EYE GAZE FOR REFERENCE RESOLUTION IN MULTIMODAL CONVERSATIONAL INTERFACES

By

Zahar Prasov

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Computer Science

2011

ABSTRACT

EYE GAZE FOR REFERENCE RESOLUTION IN MULTIMODAL CONVERSATIONAL INTERFACES

By

Zahar Prasov

Multimodal conversational interfaces allow users to carry a spoken dialogue with an artificial conversational agent while looking at a graphical display. The dialogue is used to accomplish purposeful tasks. Motivated by previous psycholinguistic findings, this dissertation investigates how eye gaze contributes to automated spoken language understanding in such a setting, specifically focusing on robust reference resolution—a process that identifies the referring expressions in an utterance and determines which entities these expressions refer to. As a part of this investigation we attempt to model user focus of attention during human-machine conversation by utilizing the users' naturally occurring eye gaze. We study which eye gaze and auxiliary visual factors contribute to this model's accuracy. Among the various features extracted from eye gaze, fixation intensity has shown to be the most indicative in reflecting attention. We combine user speech along with this gaze-based attentional model into an integrated reference resolution framework. This framework fuses linguistic, dialogue, domain, and eye gaze information to robustly resolve various kinds of referring expressions that occur during human-machine conversation. Our studies have shown that based on this framework, eye gaze can compensate for limited domain models and dialogue processing capability. We further extend this framework to handle recognized speech input acquired situated dialogue within an immersive virtual environment. We utilize word confusion networks to model the set of alternative speech recognition hypotheses and incorporate confusion networks into the reference resolution framework. The empirical results indicate that incorporating eye gaze significantly improves reference resolution performance, especially when limited domain model information is available to the reference resolution framework. The empirical results also indicate that modeling recognized speech via confusion networks rather than the single best recognition hypothesis leads to better reference resolution performance.

DEDICATION

I would like to dedicate this work to my family for motivating me and believing in me throughout the lengthy and arduous process of creating this dissertation. I am indebted to my parents, brother, grandparents, extended family, and most of all, my dearest Alexis for their unwavering love and support.

ACKNOWLEDGMENT

My sincere gratitude goes toward my advisor and dissertation committee chair, Joyce Y. Chai, for demonstrating to me how to succeed as a Ph.D. researcher. Her leadership and assistance has been instrumental throughout my tenure as a graduate student.

I would like to thank my other committee members—Rong Jin, Anil Jain, and Frank Biocca—for their valuable advice.

I would like to thank all of my lab mates in the Language And Interaction Research (LAIR) group for their constructive sharing of ideas and pizza. A special thanks goes to Shaolin Qu for being my partner in designing test-bed multimodal conversational interfaces and user studies for data collection. I appreciate every participant that was involved in the data collection user studies.

This work was supported by IIS-0347548 and IIS-0535112 from the National Science Foundation and an IGERT fellowship. A special note of appreciation goes to the IGERT program and the Cognitive Science Department at Michigan State University for inspiring me to conduct interdisciplinary research and for allowing me to use their eye tracking facilities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Introduction and Motivation

1.1.1 Conversational Interfaces

Conversational interfaces allow users to carry a dialogue with a graphical display using speech to perform a particular task or a group of related tasks. This method of interaction deviates from more traditional direct manipulation and WIMP (windows, icons, menus, and pointing devices) interfaces. Conversational interfaces are more flexible and natural to use. They allow for a better use of human communication skills, a larger diversity of users and tasks, and a faster task completion rate for visual-spatial tasks.

The back-bone of conversational interfaces is the component that supports human-machine dialogue. A dialogue system is used to interpret user input, decide which actions should be taken as a response to this input, and determine the best way to present this response to the user. Although, dialogue system architectures vary greatly in terms of the components they contain, they all revolve around completing these three tasks.

Currently, a major hurdle of conversational systems is an insufficient language understanding capability. Language understanding errors can cause miscommunications between a user and a conversational system. When this happens, the system is likely to behave in an unexpected manner from the user's perspective. This can lead to user confusion and dissatisfaction with system performance. More importantly it can be very detrimental to task success rate and task completion rate. A high communication error rate makes a conversational system unusable.

1.1.2 Language Understanding

In a conversational system, there exist three levels of language processing in order to obtain an overall interpretation of a user input utterance. First, in the automated speech recognition (ASR) stage, the speech signal is converted into a textual format. Next, in the natural language processing (NLP) stage, the recognized text is syntactically and semantically analyzed. A partial semantic representation of the spoken utterance is constructed. At this point, there may still be some ambiguity about the precise meaning of the utterance. Finally, in the discourse processing stage, this ambiguity is resolved by considering the context in which the utterance was produced. This context can consist of prior dialogue information, visual information, or general domain knowledge.

Communication errors can occur at any of the three stages of language processing. Some errors may lead to language understanding failure, while others may be resolved during a later processing stage. For example, given the utterance "the green chair" and an ASR misrecognition of the word *chair* visual context may be used to determine that there is only one green object appearing on the screen; thus enabling the system to compensate for the ASR error and correctly identify the object in question. Although conversational interfaces offer many advantages over more traditional interfaces, they suffer a major disadvantage in that spoken language input can be noisy and ambiguous.

The objective of our work is to improve language understanding capability in a multimodal conversational interface. In prior work, the general approach to improving language understanding is to incorporate knowledge into one of the three stages of language processing. We continue this trend, but from a different perspective. We study the effects of utilizing naturally occurring eye gaze during human-machine conversation as an additional knowledge source to the spoken language input. Our task becomes that of multimodal interpretation given two input modes: speech and eye gaze.

1.1.3 Multimodal Interpretation

Conversational interfaces that accept and make use of multiple modes of input (such as speech, lip movement, deictic gesture, eye gaze, etc.) are considered to be multimodal. Such systems must fuse multiple modes of input to form a cohesive interpretation of the user's actions. This process is known as *multimodal interpretation*. Multimodal interpreters often attempt to combine several noisy inputs. It has been shown that the noisy inputs can mutually disambiguate each other and achieve a more robust interpretation than either of the inputs alone [68].

While most previous work in multimodal interpretation has focused on combining speech and gesture, here we focus on combining speech and eye gaze information which is motivated by prior psycholinguistic studies that have shown a tight coupling between speech and eye gaze during language production [30, 41, 85, 28]. Eye gaze has been shown to be a window to the mind. That is, the eye fixates on symbols that are being cognitively processed. The direction of gaze carries information about the focus of a person's attention [41]. Knowing the focus of attention can be used to improve spoken language understanding in a conversational system.

Information obtained from eye gaze based focus of attention can be directly useful in one important aspect of multimodal interpretation: *reference resolution*. Reference resolution is the process of identifying the best matching referent (object in the physical or virtual world) given a spoken referring expression. For example, given the user utterance “I like **the one** next to **the green chair**” during interaction with a conversational interface, the phrases shown in bold are referring expressions that refer to two objects appearing on the interface.

Reference resolution can be improved by incorporating eye gaze information into any of the three levels of language processing. Our work is concentrated on fusing the multimodal inputs during the discourse processing stage while using relatively standard ASR and NLP techniques. The main rationale for this is that communication errors are possible even with perfect ASR and NLP. Consider the utterance in the above example. If there are multiple green chairs on the visible display, it is impossible to resolve the referring expressions in the utterance purely based on ASR and NLP results (even with perfect interpretation). Context is necessary to resolve these references. This context can come from many knowledge sources, including eye gaze. Eye gaze can be used to determine that a particular chair was recently in the focus of attention, thus disambiguating the referring expression in this example.

A major advantage of eye gaze is that it is a naturally occurring byproduct of speech. Eye gaze information is easily available with the help of eye tracking technology.

1.1.4 Eye Tracking

Eye tracking is the process of measuring eye movement, or changes in gaze direction, as a user views a real or virtual environment. The direction of eye gaze has been shown to carry information about the focus of a person’s attention [41]. Even though there may be some dissociation between covert visual attention and overt visual attention that is expressed in the form of eye movements, behavioral neurophysiological studies have shown a tight coupling between the two [30, 21].

Eye tracking is used as a mechanism to study attention in various research fields, including psychology, cognitive science, neuroscience, and computer science. Eye tracking is used in vision studies to help determine how humans perceive visual scenes. It is used in psycholinguistic studies to help determine how user comprehend and produce language [19, 85, 84, 82, 59, 28, 27]. In HCI, eye tracking is used to evaluate the usability of user interfaces [18, 35, 64]. In human-machine communication, it has been used as either a direct means of interface control [86, 33, 90] or as a supplemental input for improved interpretation [61, 78, 3, 80, 60, 54, 75, 17, 2, 20, 76].

1.2 Specific Research Questions

The general goal of this dissertation is to investigate how eye gaze can be used to improve the robustness of spoken language understanding in a multimodal conversational system. In doing this, we want to determine the relationship between speech and eye gaze in reference resolution. Our approach is to combine two pre-existing technologies and fuse the information that they provide. These technologies are automatic speech recognition (ASR) and eye tracking. It is important to stress that our goal is not to improve either ASR or eye tracking performance. Rather it is to use these two technologies to improve spoken language understanding. More specifically, this dissertation addresses the following three research questions:

1. Can eye gaze reliably predict user attention during conversation? Is this prediction influenced by other factors from the environment (e.g., visual features of the display, etc.)

2. Can eye gaze based attention prediction help to resolve linguistic referring expressions produced during conversation? Will the role of eye gaze change with different capabilities of automated language processing?

3. Can findings from conversation with simple static scenes be generalized to an immersive, situated scenario? What additional modeling is required for situated dialogue? What additional modeling is required to process recognized speech input?

To investigate the first question, we have designed an attentional salience model that identifies a user's focus of attention while he or she interacts with a conversational interface. More precisely, this model predicts which objects on the interface are likely to be referenced in speech within close temporal proximity of any given point in time during conversation. As a part of this investigation we study the temporal alignment between speech and gaze. We study the eye gaze based and auxiliary visual features that can assist attention prediction. We also discuss which evaluation metrics are appropriate for comparing different attention prediction models.

To investigate the second question, we have developed an integrated probabilistic reference resolution framework. This framework combines various input modes (particularly speech and eye gaze) to robustly resolve referring expressions belonging to various kinds of grammatical categories. Using this framework we study how improved domain modeling capability affects the role of eye gaze.

To investigate the third question category, we designed an immersive conversational interface that supports situated dialogue. Using this interface with a situated scenario, we collected speech and eye gaze data. We use this data to study the effect of ASR on reference resolution and how eye gaze can be used to alleviate speech recognition errors.

1.3 Outline

The remaining chapters of this dissertation are:

- Chapter 2 describes some related work that motivates this study.

- Chapter 3 provides a discussion of state-of-the-art eye tracking technology.

- Chapter 4 describes two domains and data sets used in this investigation.

- Chapter 5 explores the factors that can be used to predict a users focus of attention in a multimodal conversational interface.

- Chapter 6 examines the effects of incorporating eye gaze into a probabilistic reference resolution framework.

- Chapter 7 describes an enhanced framework that incorporates eye gaze with recognized speech hypotheses for reference resolution within an immersive domain.

- Chapter 8 summarizes the main contributions of this dissertation and provides a glimpse into future research directions.

Chapter 2

Related Work

In order to study the relationship between speech and eye gaze in multimodal conversational interfaces, we must first understand how these interfaces deviate from more traditional interfaces. We need to understand how these interfaces are constructed and what additional challenges they face. At the same time, we must understand how humans process multimodal information. This will give us some insight about the relationship between language and eye gaze. Once we know more about this relationship, we can exploit it when constructing computational systems. In this chapter, we first give an overview of previous work in multimodal conversational interfaces. We provide a more detailed focus on the methodology that has been used to interpret multimodal input. Then, we discuss user attention modeling, which can be used to disambiguate spoken language during multimodal conversation. Next, we discuss human eye gaze behavior during language production tasks and we explore how eye gaze has been previously used in human-machine communication.

2.1 Multimodal Conversational Interfaces

Multimodal conversational interfaces allow users to interact with computers naturally and effectively. Since their inception, many multimodal systems combining a variety of input and output modalities have been constructed. Multimodal conversational interfaces typically combine speech with one or more additional input modalities, including manual pointing gestures [4, 62], pen-based gestures [14, 39], lip movements and facial expressions [71], and eye gaze [49, 78]. Many multimodal systems enhance information presentation by combining multiple output modalities [9] such as speech, text, graphics, animation, and video. Multimodal conversational interfaces are most beneficial in visual and spatial domains. They are prevalent in a variety of applications including map-based information acquisition and manipulation systems, virtual reality systems for simulation and training, person identification systems for security purposes, web-based transaction systems, and mobile hand-held voice assistance.

While conversational interfaces have many advantages over traditional interfaces, they also face significant challenges. Development of multimodal conversational interfaces requires a significant effort in the areas of automatic speech recognition (ASR) for conversion of acoustic speech signals to machine-readable text strings, natural language processing (NLP) for interpreting the meanings and intentions of dialogue utterances, multimodal fusion for consistently combining information from various input modes, dialogue processing for maintaining a coherent collaborative dialogue, text-to-speech synthesis for responding to users with speech output, etc. Most of these technologies are error-prone and can lead to communication errors. A lot of effort has been put toward preventing and recovering from communication errors. One obvious and effective error prevention method is to improve *multimodal interpretation*. Multimodal interpretation is the process of combining ASR, NLP, and multimodal fusion techniques to generate a semantic representation from the given user inputs. We believe that the most beneficial way of utilizing eye gaze information in a multimodal conversational system is to improve multimodal interpretation.

2.2 Multimodal Interpretation

The goal of multimodal interpretation is to determine the semantic meaning of user inputs. Different input modes carry different cues about the meaning of a user's behavior. Gesture is considered an *active* input mode because a user must make a conscious decision to produce this type of input. Conversely, lip movements, facial expressions, and eye gaze are considered *passive* input modes because they are naturally occurring byproducts of interaction. Active input modes tend to be more reliable indicators of user intent [67]. However, passive input modes are unobtrusive and easily available. They are typically used to assist the prediction and interpretation of active input modes [67]—most commonly, speech.

One important aspect of multimodal interpretation is multimodal *reference resolution*: the process of identifying the best matching referent (object in the physical or virtual world) for each spoken referring expression in an utterance. This requires the use of automatic speech recognition (ASR), which may often be erroneous. Even with perfect ASR, language ambiguities make it challenging to determine which referent objects are referred to with a spoken expression. In multimodal conversational interfaces, supporting input modes to speech are often used to disambiguate these expressions. Another challenge, especially during spontaneous conversation, is that words contained in users' referring expressions are out of the vocabulary of this conversational system's domain. When such referring expressions are repeated many times throughout the conversation, it is possible for the system to learn new vocabulary using grounded language acquisition techniques.

2.2.1 Multimodal Reference Resolution

Much of the previous work on multimodal reference resolution has focused on combining speech and deictic gesture inputs. Chai et. al. used a greedy approach based on cognitive and linguistic theories to interpret speech, gesture, and visual information with a map-based user interface [10, 13]. Johnston et. al. used unification [37] and finite state [38] approaches to parse multimodal input in a map-based domain. Like deictic gestures, eye gaze provides information about a user's focus of attention. The difference is that eye fixations are produced subconsciously, there are far more eye fixations than gestures during interaction, and eye gaze data is much more noisy than deictic gesture data. This makes eye gaze more challenging to process than deictic gestures. Additionally, different cognitive factors govern the relationship between speech and gesture vs. speech and eye gaze. Thus, speech & gesture reference resolution should not be reused directly and new algorithms need to be constructed for fusing speech and eye gaze.

2.3 Modeling User Attention

Most previous work on modeling user attention in a multimodal conversational environment uses image processing techniques to predict which objects are likely to be talked about. These studies work on immersive visually rich domains. This enables objects appearing on the interface to be differentiated based on their visual properties. Eye tracking is mentioned as a potential alternative or even complement to their techniques, but no data exists supporting this possibility. Some studies do indeed use eye tracking to model user attention in a conversation system. However, these studies examine the use of eye gaze as an active mode of input that controls the navigation of the interface. Additionally, these studies use relatively simple visual displays. Instead, we utilize eye gaze as a passive input mode that supplements speech (the dominant active input mode) in a visually rich domain.

2.3.1 User Attention Modeling via Image Processing

There has been a long-standing goal in AI, HCI, and NLP research to integrate natural language and visual processing [57]. Much of this effort has focused on natural language understanding in interfaces that provide a visualized virtual space. Such interfaces allow users to manipulate objects and refer to these objects using *situated language*. Byron [6, 7] defines situated language as language that contains the following two properties: *immersion* and *mobility*. Situated language is spoken from a particular point of view within a visual environment. One of the earliest and most famous systems to integrate natural language and visual processing is Winograd's SHRDLU [87]. It is capable of having a dialogue with a user concerning the activity of a simulated robot arm in a simple blocks world. Other systems include DenK [47] and VIENA [40]. Recently, more systems have begun focusing on modeling visual context to assist natural language understanding [25, 45, 7]. Kelleher [45] and Byron [7] explicitly model visual context as user visual attention via image processing techniques.

Using image processing techniques to model visual salience has been a popular methodology. These techniques model one of the two types of attentional salience that is exhibited by humans: bottom-up image-based saliency cues [48]. These cues are independent of the nature of a particular task. They can be thought of as subconscious and pre-attentional. Factors such as visual familiarity, intentionality, object category, an object's physical characteristics relative to other objects in the scene, and the geometrical layout of the scene affect the visual salience of an object [50]. Typically, objects that deviate most from their visual context tend to be most salient [48]. These objects tend to "pop-out", making them the focus of attention in later stages of visual processing [63]. For example, a red object amongst many green objects will be visually salient. The second form of attention is a deliberate top-down form of attention that requires an effort from the user. This form of visual attention is governed by the user's goals during a particular task. For example, attempting to find a red horizontal target. The amount of time required to deploy this conscious form of visual attention is on the same order of magnitude as the time necessary to make an eye movement [48]. Both mechanisms can operate in parallel.

Thus far, modeling bottom-up attentional salience has been more fruitful than modeling the top-down more deliberate attention. Modeling image-based saliency cues has the advantage of being domain and task independent. There is also much less deviation between humans in how they exhibit this type of attention. Additionally, well established 3-D graphics techniques are good candidates for this type of modeling. Two categories of graphics models exist: ray casting and false coloring. "Ray casting can be functionally described as drawing an invisible line from one point in a 3-D simulation in a certain direction, and then reporting back all the 3-D object meshes this line intersected and the coordinates of these intersections [46]." This technique is often used in offline rendering of graphics, but it is typically too computationally expensive to use for real-time graphics rendering.

The false coloring technique is more suitable for real-time graphics rendering and is thus more useful for modeling visual attention. In this model, each object on the interface is assigned a unique color or vision ID [65]. This color differs from the color displayed on the interface and is only used for off-screen attentional salience modeling [45]. In the LIVE project Kelleher et. al. [45, 46] use the false coloring technique to model visual salience. This is done for the purpose of identifying which objects in the visual environment are likely to be referred to during speech. Their basic assumption is that an object's salience depends on its size and centrality within a scene. Thus, in their algorithm, each pixel of the false-colored image is weighted according to its centrality in the scene, with pixels closer to the center receiving higher weight. Then, all weighted pixels of the same false color, corresponding to a unique object, are summed together to construct the object's salience score. At each point in time during interaction with the LIVE system, objects can be ranked by their salience scores. Kelleher et. al. only accommodate a small portion of the visual salience factors described in [50], but they claim that this is a reasonable visual salience model that is capable of operating in real-time. Furthermore, the use of false coloring allows their algorithm to naturally account for the effects of partial object occlusion [45].

Byron et. al. [6, 7] construct a more sophisticated visual salience model for the purpose of coreference resolution in a situated dialogue utilizing the Quake II game engine. In this domain, as in the LIVE project, the field of view presents as ongoing data stream. However, Byron et. al. explicitly model the temporal features of the domain along with the visual features described in [50] and [45]. Additionally, in this domain, the object centrality feature turns out to be a poor indicator of visual salience due to some navigational control difficulties presented by the interface.

The predominant features considered in Byron's attentional salience model include *uniqueness* (U), *recency* (R), and *persistence* (P). An object's uniqueness—deviation from context—causes it to become attentionally salient. Here, the uniqueness of an object is modeled based on its frequency of occurrence in a given time window. Initially, all objects have the same uniqueness value. Over time, objects that appear in the field of view more frequently are penalized. Once an object drops out of the field of view, it is assumed that its attentional salience decays with time. This decay is modeled by a Gaussian function, which ensures exponential decay. The recency of an object partially depends on its persistence—the amount of time it was visible prior to going out of view. The computations of recency and persistence are consistent with the decay of visual memory [23].

2.3.2 User Attention Modeling via Eye Gaze

In the iTourist project, Qvarfordt et. al. [78, 77] attempt to take a step toward using eye gaze as an integrated channel in a multimodal system. They attempt to determine the visual salience of each object as a user views a map interface designed to facilitate a trip planning task. As people gaze at objects on the screen, a salience score (IScore) is calculated for each object. Once this score reaches a predefined threshold, the object becomes *activated* and the system provides information about this object to the user. In this scenario eye gaze is knowingly used by a participant as an active mode of input.

In this project, the salience score (IScore) was determined by a quadratic combination of eye gaze intensity along with a linear combination of auxiliary features extracted from eye gaze data. The most important factor in this salience score is the *absolute fixation intensity* measure, which is the amount of time object o is fixated in a given time interval. The auxiliary features considered in this work include fixation frequency, object size, the categorical relationship with the previous active object, and the IScore of the previous active object.

Qvarfordt's IScore measurement is a good starting point for modeling attentional salience using eye gaze. However, it has several deficiencies that we would like to overcome. First, it uses eye gaze as an overt conscious mode of input as opposed to our use of eye gaze as a subconscious auxiliary input mode. This represents itself in the features where the IScore depends on the previously activated object. In our scenario, these features are unavailable. Second, the domain used in this work is not very visually rich and does not demonstrate the use of situated language. Finally, the influence weight constants are empirically defined. These weights may be dependent on the given domain or particular task; thus, it is preferable to construct a domain-independent attentional salience model.

2.4 Eye Gaze

Previous psycholinguistic studies have shown that eye gaze is one of the reliable indicators of what a person is “thinking” about [30]. The direction of gaze carries information about the focus of the user’s attention [41]. Even though there may be some dissociation between *covert* visual attention and *overt* visual attention, which is expressed in the form of eye movements, behavioral neurophysiological studies have shown a tight coupling between the two [21, 30]. In this section, we discuss studies that have utilized eye gaze information during language production and human-machine communication.

2.5 Role of Eye Gaze in Language Production

Eye gaze has been continually used as a window to the mind in language comprehension and production studies. In human language processing tasks specifically, eye gaze is tightly linked to cognitive processing. Tanenhous et. al. conducted an eye tracking study to determine that the perceived visual context influences spoken word recognition and mediates syntactic processing [85]. Meyer et. al. [59] studied eye movements in an object naming task. It was shown that people consistently fixated objects prior to naming them. Griffin [28] showed that when multiple objects were being named in a single utterance, speech about one object was being produced while the next object was fixated and lexically processed.

Our work differs from these studies in several ways. First, our goal is to improve automated interpretation of utterances rather than to study the reasons that such utterances are made. Second, our domain contains a richer visual environment. Third, utterances produced in our setting are conversational in nature.

