Fusing Eye Gaze with Speech Recognition Hypotheses to Resolve Exophoric References in Situated Dialogue

Zahar Prasov and Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
{prasovza,jchai}@cse.msu.edu

Abstract

In situated dialogue humans often utter linguistic expressions that refer to extralinguistic entities in the environment. Correctly resolving these references is critical yet challenging for artificial agents partly due to their limited speech recognition and language understanding capabilities. Motivated by psycholinguistic studies demonstrating a tight link between language production and human eye gaze, we have developed approaches that integrate naturally occurring human eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue in a virtual world. In addition to incorporating eye gaze with the best recognized spoken hypothesis, we developed an algorithm to also handle multiple hypotheses modeled as word confusion networks. Our empirical results demonstrate that incorporating eye gaze with recognition hypotheses consistently outperforms the results obtained from processing recognition hypotheses alone. Incorporating eye gaze with word confusion networks further improves performance.

1 Introduction

Given a rapid growth in virtual world applications for tutoring and training, video games and simulations, and assistive technology, enabling situated dialogue in virtual worlds has become increasingly important. Situated dialogue allows human users to navigate in a spatially rich environment and carry a conversation with artificial agents to achieve specific tasks pertinent to the environment. Different from traditional telephony-based spoken dialogue systems and multimodal conversational interfaces, situated dialogue supports immersion and mobility in a visually rich environment and encourages social and collaborative language use (Byron et al., 2005; Gorniak et al., 2006). In situated dialogue, human users often need to make linguistic references, known as exophoric referring expressions (e.g., `the book to the right`), to extralinguistic entities in the environment. Reliably resolving these references is critical for dialogue success. However, reference resolution remains a challenging problem, partly due to limited speech and language processing capabilities caused by poor speech recognition (ASR), ambiguous language, and insufficient pragmatic knowledge.

To address this problem, motivated by psycholinguistic studies demonstrating a close relationship between language production and eye gaze, our previous work has incorporated naturally occurring eye gaze in reference resolution (Prasov and Chai, 2008). Our findings have shown that eye gaze can partially compensate for limited language processing and domain modeling. However, this work was conducted in a setting where users only spoke to a static visual interface. In situated dialogue, human speech and eye gaze patterns are much more complex. The dynamic nature of the environment and the complexity of spatially rich tasks have a massive influence on what the user will look at and say. It is not clear to what degree prior findings can generalize to situated dialogue. Therefore, this paper explores new studies on incorporating eye gaze for exophoric reference resolution in a fully situated virtual envi-

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 471–481,
MIT, Massachusetts, USA, 9-11 October 2010.

ronment — a more realistic approximation of real world interaction. In addition to incorporating eye gaze with the best recognized spoken hypothesis, we developed an algorithm to also handle multiple hypotheses modeled as word confusion networks.

Our empirical results have demonstrated the utility of eye gaze for reference resolution in situated dialogue. Although eye gaze is much more noisy given the mobility of the user, our results have shown that incorporating eye gaze with recognition hypotheses consistently outperform the results obtained from processing recognition hypotheses alone. In addition, incorporating eye gaze with word confusion networks further improves performance. Our analysis also indicates that, although a word confusion network appears to be more complicated, the time complexity of its integration with eye gaze is well within the acceptable range for real-time applications.

2 Related Work

Prior work in reference resolution within situated dialogue has focused on using visual context to assist reference resolution during interaction. In (Kelleher and van Genabith, 2004) and (Byron et al., 2005), visual features of objects are used to model the focus of attention. This attention modeling is subsequently used to resolve references. In contrast to this line of research, here we explore the use of human eye gaze during real-time interaction to model attention and facilitate reference resolution. Eye gaze provides a richer medium for attentional information, but requires processing of a potentially noisy signal.

Eye gaze has been used to facilitate human machine conversation and automated language processing. For example, eye gaze has been studied in embodied conversational discourse as a mechanism to gather visual information, aid in thinking, or facilitate turn taking and engagement (Nakano et al., 2003; Bickmore and Cassell, 2004; Sidner et al., 2004; Morency et al., 2006; Bee et al., 2009). Recent work has explored incorporating eye gaze into automated language understanding such as automated speech recognition (Qu and Chai, 2007; Cooke and Russell, 2008), automated vocabulary acquisition (Liu et al., 2007; Qu and Chai, 2010), attention prediction (Qvarfordt and Zhai, 2005; Fang et al., 2009).

