac{\sum_e \sum_i \text{\# words correctly acquired for } c_e^i}{\sum_e \sum_i \text{\# words acquired for } c_e^i}$$

- Recall

$$\frac{\sum_e \sum_i \text{\# words correctly acquired for } c_e^i}{\sum_e \sum_i \text{\# ground-truth[1] words of } c_e^i}$$

- F-measure

$$\frac{2 \times precision \times recall}{precision + recall}$$

The metrics of precision, recall, and F-measure are based on the n-best words acquired for the entity properties. Therefore, we have different precision, recall, and F-measure when n changes.

The metrics of precision, recall, and F-measure only provide evaluation on the top n candidate words. To measure the acquisition performance on the entire ranked list of candidate words, we define a new metric as follows:

- Mean Reciprocal Rank Rate (MRRR)

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

  where N_e is the number of all ground-truth 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.

Entities may have a different number of ground-truth words. For each entity e, we calculate a Reciprocal Rank Rate (RRR), which measures how close the ranks of the ground-truth words in the candidate word list is to the best scenario where the top N_e words are the ground-truth words for e. RRR is in the range of $(0, 1]$. The higher the RRR, the better is the word acquisition performance. The average of RRRs across all entities gives the Mean Reciprocal Rank Rate (MRRR).

Note that MRRR is directly based on the learned word-entity associations $p(w|e)$, it is in fact a measure of grounding words to entities.

8.2 Evaluation Results

To compare the effects of different speech-gaze alignments on word acquisition, we evaluate the following models:

- Model-1 – base model I without word-entity alignment (Equation (1)).
- Model-2 – base model II with positional alignment (Equation (2)).
- Model-2t – enhanced model with temporal alignment (Equation (3)).
- Model-2s – enhanced model with semantic alignment (Equation (8)).
- Model-2ts – enhanced model with both temporal and semantic alignment (Equation (10)).

To compare the different ways of incorporating semantic relatedness in word acquisition as discussed in Section 6.3.1, we also evaluate the following models:

- Model-1-r – Model-1 with semantic relatedness rescoring of word-entity association.
- Model-2t-r – Model-2t with semantic relatedness rescoring of word-entity association.

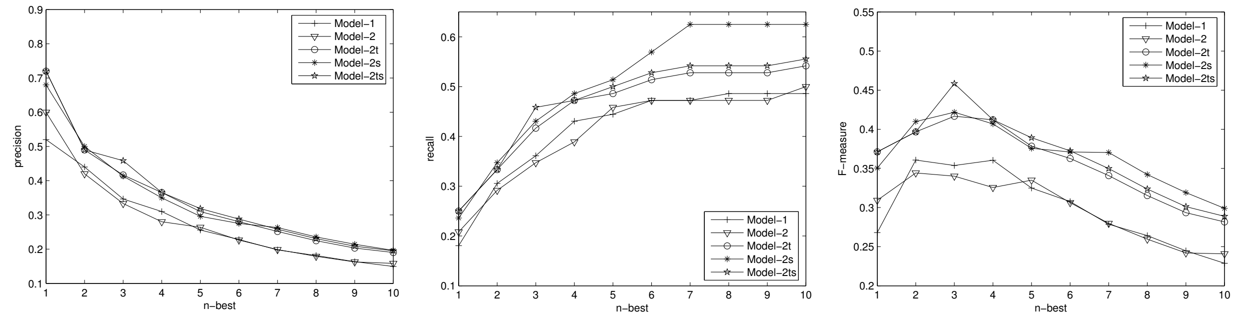
Figure 5 shows the results of models with different speech-gaze alignments. Figure 6 shows the results of models with semantic relatedness rescoring. In Figure 5 & 6, *n-best* means the top *n* word candidates are chosen as acquired words for each entity. The Mean Reciprocal Rank Rates of all models are compared in Figure 7.

8.2.1 Results of using different speech-gaze alignments

As shown in Figure 5, Model-2 does not show a consistent improvement compared to Model-1 when a different number of n-best words are chosen as acquired words. This result shows that it is not very helpful to consider the index-based positional alignment of word and entity for word acquisition.