2.6 Role of Eye Gaze in Human-Machine Communication

Initially in human-machine interaction, eye gaze was used as a pointing mechanism in direct manipulation interfaces [34] or as an auxiliary modality to speech during multimodal communication [8, 78, 43]. However, these studies have shown that people prefer not to use eye gaze as an active input mode, but do not mind using eye gaze as a passive input mode.

As eye tracking technology becomes more available, eye gaze information is being used more frequently in human-machine communication as an implicit, subconscious reflex of speech. During interaction with a multimodal conversational interface, the eye automatically moves towards the object of interest without the user needing to make a conscious decision to do so. In fact, the user is often unaware of every object that was visually attended.

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 [61, 3, 80, 60, 2]. Recent work has explored incorporating eye gaze into automated language understanding such as automatic speech recognition [75, 17], automated vocabulary acquisition [54, 76], and attention prediction [78, 20]. While, an improvement in speech recognition is likely to lead to an improvement in reference resolution and overall multimodal interpretation, it is insufficient for avoiding all possible reference resolution errors. Even with perfect speech recognition, language contains enough ambiguity to make reference resolution a challenging task. For example, given a display containing two (or more) chairs and the spoken utterance “I would remove that chair”, it is impossible to identify which chair is being referred to based solely on this utterance. We use eye gaze and visual context information to handle such ambiguities.

Chapter 3

Eye Tracking Technology

3.1 Introduction and Motivation

Major developments in eye tracking technology have been seen in recent years. These developments are making eye trackers increasingly more accurate, less intrusive, and less expensive. The tradeoff between these three criteria makes choosing an eye tracking system a non-trivial task. The following three decisions must be made when choosing an eye tracker for an experiment:

- Infrared vs. visible light data processing
- Head-mounted vs. remote
- Single camera vs. multi-camera

Eye trackers that use infrared imaging to determine eye gaze position tend to be significantly more accurate in a laboratory setting than those that use the visible light spectrum. However, infrared eye trackers are less suited for use outdoors because of interfering infrared light. Infrared eye trackers use hardware specific methods to determine eye gaze position. Conversely, eye trackers that work with the visible light spectrum employ more computationally expensive image processing techniques in software. Commonly used commercial eye trackers employ infrared imaging. There are multiple open source and research prototype eye tracking systems that employ imaging using the visible light spectrum.

Head-mounted eye trackers tend to be more accurate and more obtrusive than remote eye trackers. Here, remote eye tracking refers to "a system that operates without contact with the user and permits free head movement within reasonable limits without losing tracking [58]." The first remote eye trackers were composed of a multi-camera setup (which reduces the need for camera calibration), but most recent commercial and research technology in both head-mounted and remote eye tracking is based on a single camera setup.

People who are predominantly interested in eye tracking accuracy (e.g. vision and psycholinguistic researchers) are best served by infrared head-mounted eye trackers. People who are interested in unobtrusive, relatively high-accuracy, real-time eye tracking (e.g. HCI analysis) are best served by infrared remote eye trackers. People who are interested in inexpensive outdoor or off-line eye tracking are best served by eye tracking based on the visible light spectrum.

The research focus of this dissertation, eye gaze for reference resolution in multimodal conversational interfaces, does not directly fall into any of these three categories. However, can be best suited by an infrared remote eye tracker which allows for unobtrusive, accurate, real-time eye tracking. In our research, eye gaze is used to predict a user's focus of attention. This attention prediction model is then used to improve language understanding via improved reference resolution. Much of this work consists of offline processing of collected user speech and eye gaze data. Thus, there is potential for using inexpensive webcam based eye tracking for a portion of this research. On the other hand, part of this research is investigating fine grained temporal relationships between spoken referring expressions and corresponding eye fixations. Commercial infrared eye trackers are more suitable for this investigation. Webcam-based eye trackers are becoming more suitable for this task as more efficient algorithms for detecting eye gaze position are being developed and computer systems are gaining in processing power.

In the following sections, I will discuss the potential eye tracking technology for the aforementioned research. First, I will describe the eye tracking requirements for our investigation. Then, I will compare the available eye tracking technology. Finally, I will determine which of the eye trackers is best suited for completing this investigation.

3.2 Requirements

As a part of our investigation, we must conduct user studies to gather user interaction data with multimodal conversational interfaces. For example, one user study (described in detail in §4.3) consists of a Treasure Hunting task in an immersive, situated game environment. The user study is expected to use a 17 inch LCD monitor with 1280x1024 resolution. This is equivalent to 41.9 and 33.5 pixels per cm in the horizontal and vertical directions, respectively. The distance between the user and the screen (including the eye tracker) is expected to be approximately 60 cm. The user is expected to interact with the system for a duration of 30 minutes or until the Treasure Hunting task is finished. In the preliminary user studies, no users were able to finish the task in less than 30 minutes.

In choosing an eye tracker for any user study, there are several statistics to keep in mind. These include eye gaze position accuracy, sampling rate, and range of allowable head movement. Eye gaze position accuracy is typically expressed in terms of visual angle. It can be converted to distance on the screen if the distance between the user and the screen as well as the screen resolution are known. For a specific experiment, the required eye gaze position accuracy is dependent on the screen resolution and the size of the objects that appear on the interface. If the objects are large and only a few objects can appear on the screen at one time, fixations to the correct object can be identified even with low eye gaze position accuracy. Sampling rate refers to how often a gaze position is recorded and is typically expressed in Hertz (Hz). The required sampling rate is dependent on the temporal granularity of the user behavior to be analyzed. For example, a very high sampling rate is needed to study saccades because they constitute very rapid eye movements. The range of head movement refers to the distance (typically measured in centimeters), relative to the position the head is in during camera calibration, in each of three directions a user's head can move without a noticeable degradation in accuracy. The required head movement range is unlikely to be a bottleneck for the experiments discussed in this dissertation. Users are expected to move their head freely, but not excessively.

3.2.1 Eye Gaze Position Accuracy

Since eye gaze position is used to determine which object is fixated, the size of the objects appearing on the interface is critical. Ideally, we would like to ensure that given a gaze position for an intended object the system is 95% confident that a sampled gaze point will fall onto this object.

A rough estimate for required gaze position accuracy for the situated environment can be estimated from data for the static environment. The object sizes in these two environments are comparable. In the static environment area of the smallest object 1700 pixels when viewed on the screen. The average size for a fixated object is 18208 pixels. A rough approximation for the maximal gaze position accuracy for this setup can be computed. It is important to note that intentional fixations to an object tend to occur near its center. This is especially true for small objects. Thus, a better estimate of required gaze position accuracy is to consider the smallest object of interest in our study and assume that a user will fixate the exact center of this object. To simplify this calculation, this object is assumed to have circular shape with

$radius = \sqrt{\frac{1700}{\pi}} = 23.26$ pixels. The results for 95%, 90%, and 80% object fixation confidence for this scenario are shown in Table 3.1. These estimates are a bit optimistic for this object because a user is not guaranteed to attempt to fixate on its exact center. However, larger objects have a much higher gaze-position error tolerance even if an off-center point is fixated. Thus this is a reasonable approximation for the maximum gaze position error that can be tolerated

Confidence Level	Maximum Error Approximation (in pixels)
95%	23.86
90%	24.52
80%	26.01

Table 3.1: Maximum Allowable Gaze Position Error

3.2.2 Sampling Rate

The required eye tracking sampling rate is dependent on the temporal granularity of the phenomena to be observed. In our investigation, we are interested in observing fixations to objects that can be cognitively processed. According to psycholinguistic studies, fixations lasting as little as 20 ms can be cognitively processed. In the data we collected for the static scene scenario, the shortest fixation length is 52 ms, the mean fixation length is 356 ms with a standard deviation of 281 ms.

One important aspect of an eye tracking system is determining whether or not a gaze point constitutes a fixation. Fixations are typically detected by comparing a series of consecutive gaze points. If the gaze position is approximately the same during a series of gaze points whose total length is at least 20 ms, then this series is deemed to be a fixation. Based solely on a single gaze point, it is impossible to determine whether this gaze point constitutes a saccade or a fixation. Thus the minimum acceptable sampling rate for detecting every fixation in our data is 2 gaze points per 52 ms or, equivalently 38.46 Hz. Detection of 95% of fixations may be sufficient. In our data, fewer than 5% of fixations are 120 ms or shorter. Thus, for this data a sampling rate of 2 gaze points per 120 ms (16.67 Hz) may be acceptable.

3.3 Eye Tracker Comparison

As mentioned in §3.2, available eye trackers vary in eye gaze position accuracy, sampling rate, and range of allowable head movement. In the following sections I will compare four eye tracking systems according to these three criteria. Two of the systems are commercial systems based on infrared imaging, while two of the systems are low cost research prototypes. The Eye Link II is a commercial head-mounted eye tracker, while the OpenEyes [52] is a low cost head-mounted alternative. OpenEyes is an open-source open-hardware toolkit for low-cost real-time eye tracking. It uses a USB web camera with a resolution of 640x480 pixels and with a frame rate of 30 Hz. The Tobii eye tracker is a commercial remote eye tracker, while the web camera based eye tracker described in Hennessey, et. al. [31] is a low cost alternative. This eye tracking system also uses a camera with a resolution of 640x480 pixels and with a frame rate of 30 Hz. Other research prototypes [66, 79, 88] are available. However, none of these have reported higher accuracies or sampling rates than the two aforementioned options. Also, many of these have reported a smaller head movement range. Additionally, Pedersen et. al. [69] claim that they can use a webcam for offline processing of eye tracking data to achieve comparable performance to commercial eye trackers, but they do not provide gaze position accuracy statistics for comparison.

3.3.1 Eye Gaze Position Accuracy

Eye gaze position accuracy is typically measured by having users perform a calibration, and then looking at a number of evenly distributed points on the screen. Accuracy is reported as the average deviation between the intended and measured gaze data points. It is typically expressed in terms of error in visual angle between the actual gaze position and the gaze position predicted by the eye tracker. The calculation of visual angle can be seen in Figure 3.1. Given the visual angle θ_1, gaze position error can be expressed in terms of distance $|ba|$ using Equation 3.1. Given $|d| = 60$ cm and a 41.9 pixels/cm screen resolution, gaze position error can be expressed in terms of pixel length.

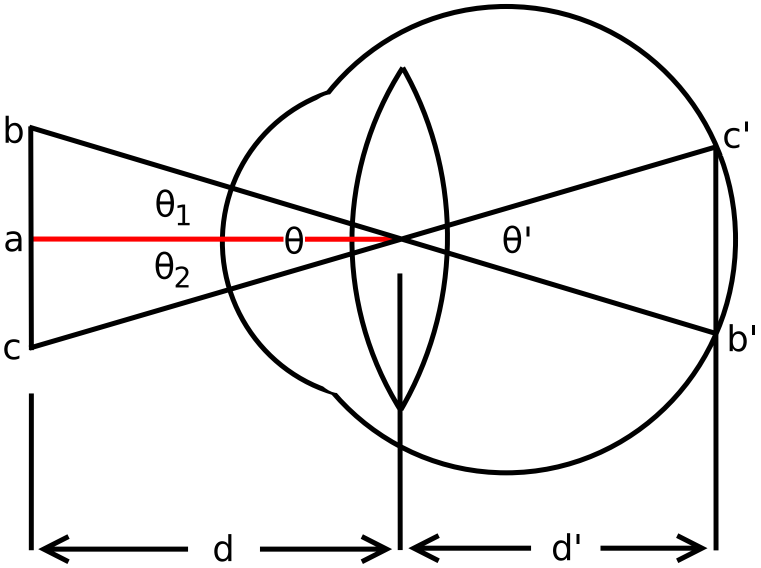


Figure 3.1: Calculation of Visual Angle — For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

$$|ba| = tan(\theta_1) \times d \tag{3.1}$$

The gaze position error for four systems is shown in Table 3.2.

	Visual Angle (degrees)	Distance (mm)	Distance (pixels)
Eye Link II	0.5	5.24	21.99
Tobii 1750	0.5	5.24	21.99
OpenEyes [52]	1.0	10.48	43.98
Remote web camera [31]	0.9	9.43	39.58

Table 3.2: Eye Gaze Position Error

Based on these findings, the two commercial eye trackers are clearly more appropriate for our investigation than the low cost research prototypes. The commercial eye trackers fall within the 23.86 pixel maximum gaze position error length (as shown in Table 3.1), while the research prototypes don't come close. This table shows the maximum allowable gaze position error in pixels for the smallest appearing object. For objects as small as 46.52 pixels in width (corresponding to the smallest object used in our study), an error of 39.58 pixels is unacceptable. For larger objects this error may be acceptable.

It is important to note that these results were calculated under ideal conditions. Head movement and systematic gaze drift can have a drastic effect on these error lengths (typically doubling them in commercial eye trackers). These types of errors can be minimized by often recalibrating the eye tracker during a user study. More accurate eye trackers are preferred because frequent recalibration can be disruptive during a user study.

3.3.2 Sampling Rate

The sampling or frame rate of a camera is typically reported in Hertz (Hz). The sampling rates of four eye tracking systems are shown in Table 3.3.

	Sampling Rate (Hz)	Sampling Rate (ms)
Eye Link II	250	4
Tobii 1750	50	20
OpenEyes [52]	30	33
Remote web camera [31]	30	33

Table 3.3: Eye Tracker Sampling Rate

Based on these finding, the Tobii 1750 eye tracker is most appropriate for this investigation. It can achieve a sampling rate higher than the conservative requirement of 38.46 Hz. The much higher sampling rate of the Eye Link II is likely unnecessary, albeit never harmful. The more lenient requirement of 16.67 Hz can be satisfied with either of the two low cost alternatives.

3.3.3 Head Movement Range

The difference between the systems in terms of head movement range is immaterial for eye tracking accuracy in the experiments presented in this dissertation. However, a non-intrusive remote eye tracker with a large range of allowable head movement is more suitable for human-computer interaction. The display-mounted Tobii 1750 eye tracker enables natural human-computer interaction during which users may even be unaware that their eye gaze is being tracked. The Eye Link II head-mounted eye trackers are less suitable for these experiments because they may make users uncomfortable, thus affecting the interaction.

3.4 Summary

The analysis in this chapter shows that the Eye Link II and the Tobii 1750 eye trackers are both adequate for our investigation. The low cost research alternatives may be adequate under ideal conditions, but are not ready to be used for our eye tracking needs. Given a choice between the two commercial eye trackers, the Tobii 1750 is preferred because it is a remote rather than head-mounted eye tracker. This has the advantage of it being unobtrusive to the user, and thus is more suitable for the experiments presented in this dissertation.

Chapter 4

Parallel Speech and Eye Gaze Data Corpus

4.1 Introduction

Currently, there is no publicly available data corpus that contains a combination of spontaneous speech and eye movement. As far as we know, only a few studies track spontaneous speech and eye movement in a visually rich domain [16] [78]. Cooke [16] has constructed a corpus containing "eye movement direction data and related spontaneous speech for an unscripted, visually oriented, task constrained, human-to-human dialogue." A map-based domain was used for this study. Eye gaze is utilized as a supplemental input mode to assist in spoken language understanding. Qvarfordt et. al. [78] also used a map-based domain, but collected speech and eye movement data during a human-computer dialogue. In this study, eye gaze indicates user attention and signifies to the computer system that more information is desired about a particular object appearing on the visual display.

Our aim is to construct a corpus of spontaneous conversational speech and eye gaze in a visually rich domain. Similarly to [16], eye gaze is not used to directly manipulate the interface, but rather as a supplemental mode of input to assist spoken language understanding. Similarly to [78], the collected speech portrays a dialogue between a human user and a computer system. Moreover, our investigation requires a richer visual domain than exemplified in these studies.

The rest of this chapter describes two domains and corresponding user studies that were used to construct a parallel speech and eye gaze data corpus consisting of two datasets. A domain with a static visual interface is described in §4.2 and the experimental design used for data collection within this domain is presented in §4.2.2. A domain with a more complex immersive visual interface is described in §4.3 and the experimental design for this domain is presented in §4.3.2. The resulting parallel data corpus is described in §4.4. With the data corpus collected in these studies, we are capable of addressing the research questions proposed in §1.2.

4.2 Static Domain

4.2.1 Interaction Environment

In the first user study, a static multimodal conversational interface is used to collect speech and gaze data. Users view a 2-dimensional snapshot of a 3-dimensional virtual bedroom (shown in Figure 4.1) and answer a series of questions about what they observe. In this figure, each object is shown with a unique ID that is hidden from the users. The scene used in this experiment contains objects such as a door, a bed, desks, chairs, etc. It contains many distinct yet similar objects, most of which users are familiar with. Some of the objects in the room are arranged in a typical expected fashion, while other objects are out of place. Many objects visually overlap (are in front or behind) other objects.

Each object is defined as a Region of Interest that the system "knows" about. Each Region of Interest belongs to a generic semantic category. The object categorizations

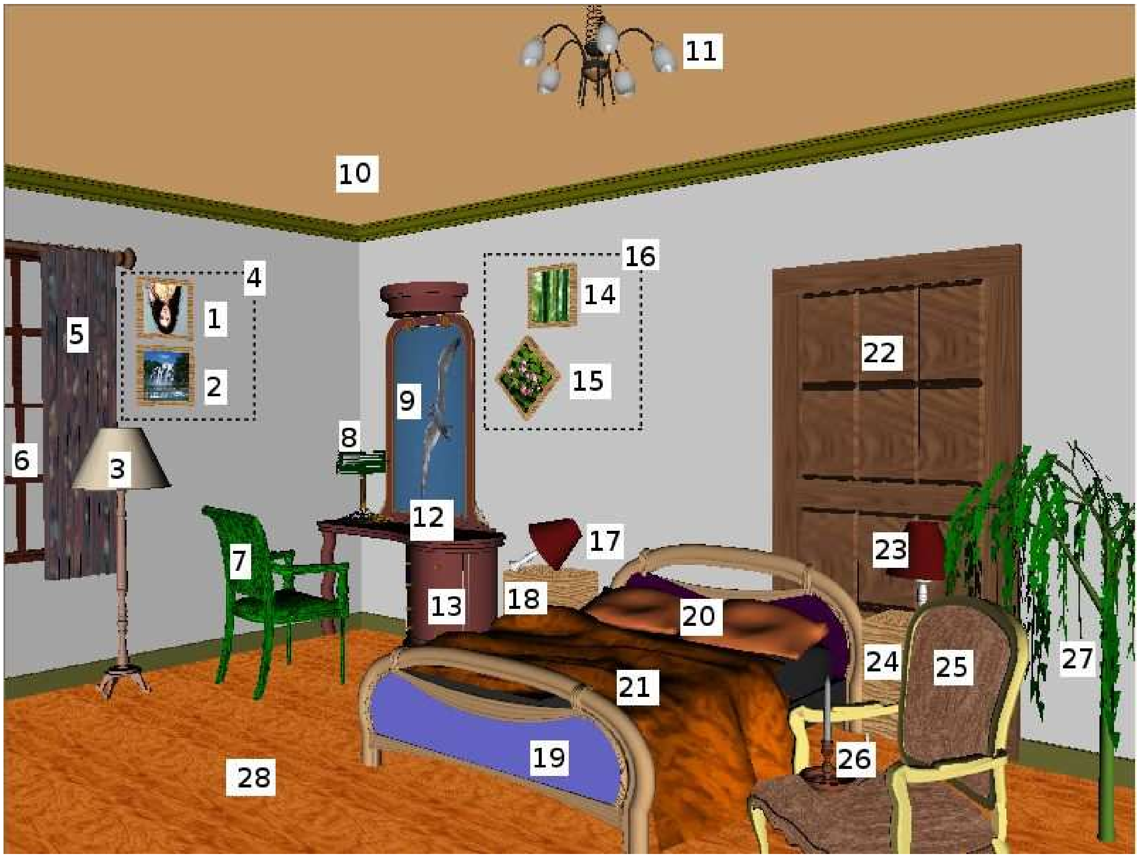


Figure 4.1: Virtual Bedroom Scene

Semantic Category	Contained Objects (Ids)
bed	{19}
blanket	{21}
candle	{26}
chair	{7, 25}
curtains	{5}
door	{22}
lamp	{3, 8, 11, 17, 23}
mirror	{9}
picture	{1, 2, 14, 15}
pillows	{20}
plant	{27}
table	{12, 18, 24}
window	{6}
OTHER	{4, 10, 13, 16, 28}

Table 4.1: Semantic Categories in Virtual Bedroom

are shown in Table 4.1. This table shows the semantic categories with corresponding objects in the virtual bedroom scene. More detailed visual and linguistic information that is made available to the system is shown in Appendix A.

The rationale behind using this scene lies particularly in that this interface exhibits the aforementioned traits. One important trait is that multiple similar objects (e.g. multiple paintings) shown on the interface. This is likely to cause users to use similar vocabulary when referring to these objects, allowing eye gaze to be used as a disambiguation mechanism. Having distinct objects in the scene allows users to unambiguously refer to a particular object. This allows us to demonstrate the usefulness of utilizing eye gaze in spoken language understanding without constraining the scenario to *require* eye gaze for language interpretation. Another important trait is allowing the scene to contain overlapping objects. This complicates the problem, but makes it more realistic. When objects overlap, the direction of eye gaze does not provide sufficient amount of information to uniquely determine the object of attention. Thus, eye gaze becomes less reliable. We believe that it is important to consider and be able to deal with this problem. One more important trait is having some unexpected objects shown on the interface. According to vision studies, users are more likely to attend to an object in a scene if it contains unexpected characteristics; that is, significantly deviates from its surrounding objects. For example, a red object among many green ones or a toaster in a barn. As discussed in [28], users are likely to attend objects as they prepare to speak about them. These two phenomena may come into conflict. Placing some unexpected objects in the visual environment allows us to study whether one of the two phenomena dominates.

4.2.2 Experimental Design

Participants

Ten participants partook in the experiment. Of these, data for 4 participants had to be thrown out. Three were thrown out because inaccurate eye gaze tracking and one because of a failed speech recording. The remaining 6 participants were comprised of a mixture of graduate students and faculty from Michigan State University. Three were male and three were female. Four were native English speakers, while two spoke English as a second language.

Procedure

Users are asked to answer a series of questions about various objects in the virtual bedroom described in §4.2. These questions range from factual questions about particular objects to open-ended questions about collections of objects. For example, the following three questions were included in this experiment:

- Is there a bed in this room?
- What do you dislike the most in this room?
- How would you like to change the furniture in the room?

SR Research Experiment Builder[1] is used to construct the experiment procedure. Our experiment involves 14 trials (one per question) that are split into 3 sessions. Each session is preceded by eye tracker calibration. Frequent re-calibration of the eye tracker ensures accuracy. Vision studies suggest re-calibrating the eye tracker at least every 10 minutes. In our study, each session takes less than 10 minutes.

A particular trial involves displaying the scene to the user, asking the user a question about the scene, and recording their response as well as eye gaze. The user decides when he or she is finished with the trial by pressing any key on the keyboard. The trial flowchart is shown in Figure 4.2.