Motivated by previous psycholinguistic findings that eye gaze is tightly linked with language processing (Just and Carpenter, 1976; Tanenhous et al., 1995; Meyer and Levelt, 1998; Griffin and Bock, 2000), our prior work incorporates eye gaze into reference resolution. Our results demonstrate that such use of eye gaze can potentially compensate for a conversational systems limited language processing and domain modeling capability (Prasov and Chai, 2008). However, this work is conducted in a static visual environment and evaluated only on transcribed spoken utterances. In situated dialogue, eye gaze behavior is much more complex. Here, gaze fixations may be made for the purpose of navigation or scanning the environment rather than referring to a particular object. Referring expressions can be made to objects that are not in the user's field of view, but were previously visible on the interface. Additionally, users may make egocentric spatial references (e.g. "the chair on the left") which require contextual knowledge (e.g. the users position in the environment) in order to resolve. Therefore, the focus of our work here is on exploring these complex user behaviors in situated dialogue and examining how to combine eye gaze with ASR hypotheses for improved reference resolution.

Alternative ASR hypotheses have been used in many different ways in speech driven systems. Particularly, in (Mangu et al., 2000) multiple lattice alignment is used for construction of word confusion networks and in (Hakkani-Tür et al., 2006) word confusion networks are used for named entity detection. In the study presented here, we apply word confusion networks (to represent ASR hypotheses) along with eye gaze to the problem of reference resolution.

3 Data Collection

In this investigation, we created a 3D virtual world (using the Irrlicht game engine[1]) to support situated dialogue. We conducted a Wizard of Oz study in which the user must collaborate with a remote artificial agent cohort (controlled by a human) to solve a treasure hunting task. The cohort is an "expert" in treasure hunting and has some knowledge regard-

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

ing the locations of the treasure items, but cannot see the virtual environment. The user, immersed in the virtual world, must navigate the environment and conduct a mixed-initiative dialogue with the agent to find the hidden treasures. During the experiments, a noise-canceling microphone was used to record user speech and the Tobii 1750 display-mounted eye tracker was used to record user eye movements.

A snapshot of user interaction with the treasure hunting environment is shown in Figure 1. Here, the user's eye fixation is represented by the white dot and saccades (eye movements) are represented by white lines. The virtual world contains 10 rooms with a total of 155 unique objects that encompass 74 different object types (e.g. chair or plant).

Figure 1: Snapshot of the situated treasure hunting environment

Table 1 shows a portion of a sample dialogue between a user and the expert. Each S_i represents a system utterance and each U_i represents a user utterance. We focus on resolving exophoric referring expressions, which are enclosed in brackets here. In our dataset, an exophoric referring expression is a non-pronominal noun phrase that refers to an entity in the extralinguistic environment. It may be an evoking reference that initially refers to a new object in the virtual world (e.g. `an axe` in utterance U_2) or a subsequent reference to an entity in the virtual world which has previously been mentioned in the dialogue (e.g. `an axe` in utterance U_3). In our study we focus on resolving exophoric referring expressions because they are tightly coupled with a user's eye gaze behavior.

From this study, we constructed a parallel spoken utterance and eye gaze corpus. Utterances, which

S_1	Describe what you're doing.
U_1	I just came out from the room that I started and i see **[one long sword]**
U_2	**[one short sword]** and **[an axe]**
S_2	Compare these objects.
U_3	one of them is long and one of them is really short, and i see **[an axe]**

Table 1: A conversational fragment demonstrating interaction with exophoric referring expressions.

	Utterance: i just came out from the room that i started and **i see [one long sword]**
H_t :	...i_5210 see_5410 [one_5630 long_6080 sword_6460]
H_1:	...icy_5210 winds_5630 along_6080 so_6460 words_68000
H_2:	...icy_5210 [wine_5630] along_6080 so_6460 words_6800
...	
H_{25}:	...icy_5210 winds_5630 [long_6080 sword_6460]
...	

Table 2: Sample n-best list of recognition hypotheses

are separated by a long pause (500 ms) in speech, are automatically recognized using the Microsoft Speech SDK. Gaze fixations are characterized by objects in the virtual world that are fixated via a user's eye gaze. When a fixation points to multiple spatially overlapping objects, only the one in the forefront is deemed to be fixated. The data corpus was transcribed and annotated with 2204 exophoric referring expressions amongst 2052 utterances from 15 users.

4 Word Confusion Networks

For each user utterance in our dataset, an n-best list (with $n = 100$) of recognition hypotheses ranked in order of likelihood is produced by the Microsoft Speech SDK. One way to use the speech recognition results (as in most speech applications) is to use the top ranked recognition hypothesis. This may not be the best solution because a large amount of information is being ignored. Table 2 demonstrates this problem. Here, the number after the underscore denotes a timestamp associated with each recognized spoken word. The strings enclosed in brackets de-

note recognized referring expressions. In this example, the manual transcription of the original utterance is shown by H_t. In this case, the system must first identify `one long sword` as a referring expression and then resolve it to the correct set of entities in the virtual world. However, not until the twenty fifth ranked recognition hypothesis H_{25}, do we see a referring expression closest to the actual uttered referring expression. Moreover, in utterances with multiple referring expressions, there may not be a single recognition hypothesis that contains all referring expressions, but each referring expression may be contained in some recognition hypothesis. Thus, it is desirable to consider the entire n-best list of hypotheses.