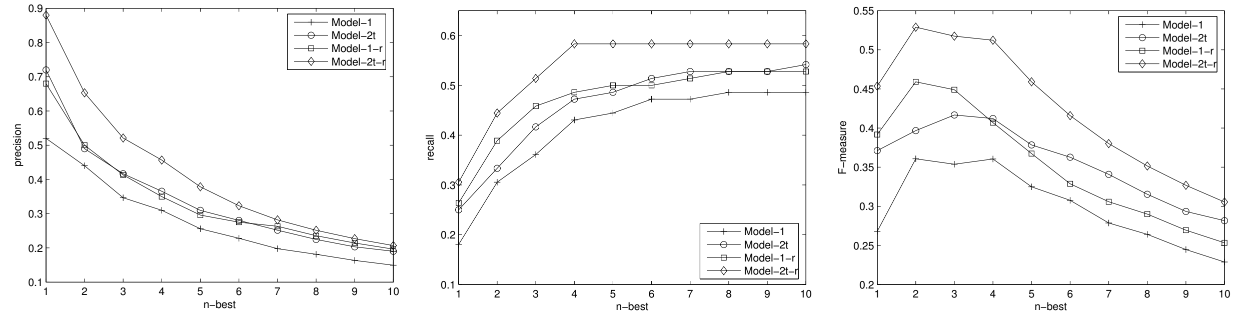
Figure 5 also shows that models considering temporal or/and semantic information (Model-2t, Model-2s, Model-2ts) consistently perform better than the models considering neither temporal nor

[1] The ground-truth words were compiled and agreed upon by two human judges.



(a) precision (b) recall (c) F-measure

Figure 5: Performance of word acquisition when different types of speech-gaze alignment are applied



(a) precision (b) recall (c) F-measure

Figure 6: Performance of word acquisition when semantic relatedness rescoring of word-entity association is applied

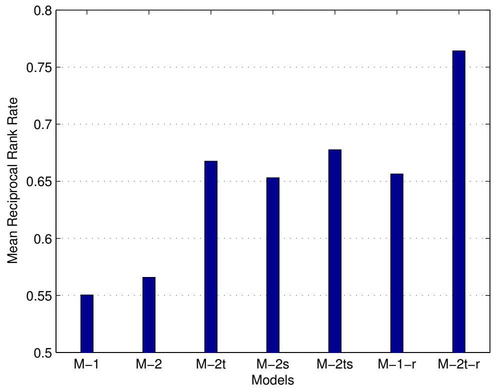


Figure 7: MRRRs achieved by different models

semantic information (Model-1, Model-2). Among Model-2t, Model-2s, and Model-2ts, it is found that they do not make consistent differences.

As shown in Figure 7, the MRRRs of different models are consistent with their performances on F-measure. A t-test has shown that the difference between the MRRRs of Model-1 and Model-2 is not statistically significant. Compared to Model-1, t-tests have confirmed that MRRR is significantly improved by Model-2t ($t = 2.27, p < 0.02$), Model-2s ($t = 3.40, p < 0.01$), and Model-2ts($t = 2.60, p < 0.01$). T-tests have shown no significant differences among Model-2t, Model-2s, and Model-2ts.

8.2.2 Results of applying semantic relatedness rescoring

Figure 6 shows that semantic relatedness rescoring improves word acquisition. After semantic relatedness rescoring of the word-entity associations learned by Model-1, Model-1-r improves the F-measure consistently when a different number of n-best words are chosen as acquired words. Compared to Model-2t, Model-2t-r also improves the F-measure consistently.