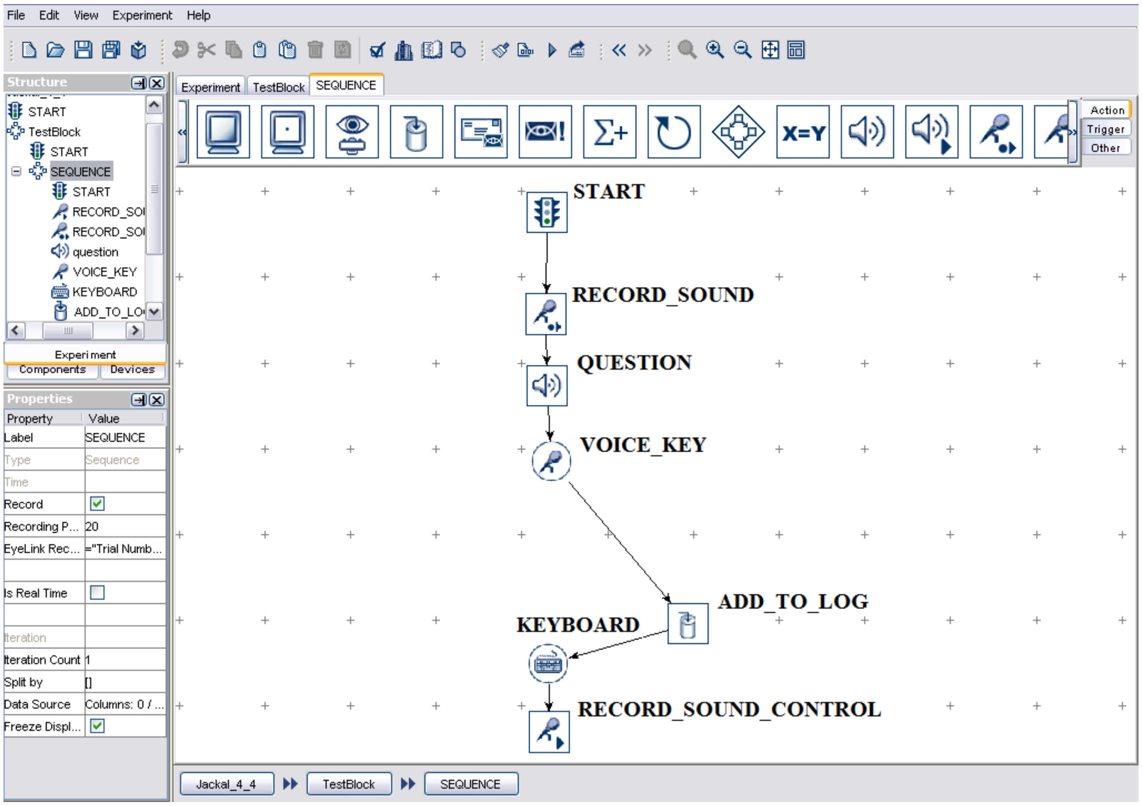
Each trial is followed by drift correction that forces the user to look at the center of the screen. This has a dual purpose. First, it acts as a mini-calibration that compensates for the drift between the eye and the crosshairs on the eye tracker. Second, it ensures that all trials begin with the user looking at the same point in the image. This eliminates any user bias of looking at a particular object before speech onset.

Apparatus

The Eyelink II head-mounted eye tracker (shown in Figure 4.3) sampled at 250 Hz is used to track user eye movements. Eye gaze points were garnered using only pupil reflection (rather than pupil and corneal reflection). The eye tracker automatically saves the screen coordinates and time-stamps of each gaze point. The collected raw gaze data is extremely noisy. The data is processed to eliminate irrelevant gaze points and smoothed to identify gaze fixations. The fixation compilation procedure is described in more detail in §4.4.1. All regions of interest circumscribing a fixation point are considered to be fixated.

Concurrently with eye tracking, user speech data is recorded using a noise-canceling

[1] http://www.sr-research.com/optns_eb.php



Property	Value
Label	SEQUENCE
Type	Sequence
Time	
Record	☑
Recording P...	20
EyeLink Rec...	="Trial Numb...
Is Real Time	☐
Iteration	
Iteration Count	1
Split by	[]
Data Source	Columns: 0 / ...
Freeze Displ...	☑

Figure 4.2: Screenshot of Trial Flowchart Constructed with SR Research Experiment Builder

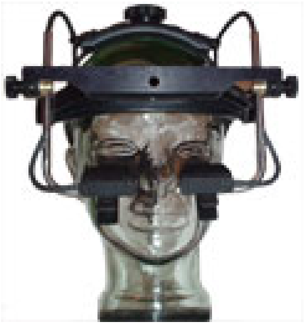


Figure 4.3: Eyelink II Head-Mounted Eye Tracker

microphone and the Audacity Sound Editor[2]. Each utterance is manually transcribed and annotated with the help of the Audacity toolkit. Each referring expression in the utterance is annotated with the correct references to either a single object (region of interest found in the eye gaze data) or multiple objects. The resulting data set is used to conduct our investigation.

4.3 Immersive Domain

4.3.1 Situated Interaction Environment

For the second user study, we created a 3D virtual world (using the Irrlicht game engine[3]) to collect data containing situated language and eye gaze. We conduct 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 regarding 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. Such an environment allows us to study situated language in a spatially rich, purposeful, and easily controllable environment as is sometimes done in video game environments [25].

A snapshot of a sample interaction with the treasure hunting environment is shown in Figure 4.4. Here, the user has just entered a computer room and is currently fixating on the file cabinet. The user's eye fixations are represented by white dots, while the user's saccades (eye movements) are represented by white lines.

The treasure hunting environment contains similar types of objects as the bedroom scene described in §4.2. A user can navigate through this environment and converse

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

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

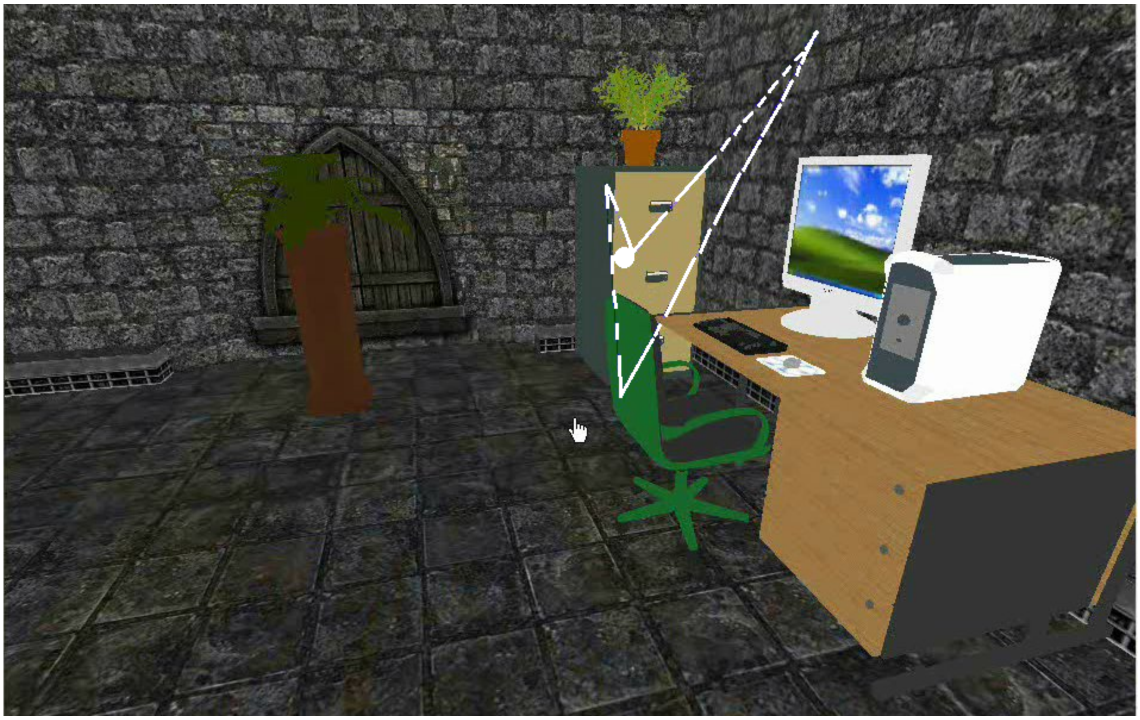


Figure 4.4: Screenshot of Situated Treasure Hunting Environment

with the system while viewing the objects from any vantage point the user prefers. This means that objects that are being referred to by the user may visually overlap (e.g. the chair and the file cabinet object in 4.4) or may no longer be in the field of view, altogether. Each object is specified by a three dimensional model whose location relative to the environment, and consequently relative to the field of view is available to the system. In total 155 unique objects that encompass 74 different semantic categories (e.g. chair, plant, cabinet, etc.) are present in the treasure hunting environment.

4.3.2 Experimental Design

Participants

Sixteen participants partook in the experiment. Data for one of the participants had to be thrown out because of a failure in training an automatic speech recognition user model. All of the participants were comprised of students from Michigan State University.

Procedure

Users are presented with a virtual castle and a treasure hunting task. They are told to navigate the castle and communicate with an artificial conversational assistant to find the treasure. The assistant has partial knowledge about where the treasure is and how to find it. The user has additional knowledge about what can be seen in the castle environment. Together they are responsible for deciphering this puzzle. In reality the conversational assistant consists of a Wizard of Oz interface operated by a human. All user speech and eye gaze is automatically recorded by the system.

Apparatus

The Tobii 1750 display-mounted eye tracker (shown in Figure 4.5) sampled at 50 Hz is used to track user eye movements. Eye tracking is integrated into the treasure hunting interface; automatically saving the screen coordinates and time-stamps of each gaze point. The noisy raw gaze coordinate data is is processed to eliminate irrelevant gaze points and smoothed to identify gaze fixations; described in more detail in §4.4.1. Each object in the line of view of a fixation constitutes a potentially fixated object; with an object in the foreground being assigned a higher fixation probability than one in the background.

Automatic speech recognition (ASR) is also integrated into the treasure hunting interface. A noise-canceling microphone is used to record user speech data. Microsoft Speech SDK is used to construct an n-best list of ASR hypotheses for each user utterance. Each utterance in this data is manually annotated to include speech transcription, identification of referring expressions, and the set of objects that the expressions refer to.

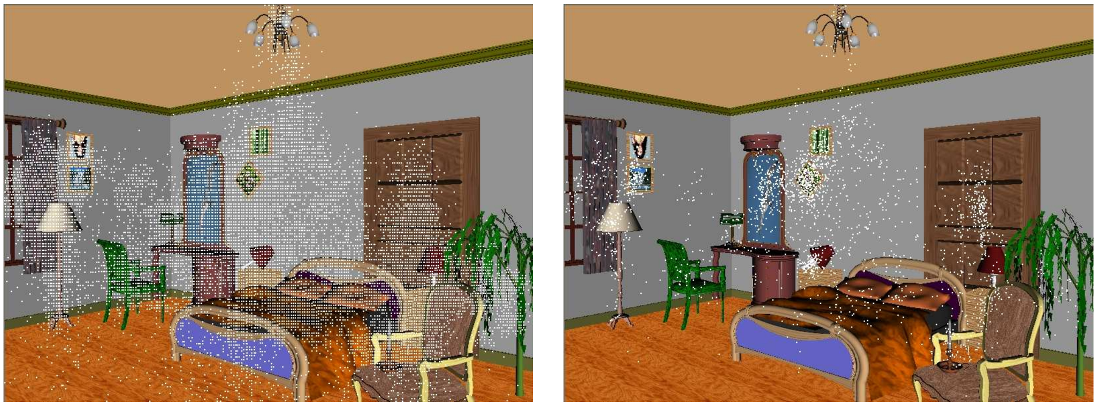
Figure 4.5: Tobii Display-Mounted Eye Tracker

4.4 Data Processing

As mentioned in §4.1, our goal is to construct a parallel corpus of spontaneous conversational speech and corresponding eye gaze. The collected data, as it is described above, cannot serve this purpose. It needs to be post-processed in order to construct the desired parallel corpus. The following sections describe this in more detail.

4.4.1 Postprocessing of Eye Movement Data

The collected raw gaze data is extremely noisy. The raw data consists of the screen coordinates of each gaze point sampled at every four milliseconds. As can be seen in Figure 4.6(a), this data is not very useful for identifying fixated objects. This figure shows all of the gaze points occurring for a particular trial of a particular user study. The raw gaze data is processed to eliminate irrelevant gaze points and smoothed to identify fixations. Irrelevant gaze points occur either when a user looks on-screen or when a user makes saccadic eye movements. Vision studies have shown that no cognitive processing occurs during saccadic eye movements; this is know as "saccadic suppression" [56]. In addition, to removing invalid gaze points the data is smoothed by aggregating short, consecutive fixations. It is well known that eyes do not stay still, but rather make small frequent jerky movements. In order to best determine fixation locations, five consecutive gaze locations are averaged together to identify fixations. The processed eye gaze data can be seen in Figure 4.6(b). Here, each point represents a gaze fixation. The gaze fixations tend to be congregated around regions of interest.



(a) Raw fixation on the display

(b) Smoothed fixation on the display

Figure 4.6: The 3D Room Scene for User Studies and Eye Fixations on the Interface

Given a domain model specifying the location of each object on the screen, the gaze fixation coordinates can be used to determine which objects are fixated at a particular time. Note that it is impossible to uniquely identify exactly which object is fixated at a given moment because several objects may overlap. The resulting eye gaze data consists of a list fixations, each of which is time-stamped and labeled with a set of potentially attended interest regions.

4.4.2 Annotation of Speech Data

Speech data acquired from user interaction with the static multimodal interface is manually transcribed and timestamped with the help of the Audacity toolkit. This toolkit is used to detect pauses (short periods of silence in the speech signal). The beginning of a pause signifies the end of a word, while the end of a pause signifies the beginning of the next word. This information is used to determine the start and end timestamp of each word in the transcript.

Speech data acquired from user interaction with the immersive multimodal interface is also manually transcribed. However, the timestamp information is automatically determined by combining the transcript with the ASR recognition hypotheses that are generated by Microsoft Speech SDK.

In both data sets, each utterance is manually annotated with all referring expressions. Each exophoric referring expression (one that refers to an object appearing in the virtual environment) is labeled as either a definite or pronominal noun phrase. Additionally, it is labeled with all of the objects that are being referred to by the expression.

4.5 Summary

This chapter documents the design and collection of a parallel speech and eye movement corpus within two domains. A user study was conducted for each domain, resulting in two datasets. Dataset 1 is created via user interaction with the static multimodal interface and is used for constructing attention prediction and reference resolution models described in Chapters 5 and 6, respectively. Dataset 2 is created via user interaction within the immersive virtual world and is used for constructing reference resolution models described in Chapter 7. Both datasets contain annotated spontaneous conversational speech with references to objects in the extralinguistic environment along with eye fixations corresponding to the referenced objects.

Chapter 5

Eye Gaze in Attention Prediction

5.1 Introduction

In a conversational system, determining a user's focus of attention is crucial to the success of the system. Motivated by previous psycholinguistic findings, we are currently examining how eye gaze contributes to automated identification of user attention during human-machine conversation. As part of this effort, we investigate the contributing roles of various features that are extracted from eye gaze and the visual interface. More precisely, we conduct a data-driven evaluation of these features and propose a novel evaluation metric for performing such an investigation. In this chapter, we describe this empirical investigation and discuss further implications of eye gaze based attention prediction for language understanding in multimodal conversational interfaces.

This investigation differs from previous work in two aspects. First, previous studies examine the use of eye gaze as an active mode of input that controls the navigation of the interface. This work has shown that gaze *fixation intensity* (commonly referred to as *dwell time*) is an important feature for predicting user attention [78, 77]. More precisely, the dwell time is the total amount of time that a particular object is fixated during a given time window, while the fixation intensity is the ratio betwen the dwell time and the length of the time window (which is typically kept constant). The work reported here addresses a different scenario, where speech is the main mode of interaction, while eye gaze is a naturally occurring byproduct. Second, unlike previous investigation focusing on the role of eye gaze in language production [59, 27], our work is conducted in a conversational setting that involves interaction between a user and a machine. Additionally, as described in Chapter 4, a more visually rich domain is used than in typical eye tracking studies. These unique settings, which have received less attention, apply to a range of realistic and important problems that involve speech communication between a user and a graphical display. This chapter investigates the role of eye gaze in this important setting. Specifically, we attempt to address the following research questions:

1. How are pertinent eye gaze fixations temporally distributed relative to spoken utterances?

2. What effect do eye gaze based features have on performance of the user attention prediction task?

3. Can auxiliary visual features further enhance the reliability of the prediction? If so, what effect do different representations of these auxiliary visual features have on performance of the attention prediction task?

4. What are appropriate evaluation metrics for measuring the reliability of different features in the attention prediction task?

In the rest of the chapter we describe the problem in more detail and attempt to answer these research questions. The attention prediction problem is formally defined in §5.2. Sections §5.3 - §5.6 address the four research questions posed above. Specifically, §5.3 explores the temporal alignment between speech and eye gaze, §5.4

discusses the evaluation metrics used to conduct our evaluation, §5.5 describes the features that are considered in our study, and §5.6 presents the experimental results of our feature evaluation. Finally, §5.7 provides the highlights of what has been learned from this investigation.

5.2 Problem Definition

We formulate the attention prediction task as an *object activation* problem. This task involves identifying whether at a given time a particular object on the graphic display is activated or not. An object is considered to be activated if it is, indeed, the focus of attention. This problem can be modeled as a binary classification problem. However, the binary decision of whether or not an object is activated is too coarse to portray a realistic model of human focus of attention. Instead, we determine the likelihood that each object is activated and rank the objects in descending order based on their *activation score* (also called *attentional salience score*). This method can be used to model a gradual change of the focus of attention with respect to time. An object's likelihood of activation, unlike its class label, cannot jump from 1 to 0 instantaneously. Additionally, this method allows us to make a more fine-grained evaluation of our feature set because an object may have more possible ranking scores compared to just two boolean values.

5.2.1 Object Activation Model

To serve our purpose, we chose the logistic regression model in our investigation. This model can be used combine multiple continuous numerical features for predicting object activation and because can directly compute the probability of an object to be activated—this value can be used to rank object of activations. This approach computes a model that best describes the data while minimizing assumptions made about how the data is generated (maximizing entropy). It can be used to objectively determine the reliability of various features for the object activation task. Features that consistently cause the model to achieve higher performance can be considered more reliable.

Logistic regression uses a well-known objective function to determine the likelihood that a given data instance belongs to a particular class [81]. This model assumes that the log-ratio of the positive class to the negative class can be expressed as a linear combination of features as in the following equation:

$$log\left(\frac{p(y^{+}|\overrightarrow{x})}{p(y^{-}|\overrightarrow{x})}\right) = \overrightarrow{x}\overrightarrow{w} + c \tag{5.1}$$

where the following constraint holds:

$$p(y^{+}|\overrightarrow{x}) + p(y^{-}|\overrightarrow{x}) = 1 \tag{5.2}$$

Here, y refers to the class label (in our case, y^{+} means activated and y^{-} means not activated), $\overrightarrow{x}$ refers to the feature vector, and $\overrightarrow{w}$ and c are parameters to be learned from the data. Thus, the activation score can be formally defined in Equation (5.3):

$$AS(\overrightarrow{x}) = p(y^{+}|\overrightarrow{x}) = \frac{1}{1 + e^{-\overrightarrow{x}\overrightarrow{w}-c}} \tag{5.3}$$

The logistic regression model was chosen for this study because it can be used to combine continuous numerical features in order to directly compute the probability of an object activation, which can be used to rank objects according to the activation score. Additionally, logistic regression has been shown to be effective for constructing a ranking model within a small data set [70]. It would be useful to compare the logistic regression results with other machine-learned ranking models. Ranking Support Vector Machines (SVM) [32, 36] and Naive Bayes [91] models have been shown to be particularly effecting for creating ranking models for information retrieval. Decision tree induction, classification rules, and neural networks have also been shown to create competitive ranking models for certain domains. A comparison of these models is out of the scope of this work. For a rigorous survey on ranking models, see the paper "Learning to Rank for Information Retrieval" [53].

5.2.2 Dataset

The dataset used to develop the object activation model is derived from the Parallel Speech and Eye Gaze Data Corpus described in Chapter 4. The collected eye gaze data consists of a list fixations, each of which is time-stamped and labeled with a set of interest regions. Speech data is time-stamped and each referring expression in the speech utterance is manually annotated with the correct references to either a single object (region of interest found in the eye gaze data) or multiple objects. The data instances used to train and evaluate the various configurations of the object activation model is constructed by segmenting the parallel corpus into frames.

Here a frame denotes a list of data instances occurring during a particular time window W. Each frame corresponds to a spoken reference, while each data instance corresponds to an interest area that is fixated during W. For each object that is deemed to be attended (according to the human annotator), the corresponding data instance is assigned the positive (y^+) class label. For each object that is fixated, but not referenced, the corresponding data instance is assigned the negative (y^-) class label. The feature vector for each data instance is computed from the eye gaze data. The features are described in more detail in §5.5. In total, 449 frames containing 1586 data instances were used in the reported studies.

Currently, we have set W to [-1500..0] ms relative to the onset of a spoken referent to an on-screen interest area. Other time windows are possible, but this one seemed to achieve the best empirical results. This time window is comparable to prior language production studies, which have shown that users typically fixate an object between 0 and 1200 ms prior to naming the corresponding object in speech. In our case, a slightly wider range for W is reasonable because our display is visually richer than the one used in the aforementioned studies.

5.2.3 Activation Model Training

The Bayesian Logistic Regression Toolkit [22] provided by Rutgers University is used to create computational models that rank objects of interest in a given time window W based on their likelihood of activation. In the training phase the system automatically learns the influence weights of our various features that maximize this function's correspondence with our data (or equivalently, minimizes deviation from our data). Given an unknown data instance, the resulting model provides the likelihood that this data instance belongs to the *activated* class.

5.2.4 Application of Activation Model

Here we discuss how to apply an activation model to rank objects of interest in a given time window (represented as a frame of unclassified data instances). As we have already mentioned, an activation model can provide the likelihood that a particular data instance belongs to the *activated* class. Given a data frame associated with a particular time window, the model is applied to each instance in the frame. The data instances are then ranked in descending order based on their likelihood of activation as determined by the model. Note that our data collection scheme guarantees that each data instance in a particular data frame must correspond to a unique object of interest. Thus, the result is a ranked list of objects.

A single set of data was used for both training and testing in this evaluation. A rigorous evaluation should use other classification models as comparisons to logistic regression as well as a dedicated testing set for evaluation. However, this was not done because this was a preliminary and exploratory study designed to determine which features must be explored in more detail. The goal of this study was feature construction rather than a rigorous feature evaluation. Moreover, the data set that was used is very small: there are 449 time periods sampled; each time period is 1500-ms long and contains a mixture of 3.5 (on average) activated and non-activated objects.

5.3 Temporal Alignment between Speech and Eye Gaze

As a preliminary, to gain a better understanding of our problem and our dataset, we investigate the alignment between the speech and eye gaze data streams. Previous psycholinguistic studies have shown that on average eye gaze fixations to an object in a visual scene occur 932 ms before the onset of a spoken reference in a language production task [27]. Pertinent eye fixations—that is, those to objects that will be referenced—can range anywhere from 1500 ms before onset up until onset of a spoken reference. Knowing the range and the mean does not provide sufficient information about the nature of pertinent eye fixations in our complex scenes. We also need to know how eye gaze fixations are distributed relative to the onset of a spoken reference. This will allow us to have a better understanding of which eye fixations are indicative of attentional salience and which are irrelevant. Fixations that are deemed to be pertinent can be given a higher weight in our object activation model. Additionally, this experiment will allow us to validate our choice for time window W.