To address this issue, we adopted the word confusion network (WCN): a compact representation of a word lattice or n-best list (Mangu et al., 2000). A WCN captures alternative word hypotheses and their corresponding posterior probabilities in time-ordered sets. In addition to being compact, an important feature for efficient post-processing of recognition hypotheses for real-time systems, WCNs are capable of representing more competing hypotheses than either n-best lists or word lattices. Figure 2 shows an example of a WCN for the utterance "...I see one long sword" along with a timeline (in milliseconds) depicting the eye gaze fixations to potential referent objects that correspond to the utterance. The confusion network shows competing word hypotheses along with corresponding probabilities in log scale.

Using our data set, we can show that word confusion networks contain significantly more words that can compose a referring expression than the top recognition hypothesis. The confusion network keyword error rate (KWER) is 0.192 compared to a 1-best list KWER of 0.318, where a keyword is a word that can be contained in a referring expression. The overall WER for word confusion networks and 1-best lists are 0.315 and 0.460, respectively. The reported WCN word error rates are all oracle word error rates reflecting the best WER that can be attained using any path in the confusion network. One more important feature of word confusion networks is that they provide time alignment for words that occur at approximately the same time interval in competing hypotheses. This is not only useful for efficient syntactic parsing, which is necessary for identifying referring expressions, but also critical for integration with time aligned gaze streams.

5 Reference Resolution Algorithm

We have developed an algorithm that combines an n-best list of speech recognition hypotheses with dialogue, domain, and eye-gaze information to resolve exophoric referring expressions. There are three inputs to the multimodal reference resolution algorithm for each utterance: (1) an n-best list of alternative speech recognition hypotheses ($n = 100$ for a WCN and $n = 1$ for the top recognized hypothesis), (2) a list of fixated objects (by eye gaze) that temporally correspond to the spoken utterance and (3) a set of potential referent objects. Since during the treasure hunting task people typically only speak about objects that are visible or have recently been visible on the screen, an object is considered to be a potential referent if it is present within a close proximity (in the same room) of the user while an utterance is spoken.

The multimodal reference resolution algorithm proceeds with the following four steps:

Step 1: construct word confusion network A word confusion network is constructed out of the input n-best list of alternative recognition hypotheses with the SRI Language Modeling (SRILM) toolkit (Stolcke, 2002) using the procedure described in (Mangu et al., 2000). This procedure aligns words from the n-best list into equivalence classes. First, instances of the same word containing approximately the same starting and ending timestamps are clustered. Then, equivalence classes with common time ranges are merged. For each competing word hypothesis its probability is computed by summing the posteriors of all utterance hypotheses containing this word. In our work, instead of using the actual posterior probability of each utterance hypothesis (which was not available), we assigned each utterance hypothesis a probability based on its position in the ranked list. Figure 2 depicts a portion of the resulting word confusion network (showing competing word hypotheses and their probabilities in log scale) constructed from the n-best list in Table 2.