Comparing the two ways of using semantic relatedness for word acquisition, it is found that rescoring word-entity association with semantic relatedness works better. When semantic relatedness is used together with temporal information to constrain word-entity alignments in Model-2ts, word acqui-

Model	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
M-1	**table**(0.173)	**dresser**(0.067)	area(0.058)	picture(0.053)	dressing(0.041)
M-2t	**table**(0.146)	**dresser**(0.125)	dressing(0.061)	**vanity**(0.051)	fact(0.050)
M-2t-r	**table**(0.312)	**dresser**(0.241)	**vanity**(0.149)	**desk**(0.047)	area(0.026)

Table 1: N-best candidate words acquired for the entity *dresser_1* by different models

sition performance is not improved compared to Model-2t. However, using semantic relatedness to rescore word-entity association learned by Model-2t, Model-2t-r further improves word acquisition.

As shown in Figure 7, the MRRRs of Model-1-r and Model-2t-r are consistent with their performances on F-measure. Compared to Model-2t, Model-2t-r improves MRRR. A t-test has confirmed that this is a significant improvement ($t = 1.97, p < 0.03$). Compared to Model-1, Model-1-r significantly improves MRRR ($t = 2.33, p < 0.02$). There is no significant difference between Model-1-r and Model-2t/Model-2s/Model-2ts.

In Figures 5&6, we also notice that the recall of the acquired words is still comparably low even when 10 best word candidates are chosen for each entity. This is mainly due to the scarcity of those words that are not acquired in the data. Many of the words that are not acquired appear less than 3 times in the data, which makes them unlikely to be associated with any entity by the translation models. When more data is available, we expect to see higher recall.

8.3 An Example

Table 1 shows the 5-best words acquired by different models for the entity *dresser_1* in the 3d room scene (see Figure 1). In the table, each word is followed by its word-entity association probability $p(w|e)$. The correctly acquired words are shown in bold font.

As shown in the example, the baseline Model-1 learned 2 correct words in the 5-best list. Considering speech-gaze temporal information, Model-2t learned one more correct word *vanity* in the 5-best list. With semantic relatedness rescoring, Model-2t-r further acquired word *desk* in the 5-best list because of the high semantic relatedness of word *desk* and the type of entity *dresser_1*. Although neither Model-1 nor Model-2t successfully acquired the word *desk* in the 5-best list, the rank (=7) of the word *desk* in Model-2t's n-best list is much higher than the rank (=21) in Model-1's n-best list.

9 Conclusion

Motivated by the psycholinguistic findings, we investigate the use of eye gaze for automatic word acquisition in multimodal conversational systems. Particularly, we investigate the use of speech-gaze temporal information and word-entity semantic relatedness to facilitate word acquisition. Our experiments show that word acquisition is significantly improved when temporal information is considered, which is consistent with the previous psycholinguistic findings about speech and eye gaze. Moreover, using temporal information together with semantic relatedness rescoring further improves word acquisition.

Eye tracking systems are no longer bulky systems that prevent natural human machine communication. Display mounted gaze tracking systems (e.g., Tobii) are completely non-intrusive, can tolerate head motion, and provide high tracking quality. Integrating eye tracking with conversational interfaces is no longer beyond reach. Recent works have shown that eye gaze can facilitate spoken language processing in conversational systems (Qu and Chai, 2007; Prasov and Chai, 2008). Incorporating eye gaze with automatic word acquisition provides another potential approach to improve the robustness of human machine conversation.

Acknowledgments

This work was supported by IIS-0347548 and IIS-0535112 from the National Science Foundation. The authors would like to thank Zahar Prasov for his contribution on data collection. The authors would also like to thank anonymous reviewers for their valuable comments and suggestions.

References

Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. 2003.

Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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

John M. Henderson and Fernanda Ferreira, editors. 2004. *The interface of language, vision, and action: Eye movements and the visual world*. New York: Taylor & Francis.

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

Yi Liu, Joyce Y. Chai, and Rong 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)*.

E. Matin. 1974. Saccadic suppression: a review and an analysis. *Psychological Bulletin*, 81:899–917.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*.

Zahar Prasov and Joyce Y. 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)*.

Shaolin Qu and Joyce Y. Chai. 2007. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *Proceedings of the Conference of the North America Chapter of the Association of Computational Linguistics (NAACL)*.

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

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

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