Figure 5.1 shows a histogram reflecting the percentage of pertinent fixations based on the time difference between the onset of a fixation and the corresponding spoken referring expression. Here the X-axis represents the starting point of a fixation within time window W, occurring prior to the onset of a reference to an object. This value ranges from 1500 ms before reference onset until precisely at onset. Fixations are aggregated into interval bins of 100 ms. The Y-axis represents the proportion of fixations that contain an object matching the referenced object. This proportion was calculated with the following procedure:

1. Each fixation is classified as pertinent or irrelevant. Irrelevant fixations are those that do not contain an object that is reference within time window W. Note that a several objects may be fixated during a single fixation. Also, note that this classification occurs relative to a particular spoken reference. Thus, a particular fixation can be classified as pertinent for one reference, but irrelevant for another.

2. Each fixation is classified into a bin of length 100 ms. The bins represent the amount of time that passes between the start of an eye fixation and an object reference. Thus, for example, the bar between 300 and 400 represents all fixations that begin occurring in the time range [-400..-300] relative to an object reference.

3. To calculate the percentage, the number of pertinent fixations in each bin is divided by the total number of fixations in this bin.

It is important to determine which time periods are most likely to generate a pertinent fixation. However, the large spread indicates that pertinent fixations are fairly evenly distributed during the 1500 ms time interval prior to a spoken reference. Nevertheless, as can be seen from Figure 5.1, there is a general trend that a fixation is more likely to be pertinent if it occurs close center of the 1500 ms time range rather than on the outskirts. The fixation data is shown to have a negative skew. That is, the left (lower value) tail of the graph is longer. Thus, a fixation is more likely to be

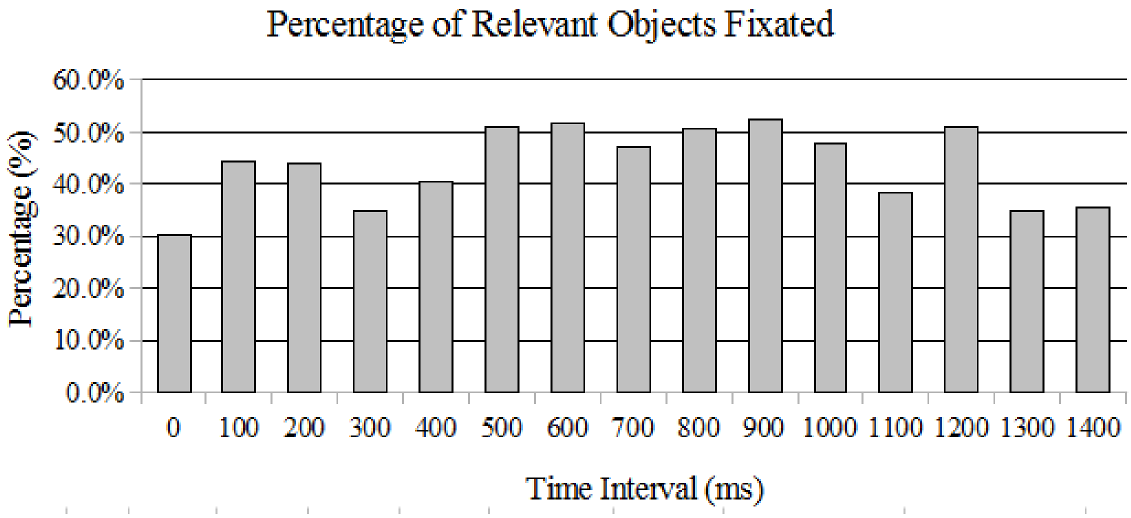


Figure 5.1: Percentage of Pertinent Eye Fixations Grouped by Onset

pertinent if its to the right of the mean—further from the spoken reference, but still within the 1500 ms time range—than to the left.

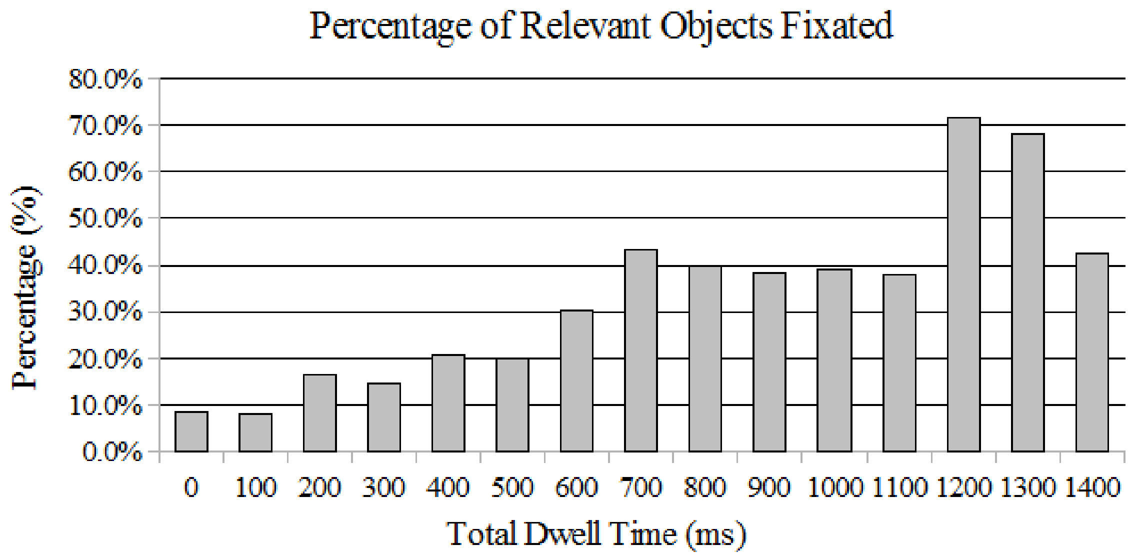


Figure 5.2: Percentage of Pertinent Fixated Objects Grouped by Dwell Time

It is also important to identify which fixation lengths are most likely to generate a pertinent fixation. The length of any one particular fixation is not important as the total dwell time for each object during the 1500 ms time window. The dwell time is the sum of the lengths of all fixations to a particular object occurring during the given time window. Figure 5.2 shows the relationship between an object's total dwell time and likelihood of being referenced during speech. The total dwell time is aggregated into 100 ms interval bins. For example, the bin marked 0 contains all of the instances where an object's dwell time was in the range $(0..100]$ ms. This figure shows a general trend that an object with a larger dwell time (or fixation intensity) is more likely to be referenced than an object with a small dwell time. However, this relationship is not completely monotonic.

5.4 Evaluation Metrics

To evaluate the impact of different features on attention prediction, we borrowed the evaluation metrics used in the Information Retrieval (IR) and Question Answering (QA) fields. In these fields, three key metrics have been widely used to assess system performance are Precision, Recall, and Mean Reciprocal Ranking (MRR). In IR, Precision measures the percentage of retrieved relevant documents out of the total number of retrieved documents and Recall measures the percentage of retrieved relevant document out of the total number of relevant documents. In QA, MRR measures the average reciprocal rankings of the first correct answers to a set of questions. For example, given a question, if the rank of the first correct answer retrieved is N, then the reciprocal ranking is $1/N$. MRR measures the average performance across a set of questions in terms of their reciprocal rankings.

Given these metrics, we examined whether they could be applied to our problem of attention prediction. In the context of attention prediction, the *document retrieved* or *answer retrieved* should be replaced by *objects activated*. For example, the precision would become the percentage of the number of correctly identified activated objects (i.e., those objects are indeed the attended objects) out of the total number of activated objects. The reciprocal ranking measurement would become the reciprocal ranking of the first correctly activated object.

Since the result from the logistic regression model is the likelihood for an object to be activated, it is difficult to precisely determine the number of objects that are activated based on the likelihood. Presumably, we can set up a threshold and consider all the objects with the likelihood above that threshold activated. However, it will be difficult to determine such a threshold. Nevertheless, the likelihood of activation can lead to the ranking of the objects that are likely to be activated. Thus, the desired evaluation metric for this investigation should determine how well our model ranks the objects in terms of possible activation of those objects. For these reasons, we decide that the precision and recall metrics are not suitable for our problem and MRR seems more appropriate.

Even with the MRR measurement, there is still a problem. MRR in QA is concerned with the reciprocal ranking of the first correct answer; but, here in attention prediction, multiple objects could be simultaneously attended to (i.e., activated) in a given frame. We need to consider the reciprocal ranking for each of these objects. Therefore, we extended the traditional MRR to a *normalized* MRR (NMRR) which takes all attended objects into consideration. The normalized MRR is defined in Equation (5.4)

$$NMRR = \frac{MRR}{UpperBoundMRR} \tag{5.4}$$

where the upper bound MRR represents the MRR of the best possible ranking.

For example, suppose the logistic regression model ranks the likelihood of activation in a frame of four objects as shown in Table 5.1. Among these objects, only *lamp* and *bed lamp* are referenced in user speech within this frame. In this case,

$$NMRR = \frac{1/2 * (1/2 + 1/3)}{1/2 * (1 + 1/2)} = 0.556 \tag{5.5}$$

Here, the numerator represents the mean reciprocal ranking of the predicted acti-

Object		Class Label	Rank
ID #	Description		
12	dresser	-	1
3	floor lamp	+	2
8	bed lamp	+	3
19	bed	-	4

Table 5.1: Sample Test Data Frame with Four Ranked Instances

vations in this frame. The denominator represents the MRR of the best possible ranking, which in this case would rank lamp and bed lamp as the top two ranked objects.

With the evaluation metrics defined, proceed to the feature evaluation.

5.5 Feature Evaluation

5.5.1 Eye Gaze Features

Gaze fixation has been shown to be closely tied to user attention. Fixation intensity can be crudely defined as the length of a fixation upon an object in a visual scene. Generally, long fixations signify that a user is attending to this object. This increase in likelihood of attention increases the likelihood that this object will be referenced in the user's speech. Here, we describe four different representations of the fixation intensity measure. The reason to consider these variations is to evaluate their potentially different impact and reliability for prediction of user attention. We identified the following variations:

- Absolute Fixation Intensity (AFI): the proportion of time spent fixating on an object during a particular time window W. The time window considered in this study ranges from onset of a spoken reference to 1500 ms prior to this reference. Objects that are fixated for a long period of time are considered to be more likely to be activated than those fixated for a short period of time.

- Relative Fixation Intensity (RFI): the ratio between the AFI of a candidate object in time window W and the maximal AFI in W. An object may have a low AFI, but still have a high RFI.

- Weighted Absolute Fixation Intensity (WAFI): As discussed in §5.3, not all fixations are equally important in predicting user focus of attention. Presumably, fixations appearing closer to the reported mean (758 ms prior to onset of a spoken referent) are more indicative of the user's true focus of attention. The WAFI measure takes this factor into consideration by weighting the fixation durations based on a skew-normal distribution. The weighting is done based on Figure 5.1, which shows that this distribution seems to accurately describe the eye gaze data.

- Weighted Relative Fixation Intensity (WRFI): ratio between WAFI of an object and the maximum WAFI in a particular window W.

While fixation intensity is likely to be the most important factor, we hypothesize that other auxiliary features can also contribute to the eye gaze behavior and thus the prediction of attention. We discuss those features in §5.5.2.

5.5.2 Visual Features

Object Size

When users interact with a graphic display, the size of the object on the display can potentially affect user eye gaze behavior, and thus affect the prediction of attention. For example, it is difficult to fixate on small objects for long periods of time. People instinctively make small jerky eye movements. Large objects are unaffected by these movements because these movements are unlikely to escape the object boundary. Thus our hypothesis is that small objects with a lower fixation intensity are still likely to be the attended objects (i.e., be activated by the eye gaze). To take this effect into consideration, we use the object size to represent the area of a candidate object relative to a baseline object (e.g., the largest object in the scene). To obtain the value for this feature, each object is specified by a list of (x, y) scene coordinates. An object's area is represented by the bounding box considering only the minimal and maximal x and y coordinates. The object size feature is normalized by taking a ration of a candidate object to a baseline object (the largest object in the visual scene).

Visual Occlusion

In a graphical display, it is likely that objects overlap with one another. The visual occlusion of an object represents how much of this object is obstructed (by other objects) from the user's viewpoint. We hypothesize that when user eye gaze happens to simultaneously fixate on two overlapping objects, the user is likely to be more interested in the object appearing in front. Objects in the back are less likely to be attended. This aspect can be considered on either a fixation level or a time window level.

When considering visual occlusion on a fixation level, this feature can be used to clarify ambiguous fixations. Ambiguous fixations are those for which a single unique object cannot be determined. Fixations are partially disambiguated by removing all objects that are visually occluded by another object within the duration of this fixation. Unfortunately, this does not completely disambiguate fixations because a fixation may cover the area of two non-overlapping objects if these two objects are very close together.

When considering visual occlusion on a time window level—as a group of consecutive fixations in a particular time window—we calculate the amount of visual occlusion for each object during a particular time window. To take this aspect into consideration, we can use the following two variations of this feature:

- Absolute Visual Occlusion: the area of an object that is obstructed from vision by objects that were fixated during time window W.
- Relative Visual Occlusion: the percentage of an object that is obstructed from vision. This value is equivalent to the ratio between Absolute Visual Occlusion and Size for each object.

It is important to note the difference between representing visual occlusion on a fixation level vs. a time window level. First, two objects may be considered visually occluded during the entire time window, but not during any fixation in W. This occurs, when two overlapping objects are fixated during W, but never at the same time. Conversely, if an object is visually occluded during any fixation in W, its Absolute Visual Occlusion and Relative Visual Occlusion measures must have non-zero values. The other important difference is how this feature is used in our model. When considering visual occlusion on a time window level, this feature is one of the features used to create models for attention prediction. However, when considering visual occlusion on a fixation level, it is not directly used in model creation. Instead, it is used to preprocess the data, disambiguating indistinct fixations supplied to the logistic regression framework.

Fixation Frequency

According to Qvarfordt [78], fixation frequency may be an important feature to consider. This feature represents the number of times an object was fixated in W. For example; if a user looks at an object, then looks away from it, and then looks back; the fixation frequency for this object will be 2. According to Qvarfordt [78], when a user looks back and forth toward an object, it seems likely that the user is interested in this object.

5.6 Experimental Results

To evaluate the role of fixation intensity (i.e., Question 2) and auxiliary visual features (i.e., Question 3) in attention prediction, we conducted a five set cross-validation. More specifically, the collected data are randomly divided into five sets. Four of these sets are used for training, while the remaining set is used for testing. This procedure is repeated five times, with alternating test sets, and the averaged results are reported in the following sections. The object activation models were created using two data sets. The first is the original data set described in §5.2.2, while the second is a version of this same data set that is preprocessed to partially disambiguate fixations to multiple objects. Fixations in the preprocessed data set are disambiguated using the visual occlusion feature considered on a fixation level. That is, each fixation is considered to correspond to a unique object appearing on the interface, namely the frontmost object that is fixated.

5.6.1 Experiment 1: Eye Gaze Feature Evaluation

In this section we compare object activation models created by using each of the four variations of the fixation intensity measure. The goal here is to determine the effect of weighting fixations based on their distributions of starting times relative to a spoken reference versus treating every fixation equally. First, we discuss the methodology for creating weighted fixation intensity measures. Then we present results comparing the various object activation models and discuss their implications.

To create our two weighted fixation intensity measures (WAFI and WRFI) we use the statistics acquired about the distribution of fixations starts. More precisely, we weight each fixation by a skew-normal density function [1] with mean, standard deviation, and skewness discovered while addressing Question 1.

The results of each model constructed from its corresponding variation of the fix-

Fixation Intensity Variation	NMRR Evaluation	
	Original Data Set	Preprocessed Data Set (Disambiguated Fixations)
AFI	0.661	0.752
WAFI	0.656	0.745
RFI	0.652	0.754
WRFI	0.650	0.750

Table 5.2: Evaluation of Fixation Intensity Weighting using the NMRR Metric

ation intensity feature are are shown in Table 5.2. These results clearly indicate that there is very little variation among the different representations of fixation intensity across each of the two datasets. The first thing to note is that this lack of variation is to be expected between relative and absolute versions of the same fixation intensity measurement (AFI vs. RFI and WAFI vs. WRFI). This is because the evaluation conducted here is based on mean reciprocal ranking. Given two objects in a single time frame W, the one with the higher absolute fixation intensity is guaranteed to have a higher relative fixation intensity. Thus, the object ranking remains unchanged.

The effect of weighting fixation intensity seems to decrease performance on the object activation task. This decrease is virtually negligible. Nonetheless, this result is somewhat vexing as we expected that weighting the fixation intensity would improve prediction of object activation. One possible explanation for this lack of improvement is that fixations are fairly evenly distributed during each time frame W. This makes the weight function very flat and virtually insignificant. Another possibility is that the distribution of fixation starts has multiple peaks rather than a single peak at the mean as is the assumption of the normal distribution. Thus, neither the normal distribution nor the skew-normal distribution accurately models the distribution of eye fixation starts relative to spoken object references.

Row	Features	Original Data Set		Disambiguated Data Set	
		AFI	RFI	AFI	RFI
1	Fixation Intensity Alone	0.661	0.652	0.752	0.754
2	Fixation Intensity + Absolute Visual Occlusion	0.667	0.666	0.758	0.754
3	Fixation Intensity + Relative Visual Occlusion	0.667	0.669	0.762	0.751
4	Fixation Intensity + Size	0.656	0.657	0.763	0.759
5	Fixation Intensity + Fixation Frequency	0.653	0.644	0.743	0.756
6	All features (Absolute Visual Occlusion)	0.662	0.663	0.768	0.760
6	All features (Relative Visual Occlusion)	0.660	0.669	0.764	0.768

Table 5.3: Evaluation of Visual Features using the NMRR Metric

5.6.2 Experiment 2: Visual Feature Evaluation

In this section we evaluate the performance of auxiliary visual features on the object activation task. Configurations of various combinations of these features with fixation intensity are examined. Given that the effect of weighting fixation intensity is insignificant, only AFI and RFI are considered. The results are shown in Table 5.3 and discussed separately for each feature.

Visual Occlusion

As Table 5.3 shows, all configurations that use the preprocessed data set, which augments the fixation intensity measurement with a fixation-level account of visual occlusion, perform significantly better than their counterparts that use the original data set. The only difference between the preprocessed and original data sets is the incorporation of the fixation-level visual occlusion feature. This clearly means that visual occlusion is a reliable feature for the object activation prediction task. However, it is also clear that the representation of visual occlusion is very important. Adding the frame-based visual occlusion feature to the logistic regression model (rows 2 and 3) has almost no effect. It may be possible that a better representation for visual occlusion remains unexplored.

On average, the effect of both absolute and visual occlusion is more significant for the original data set (especially when RFI is used). This is not surprising because the preprocessed data set partially incorporates the visual occlusion feature, so the logistic regression model does not get an added bonus for using this feature twice.

Object Size

The object size feature seems to be a weak predictor of object activation. Using only the fixation intensity and object size features (row 4 of Table 5.3), the logistic regression model tends to achieve approximately the same performance as when the object size feature is excluded.

This result is quite unexpected. As we have already mentioned, human eye gaze is very jittery. Our expectation is that small objects can have a low fixation intensity and still be activated. Thus, in our model small objects should need a lower fixation intensity to be considered as activated than do large objects. Our results do not support this general trend. A possible explanation is that this trend should only be apparent when using a visual interface with a mixture of large and small objects. In our interface, most of the objects are fairly large. For large object, jittery eye movements do not alter fixation intensity because the eye jitter does not cause fixations to occur outside of an object's interest area boundary. Even when some objects are smaller than others, they are not sufficiently small to be affected by eye jitter. Thus, it is likely that the size feature should only be considered when comparing fixation intensities of sufficiently small objects to larger counterparts. At this point, it is unclear how small is sufficiently small.

Fixation Frequency

Fixation frequency is a weak predictor of object activation. Incorporating the fixation frequency feature into the Bayesian Logistic Regression framework creates models that tend to achieve worse performance than when this feature is left out. At best, models using fixation frequency achieve a comparable performance to those not using it. According to Qvarfordt [78], fixation frequency is an important feature to consider because one way of signifying interest in objects is looking back and forth between two or more objects. In this case, each of these objects would have a fairly low fixation intensity as time is spent across multiple objects, but each of the objects should be considered activated. In our user studies, however, we did not find this user behavior and our results are consistent with this observation. This behavior is likely to be specific to the map-based route planning domain where users often need to look back and forth between their starting and destination location.

5.7 Discussion

We have shown that fixations that are pertinent in the object activation problem are fairly evenly distributed between the onset of a spoken reference and 1500 ms prior. Fixation intensity can be used to predict object activation. Weighting fixations based on a skew-normal distribution does not improve performance on the object activation task. However, preprocessing our fixation data by including the fixation-level visual occlusion feature considerably improves reliability of the fixation intensity feature. Moreover, since performance is so sensitive to feature representation, there is much potential for improvement. We have also presented the NMRR evaluation metric that can be used to evaluate the quality of a ranked list.

This work can be extended to combine our activation model with spoken language processing to improve interpretation. This question can be addressed by constructing an N-best list of spoken input with an Speech Recognizer (ASR). The speech-based ranked lists of utterances and the gaze-base ranked lists of activations can be used to mutually disambiguate [68] each other in order to more accurately determine the object(s) of interest given an utterance and a graphical display. This knowledge can be used to plan dialogue moves (e.g. detect topic shifts, detect low-confidence interpretations, determine the need for confirmation and clarification sub-dialogs, etc.) as well as to perform multimodal reference resolution [11]. We believe that this work will open new directions for using eye gaze in spoken language understanding.

5.8 Summary

In this chapter, we present an eye gaze based approach for modelling user focus of attention. We use the collected data that is described in Chapter 4 along with the Bayesian Logistic Regression Toolkit to construct these models. These models incorporate various eye gaze and auxiliary visual features. Our feature evaluation, using the novel Normalized Mean Reciprocal Rank metric, indicates that fixation intensity–the amount of time that an object is fixated during a given time window–is by far the most important factor in predicting a user's focus of attention. The auxiliary visual features do not produce a significant improvement in attention prediction accuracy. Neither does weighting the fixation intensity feature based on the temporal locations of eye fixations relative to the target spoken utterance. These results have important implications for language understanding in visual and spatial environments. The ability to reliably predict user focus of attention will allow for more effective reference resolution. This will, in turn, improve language understanding and overall usability of multimodal conversational interfaces.

Chapter 6

Eye Gaze in Reference Resolution within a Static Domain

6.1 Introduction

Our hypothesis is that exploiting user eye gaze can help alleviate language processing problems in conversational interfaces. There has been a significant body of work in the field of psycholinguistics studying the link between eye gaze and speech [28, 30] [41, 85]. Eye gaze has been shown to be a window to the mind. That is, the eye fixates on symbols that are being cognitively processed. The direction of gaze carries information about the focus of a person's attention [41].

Eye-tracking technology is being incorporated into speech-driven computer interfaces more and more frequently [43, 77, 78]. However, it remains unclear exactly to what degree eye gaze is capable of helping automated language processing. In this work, we investigate what information can be obtained from eye gaze as it relates to reference resolution in multimodal conversational interfaces. Reference resolution is the process of identifying application specific entities that are referred to by linguistic expressions. For example, identifying a picture on the interface that is referenced by the expression “the painting of a waterfall”.