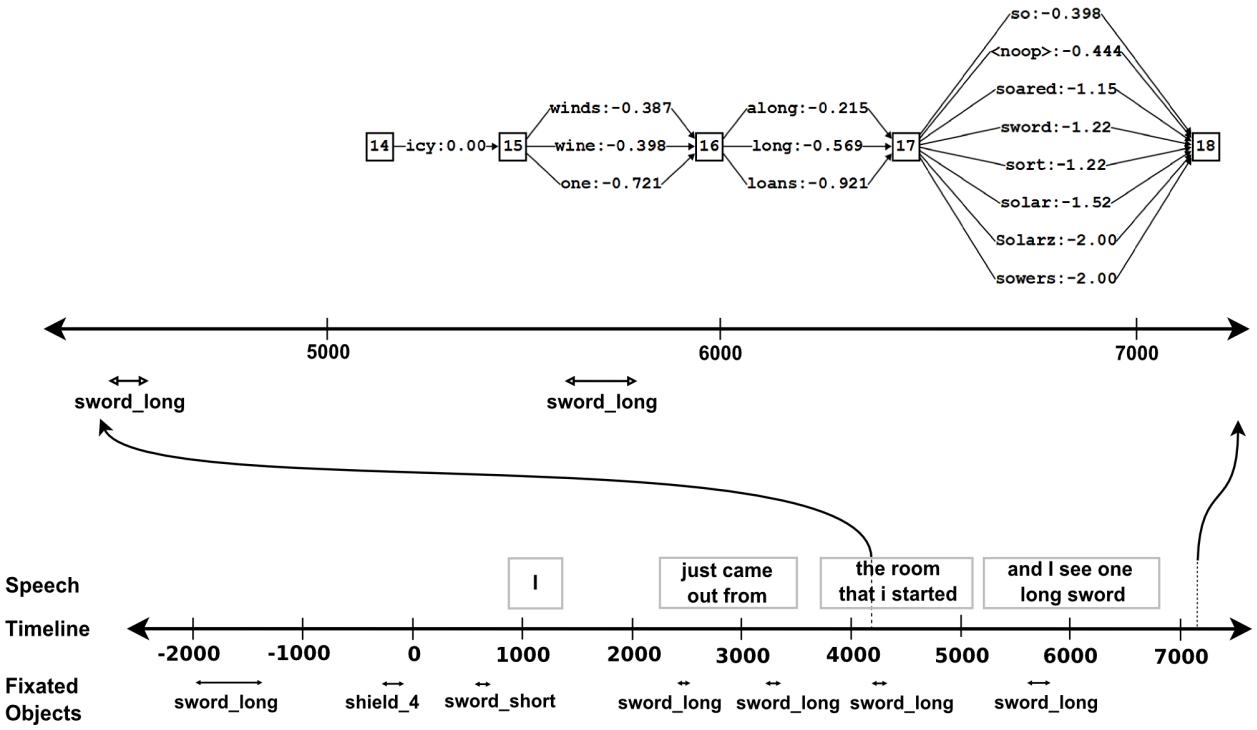

Figure 2: Sample parallel speech and eye gaze data streams, including a portion of the sample WCN

Step 2: extract referring expressions from WCN The word confusion network is syntactically parsed using a modified version of the CYK (Cooke and Schwartz, 1970; Kasami, 1965; Younger, 1967) parsing algorithm that is capable of taking a word confusion network as input rather than a single string. We call this the CYK-WCN algorithm. To do the parsing, we applied a set of grammar rules largely derived from a different domain in our previous work (Prasov and Chai, 2008). A parse chart of the sample word confusion network is shown in Table 3. Here, just as in the CYK algorithm the chart is filled in from left to right then bottom to top. The difference is that the chart has an added dimension for competing word hypotheses. This is demonstrated in position 15 of the WCN, where `one` and `wine` are two nouns that constitute competing words. Note that some words from the confusion network are not in the chart (e.g. `winds`) because they are out of vocabulary. The result of the syntactic parsing is that the parts of speech of all sub-phrases in the confusion network are identified. Next, a set of all exophoric referring expressions (i.e. non-pronominal noun phrases) found in the word confusion network are extracted. Each referring expression has a corresponding confidence score, which can be computed in many many different ways. Currently, we simply take the mean of the probability scores of the expression's constituent words. The sample WCN has four such phrases (shown in bold in Table 3): `wine` at position 15 with length 1, `one long sword` at position 15 with length 3, `long sword` at position 16 with length 2, and `sword` at position 17 with length 1.

Step 3: resolve referring expressions Each referring expression r_j is resolved to the top k potential referent objects according to the probability $P(o_i|r_j)$, where k is determined by information from the linguistic expressions. $P(o_i|r_j)$ is determined using the following expression:

$$P(o_i|r_j) = \frac{AS(o_i)^{\alpha} \times Compat(o_i, r_j)^{1-\alpha}}{\sum_i AS(o_i)^{\alpha} \times Compat(o_i, r_j)^{1-\alpha}} \quad (1)$$

In this equation,

- AS: Attentional salience score of a particu-

length	...						
	5						
	4						
	3		**NP → NUM Adj-NP**				
	2			Adj-NP → ADJ N, **NP → Adj-NP**			
	1		(1) N → wine, **NP → N** (2) NUM → one	ADJ → long	N → sword, **NP → N**		
	...	14	15	16	17	18	...
	WCN position						

Table 3: Syntactic parsing of word confusion network

lar object o_i, which is determined based on the gaze *fixation intensity* of an object at the start time of referring expression r_j. The fixation intensity of an object is defined as the amount of time that the object is fixated during a predefined time window W (Prasov and Chai, 2008). As in (Prasov and Chai, 2008), we set $W = [-1500..0]$ ms relative to the beginning of referring expression r_j.

- $Compat$: Compatibility score, which specifies whether the object o_i is compatible with the information specified by the referring expression r_j. Currently, the compatibility score is set to 1 if referring expression r_j and object o_i have the same object type (e.g. chair), and 0 otherwise.