When a referring expression is uttered by a user, according to Conversation Implicature [26], users will not intentionally make ambiguous expressions to confuse the system. The reason that the system fails to correctly resolve some references is because the system lacks the adequate domain models or linguistic processing capability that a human may have. More referring expressions can be resolved with more sophisticated domain modeling and corresponding linguistic processing. However, sufficiently representing domain knowledge and processing human language has been a very difficult problem. Moreover, some domain information is very difficult to encode during design time. This is especially true for dynamic information that can change as the user manipulates the interface. For example, an object—presented via the user interface—that is perceived to be upside-down can be rotated by the user to no longer have this property. The rotation operation is likely to change the language used to describe this object. While encoding the orientation information of each object is relatively simple, encoding this potential change in language is difficult. This domain information is often necessary to disambiguate spoken references. Thus our investigation considers whether the use of eye gaze can compensate for some of these limitations. We hypothesize that the use of eye gaze has the capacity to eliminate the need to encode some complex domain information without sacrificing reference resolution performance.

To investigate the role of eye gaze in reference resolution in multimodal conversational interfaces we proceed with two primary research objectives. First, we construct a general probabilistic framework that combines speech and eye gaze information for robust reference resolution. This framework is capable of independently encoding linguistic, dialogue, domain, and eye gaze information and combining these various information sources to resolve references to objects on the interface. Second, we attempt to validate our hypothesis that utilizing eye gaze in reference resolution can compensate for a lack of sophisticated domain modeling. Specifically, we compare the amount of information provided by utilizing eye gaze relative to that provided by increasing levels of encoded domain complexity (along with linguistic processing capable of handling this new domain information). To pursue these objectives, we use transcribed speech rather than automatically recognized speech data as a first step in our investigation. This allows us to focus on studying the relationship between eye gaze and linguistic processing without additional confounding variables. The findings from this initial investigation will help us better understand the role of eye gaze in spoken language understanding and establish an upper bound performance for our next step of processing recognized speech.

In the following sections we discuss our investigation in more detail. We describe our integrated reference resolution framework in §6.2. We proceed to discuss the experiments we conducted using the integrated framework to process our collected data in §6.3. We conclude with some implications of this investigation and provide some potential research directions for the future 6.4.

6.2 Multimodal Reference Resolution

Reference resolution in a Multimodal Conversational Interface is the process of identifying the intended objects given a user's spoken utterance. Figure 6.1 shows a sample dialogue in our conversational interface with the referring expressions to concrete objects marked in bold. Each S_i represents a system utterance and each U_i represents a user utterance. In this scenario, referring expressions tend to be definite noun phrases, such as "the bed", or pronouns, such as "it" or "that". Note that some pronouns may have no antecedent (e.g. "it" in utterance U_2) or an abstract antecedent [5], such as "democracy". Some expressions, (i.e. *evoking* references) initially refer to a new concept or object. Conversely, *anaphoric* references subsequently

S_1	What is your favorite piece of furniture?
U_1	I would say my favorite piece of furniture in the room would be **the bed**
U_2	It looks like **the sheet** is made of leopard skin
U_3	And I like **the cabinets** around **the bed**
U_4	Yeah I think **that's** is my favorite
S_2	What is your least favorite piece of furniture?
U_5	**The brown chair** toward the front with **the candlestick** on **it**
U_6	**It** seems pretty boring

Figure 6.1: Conversational Fragment with Different Types of Referring Expressions

refer back to these objects. Anaphoric referring expressions are the most common type of *endophoric* referring expressions—referring to an entity in the dialogue. Expressions that are not endophoric, must be *exophoric*—referring to an entity in the extralinguistic environment. Evoking references are typically exophoric.

A robust reference resolution algorithm should be capable of handling referring expressions that belong to a multitude of grammatical categories, as described above. We focus on resolving the most common referring expressions: definite noun phrases, which can be either exophoric or endophoric, and anaphoric pronouns whose antecedents are entities in the extralinguistic environment. These referring expressions compose a grave majority of the references found in our collected dialogs.

We have constructed an integrated probabilistic framework for fusing speech and eye gaze for the purpose of conducting reference resolution. An *integrated* framework achieves two important objectives: (1) a single framework is capable of combining linguistic, dialogue, domain, and eye gaze information to resolve references and (2) a single framework is capable of resolving both exophoric and endophoric references. In fact, the system does not need to know whether an expression is exophoric or endophoric to reliably resolve it. This framework contains parameters that can be determined empirically or through machine learning approaches. The following section describes the framework.

6.2.1 Integrated Framework

Given an utterance with n referring expressions $r_1, r_2, \ldots, r_n$, the reference resolution problem can be defined as the process of selecting a sequence of sets of objects $O_1, O_2, \ldots, O_n$ that best matches the referring expressions in the given utterance. Directly determining the best match is very difficult and potentially intractable.

To make this problem tractable we simplify it in two ways. First, we make the assumption that the occurrence of one referring expression is independent of another. Thus, the matching between O_j and r_j can be done individually for each expression r_j. Next, we construct each object set O_j using a two phase process. In the first phase, the likelihood $P(o_i|r_j)$ that r_j refers to object o_i is determined for each object in O, which is the set of all possible objects in the domain. In the second phase, O_j is constructed by combining the top k ranked o_i's, where k depends on whether referring expression r_j is singular or plural.

Previous studies have shown that user referring behavior during multimodal conversation does not occur randomly, but rather follows certain linguistic and cognitive principles [13, 44]. Different referring expressions signal the cognitive status of the intended objects. In our investigation, we model the user's cognitive status by considering two sources of potential referents: Visual History (VH) and Dialogue History (DH). Each source reflects a different cognitive status, which is often associated with different types of referring expressions. Visual History represents the visible display information available on the time scale of the currently processed spoken utterance. It is usually associated with exophoric references. Dialogue History represents the stack of previously resolved spoken utterances in the current conversation. It is often associated with anaphoric references.

We can use Equation 6.1 to determine the probability of a particular object being referenced given a particular referring expression. Here, we introduce a hidden variable T, which represents the cognitive status model of the intended object (either VH or DH):

$$\begin{aligned} P(o_i|r_j) &= \sum_{T \in \{VH, DH\}} P(o_i|T, r_j)P(T|r_j) \\ &= P(o_i|VH, r_j)P(VH|r_j) + \\ &\quad P(o_i|DH, r_j)P(DH|r_j) \end{aligned} \tag{6.1}$$

Equation 6.1 first evaluates the likelihood of each cognitive status associated with the given referring expression. Then, given the cognitive status and the referring expression, it determines the probability that a particular object is referenced. Linguistic and domain information can be extracted for each referring expression r_j while dialogue and eye gaze information is encoded into the DH and VH, respectively. Algorithm 1 demonstrates the procedure to resolve a given referring expression r_j using Equation 6.1. This algorithm allows us to achieve the objectives of integrating linguistic, dialogue, domain, and eye gaze information into a single probabilistic framework that is capable of resolving both exophoric and endophoric references. In the following sections we show the specifics of how this information is integrated into the framework.

6.2.2 Likelihood of Cognitive Status

In Equation 6.1, $P(VH|r_j)$ and $P(DH|r_j)$ represent the likelihood of the cognitive status given a referring expression. These quantities are determined based on the grammatical category of the referring expression (pronoun, definite noun, etc.). For example, if r_j is a definite noun phrase, it is likely referring to an element from the visual history. This is because definite noun phrases are often made as evoking refer-

Algorithm 1 Resolve Referring Expression

```
Input: $r_j$ {referring expression to be resolved}
Input: $\overrightarrow{O}$ {set of potential referent objects}
Input: $\overrightarrow{F}$ {list of eye fixations}
$VH \leftarrow$ constructVH($\overrightarrow{F}$)
for all $o_i \in \overrightarrow{O}$ do
    compute $P(o_i|VH, r_j)$
end for
for all $o_i \in \overrightarrow{O}$ do
    compute $P(o_i|DH, r_j)$
    compute $P(o_i|r_j) \leftarrow P(o_i|VH, r_j)P(VH|r_j)$
                               $+P(o_i|DH, r_j)P(DH|r_j)$
end for
sort $\overrightarrow{O}$ in descending order according to $P(o_i|r_j)$
if numberStatusKnown(r) then
    $m \leftarrow$ number$(r_j)$
else
    find the smallest $i$, such that $P(o_i|r_j) > P(o_{i+1}|r_j)$
    $m \leftarrow i$
end if
$\overrightarrow{o} = \emptyset$
for $i = 0$ to $m$ do
    $\overrightarrow{o} \leftarrow \overrightarrow{o} \cup o_i$
end for
return $\overrightarrow{o}$
```

ences, mentioning objects that do not appear in the dialogue history. Those definite noun phrase references that are subsequent references to already mentioned objects are unlikely to have recently occurred in the dialogue history because the more recent mentions tend to be pronouns. Thus, in this situation $P(VH|r_j)$ will be high and $P(DH|r_j)$ will be relatively low.

In the case expression r_j is a pronoun, we consider the VH and DH equally important in resolving pronominal referring expressions. This is because we expect the intended object to be recently present in the VH as well as the DH. Given that in our domain referent objects rarely go out of view, users are likely to gaze at these objects, thereby putting them in the recent visual history. Thus, we set

$$P(VH|r_j) = P(DH|r_j) = 0.5$$

If r_j is not a pronoun, it is typically a definite noun phrase given our domain. In this case the DH is unlikely to contain useful information. Thus we set

$$\begin{aligned} P(VH|r_j) &= 1.0 \\ P(DH|r_j) &= 0.0 \end{aligned}$$

This means that these phrases take into account only the currently processed utterance.

Although we empirically set these parameters, it is possible to directly learn them from the data. It is important to note that small deviations in these probabilities have a minuscule effect on our reference resolution results.

6.2.3 Likelihood of Objects

Given the cognitive status, the likelihood that an object is referred to can be represented by the following equation:

$$P(o_i|T, r_j) = \frac{AS(o_i, T)^{\alpha(T)} \times Compat(o_i, r_j)^{1-\alpha(T)}}{\sum_i AS(o_i, T)^{\alpha(T)} \times Compat(o_i, r_j)^{1-\alpha(T)}} \tag{6.2}$$

where

- AS: Attentional salience score of a particular object. This score estimates the salience of object o_i given a cognitive status. For example, if the modeled cognitive status is VH, the salience of the object can be determined based on eye fixations that occur at the time the user utters referring expression r_j along with a combination of the object's visual properties. If the modeled cognitive status is DH, the score is based on how recently in the dialogue o_i has been mentioned.

- $Compat$: Semantic compatibility score. This score represents to what degree that a particular object is semantically compatible with the referring expression. Factors such as the object's semantic type, color, orientation, and spatial location should be considered. Here, linguistic and domain information are brought together. For example, every color that is contained in the domain model must be contained in the system's lexicon and must be identified as a color when spoken by a user.

- $\alpha(T)$: Importance weight of Attentional Salience relative to Compatibility. This weight depends on whether the cognitive status T is modeled as the visual history or dialogue history.

The details of these functions are explained next.

Attentional Salience with Cognitive Status Modeled as Visual History When the cognitive status is modeled as $T = VH$, $AS(o_i, VH)$ represents an object's visual salience. There are many different ways to model the attentional salience of objects on the graphic display. For example, visual interface features, such as object centrality and size, can be extracted using image processing [46]. In continuously changing domains temporal features such as prominence and recency of an object appearing on the visual interface [7] can be used.

Instead of using visual features, we compute the attentional salience score with the use of eye gaze. One method of doing this is by using our previously developed attention prediction algorithm [74]. In this work the attentional salience score of an object is represented by the Relative Fixation Intensity of this object. That is,

$$AS(o_i, VH) = RFI(o_i) \in [0..1]$$

We have previously shown that the attentional salience of an object can be reliably modeled by the Relative Fixation Intensity (RFI) feature obtained from collected eye gaze data. Fixation Intensity represents the amount of time that an object is fixated in a given a time window W. The RFI of an object is defined as the ratio between the amount of time spent fixating a candidate object during W and the amount of time spent fixating the maximally long fixated object during W. The intuition here is that objects that are fixated for a long period of time are considered to be more salient than those fixated for a short period of time. However, long period of time should be defined relative to other fixated objects during the same time period. A user may look at an object for a short period of time relative to W, but if this is the only object fixated during this time, it is likely to be salient. In our previous study [74], a good time range for W was empirically found be [-1500..0] ms relative to the onset of a spoken reference. This range conforms to prior psycholinguistic studies that show that on average eye fixations occur 932 ms (with sizable variance) prior to the onset of a referring expreession [27, 28].

We chose to use the Attentional Salience model as defined here for three reasons. First, we believe that models based on eye-tracking data are more reliable than models based on image processing. Second, the model based on the RFI feature is the simplest eye gaze based model. Although, auxiliary gaze-based features (e.g. fixation frequency) and visual interface-based features (e.g. object occlusion and object size) can potentially improve the AS model, [74] shows that consideration of these features provide only a marginal improvement in the reliability of the attention prediction model given the domain described in the Data Collection section. Finally, the RFI-based model can generalize to any visual domain and makes for a good baseline.

Attentional Salience with Cognitive Status Modeled as Dialogue History When cognitive status is modeled as $T = DH$, an object's salience within the dialogue is represented by
$AS(o_i, DH)$. In our investigation, this score indicates how recently (on a dialogue history time scale) a particular object has been referred to. An object's salience within a dialogue decays exponentially over time as shown in Equation 6.3

$$AS(o_i, DH) = \frac{(1/2)^{n_i}}{Z} \in [0..1] \tag{6.3}$$

where, n_i is the number of referring expressions that have been uttered since the last time object o_i has been referred to and Z is the normalization factor $\sum_i (1/2)^{n_i}$.

The dialogue history information is represented as a table of such salience scores. Initially each object's score is 0. When an expression referring to an object is uttered, its score is set to 1. At the same time, the scores of each other object are divided by 2.

Referring Expression		Object				
		o_{18}	o_{19}	o_{21}	o_{24}	$\cdots$
U_1	the bed	0	1	0	0	0
U_2	the sheet	0	1/2	1	0	0
U_3	the cabinets	1	1/4	1/2	1	0
U_3	the bed	1/2	1	1/4	1/2	0
U_4	that's	1/4	1	1/8	1/4	0

Table 6.1: Sample Dialogue History Table

For example, the first four utterances of the dialogue in Figure 6.1 would result in the scores shown in Table 6.1. Here, the first two columns represent a time step during which a referring expression is uttered. The remaining columns represent all of the objects that could possibly be mentioned. The object IDs correspond to those shown in Figure 4.1. Each cell represents the corresponding object's salience score within the dialogue. In this example, after U_2 is uttered and its only referring expression is resolved to object o_{21}, this object's score is set to 1. The salience scores of all other objects are divided by 2. Thus, when U_3 is uttered,

$$AS(o_{21}, DH) = \frac{1}{1 + 1/2} = 2/3$$

making o_{21} the most salient object. Utterance U_3 contains two referring expression which are processed in the order of being spoken. Given that the referring expression "the cabinets" is correctly resolved to the two cabinets o_{18} and o_{24}, the salience scores of each of these two objects are set to 1 and once again all other scores are discounted.

Semantic Compatibility The semantic compatibility score $Compat(o_i, r_j)$ represents to what degree object o_i is semantically compatible with referring expression r_j. The compatibility score is defined in a similar manner as in our previous work [13, 11]:

$$Compat(o_i, r_j) = Sem(o_i, r_j) \times \prod_k Attr_k(o_i, r_r) \tag{6.4}$$

In this equation:

- $Sem(o_i, r_j)$ captures the coarse semantic type compatibility between o_i and r_j. It indicates that the semantic type of a potential referent should correspond to the semantic type of the expression used to refer to it. Consequently, in our investigation, $Sem(o_i, r_j) = 0$ if the semantic types of o_i and r_j are different and $Sem(o_i, r_j) = 1$ if they are the same or either one is unknown.

- $Attr_k(o_i, r_r)$ captures the object-specific attribute compatibility (indicated by the subscript k) between o_i and r_j. It indicates that the expected features of a potential referent should correspond to the features associated with the expression used to refer to it. For example, in the referring expression "the brown chair", the color feature is *brown* and therefore, an object can only be a possible referent if the color of that object is *brown*. Thus, we define $Attr_k(o_i, r_r) = 0$ if both o_i and r_j have the feature k and the values of this feature are not equal. Otherwise, $Attr_k(o_i, r_r) = 1$.

Attentional Salience vs. Semantic Compatibility In Equation 6.2, $\alpha(T)$ represents the importance tradeoff between attentional salience and semantic compatibility. A high value of $\alpha(T)$ indicates that the AS score is more important for reference resolution than the semantic compatibility score. A low value indicates that the opposite is true. For different kinds of interfaces $\alpha(T)$ may change. For example, if an interface is composed of many very small objects or many overlapping objects, eye gaze becomes less reliable. In this case, $\alpha(T)$ should be fairly low. In our investigation, we wanted both quantities to have equal importance. However, semantic compatibility should win out in case of a tie. A tie can occur if all elements in T are semantically incompatible with referring expression r_j and no compatible elements are salient. Thus, we set $\alpha(T) = 0.49$ for both VH and DH. If we do not want to integrate eye gaze in reference resolution, we can set $\alpha(T) = 0.0$. In this case, reference resolution will be purely based on compatibility between visual objects and information specified via linguistic expressions.

6.2.4 Selection of Referents

Once the likelihoods for each candidate object are computed according to Equation 6.1 we can proceed to selecting the most probable referenced-object set O_j given a referring expression r_j. First, objects are ranked according to the probabilistic framework. Next, the number of objects being referenced is considered. Each reference can be singular or plural. If it is plural, the exact number of objects being referenced may be known (e.g. "these two pictures") or unknown (e.g. "these pictures"). If r_j is singular, the highest ranked object is selected. If more than one object has the highest score, there is an ambiguity and reference resolution fails. If r_j is plural and the number of referenced objects is known to be k, then the top k objects are selected. If r_j is plural, but the number of referenced objects is unknown, then all of the objects that have the top score are selected.

6.3 Experimental Results

As we have already mentioned, a major goal of this work is to investigate whether or not the use of eye-tracking data can compensate for an incomplete domain model without sacrificing reference resolution performance. We categorize four different levels of domain models. These models vary in complexity such that a complex model completely entails a simpler model as shown in Figure 6.2. In addition to these four levels, we consider an *empty* domain model that contains no domain information or linguistic processing.

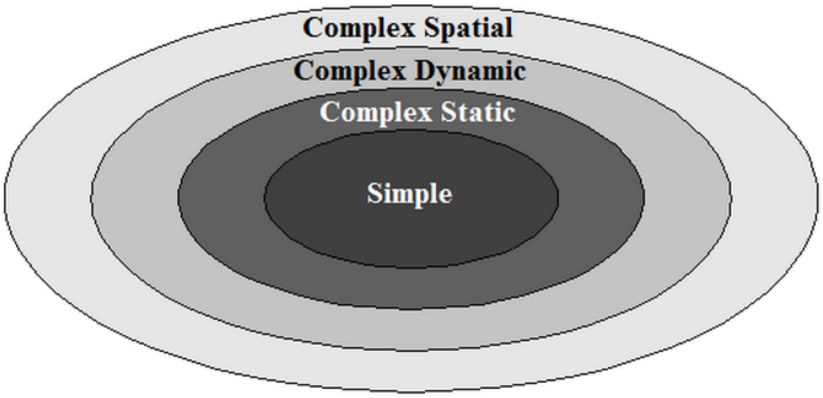


Figure 6.2: Domain Hierarchy

- Empty (∅): This model makes no domain or linguistic information available. When it is employed, referents are determined purely based on eye gaze.

- Simple (*Simp*): Domain and linguistic information is limited to a coarsely defined semantic type of each object (e.g. *lamp*, *chair*, *painting*, etc.).

- Complex Static (*CStat*): Various object attributes are encoded. These attributes satisfy the requirement of being inherent to a particular object. More precisely, these attributes cannot be changed through interaction with the interface (e.g. *floor* lamp, painting of a *waterfall*).

- Complex Dynamic (*CDyn*): Object attributes that can potentially change are encoded (e.g. *green* chair, *crooked* lamp). For example, a user may say "paint the chair red" when talking about the green chair. After this operation, the chair is no longer green.

- Complex Spatial (*CSpat*): Spatial relationships between objects are encoded (e.g. the chair *next to* the bed, the one *above* the waterfall). Given that a user can move around in the interface, spatial relationships are clearly dynamic attributes.

The domain model variation is incorporated into our probabilistic framework via the semantic compatibility function shown in Equation 6.4. The Simple domain encodes information necessary to make coarse semantic type compatibility judgments $Sem(o_i, r_j)$ while each of the more complex domains add more attributes to be considered by $Attr_k(o_i, r_r)$. It is important to note that an increasing level of linguistic processing capacity is associated with an increasing level of domain model complexity. For example, if the domain model contains the notion of a painting of a waterfall (o_2 in Figure 4.1), the token *waterfall* must be present in the system's lexicon and the system must be able to determine that this token is an attribute of object o_2.

The *Simp* domain model has the advantage of being completely independent of the specific tasks that can be performed by the conversational interface. That is, this domain model can be constructed by simply knowing the semantic type of the objects present in the interface. The more complex domain models attempt to encode information that uniquely distinguishes objects from one another. This require knowledge about the similarity and difference between these objects, the kinds of references people tend to make to various objects, as well as information about the visual layout of these objects. This information is highly dependent on the design of the interface as well as the task being performed.

Here, we aim to compare the reference resolution performance of the speech & eye gaze algorithm to a speech-only algorithm. In this comparison the inclusion or exclusion of eye gaze is the only variable. As we have shown, eye gaze can be encoded into our reference resolution framework as $AS(o_i, VH)$. Eye gaze information can be removed by changing the $AS(o_i, VH)$ score to be uniformly distributed over all objects o_j. Next, we proceed to analyze the impact of eye gaze on reference resolution, we use the 371 collected multimodal inputs described in the Data Collection section. Note that in our current work, most of the parameters were empirically determined. The results presented here should not be viewed as a formal evaluation of the approach, but rather an analysis of the role of eye gaze in compensating for insufficient domain modeling. We are in the process of collecting more data and will do a formal evaluation in the near future.

Figure 6.3 shows the performance comparison of the reference resolution algorithm that incorporates eye gaze data vs. the speech-only algorithm. The comparison is performed using each of the aforementioned domain complexities in terms of reference resolution accuracy. This chart shows that the speech & eye gaze algorithm performs significantly better than the speech-only algorithm for each of the domain variations. The best performance of approximately 73% is reached when eye gaze is combined with the most complex level of domain (*CSpat*).

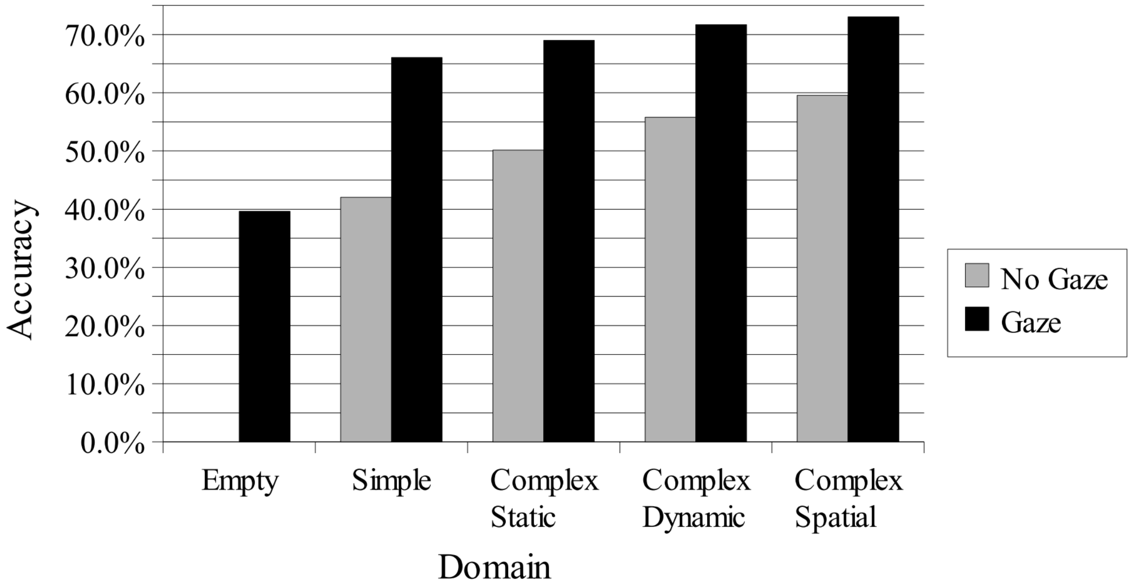


Figure 6.3: Overall Reference Resolution Accuracy

Table 6.2 shows a more detailed picture of the effect of eye gaze in multimodal reference resolution for each level of domain complexity. References are divided into pronouns and definite noun phrases. In our data set, all referring expressions to concrete objects are comprised of these two grammatical categories. Here, we investigate whether the effect of eye gaze varies depending on the grammatical category of the referring expression. Of the 371 total inputs (Table 6.2a), 84 (22.64%) are pronominal references (Table 6.2b) and 287 (77.36%) are definite nominal references (Table 6.2c). Each category of expression exhibits a significant[1] improvement ($p < 0.05$) in reference resolution accuracy when eye gaze is used. Additionally, they both exhibit diminishing returns with increasing domain complexity. That is, the improvement gained by using eye gaze is larger when the domain is simpler. This effect appears to be slightly more evident in the case of pronominal noun phrases, but there is insufficient data to make this conclusion.

Interestingly, the addition of eye gaze information to the simple domain has a larger affect on reference resolution performance than the addition of domain information. In Table 6.2a, compare the top row (*Simp*) with gaze to the bottom row (CSpat) without gaze. The former outperforms the latter 245 to 221, for a 57.05% vs. 41.61% increase over the baseline case (*Simp* without gaze). This result implies that gaze-data can potentially provide information toward the reference resolution task that isn't provided by a fairly complex domain model.

In addition to considering the orthogonal information provided by eye gaze relative to domain modeling, we want to consider the amount of common information. Our objective is to determine the coverage of eye gaze relative to domain model complexity as it pertains to the reference resolution task. Here, *coverage* refers to the percentage of references that require a complex domain definition to be resolved by the speech-only algorithm that can instead be resolved by resorting to eye gaze with the *Simp* domain model. Table 6.3 displays this comparison. Here references are subdivided by according to the minimal domain information necessary to resolve the particular referent. For example, the second row of Table 6.3a shows the number of references that cannot be resolved by the *Simp* domain, but can be resolved by adding static

[1] The significance tests were conducted using a paired t-test comparing each user's reference resolution accuracy with and without the use of eye gaze information.

	Domain Type	Without Gaze	With Gaze	t-value	P-value	Improvement
(a)	∅	0	147	—	—	—
Total	*Simp*	156	245	11.87	0.0001	57.05%
	CStat	186	256	11.80	0.0001	37.63%
	CDym	207	266	10.81	0.0001	28.50%
	CSpat	221	271	5.48	0.0028	22.62%
(b)	∅	0	30	—	—	—
Pronoun	*Simp*	21	43	3.99	0.0104	104.76%
	CStat	26	44	3.22	0.0234	69.23%
	CDym	33	47	3.80	0.0127	42.42%
	CSpat	33	47	3.80	0.0127	42.42%
(c)	∅	0	117	—	—	—
Definite	*Simp*	135	202	11.39	0.0001	49.63%
	CStat	160	212	10.80	0.0001	32.50%
	CDym	174	219	9.82	0.0002	26.86%
	CSpat	188	224	3.99	0.0105	19.15%

Table 6.2: Reference Resolution Improvement Due to Eye Gaze

attribute information to the domain model.

Several implications can be made from these results. The first is that eye gaze can be used to resolve a significant number of references that require a complex domain when eye gaze is not used. Eye gaze is most beneficial when little domain information is available. Additionally, eye gaze does not completely eliminate the need for domain modeling. As shown in table 6.2, eye gaze alone can be used to resolve only about 40% (147 out of 371) references. Rather some of the errors that are caused by having an insufficient domain model are alleviated when eye gaze is used. The second thing to note is that the *Simp* domain with eye gaze information can resolve almost all (98.72%) of the references that can be resolved with the same domain, but without eye gaze information. While this result is not very surprising, one may have expected that the coverage would be 100% in this case, but it is apparent that eye gaze can introduce a small amount of noise into the reference resolution process. The Attention Prediction model is not perfect and eye gaze is well known to be noisy.

	Domain Type	Resolved References	Coverage by *Simp* Domain with Eye Gaze	
(a) Total	*Simp*	156	154	98.72%
	CStat–Simp	30	19	63.33%
	CDyn–CStat	21	9	42.86%
	CSpat–CDyn	14	6	42.86%
(b) Pronoun	*Simp*	21	20	95.24%
	CStat–Simp	5	5	100.00%
	CDyn–CStat	7	2	28.57%
	CSpat–CDyn	0	0	—
(c) Definite	*Simp*	135	134	99.26%
	CStat–Simp	25	14	56.00%
	CDyn–CStat	14	7	50.00%
	CSpat–CDyn	14	6	42.86%

Table 6.3: Reference Resolution Coverage via Eye Gaze

6.4 Discussion

Given that the best performing reference resolution algorithm achieves about 73% accuracy, it is important to consider the sources that cause the errors in the remaining 27% of referring expressions. Typically, errors occur when neither domain information nor gaze is capable of the resolving a referring expression alone. Several factors can cause errors in each of the modalities.

First, the current framework is only capable of handling anaphoric pronouns (those that refer back to an entity in the dialogue), but not cataphoric pronouns (those that refer to an entity in the forthcoming speech). Additionally, the language processing component has some limitations. For example, the phrase "the chair to the left of the bed" can be interpreted, but "the bed with a chair to the left" cannot. This example demonstrates the difficulty of linguistic processing and provides more motivation for using eye gaze to compensate for such deficiencies.

Second, as has been noted, eye gaze data is very noisy and eye gaze is far from a perfect predictor of user focus of attention. Some of this noise comes from inaccurate eye tracker readings. Additionally, errors in visual attention prediction can arise from an insufficiently accurate temporal mapping between speech and eye gaze. For example, a user may look at an object and then wait a significant amount of time before referring to it. Thus, this object's salience will be far lower than intended by the user. Alternatively, a user's overt attention (eye gaze) may be disassociated from the user's covert attention. This may happen because the user has become very familiar with the visual interface and no longer needs to fixate an object to be able to refer to it. Further investigation is necessary to pinpoint exactly what causes these eye gaze errors and how much reference resolution suffers because of them.

Our current investigation has focused on processing transcribed linguistic expressions. When real-time speech is used, an important issue that needs further investigation is the effect of eye gaze when automatic speech recognition is considered. Speech recognition errors cause even more difficulty for linguistic processing. For example, even if there exists a sophisticated spatial model of the objects appear on the interface, a failure in linguistic processing can cause these domain models to become useless. Eye gaze is likely to compensate for some of these failures in the same way as it compensates for insufficient domain modeling.

6.5 Summary

In this chapter, we present an integrated approach that incorporates eye gaze for reference resolution. The empirical results show that the use of eye gaze improves interpretation performance and can compensate for some problems caused by the lack of domain modeling. These results have further implications for constructing conversational interfaces. The improvements in interpretation of object references gained via incorporating eye gaze will allow systems to make fewer unexpected and erroneous responses. Additionally, reducing the reliance on complex domain modeling will make multimodal conversational interfaces easier to design.

Chapter 7

Eye Gaze in Reference Resolution within an Immersive Domain

7.1 Introduction

In our quest to improve spoken language understanding, we have been studying how eye gaze can be utilized for reference resolution in multimodal conversational systems. As a first step in this investigation, we have developed salience models that can be used to predict a user's focus of visual attention (as discussed in Chpater 5). These models incorporate features derived from the user's eye gaze along with features derived from the visual environment. We have shown that *fixation intensity*—the amount of time a user gazes on a visual object during a given time window—is the key feature necessary to model a user's visual focus of attention.

In the next part of our investigation, we developed an integrated framework that is capable of resolving spoken references to objects that appear on the interface. This framework seamlessly combines linguistic, dialogue, domain, and eye gaze information. Using these models, we are capable of studying which factors have a significant impact on robust reference resolution in a multimodal conversational interface. We have shown that using this gaze-based attentional salience model significantly improves reference resolution performance. Moreover, we have shown that using eye gaze outperforms using complex domain modeling techniques.

However, these investigations were 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. It is also unclear how effectively our integrated reference resolution framework can accommodate automatic speech recognition (ASR) input and whether this can be done in real-time. In this chapter we provide analysis of our investigation to address these questions.

The treasure domain, presented in §4.3, provides an immersive environment to study situated language. The study of situated language "focuses primarily on investigations of interactive conversation using natural tasks, typically in settings with real-world referents and well-defined behavioral goals [84]." Situated language is spoken from a particular point of view within a physical or simulated context [6]. An interface that supports situated language has the following requirements: a spatially (visually) rich environment [24] that allows for immersion and mobility [7] and a purposeful dialogue that encourages social and collaborative language use [24]. Collaborative dialogue can have properties that are distinct from other types of dialogue. For example, Piwek et. al. [72] found that the cognitive modeling of one's partner in a collaborative dialogue in Dutch can produce dialogue patterns that are contrary to the Givenness Hierarchy [29]. In this setting, users were pairs of partners and the speaker directs his or her partner to construct a small object. Speakers in this setting tended to use distal expressions ("this/these") for referents already known to the hearer, and proximal expressions ("that/those") for new items that the hearer was being instructed to locate. We believe that an immersive, situated scenario will allow us to extend our study of eye gaze in a multimodal conversational interface.

It is important to consider such a scenario when studying spoken language understanding. One reason is that many conversational utterances can only be understood with respect to the context of the language use. This context includes a shared visual space along with shared conversational goals which exhibit collaborative processes intrinsic to conversation. Many conversational characteristics emerge only when both conversational parties have common goals and each party participates as both a speaker and listener [84]. Another reason for considering this scenario is that it more accurately represents real-world conversation. Immersion and mobility in a visual environment are necessary for users to produce realistic spatial references, such as “the book to the right” or “the plant near the bed”. These qualities are also necessary for the production of references to objects that were at one time, but are no longer, visible on the user’s display. The static multimodal conversational interface is insufficient for modeling natural real-world conversation as it does not support immersion and mobility, nor does it allow for collaborative language use.

Using an immersive, situated scenario rather than a static visual scene requires additional investigation. We proceed with the research objective of integrating recognized speech input with the reference resolution framework while maintaining real-time computational performance. We augment the integrated reference resolution framework presented in Chapter 6 to allow for ASR rather than transcribed speech input. We present various models that can support recognized speech input. Specifically, we attempt to use information from the entire n-best list of recognition hypotheses for each utterance to identify the referring expressions that are present in this utterance and to resolve each referring expression to the corresponding set of reference objects. We focus on resolving exophoric referring expressions, which are non-pronominal noun phrases that refer to entities in the extralinguistic environment (e.g. “the book to the right” or “that one plant”) because they are tightly coupled with a user’s eye gaze behavior. We present several reference resolution algorithm variations that use a confusion network to represent an n-best list of recognition hypotheses. We evaluate the use of confusion networks as models of ASR input in terms of reference resolution accuracy and computation speed. Finally, we attempt to determine to what degree the results from the investigation of user interaction with a static multimodal conversational interface can generalize to interaction with an immersive multimodal conversational interface.

In the remaining sections of this chapter we present a detailed description of the process and effects of integrating recognized speech input into our reference resolution framework within situated dialogue. In §7.2 we discuss the challenges of using recognized speech input for reference resolution. We also describe the modifications that are made to the integrated reference resolution framework presented in Chapter 6 to allow for recognized rather than transcribed speech input. In §7.3 we describe how eye gaze can be used to improve reference resolution performance despite a significant number of speech recognition errors. In §7.5 we present the computational complexity of the resulting algorithm and evaluate its potential to be used for real-time processing. In §7.4 we present the experiments we conducted using the integrated framework focusing on the effect of ASR and eye gaze input on reference resolution. In §7.6 we discuss the steps that should be taken to utilize the presented reference resolution algorithm in a different domain. Finally, in §7.7 we discuss the implications of the empirical results and present a reference resolution error analysis.

7.2 Incorporating Recognized Speech

Chapter 6 presents our investigation of reference resolution with speech input that has been transcribed by a human annotator. This is useful for obtaining an estimate of the upper bound performance of our reference resolution algorithms. However, direct speech input must be considered so that we have a clear picture of reference resolution performance in a realistic, fully automated multimodal conversational system. Such a system must contain an automatic speech recognition (ASR) component. An ASR component is responsible for converting acoustic speech signals into machine-readable text strings. It has been well documented that ASR can be very error-prone. Speech recognition errors can have a significantly detrimental effect on reference resolution. In fact, erroneous recognition of spoken references are the leading cause of incorrectly resolved references [12].

Given that the purpose of our research is to improve spoken language understanding via eye gaze rather than ASR improvement, we use an off-the-shelf commercial ASR product. Specifically we use the Microsoft Speech SDK, which takes an acoustic speech signal as input and produces a transcript for each spoken utterance. This results in an n-best (with $n = 100$) list of recognition hypotheses that are ranked in order of likelihood. We use this n-best list of recognition hypotheses to determine and resolve all referring expressions that are present in a given utterance. In this section, we describe how to mitigate the adverse effect of ASR on exophoric reference resolution.

7.2.1 Problem Formulation

Table 7.1 shows a portion of a sample dialogue between a user U and an artificial agent S within the Treasure Hunting immersive virtual world (described in §4.3). Each U_i represents a user utterance and and each S_i represents an artificial agent (i.e. system) utterance. Exophoric referring expressions 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, which are tightly coupled with a user's eye gaze behavior. Here, we present an approach for using a list of recognition hypotheses in exophoric reference resolution.

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**]
S_3	Do you see another one.
U_4	yeah I see a
U_5	[**three shields**] and [**three brown arrows**]
U_6	and [**three green arrows**]

Figure 7.1: Conversational Fragment with Exophoric Referring Expressions

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 it leads to a large amount of information being ignored. Table 7.1 demonstrates this problem. Here, the number after the underscore denotes a timestamp associated with each recognized spoken word. The strings enclosed in brackets denote 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, there may not be a single recognition hypothesis that contains the entire referring expression, but each constituent word from the referring expression may be contained in some recognition hypothesis. Thus, for the purpose of resolving referring expressions, it is desirable to consider the entire n-best list of recognition hypotheses.

Hypothesis	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
$\vdots$	
H_9:	. . . icy_5210 winds_5630 along_6080 [sword_6460]
$\vdots$	
H_{25}:	. . . icy_5210 winds_5630 [long_6080 sword_6460]
$\vdots$	

Table 7.1: Portion of Sample n-best List of Recognition Hypotheses

Consequently, we need to augment our integrated reference resolution framework (described in Chapter 6) to use the entire n-best list of recognition hypotheses as input instead of transcribed speech input. As we have said, the goal of reference resolution is to find a mapping between a sequence of referring expressions $r_1, r_2, \ldots, r_n$ and a sequence of object sets $O_1, O_2, \ldots, O_n$. Directly finding the probability $P(O_1, O_2, \ldots, O_n | r_1, r_2, \ldots, r_n)$ is computationally intractable in a real-time system. Instead, our current approach is to calculate $P(o_i | r_j)$ for each referring expression r_j and object o_i independently.

This presents several challenges. First, given a list of recognition hypotheses, it may be difficult to determine the number of referring expression n that were produced in the given utterance. This problem can be seen in Table 7.2. Based on the shown ASR hypotheses, it is not clear whether this utterance contains zero, one, two, or more referring expression. Here, the system must identify `the lamp` and `the desk` as referring expressions in order to have a chance to resolve them. However, no recognition hypothesis contains the correct number of referring expression, which in this case is two. Nonetheless, it is clear that both referring expressions are present in the n-best list. In addition to knowing how many referring expressions are present in an utterance, it is necessary for the system to know their temporal alignment. Even if there is a way to identify the presence of referring expressions, it is not trivial to determine which recognition hypotheses should be used to resolve these references. Viewing Table 7.2 again as an example, lets assume that the reference resolution algorithm has determined that the the spoken utterance contains a referring expression that begins at timestamp 1540. There are two potential referring expressions that fit this criteria: `the plants` and `the lamp`. The algorithm could naively take the highest ranked referring expression and resolve it. However, there may be more information available to the system. For example, lets consider the situation in which the field of view contains a lamp but no plants, and the user's gaze if fixated on this lamp. This can indicate that `the lamp` is the more likely referring expression and that it should be resolved to the lamp object that is being viewed.

Utterance:	I like [the lamp] and [the desk]
H_t:	I_1020 like_1160 [the_1540 lamp_1670] and_1860 [the_2130 desk_2310]
H_1:	I_1020 like_1160 camping_1580 [the_2130 desk_2310]
⋮	
H_6:	I_1020 bike_1160 [the_1540 plants_1670] until_1860 dusk_2310
H_7:	I_1020 like_1160 to_1540 camp_1670 until_1860 dusk_2310
H_8:	[light_1050] [the_1540 lamp_1670] and_1860 [the_2130 desk_2310]
⋮	

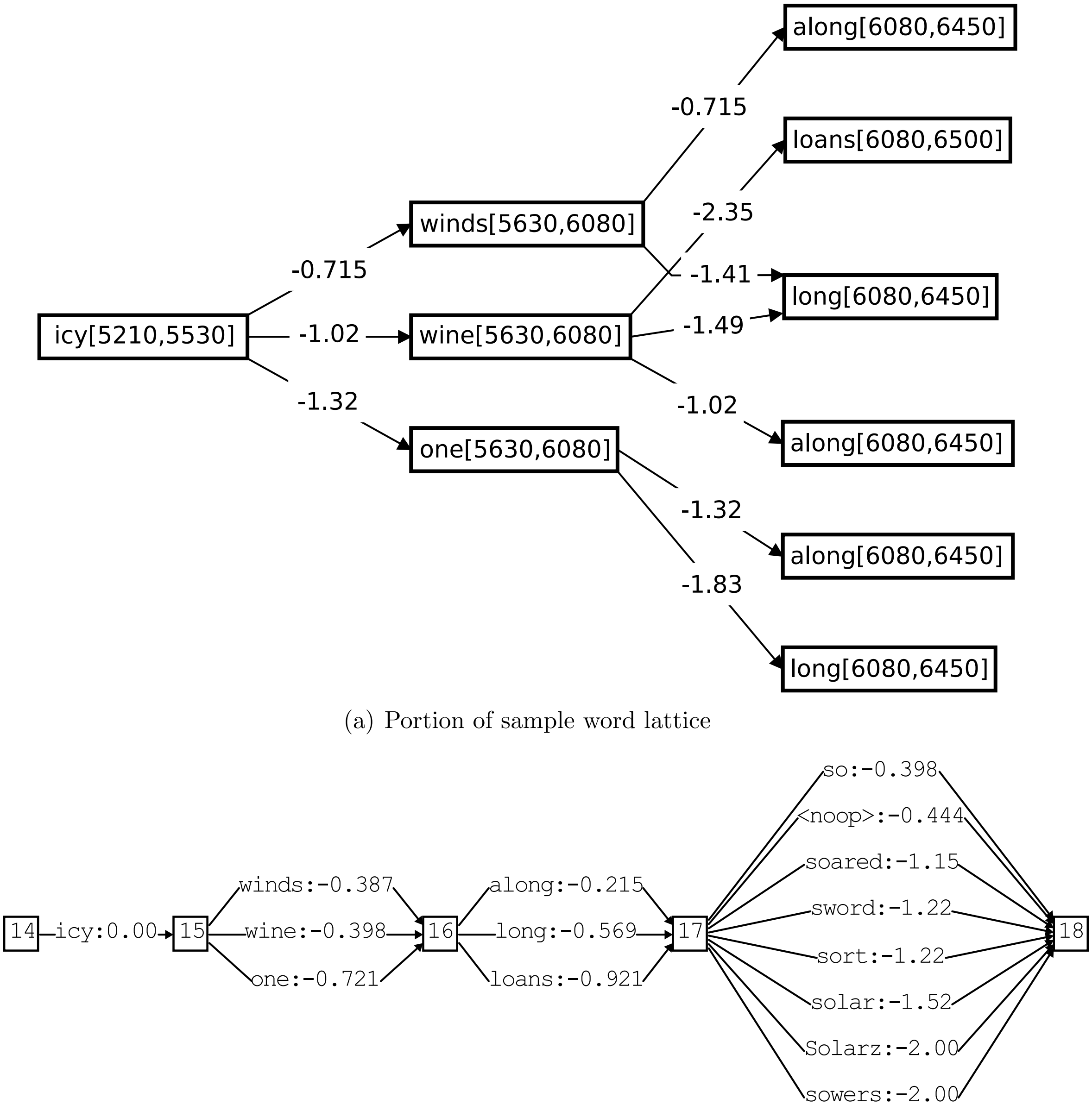
Table 7.2: n-best List of Recognition Hypotheses with Multiple Referring Expressions

We attempt to address the aforementioned difficulties by developing algorithms that combine an n-best list of ASR hypotheses with, dialogue, domain, and eye gaze information. This is done by modifying the reference resolution framework described in Chapter 6 to employ word confusion networks, which are described in the following section.

7.2.2 Word Confusion Networks

A word confusion network (WCN) is a compact representation of a word lattice or n-best list [55]. Word lattices and word confusion networks are both directed graph representations of alternative ASR hypotheses, but with different topological structure. Figure 7.2 depicts the topology of portions of a sample word lattice and word confusion network for the utterance "...I see one long sword", both corresponding to the n-best list of ASR hypotheses shown in Table 7.1. In the word lattice, each vertex represents a word occurring during a particular time interval and the probability that it was uttered during this time interval. Each incoming edge represents the word's language model scores. In the word confusion network, each vertex represents a point in time. Each edge represents a word and the probability that this word was uttered in the time interval specified by the adjoining vertices. Here, all probability values are represented in log scale. Typically an acoustic score is incorporated into the word probability score calculation for word lattices and word confusion network. For the purpose of this work, however, acoustic information is ignored because it has a negligible effect on WCN word error rate when timing information is available [55].

The word confusion network representation of ASR hypotheses has several advantages over the n-best list and word lattice representations for the purpose of reference resolution. Word confusion networks are capable of representing more alternative hypotheses in a more compact manner than either n-best lists or word lattices. It has been shown that word confusion networks constructed from only the top 5 % of word lattice links can represent more hypothesis paths than the original word lattice [55]. Moreover, these word confusions networks outperformed both n-best list ($n = 300$) and word lattice representations in terms of word recognition accuracy. This can be attributed to two key factors: (1) a larger hypothesis search space and (2) a more direct estimation of word recognition accuracy—rather than sentence recog-



(a) Portion of sample word lattice

(b) Portion of sample word confusion network

Figure 7.2: Word Lattice and Confusion Network Topology

nition accuracy—using word posterior probabilities [55]. A compact representation is critical for efficient post-processing of ASR hypotheses, especially for a real-time conversational system. A large hypothesis space is important because an uttered referring expression may not be present in the top few recognition hypotheses, as exemplified in Table 7.1 and Table 7.2. In this case, the presence of a salient potential referent that is very compatible with a low probability referring expression may be able to provide enough evidence to determine that the referring expression was indeed uttered. Another feature of confusion networks is that they provide time alignment for words that occur at approximately the same time interval in competing hypotheses as shown in Figure 7.3. The figure shows an portion of the 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. Time alignment is useful for efficient syntactic parsing as well as for approximating the timing information of each uttered word, which is necessary for modeling attentional salience of potentially referenced objects.