- α: Importance weight, in the range $[0..1]$, of attentional salience relative to compatibility. A high α value indicates that the attentional salience score based on eye gaze carries more weight in deciding referents, while a low α value indicates that compatibility carries more weight. In this work, we set $\alpha = 0.5$ to indicate equal weighting between attentional salience and compatibility. If we do not want to integrate eye gaze in reference resolution, we can set $\alpha = 0.0$. In this case, reference resolution will be purely based on compatibility between visual objects and information specified via linguistic expressions.

Once all probabilities are calculated, each referring expression is resolved to a set of referent objects. Finally, this results in a set of *(referring expression, referent object set)* pairs with confidence scores, which are determined by two components. The first component is the confidence score of the referring expression, which is explained in the Step 1 of the algorithm. The second component is the probability that the referent object set is indeed the referent of this expression (which is determined by Equation 1). There are various ways to combine these two components together to form an overall confidence score for the pair. Here we simply multiply the two components. The confidence score for the pair is used in the following step to prune unlikely referring expressions.

Step 4: post-prune The resulting set of *(referring expression, referent object set)* pairs is pruned to remove pairs that fall under one of the following two conditions: (1) the pair has a confidence score equal to or below a predefined threshold ϵ (currently, the threshold is set to 0 and thus keeps all resolved pairs) and (2) the pair temporally overlaps with a higher confidence pair. For example, in Table 3, the referring expressions `one long sword` and `wine` overlap in position 15. Finally, the resulting *(referring expression, referent object set)* pairs are sorted in ascending order according to their constituent referring expression timestamps.

6 Experimental Results

Using our data, described in Section 3, we applied the multimodal reference resolution algorithm described in Section 5. All of the data is used to report the experimental results. Reference resolution model parameters are set based on our prior work in a different domain (Prasov and Chai, 2008). For each utterance we compare the reference resolution performance with and without the integration of eye gaze information. We also evaluate using a

word confusion network compared to a 1-best list to model speech recognition hypotheses. For perspective, reference resolution with recognized speech input is compared with transcribed speech.

6.1 Evaluation Metrics

The reference resolution algorithm outputs a list of *(referring expression, referent object set)* pairs for each utterance. We evaluate the algorithm by comparing the generated pairs to the annotated "gold standard" pairs using F-measure. We perform the following two types of evaluation:

- Lenient Evaluation: Due to speech recognition errors, there are many cases in which the algorithm may not return a referring expression that exactly matches the gold standard referring expression. It may only match based on the object type. For example, the expressions `one long sword` and `sword` are different, but they match in terms of the intended object type. For applications in which it is critical to identify the objects referred to by the user, precisely identifying uttered referring expressions may be unnecessary. Thus, we evaluate the reference resolution algorithm with a lenient comparison of *(referring expression, referent object set)* pairs. In this case, two pairs are considered a match if at least the object types specified via the referring expressions match each other and the referent object sets are identical.

- Strict Evaluation: For some applications it may be important to identify exact referring expressions in addition to the objects they refer to. This is important for applications that attempt to learn a relationship between referring expressions and referenced objects. For example, in automated vocabulary acquisition, words other than object types must be identified so the system can learn to associate these words with referenced objects. Similarly, in systems that apply priming for language generation, identification of the exact referring expressions from human users could be important. Thus, we also evaluate the reference resolution algorithm with a strict comparison of *(referring expression, referent object set)* pairs. In this case, a referring expression from the system output needs to exactly match the corresponding expression from the gold standard.

6.2 Role of Eye Gaze

We evaluate the effect of incorporating eye gaze information into the reference resolution algorithm using the top best recognition hypothesis (*1-best*), the word confusion network (*WCN*), and the manual speech transcription (*Transcription*). Speech transcription, which contains no recognition errors, demonstrates the upper bound performance of our approach. When no gaze information is used, reference resolution solely depends on linguistic and semantic processing of referring expressions. Table 4 shows the lenient reference resolution evaluation using F-measure. This table demonstrates that lenient reference resolution is improved by incorporating eye gaze information. This effect is statistically significant in the case of transcription and 1-best ($p < 0.0001$ and $p < 0.009$, respectively) and marginal ($p < 0.07$) in the case of WCN.

Configuration	Without Gaze	With Gaze
Transcription	0.619	0.676
WCN	0.524	0.552
1-best	0.471	0.514

Table 4: Lenient F-measure Evaluation

Configuration	Without Gaze	With Gaze
Transcription	0.584	0.627
WCN	0.309	0.333
1-best	0.039	0.035

Table 5: Strict F-measure Evaluation

Table 5 shows the strict reference resolution evaluation using F-measure. As can be seen in the table, incorporating eye gaze information significantly ($p < 0.0024$) improves reference resolution performance when using transcription and marginally ($p < 0.113$) in the case of WCN optimized for strict evaluation. However there is no difference for the 1-best hypotheses which result in extremely low performance. This observation is not surprising since 1-best hypotheses are quite error prone and less likely to produce the exact expressions.

Since eye gaze can be used to direct navigation in a mobile environment as in situated dialogue, there could be situations where eye gaze does not reflect the content of the corresponding speech. In such situations, integrating eye gaze in reference resolution could be detrimental. To further understand the role of eye gaze in reference resolution, we applied our reference resolution algorithm only to utterances where speech and eye gaze are considered closely coupled (i.e., eye gaze reflects the content of speech). More specifically, following the previous work (Qu and Chai, 2010), we define a *closely coupled* utterance as one in which at least one noun or adjective describes an object that has been fixated by the corresponding gaze stream.

Table 6 and Table 7 show the performance based on closely coupled utterances using lenient and strict evaluation, respectively. In the lenient evaluation, reference resolution performance is significantly improved for all input configurations when eye gaze information is incorporated ($p < 0.0001$ for transcription, $p < 0.015$ for WCN, and $p < 0.0022$ for 1-best). In each case the closely coupled utterances achieve higher performance than the entire set of utterances evaluated in Table 5. Aside from the 1-best case, the same is true when using strict evaluation ($p < 0.0006$ for transcription and $p < 0.046$ for WCN optimized for strict evaluation). This observation indicates that in situated dialogue, some mechanism to predict whether a gaze stream is closely coupled with the corresponding speech content can be beneficial in further improving reference resolution performance.

Configuration	Without Gaze	With Gaze
Transcription	0.616	0.700
WCN	0.523	0.570
1-best	0.473	0.537

Table 6: Lenient F-measure Evaluation for Closely Coupled Utterances

6.3 Role of Word Confusion Network

The effect of incorporating eye gaze with WCNs rather than 1-best recognition hypotheses into reference resolution can also be seen in Tables 4 and 5. Table 4 shows a significant improvement when using WCNs rather than 1-best hypotheses for both with ($p < 0.015$) and without ($p < 0.0012$) eye gaze configurations. Similarly, Table 5 shows a significant improvement in strict evaluation when using WCNs rather than 1-best hypotheses for both with ($p < 0.0001$) and without ($p < 0.0001$) eye gaze configurations. These results indicate that using word confusion networks improves both lenient and strict reference resolution. This observation is not surprising since identifying correct linguistic expressions will enable better search for semantically matching referent objects.

Configuration	Without Gaze	With Gaze
Transcription	0.579	0.644
WCN	0.307	0.345
1-best	0.045	0.038

Table 7: Strict F-measure Evaluation for Closely Coupled Utterances

Although WCNs lead to better performance, utilizing WCNs is more computationally expensive compared to 1-best recognition hypotheses. Nevertheless, in practice, WCN depth, which specifies the maximum number of competing word hypotheses in any position of the word confusion network, can be limited to a certain value $|d|$. For example, in Figure 2 the depth of the shown WCN is 8 (there are 8 competing word hypotheses in position 17 of the WCN). The WCN depth can be limited by pruning word alternatives with low probabilities until, at most, the top $|d|$ words remain in each position of the WCN. It is interesting to observe how limiting WCN depth can affect reference resolution performance. Figure 3 demonstrates this observation. In this figure the resolution performance (in terms of lenient evaluation) for WCNs of varying depth is shown as dashed lines for with and without eye gaze configurations. As a reference point, the performance when utilizing 1-best recognition hypotheses is shown as solid lines. It can be seen that as the depth increases, the performance also increases until the depth reaches 8. After that, there is no performance improvement.

7 Discussion

In Section 6.2 we have shown that incorporating eye gaze information improves reference resolution performance. Eye-gaze information is particu-

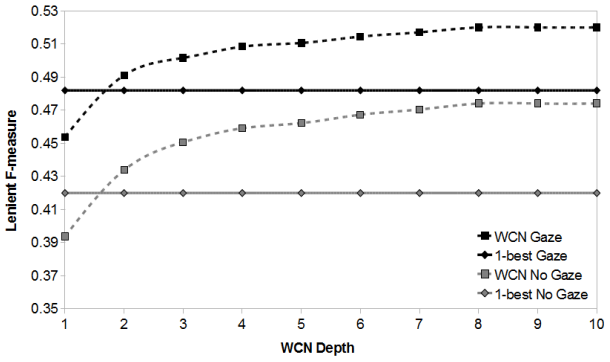

Figure 3: Lenient F-measure at each WCN Depth

larly helpful for resolving referring expressions that are ambiguous from the perspective of the artificial agent. Consider a scenario where the user utters a referring expression that has an equivalent semantic compatibility with multiple potential referent objects. For example, in a room with multiple books, the user utters "the open book to the right", but only the phrase "the book" is recognized by the ASR. If a particular book is fixated during interaction, there is a high probability that it is indeed being referred to by the user. Without eye gaze information, the semantic compatibility alone could be insufficient to resolve this referring expression. Thus, when eye gaze information is incorporated, the main source of performance improvement comes from better identification of potential referent objects.