7.3 Multimodal Reference Resolution

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. In this section, we describe the algorithm, starting with the input and output and proceding with the detailed algorithm.

7.3.1 Input

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

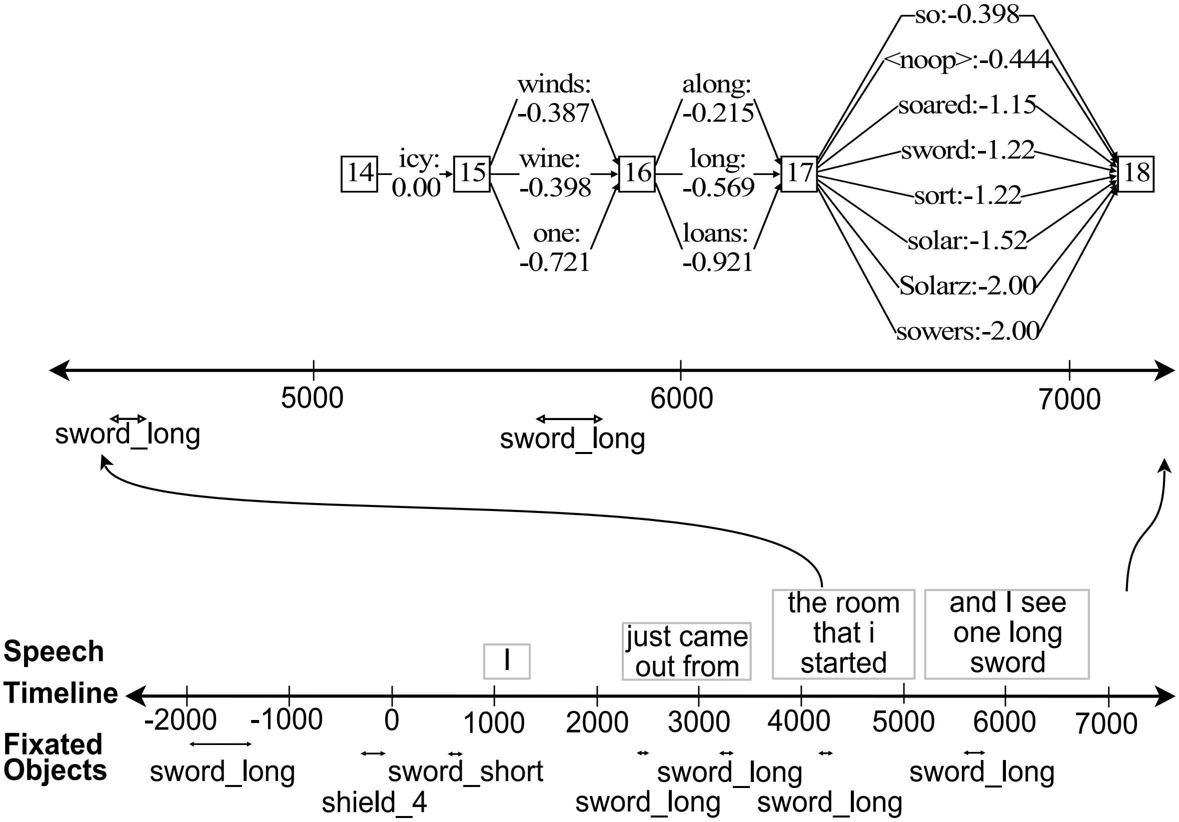


Figure 7.3: Parallel Speech and Eye Gaze Data Streams, including WCN

objects (by eye gaze) that temporally correspond to the spoken utterance and (3) a set of potential referent objects. The set of potential referent objects can vary from utterance to utterance. 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. This is because people typically only speak about objects that are visible or have recently been visible on the screen during the treasure hunting task.

7.3.2 Output

The objective of this reference resolution algorithm is to identify the referring expressions in a given utterance and to resolve each to a set of potential referent objects. Thus, the desired output of the algorithm is a temporally ordered list of (*referring expression*, *referent object set*) pairs. For example, the utterance "I see axes and one long sword" should produce the following output:

1. (`axes`, $\{$*axe_big*, *axe_small*$\}$)
2. (`one long sword`, $\{$*sword_long*$\}$)

where *axe_big*, *axe_small*, and *sword_long* are potential referent objects for this utterance.

7.3.3 Algorithm

The pseudo code for our Multimodal Reference Resolution algorithm is shown in Algorithm 2. The algorithm takes three inputs: (1) an n-best list of ASR hypotheses, $\overrightarrow{H}$; (2) a set of potential referent objects, $\overrightarrow{O}$; and (3) a list of eye fixations to potential referent objects, $\overrightarrow{F}$. The algorithm proceeds with the following four steps:

Step 1: Construct Word Confusion Network In line 1, a word confusion network WCN is constructed from the input n-best list of alternative recognition

Algorithm 2 Multimodal Reference Resolution

```
           Input: $\overrightarrow{H}$ {n-best list of ASR hypotheses}
           Input: $\overrightarrow{O}$ {set of potential referent objects}
           Input: $\overrightarrow{F}$ {list of eye fixations}
$Step1$ {   1: $WCN \leftarrow$ constructWordConfusionNetwork($\overrightarrow{H}$)
$Step2$ {   2: parse(WCN)
        {   3: $\overrightarrow{R} \leftarrow$ extractReferringExpressions(WCN)
$Step3$ {   4: let $\overrightarrow{RR} = \emptyset$
        {   5: for all $r_j \in \overrightarrow{R}$ do
        {   6:    {Ensure: $\overrightarrow{o_i} \subset \overrightarrow{O}$}
        {   7:    $\overrightarrow{o_i} \leftarrow$ resolveReferringExpression($r_r$, $\overrightarrow{O}$, $\overrightarrow{F}$)
        {   8:    $\overrightarrow{RR} \leftarrow \overrightarrow{RR} \cup (r_j, \overrightarrow{o_i})$
        {   9: end for
$Step4$ {  10: prune($\overrightarrow{RR}$)
        {  11: sort $\overrightarrow{RR}$ in ascending order according to timestamp
           12: return $\overrightarrow{RR}$
```

hypotheses with the SRI Language Modeling (SRILM) toolkit [83] using the procedure described in [55]. 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 7.2(b) 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 7.1.

Step 2: Extract Referring Expressions from WCN In line 2, the word confusion network is syntactically parsed using a modified version of the CYK [15, 42, 89] 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 which are shown in Appendix B. A parse chart of the sample word confusion network is shown in Table 7.3. 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, in line 3, a set of all exophoric referring expressions $\overrightarrow{R}$ (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 7.3): `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.

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 7.3: Syntactic Parsing of Word Confusion Network

Step 3: Resolve Referring Expressions In lines 4-9, each referring expression r_j is resolved to the top k potential referent objects, where k is determined by information from the linguistic expressions. In line 7, the procedure for resolving a single referring expression described Chapter 6 Algorithm 1 is utilized. Resolving each referring expression results in a list of *(referring expression, referent object set)* pairs $\overrightarrow{RR}$ with confidence scores, which are determined by two components. The first component is the confidence score of the referring expression, which is explained in 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 6.2). 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: Select References The last step of the algorithm is to select and return the final list of *(referring expression, referent object set)* pairs. In line 10, the list 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 7.3, the referring expressions `one long sword` and `wine` overlap in position 15. Finally, in line 11, the resulting *(referring expression, referent object set)* pairs are sorted in ascending order according to their constituent referring expression timestamps.

7.3.4 Algorithm Variations

Referring Expression Start Boundary Detection

One variation of Algorithm 2 is to use referring expression start boundary detection to determine the order in which the referring expressions should be resolved. This is in contrast to resolving all referring expressions and then pruning the resultant set. Instead, a greedy method is used to iteratively resolve and prune referring expressions. Specifically, the following procedure is used in place of lines 4 through 10 of Algorithm 2: (1) the position in the word confusion network that is most likely to contain a referring expression start boundary is determined; (2) out of the referring expressions that share this start boundary, the one with the highest confidence score is resolved; and (3) all temporally overlapping referring expressions are pruned (i.e. removed from consideration for resolution).

Noting that each potential referring expression corresponds to some position in the word confusion network, the crux of the above procedure is determining which confusion network position corresponds to the start boundary of a referring expression that has indeed been uttered. There are many possible ways to accomplish this. In the work presented here, semantic compatibility scores between the potential referent objects and the potential referring expressions corresponding to each confusion network position are calculated. For each position, the most salient potential referent object is compared with each referring expression and the mean compatibility score is calculated. The position with the highest score is determined to be the most likely to contain a referring expression start boundary.

Consideration of Referring Sub-expressions

Another variation of Algorithm 2 is to allow the `extractReferringExpressions(WCN)` function (in line 3) to return a referring expression if it is contained in another referring expression found in the word confusion network. By default, a referring expression is returned only if it is not part of another referring expression in the confusion network. For example, given that the referring expression “the big red chair” is a potential referring expression to be resolved; then, unlike the basic algorithm, this variation would allow the sub-expression “big chair” to be a potential referring expression as well. This variation of the algorithm has the potential to improve reference resolution performance by considering a larger variety of potential referring expressions with a larger variety of start boundaries. The tradeoff is that more potential referring expressions must be considered (which will increase processing time) and that many of these potential referring expressions may be incomplete.

7.4 Experimental Results

Using the data corpus presented in §4.3, we applied the multimodal reference resolution algorithm described above on the 2052 annotated utterances (from 15 users) consisting of 2204 exophoric referring expressions. The reference resolution model parameters are set according to those empirically determined using the static domain (described in §6.2. 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.

7.4.1 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 and their individual components to the annotated “gold standard” 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.

7.4.2 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. For each evaluation metric (lenient and strict) the best performing WCN variation is shown.

Table 7.4 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 7.4: 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 7.5: Strict F-measure Evaluation

Table 7.5 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 [76], 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.

Tables 7.6 and 7.7 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 7.5. 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 7.6: Lenient F-measure Evaluation for Closely Coupled Utterances

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

Table 7.7: Strict F-measure Evaluation for Closely Coupled Utterances

7.4.3 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 7.4 and 7.5. These tables show results for the best performing WCN variation, which allows for referring-subexpressions and considers referring expression start boundaries (described in more detail in §7.3.4). Table 7.4 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 7.5 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.

Effect of using Referring Expression Start Boundary Detection

In this section, we compare two versions of the reference resolution algorithm both of which operate on a word confusion network. One variation employs referring expression start boundary detection as described in §7.3.4, while the other employs the basic algorithm as described in §7.3.3.

Table 7.8 shows this comparison. As can be seen in this table, the reference resolution algorithm with boundary detection outperforms the algorithm without boundary detection when evaluated on each evaluation metric regardless of whether

Evaluation Metric	Without Boundary Detection		With Boundary Detection	
(F1-measure)	Without Gaze	With Gaze	Without Gaze	With Gaze
Lenient	0.484	0.533	0.508	0.544
Strict	0.309	0.333	0.319	0.337

Table 7.8: Evaluation of Algorithm using Referring Expression Start Boundary

eye gaze information is available to the system. Although the improvement is not significant, a marginal improvement is noticeable according to the lenient evaluation metric when eye gaze information is incorporated ($p < 0.14$). In this case the F-measure improves from 0.484 to 0.508 $(+0.024)$.

Effect of permitting Referring Sub-expressions

In this section, we again compare two variations of the multimodal reference resolution algorithm both of which operate on a confusion network. Both algorithms take a word confusion network as input and extract potential referring expressions from it. One version permits potential referring expressions to be sub-expressions of other potential referring expressions as described in §7.3.4, while the other does not.

Table 7.9 compares these two variations. As can be seen in this table, permitting referring sub-expression consideration marginally ($p < 0.124$) improves reference resolution performance according to the lenient evaluation metric when eye gaze is incorporated. The effect is insignificant when eye gaze is used. However, the performance is significantly ($p < 0.0001$) diminished according to the strict evaluation metric regardless of whether eye gaze is incorporated. Permitting sub-expressions allows the algorithm consider a larger variety of potential referring expressions with a larger variety of start boundaries which makes it more likely to find a match with a potential referent object. This makes it more likely that a resolved referring expression is correct in terms of lenient evaluation, but is incomplete and therefore is incorrect in terms the strict evaluation. Thus, systems for which it may be vital to resolve exact referring expressions should not permit referring sub-expressions.

Evaluation Metric	Don't permit Sub-expressions		Permit Sub-expressions	
(F1-measure)	Without Gaze	With Gaze	Without Gaze	With Gaze
Lenient	0.484	0.533	0.509	0.546
Strict	0.309	0.333	0.251	0.269

Table 7.9: Evaluation of Algorithm allowing Referring Sub-expressions

However, systems that require referent object identification with an inexact referring expression may benefit from allowing referring subexpressions, especially in combination with incorporating referring expression start boundary information (as shown above in §7.4.3).

7.4.4 Role of Word Confusion Network Depth

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 7.3 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 7.4 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.

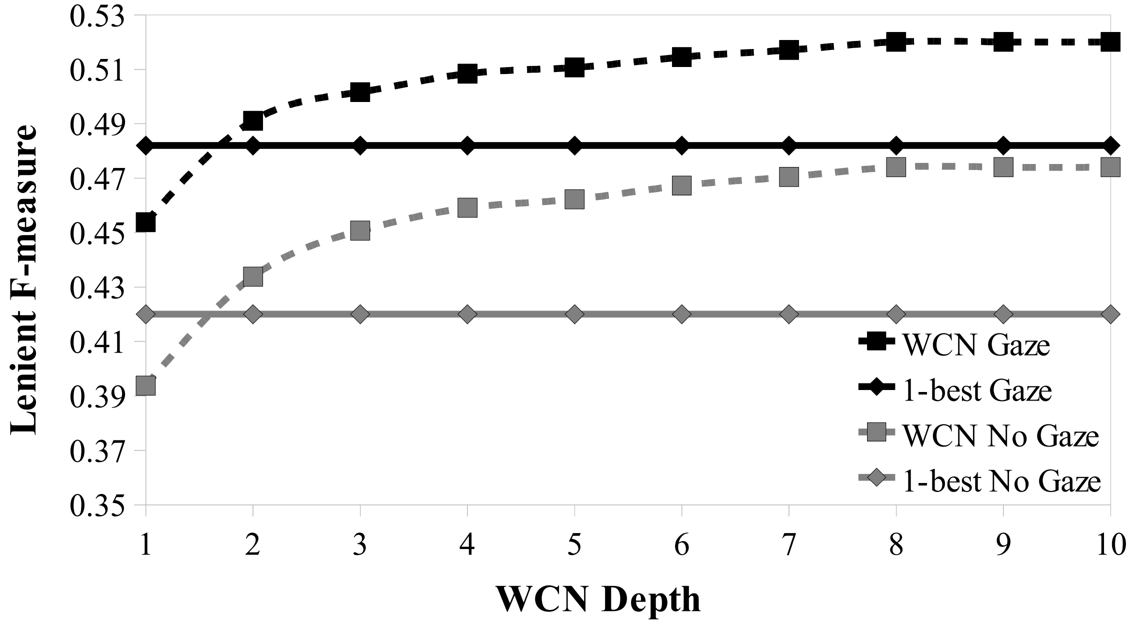


Figure 7.4: Lenient F-measure at each WCN Depth

7.5 Computational Complexity Analysis

It is our claim that the multimodal reference resolution algorithm presented in Algorithm 2 is acceptable for real-time processing of spoken utterances. The multimodal reference resolution procedure takes the following three inputs: (1) an n-best list of ASR hypotheses, (2) a set of potential referent objects, and (3) a list of eye fixations to potential referent objects. Several steps of this algorithm have an asymptotic execution time complexity of $O(N^3)$ with respect to the size of one of these inputs. This is seemingly problematic for this algorithm's use in a real-time multimodal conversational system. However, in practice the size of each of these inputs is kept small, thereby enabling real-time multimodal reference resolution. The rest of this section provides a detailed computational complexity analysis to illustrate this point.

7.5.1 Confusion Network Construction

In line 1 of Algorithm 2, a word confusion network is constructed from an n-best list of ASR hypotheses. More generally, a word lattice is used instead of an n-best list. An n-best list, however, can be viewed as a word lattice with each node having exactly one incoming and one outgoing word link. A word confusion network is constructed by finding an acceptable multiple alignment of word lattice hypotheses [55]. Since there is no known efficient solution for optimal lattice alignment, in [55] Mangu, et. al. present a heuristic approach based on lattice topology, word timing information, and phonetic similarity between words in competing recognition hypotheses. This procedure consists of three steps: (1) word lattice pruning; (2) intra-word clustering; (3) and inter-word clustering. In the lattice pruning step, all word links whose probability is below an empirically defined threshold are removed from the lattice; 95% of word links can be removed without increasing the resulting word error rate [55]. In the intra-word clustering step, identical words that overlap in time are grouped into equivalence classes. In the inter-word clustering step, similar equivalence classes are combined. The final equivalence classes make up a confusion network which can represent all hypotheses in the original (pruned) lattice.

The lattice alignment algorithm, specifically the inter-word clustering and intra-word clustering steps, has an asymptotic computational complexity of $O(|L|^3)$, where $|L|$ is the number of links in the word lattice. Typically, $|L|$ is kept small via the pruning step, which can remove 95% of the lattice links without sacrificing word recognition performance. According to [55], the lattice alignment algorithm resulted in average runtimes with a real time factor of 0.55 when processing lattices in the Switchboard corpus (on a 400 MHz Pentium-II processor). Real time factor (RTF) is defined as the ratio between the time it takes to process an input speech signal and the duration of the input speech signal. It is a hardware dependent metric that is commonly used to measure the speed of an ASR system. An RTF of 0.55 is under the 1.0 RTF threshold that is commonly used to determine whether or not an ASR system is acceptable for real-time processing.

Additionally, in our study an n-best list is used rather than a word lattice as input to the confusion network construction algorithm. This has two desired effects. First, it simplifies the speech input model by eliminating acoustic scores from the ASR hypotheses. In our study, we focus on post-processing out-of-the-box ASR hypotheses. Moreover, it has been shown in [55] that the availability of an acoustic similarity between words has no effect of the quality of confusion network construction (measured by word error rate) when timing information for each word is present. Second, the number of word links in the n-best list is greatly reduced compared to the word lattice. The number of word links in the n-best list is bounded by $O(n{\cdot}|w|)$, where n is the number of hypotheses in the list and $|w|$ 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. In our study, confusion networks were constructed using the SRI Language Modeling Toolkit (SRILM) [83]. As n-best lists were used as input rather than complete word lattices, word lattice pruning was unnecessary to ensure real-time performance. Real-time performance is discussed further in §7.5.5.

7.5.2 Confusion Network Parsing

In lines 2 and 3 of Algorithm 2, a word confusion network is syntactically parsed in order to extract all referring expressions (noun phrases) that occur in it. Here, the CYK algorithm is modified to the CYK-WCN algorithm, which accepts a word confusion network as input instead of a single string and stores all noun phrase substrings during parsing. The standard CYK algorithm (using an efficient normal form for the grammar) has an asymptotic computational complexity of $O(|G| \cdot |w|^3)$ [51], where $|G|$ is the size of the grammar and $|w|$ is the number of words in the the input string.

When using a confusion network as input, each of $|c|$ positions—also called *confusion sets*—in the confusion network must be processed. Each confusion set consists of at most $|d|$ competing words, which we refer to as the depth of the confusion network. The number of confusion sets $|c|$ in a particular WCN is proportional to the number of words $|w|$ in the longest ASR hypothesis used to construct this WCN. Parsing a WCN of depth 1 is identical to parsing a single string. When parsing a WCN of depth d, the step that checks for production rule applications must be altered. For each binary production rule, every word pair in two confusion sets is compared in this step. This results in $|G| \cdot |d|^2$ comparisons for each pair of confusion sets. Therefore, the overall computational complexity for the CYK-WCN algorithm becomes $O(|G| \cdot |d|^2 \cdot |w|^3)$.

7.5.3 Reference Resolution

In lines 4 through 9 of Algorithm 2, all $|r|$ referring expressions that are extracted from the word confusion network are resolved to a set of referent objects. The loop is executed $|r|$ times and is dominated by the `resolveReferringExpression` procedure in line 7. This procedure is shown in more detail in Algorithm 1.

In this algorithm every operation can be performed in linear time except sorting the potential referent objects $\overrightarrow{O}$, which is shown in line 8. In lines 1 through 3, the system must computes the visual history VH of each of $|O|$ potential referent objects. In order to accomplish this, $|f|$ gaze fixations that occur during the given utterance must be processed. Each potential referent object that is not processed receive a zero (or a very small uniform non-zero) visual salience score. In lines 4 through 7, the likelihood of each potential referent object given a referring expression is computed. This computation requires that all $|O|$ potential referent objects must be processed. In line 8, the potential referent objects are sorted in $O(|O| \cdot log(|O|))$ time. To finish resolving the given referring expression, in lines 9 through 18, the most salient m potential referent objects are selected. In the worst case, when the referring expression's number status is unknown and all potential referent object have equivalent salience values, $m = |O|$. This results in a $O(|f| + |O| \cdot log(|O|))$ asymptotic computational complexity for the entire `resolveReferringExpression` procedure.

Consequently, lines 4 through 9 of Algorithm 2 result in a $O(|r| \cdot |f| + |r| \cdot |O| \cdot log(|O|))$ complexity. Typically, including in our work, users produce short utterances during interaction with a conversational system. Noting that the number of gaze fixations $|f|$ during a given utterance is proportional to the length of the utterance (in terms of time), $|f|$ can be eliminated from the complexity calculation; resulting in a $O(|r| \cdot |O| \cdot log(|O|))$ asymptotic computational complexity.

7.5.4 Reference Selection

In lines 10 and 11 of Algorithm 2, the final *(referring expression, referent object set)* pairs are selected for the given utterance. In line 10, low-probability and temporally overlapping pairs are pruned. This is a $O(|r|^2)$ operation because each pair must be compared with each other pair to determine if they temporally overlap. In line 11, the pairs are sorted according to timestamp. This is a $O(|r| \cdot log(|r|)$ operation.

7.5.5 Applicability as a Real Time System

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 an utterance, using the multimodal reference resolution algorithm presented in Algorithm 2 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. In order for this algorithm to be acceptable for real time processing the input sizes must be kept small.

One way to achieve fast processing time is to keep user utterances short. This is because 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, using the Microsoft Speech SDK, utterance boundaries are determined by a 500 ms period of silence; which typically correspond to intentional pauses in user speech. This typically results in very short utterances; with a mean length of 6.41 words and standard deviation of 4.35 words. In fact, of the 2052 utterances in our evaluation data set, none exceed 31 words. Moreover, if computation speed is of the utmost importance, utterances can be made arbitrarily small by bounding an utterance based on a maximum number of words in addition to bounding the utterance by a period of silence. However, this could hinder reference resolution (and other utterance processing) performance because utterances would often be unnaturally split.