In Section 6.3 we have shown that incorporating multiple speech recognition hypotheses in the form of a word confusion network further improves reference resolution performance. This is especially true when exact referring expression identification is required (F-measure of 0.309 from WCNs compared to F-measure of 0.039 from 1-best hypotheses). Using a WCN improves identification of low-probability referring expressions. Consider a scenario where the top recognition hypothesis of an utterance contains no referring expressions or an incorrect referring expression that has no semantically compatible potential referent objects. If a referring expression with a high compatibility value to some potential referent object is present in a lower probability hypothesis, this referring expression can only be identified when a WCN rather than a 1-best hypothesis is utilized. Thus, when word confusion networks are incorporated, the main source of performance improvement comes from better referring expression identification.

7.1 Computational Complexity

One potential concern of using word confusion networks rather than 1-best hypotheses is that they are more computationally expensive to process. The asymptotic computational complexity for resolving the referring expressions using the algorithm presented in this work with a WCN is the summation of three components: (1) $O(|G| \cdot |d|^2 \cdot |w|^3)$ for confusion network construction and parsing, (2) $O(|r| \cdot |O| \cdot log(|O|))$ for reference resolution, and (3) $O(|r|^2)$ for selection of *(referring expression, referent object set)* pairs. Here, $|w|$ is the number of words in the input speech signal (or, more precisely, the number of words in the longest ASR hypothesis for a given utterance); $|G|$ is the size of the parsing grammar; $|d|$ is the depth of the constructed word confusion network; $|O|$ is the number of potential referent objects for each utterance; and $|r|$ is the number of referring expressions that are extracted from the word confusion network.

The complexity is dominated by the word confusion network construction and parsing. Also, both the number of words in an input utterance ASR hypothesis $|w|$ and the number of referring expressions in a word confusion network $|r|$ are dependent on utterance length. In our study, interactive dialogue is encouraged and, thus, utterances are typically short; with a mean length of 6.41 words and standard deviation of 4.35 words. The longest utterances in our data set has 31 words. WCN depth $|d|$ has a mean of 10.1, a standard deviation of 8.1, and a maximum 89 words. In practice, as shown in Section 6.3, limiting $|d|$ to 8 words achieves comparable reference resolution results as using a full word confusion network.

To demonstrate the applicability of our reference resolution algorithm for real-time processing, we applied it on the data corpus presented in Section 3. This corpus contains utterances with a mean input time of 2927.5 ms and standard deviation of 1903.8 ms. On a 2.4 GHz AMD Athlon(tm) 64 X2 Dual Core Processor, the runtimes resulted in a real time factor of 0.0153 on average. Thus, on average, an utterance from this corpus can be processed in just under 45 ms, which is well within the range of ac-

ceptable real-time performance.

7.2 Error Analysis

As can be seen in Section 6, even when using transcribed data, reference resolution performance still has room for improvement (achieving the highest lenient F-measure of 0.700 when eye gaze is utilized for resolving closely coupled utterances). In this section, we elaborate on the potential error sources. Specifically, we discuss two types of error: (1) a referring expression is incorrectly recognized or (2) a recognized referring expression is not resolved to a correct referent object set.

Given transcribed data, which simulates perfectly recognized utterances, all referring expression recognition errors arise due to incorrect language processing. Most of these errors occur because an incorrect part of speech (POS) tag is assigned to a word, or an out-of-vocabulary (OOV) word is encountered, or the parsing grammar has insufficient coverage. A particularly interesting parsing problem occurs due to the nature of spoken language. Since punctuation is sometimes unavailable, given an utterance with several consecutive nouns, it is unclear which of these nouns should be treated as head nouns and which should be treated as noun modifiers. For example, in the utterance "there is *a desk lamp* table and two chairs" it is unclear if the italicized expression should be parsed as a single phrase or as a list of (two) phrases `a desk` and `lamp`. Thus, some timing information should be used for disambiguation.

Object set identification errors are more prevalent than referring expression recognition errors. The majority of these errors occur because a referring expression is ambiguous from the perspective of the conversational system and there is not enough information to choose amongst multiple potential referent objects due to limited speech recognition and domain modeling. One reason for this is that a referring expression may be resolved to an incorrect number of referent objects. Another reason is that a pertinent object attribute or a distinguishing spatial relationship between objects specified by the user cannot be established by the system. For example, during the utterance "I see *a vase* left of the table" there are two vases visible on the screen creating an ambiguity if the phrase `left of` is not processed correctly. This is caused by an inadequate representation of spatial relationships and processing of spatial language. One more reason for potential ambiguity is the lack of pragmatic knowledge that can support adequate inference. For example, when the user refers to two *sofa* objects using the phrase "*an armchair* and *a sofa*", the system lacks pragmatic knowledge to indicate that `arm chair` refers to the smaller of the two objects. Some of these errors can be avoided when eye gaze information is available to the system. However, due to the noisy nature of eye gaze data, many such referring expressions remain ambiguous even when eye gaze information is considered.

8 Conclusion

In this work, we have examined the utility of eye gaze and word confusion networks for reference resolution in situated dialogue within a virtual world. Our empirical results indicate that incorporating eye gaze information with recognition hypotheses is beneficial for the reference resolution task compared to only using recognition hypotheses. Furthermore, using a word confusion network rather than the top best recognition hypothesis further improves reference resolution performance. Our findings also demonstrate that the processing speed necessary to integrate word confusion networks with eye gaze information is well within the acceptable range for real-time applications.

Acknowledgments

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

References

N. Bee, E. André, and S. Tober. 2009. Breaking the ice in human-agent communication: Eye-gaze based initiation of contact with an embodied conversational agent. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents (IVA'09)*, pages 229–242. Springer.

T. Bickmore and J. Cassell, 2004. *Social Dialogue with Embodied Conversational Agents*, chapter Natural, In-

telligent and Effective Interaction with Multimodal Dialogue Systems. Kluwer Academic.

D. K. Byron, T. Mampilly, and T. Sharma, V.and Xu. 2005. Utilizing visual attention for cross-modal coreference interpretation. In *Spring Lecture Notes in Computer Science: Proceedings of CONTEXT-05*, pages 83–96.

N. J. Cooke and M. Russell. 2008. Gaze-contingent automatic speech recognition. *IET Signal Processing*, 2(4):369–380, December.

J. Cooke and J. T. Schwartz. 1970. Programming languages and their compilers: Preliminary notes. Technical report, Courant Institute of Mathematical Science.

R. Fang, J. Y. Chai, and F. Ferreira. 2009. Between linguistic attention and gaze fixations in multimodal conversational interfaces. In *The 11th International Conference on Multimodal Interfaces (ICMI)*.

P. Gorniak, J. Orkin, and D. Roy. 2006. Speech, space and purpose: Situated language understanding in computer games. In *Twenty-eighth Annual Meeting of the Cognitive Science Society Workshop on Computer Games*.

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

D. Hakkani-Tür, F. Béchet, G. Riccardi, and G. Tur. 2006. Beyond asr 1-best: Using word confusion networks in spoken language understanding. *Computer Speech and Language*, 20(4):495–514.

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

T. Kasami. 1965. An efficient recognition and syntax-analysis algorithm for context-free languages. Scientific report AFCRL-65-758, Air Force Cambridge Research Laboratory, Bedford, Massachusetts.

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

Y. Liu, J. Y. 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)*.

L. Mangu, E. Brill, and A. Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400.

A. S. Meyer and W. J. M. Levelt. 1998. Viewing and naming objects: Eye movements during noun phrase production. In *Cognition*, volume 66, pages B25–B33.

L.-P. Morency, C. M. Christoudias, and T. Darrell. 2006. Recognizing gaze aversion gestures in embodied conversational discourse. In *International Conference on Multimodal Interfaces (ICMI)*.

Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*, pages 553–561.

Z. Prasov and J. Y. Chai. 2008. What's in a gaze? the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of 13th International Conference on Intelligent User interfaces (IUI)*, pages 20–29.

S. Qu and J. 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)*.

S. Qu and J. Y. Chai. 2010. Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research*, 37:347–377, March.

P. Qvarfordt and S. Zhai. 2005. Conversing with the user based on eye-gaze patterns. In *Proceedings Of the Conference on Human Factors in Computing Systems*. ACM.

C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh. 2004. Where to look: A study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent User Interfaces (IUI'04)*, pages 78–84. ACM Press.

A. Stolcke. 2002. SRILM an extensible language modeling toolkit, confusion network. In *International Conference on Spoken Language Processing*.

M. K. Tanenhous, M. Spivey-Knowlton, E. Eberhard, and J. Sedivy. 1995. Integration of visual and linguistic information during spoken language comprehension. In *Science*, volume 268, pages 1632–1634.

D. H. Younger. 1967. Recognition and parsing of context-free languages in time n^3. *Information and Control*, 10(2):189–208.