Another way to improve processing speed is to use a depth-limited word confusion network. Limiting the depth $|d|$ of a confusion network increases WCN parsing speed. Given that a WCN is constructed from an n-best list in our study, its depth is limited by $|n|$ and thus can be no greater than 100 words. In practice; $|d|$ has a mean of 10.1, a standard deviation of 8.1, and a maximum 89 words. In practice, as shown in Figure 7.4, limiting $|d|$ to 8 words achieves comparable reference resolution results to using a full word confusion network.

One more way to improve processing speed is to limit the number of potential referent objects $|O|$ for each utterance. In a large system the number of potential referent objects is likely to be the bottleneck for utterance processing speed. Fortunately, users almost always refer to objects that either have recently been visible on the screen or that have recently been referred to in the dialogue. Thus, these are the only objects that need to be considered when resolving a given utterance. In the immersive treasure hunting domain, which contains 155 unique objects, an object constitutes a potential referent if it is present in the same room as the user or in the dialogue history.

To demonstrate the applicability of our multimodal reference resolution algorithm for real-time processing we applied it on the *Parallel Speech and Eye Gaze Data Corpus in an Immersive Domain* presented in §4.3. This corpus contains utterances with a mean input time of 2927.5 ms and standard deviation of 1903.8 ms. The most complex (and slowest) variation of the reference resolution algorithm was considered. 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 acceptable real-time performance.

7.6 Domain Customization

Domain modeling can be a time-consuming and costly endeavor. Thus, a major aim of this dissertation is to design and present a multimodal reference resolution algorithm that is domain-independent in as much as possible. We explore the use of eye gaze for reference resolution in a situated dialogue with minimal domain modeling. Particularly, we investigate to what extent eye gaze information can compensate for missing domain information. Despite our best effort to minimize the algorithm's reliance on domain-specific information, a minimal domain model must be constructed in order to transfer our multimodal reference resolution algorithm to a new domain. Domain customization requires modification of the following three domain aspects: (1) potential referent object model, (2) language processing capabilities, and (3) reference resolution algorithm parameters. Here, we discuss the steps that must be taken in order to successfully complete this migration.

Several attributes of each potential referent object in the domain must be specified. First, a unique object ID must be assigned so that the reference resolution algorithm is able to identify referred-to objects. Then, one or more semantic object types must be encoded for the purpose of grouping objects together into categories and producing competing potential referents for each referring expression. Next, objects must be grouped in terms of their location in the visual environment in order to determine the proximity of an object to the user, which is used to determine the list of potential referent objects for a given utterance. In the Treasure Hunting domain, objects are grouped in terms of which room they are contained in. That is, when the user is in a particular room, all of the objects in this room are potential referent objects. Alternatively, the set of potential referent objects for a given utterance can be determined by the recency of the object being visible on the screen. In this case, a temporal recency cutoff value must be established. The final domain model specification is to encode any important visual attributes. These attributes include size, color, shape, material, and any other distinguishing object characteristics that may be used by users to uniquely identify each object. Depending on the desired reference resolution accuracy, this step may be ignored because, in many cases, eye gaze information is sufficient to uniquely identify a potential referent object. However, including more information will improve reference resolution performance.

Some language processing capabilities must be modified to complete a domain customization. Particularly, a domain specific lexicon and parsing grammar must be considered. The lexicon must specify the nouns and adjectives that can be used to refer to each potential referent object and its visual attributes. For example, if a large blue sofa is present in the domain model, the following terms associated with this object should be present in the lexicon: sofa, couch, lounge, big, large, huge, blue, dark, etc. Instead of specifying each word that can be associated with a potential referent object, a semantic database tool such as WordNet along with several seed words can be used to garner synonyms, hypernyms, hyponyms, and other semantic relationships between the lexical entries and potential referent objects. A portion of the lexicon does not relate to any particular potential referent object, and thus are domain independent. For example, articles, pronouns, numeric phrases, etc. can be reused without modification. In addition to the lexicon, the parsing grammar must be modified to conform to a new domain. Specifically, part of speech (POS) tags must be assigned for each term in the lexicon. The POS tags do not have to be manually specified and can be obtained using an automatic POS tagger with high accuracy. The remaining parsing grammar needs to undergo only minor modifications to account for domain-specific idioms and may largely remain unaltered from the parsing grammar shown in Appendix B.

For optimal reference resolution performance, moving to a new domain requires some reference resolution parameter tuning. These parameters include those involved in referring expression recognition (e.g. maximum WCN depth) and those involved in resolving a given referring expression (e.g. likelihood of cognitive status $P(T|r_j)$, importance weight of Attentional Salience relative to compatibility $\alpha(T)$).

7.7 Discussion

In Section 7.4.2 we have shown that incorporating eye gaze information improves reference resolution performance. Eye gaze information is particularly 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 referring expression `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 7.4.3 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 0.333 from WCNs compared to F-measure of 0.035 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 only a 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. Clearly, reference resolution performance still suffers from sub-optimal speech recognition (in comparison to manual transcriptions). However, it is important to note that it is not the goal of this paper to address speech recognition performance, but rather to use available speech recognition results to improve reference resolution performance.

The best reference resolution performance is achieved when considering only closely coupled utterances and evaluating the reference resolution algorithm on transcribed data. As can be seen in Section 7.4, even in this case, reference resolution results are still relatively low, achieving the highest F-measure of 0.7 using the Resolved Concept evaluation. Reference resolution can be further improved by incorporating more domain model information, such as object attributes (e.g. color, shape, size) that can be used to uniquely identify potential referent objects. Improved processing of complex referring expressions, such as spatial referring expressions (e.g. `the book to the right`), can also be used to uniquely identify potential referent objects. The following section describes in more detail the error sources that must be considered for improved reference resolution performance.

7.7.1 Error Analysis

A reference resolution error in an utterance can occur for two reasons (1) a referring expression is incorrectly recognized or (2) a recognized referring expression is not resolved to a correct referent object set. The closely coupled transcribed configuration produces an F1-measure of 0.941 and 0.710 according to the Recognized Concept and Identified Object Set evaluation metrics, respectively. This indicates that a small portion of the reference resolution error is caused by incorrect referring expression recognition and a larger portion of the error is caused by incorrect referent object set identification. In this section, we elaborate on the potential error sources that diminish reference resolution performance.

Given transcribed data which simulates perfectly recognized utterances, all referring expression recognition errors must arise due to incorrect language processing. Most of these errors occur because the parsing grammar has insufficient coverage. These errors fall into the following three categories (shown with examples containing incorrectly recognized referring expressions marked in bold):

1. Incorrect part of speech (POS) interpretation of a word. An example of this is when a word that has a homonym is encountered, e.g. "**can** I pick up the statue" where *can* is a verb that is treated as a referring expression (noun). Another example is when a word has multiple parts of speech, e.g. "the computer is **electronic**" where *electronic* is an adjective that is treated as a referring expression (noun).

2. A noun modifier is treated as a head noun. For example, in the utterance "there is **a desk lamp**" the referring expression shown in bold is parsed as two consecutive referring expressions *a desk* and *lamp* rather than a single referring expression because the word desk is treated as the head noun of a phrase rather than a noun modifier.

3. Out-of-vocabulary (OOV) words. For example, in the utterance "it is a **fern** in a potted plant" the word *fern* is OOV and cannot be used to distinguish between two potted plant potential referent objects.

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. There are several potential reasons for this which are described below (shown with examples containing incorrectly resolved referring expressions marked in bold):

1. A distinguishing object attribute is not encoded into the domain knowledge. For example, in the utterance "there is **a small gold ring**" the gold color of the ring is the attribute that distinguishes this ring from another visible ring and without this knowledge the system is faced with an ambiguity.

2. An explicit distinguishing spatial relationship between objects cannot be estab-

lished by the system. For example, in the utterance “I see **pottery** on a table” without the spatial relationship between the objects of pottery and the table, the system is unable to distinguish between these pottery that’s on the table and the other visible objects of pottery.

3. An implicit distinguishing spatial relationship between objects cannot be established by the system. Sometimes users do not linguistically specify that the spatial location of an object is what uniquely identifies it. For example, during the utterance “I see a picture of a lighthouse and **a chair**” there are two chairs visible on the screen creating an ambiguity. However, only one of the chairs is near the picture, and thus the user is implicitly using this information to uniquely identify which chair is being referred to.

4. The system lacks pragmatic knowledge that can be used to distinguish between multiple objects that can be called by the same name. For example, in the phrase “a armchair and **a sofa**” which is uttered by a user viewing a small and a large sofa, the word *sofa* can refer to either of the two objects. Here, the system lack pragmatic knowledge necessary to discern that if one of the sofas is referred to as an *armchair* and the other a *sofa*, the latter expression refers to the larger one.

5. A singular noun phrase refers to multiple objects. For example, the utterance “i see **a computer**” refers to a computer body, monitor, keyboard, and mouse. Since this expression is singular, the system resolves it to a single object.

6. A plural referring expression is resolved to an incorrect number of referent objects. For example, the utterance “**the bowls** have some blue” refers to four bowls, but is resolved to fewer than four bowls because those are more salient than the remaining bowls.

Some of these errors can be avoided when eye gaze information is available to the system because it can be used to disambiguate referring expressions. However, due to the noisy nature of eye gaze data, many such referring expressions remain ambiguous even when eye gaze information is considered.

One additional error source that impacts our reference resolution evaluation is annotation error. Some ambiguous referring expressions are misinterpreted by the annotator because our annotators were not experts in the Treasure Hunting domain within the immersive environment. Ideally, the data should be annotated by multiple human annotators and only utterances with high agreement should be evaluated. This was not done because annotation errors constitute a minor portion of the overall reference resolution error and because we did not have the resources to complete this labor intensive and tedious task.

7.8 Summary

In this chapter, we have examined the utility of eye gaze ans word confusion networks for reference resolution in situated dialogue within a virtual world. We have presented an implementation of a real-time multimodal reference resolution algorithm. For each utterance, the algorithm takes automatic speech recognition and eye gaze fixation information as input and returns a list of *(referring expression, referent object set)* pairs. The 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.

Combining eye gaze with word confusion networks can overcome a large portion of speech recognition errors for the reference resolution task. Eye tracking technology has improved significantly in the past decade. Integrating non-intrusive (e.g., Tobii system) and high performance eye trackers with conversational interfaces is becoming more feasible. The results from this work provide a step towards building the next generation of intelligent conversational interfaces.

Chapter 8

Conclusions

8.1 Contributions

This dissertation presents investigations on the role of eye gaze for reference resolution in multimodal conversational systems. We address the following three research questions (posed in Chapter 1):

1. Can eye gaze reliably predict user attention during conversation?

2. Can eye gaze based attention prediction help to resolve linguistic referring expressions produced during conversation?

3. What additional modeling is required for reference resolution within an immersive, situated scenario? Particularly, what additional modeling is required to process recognized speech input?

The empirical results have demonstrated that incorporating eye gaze can potentially improve modeling attentional salience and understanding spontaneous speech within a multimodal conversational interface. The following sections highlight specific research contributions.

8.1.1 Eye Gaze in Attention Prediction

We have systematically analyzed the role of eye gaze for attention prediction in a multimodal conversational system. Specifically, we developed several variations of eye gaze based attentional salience models. The empirical results indicate that fixation intensity—the amount of time that an object is fixated during a given time window—is the most important factor in predicting a user's focus of attention. Modeling a fixation to an area on the screen that includes full or partial object occlusion as a fixation to only the fore-front object (rather than considering all objects that overlap with the fixation point) produced a significant improvement in attention prediction results. Other visual features do not produce a significant improvement in attention prediction accuracy when eye gaze information is present.

8.1.2 Eye Gaze in Multimodal Reference Resolution

We have systematically analyzed the role of eye gaze for reference resolution in a multimodal conversational system. We developed an integrated reference resolution framework that combines linguistic, dialogue, domain, and eye gaze information to robustly resolve referring expressions. These referring expressions are found in spontaneous spoken utterances that are collected during human interaction with a multimodal conversational interface. We used the integrated framework to evaluate the effect of eye gaze on multimodal reference resolution. The empirical results show that eye gaze can significantly improve reference resolution performance, especially when little domain information is available to the system. Moreover, eye gaze can partially compensate for a lack of domain model information. Referring expressions that require complex domain modeling to be resolved without the aid of eye gaze, can be correctly resolved at a 43 % rate with the use of a minimal domain model (only taking into account the semantic type of objects and terms that refer to them) when eye gaze is present.

8.1.3 Integrating Eye Gaze with Automatically Recognized Speech in Multimodal Reference Resolution

We have systematically analyzed the role that eye gaze has for overcoming speech recognition errors for the purpose of improving real-time reference resolution. We have investigated various models of automatically recognized speech as input to our integrated multimodal reference resolution framework as well as various ways to incorporate eye gaze and recognized speech into the reference resolution framework. The empirical results show that using a confusion network structure to model alternative recognized speech hypotheses outperforms using only the top recognition hypothesis on the reference resolution task. Using eye gaze to post-process referring expressions extracted from the confusion network further improves reference resolution performance.

8.1.4 Immersive Multimodal Conversational Interface

We have developed an immersive multimodal conversational interface that supports spontaneous speech and eye gaze input. It supports system and user initiated multimodal conversation and can be used to exhibit situated language. The interface contains 188 3D models of objects that can be viewed from any direction (allowed by the physics of the virtual world). This system can be used as a test-bed for future study of multimodal conversation as well as for development of real-world multimodal applications.

8.1.5 Parallel Speech and Eye Gaze Data Corpus

We have constructed a parallel speech and eye gaze data corpus that has been collected via two user studies. The corpus contains time-synchronous multimodal data. That is, each word and each eye gaze fixation has a corresponding start and end time. The data has been annotated to include transcript of speech, as well as marking of all occurring referring expressions and the objects that are being referred to by the expression. This data corpus will be made available to the research community.

8.2 Future Directions

The goal of this dissertation is to study the affect of eye gaze on language understanding in multimodal conversational systems. We focus on studying eye gaze as a subconscious information channel that can be used as an indicator of the focus of attention. However, realistic conversation exhibits social interaction, during which eye gaze may be used in a variety of other ways.

During social interaction eyes are used for seeing (i.e. strictly for vision), looking (i.e. intentional eye direction for gathering visual information), thinking (i.e. raising or closing of eyes to aid the thinking process), and communicating information [73]. Seeing and looking behaviors can be used to identify the focus of attention. A thinking behavior, such as an aversion (i.e. look away) gesture, can be used to indicate a period of cognitive load and that the per is thinking and has not simply stopped speaking [60]. Communicating information is not the intent of these eye behaviors, but rather a byproduct. The most common intentionally communicative act is mutual gaze, which involves fixating on a conversational partner and delineates attentiveness to the conversational partner. Breaks in mutual gaze are important signals used to manage the turn-taking sequence [3]. Other communicative gaze acts include the exposure of emotional, belief, and goal states, and requests for feedback [73, 3]. In future work, we intend to consider the impact of such eye gaze behavior during social interaction on language understanding.

APPENDICES

Appendix A

Regions of Interest

Object Id	Semantic Category	Attributes	
		Static	Dynamic
1	picture	{girl}	{upsidedown}
2	picture	{waterfall}	
3	lamp	{floor}	
4	-		
5	curtains		
6	window		
7	chair		{green}
8	lamp	{desk}	
9	mirror		
10	-		
11	lamp	{chandalier}	
12	table	{desk}	
13	-		
14	picture	{forest}	
15	picture	{flowers}	{crooked}
16	-		

17	lamp	{bedside}	{crooked}
18	table	{cabinet}	
19	bed		
20	pillows		
21	blanket		
22	door		
23	lamp	{bedside}	
24	table	{cabinet}	
25	chair		{brown}
26	candle		
27	plant		
28	-		

Appendix B

Parsing Grammar

$$
\begin{array}{rcl}
\textit{NP} & \rightarrow & \textit{Adj-NP} \\
 & | & \textit{Base-NP} \\
 & | & \textit{Dem} \\
 & | & \textit{Num-NP} \\
 & | & \textit{Pron} \\
 & | & \textit{Dem Adj-NP} \\
 & | & \textit{Dem Base-NP} \\
 & | & \textit{Dem Num-NP} \\
 & | & \textit{Dem One} \\
 & | & \textit{Def-art Adj-NP} \\
 & | & \textit{Def-art Base-NP} \\
 & | & \textit{Def-art Num-NP} \\
 & | & \textit{Def-art One} \\
 & | & \textit{Indef-art Adj-NP} \\
 & | & \textit{Indef-art Base-NP}
\end{array}
$$

$$
\begin{array}{rcl}
\textit{Adj-NP} & \rightarrow & \textit{Adj Base-NP} \\
 & | & \textit{Adj One} \\
 & & \\
\textit{Base-NP} & \rightarrow & \textit{Base} \\
 & | & \textit{Base PrepBase} \\
 & & \\
\textit{Num-NP} & \rightarrow & \textit{Num Base-NP} \\
 & | & \textit{NumAdj Base-NP} \\
 & | & \textit{NumAdj One} \\
 & & \\
\textit{NumAdj} & \rightarrow & \textit{Num Adj} \\
 & & \\
\textit{PrepBase} & \rightarrow & \textit{Prep Base} \\
 & & \\
\textit{Adj} & \rightarrow & \textit{Adj Adj} \\
 & & \\
\textit{Num} & \rightarrow & \textit{One}
\end{array}
$$

Appendix C

Interior Decoration Data Documentation

C.1 Domain Knowledge Files

The following are user-independent files that are used to encode knowledge in the Interior Decoration domain:

Interest Area Definition (interestAreas.ias) File:

The interest area definition file specifies the location of each interest area (object) in the interior decoration scene, one per line. Each line, excluding comments (lines that start with the pound symbol (`#`)), is a tab delimited list with the following fields:

1. shape: `RECTANGLE` or `FREEHAND`
2. object id: integer value
3. coordinate list:
   - for `RECTANGLE` shape: tab separated list consisting of the upper left x, upper left y, lower right x, and lower right y coordinates defining the rectangle

- for `FREEHAND` shape: list of tab separated x,y coordinate pairs defining a polygon

4. object name: string value

objectOverlapMatrix.csv:

The semicolon-separated (*.csv) file contains a square matrix specifying the pairwise area of visual overlap (in pixels squared) between objects. Each row and column corresponds to an object ID number. The matrix diagonal represents an object's size (overlap with itself), with a value of 0 indicating an invalid object that is not considered in any further processing. A negative value indicates that the object in row y is behind the object in column x. For example, `window` (id=6) is behind `curtain` (id=5) and they overlap by $|14942|$ pixels squared.

semanticDomainKnowledge.xml:

The domain model file defines the semantic properties and terminology that can be used to refer to each object in the domain.

Each object must have a `type` property and may have additional properties (e.g., classification, color, size, location, orientation). Each object property is annotated with a concept (WordNet synset in the format of "`<word>#<pos-tag>#<sense-id>`". Each property has a list of *gold standard* attributes that correspond to the object. The *gold standard* terminology is manually compiled from all users' speech transcripts. An example object is shown in Figure C.1, where:

- Object `lamp_bedleft` (id=17) has three properties.
- The concept of property `type` is "`lamp#n#2`".
- The concept of property `classification` is "`bedside#n#1`".
- The concept of property `orientation` is "`crooked#a#1`".

- The *gold standard* terms that can refer to this object are “lamp”, “lampshade”, “bedside”, and “crooked”.

```
<object name="lamp_bedleft" id="17">
  <properties>
    <type text="lamp#n#2">lamp lampshade</type>
    <classification text = "bedside#n#1">bedside</classification>
    <orientation text="crooked#a#1">crooked</orientation>
  </properties>
</object>
```

Figure C.1: Example `object` in `semanticDomainKnowledge.xml` domain model file

Lexicon Files:

Two alternative lexicon files specify valid lexical entries and their part of speech (POS) tags in this domain. Each line, excluding comments (lines that start with the pound symbol (`#`)), has the format of “`<lexical entry>:<POS>`”. The file `lexiconSpatial.txt` contains the same lexical entries as `lexiconNonSpatial.txt` plus additional spatial language terminology.

Parse Rule (lexiconSpatial.txt) File:

This file defines the parsing grammar. Each line, excluding comments (lines that start with the pound symbol (`#`)), specifies a production rule consisting of two elements—(1) a list of non-terminal symbols that can be generated from (2) another non-terminal symbol—separated by the colon symbol (`:`).

Canonicalization Files:

Three alternative canonicalization files specifies a list of lexical entries with corresponding (attribute, value) pairs that encode a semantic interpretation in the interior decoration domain. The files are formatted as follows:

@<lexical entry>

$attribute_1 : value_1$

$attribute_2 : value_2$

...

$attribute_n : value_n$

`-------`

@<lexical entry>

...

For example, the lexical entry `them` has attributes `NUMBER` and `TOPIC` with values `Multiple` and `Collection`, respectively—indicating that this lexical entry refers to an unknown-sized collection of objects.

The three alternative canonicalization files are supersets of each other, with the file `canonSimple.txt` defining only attributes pertaining to semantic types; the file `canonComplexStatic.txt` additionally defining attributes pertaining to static (unchangeable) object properties; and the file `canonComplexDynamic.txt` defining additional attributes pertaining to dynamic object properties.

C.2 User Data Files

The following are data files that were constructed via user studies in the Interior Decoration domain:

user<id>_speech_sentences.xml:

This user speech log file contains a list of "trials", indicating a user's answer to a question posed about the interior decoration scene. Each trial consists of one or more sentences. Each sentence consists of words or references (referring expressions), which themselves consist of words. Each referring expression is classified as type “1” or “2” indicating a definite noun phrase or a pronominal expression, respectively. A sample trial is shown in Figure C.2, where:

- Each word contains the attributes `token` (which indicates the uttered word) and `timestamp`.
- The `<reference/>` tag contains the referring expression “the desk”.
- This definite noun phrase (type=“1”) refers to object with id=“12” and starts 9448 ms after beginning of the trial.

```
<trial id="12">
  <sentence>
    <word token="my" timestamp="7240"/>
    <word token="favorite" timestamp="7344"/>
    <word token="piece" timestamp="7608"/>
    <word token="of" timestamp="7776"/>
    <word token="furniture" timestamp="7893"/>
    <word token="is" timestamp="8610"/>
    <reference token="the desk" type="1" id="12" timestamp="9448">
      <word token="the" timestamp="9448"/>
      <word token="desk" timestamp="9792"/>
    </reference>
  </sentence>
    ...
  </sentence>
  ...
</trial>
```

Figure C.2: Example trial in `user<id>_speech_sentences.xml` file

fixation_<id>.csv:

The semicolon-separated (*.csv) file contains a log of user gaze fixations produced by the EyeLink II eye tracker. Each row in the file contains the following columns:

- user ID (`RECORDING_SESSION_LABEL`)
- trial number (`TRIAL_LABEL`)
- fixation start, end, and duration times in ms (`CURRENT_FIX_START`, `CURRENT_FIX_END`, `CURRENT_FIX_DURATION`, respectively)
- a set of fixated objects identified by their ID numbers (`CURRENT_FIX_INTEREST_AREAS`)
- the user's pupil size (`CURRENT_FIX_PUPIL`) in arbitrary units that vary with subject setup

Appendix D

Treasure Hunting Data Documentation

D.1 Domain Knowledge Files

The following are user-independent files that are used to encode knowledge in the Treasure Hunting domain:

semanticDomainKnowledge.xml:

The domain model file defines (1) the semantic properties and terminology that can be used to refer to each object in the domain and (2) the semantic properties of object-independent terminology that can be used to refer to any object in the domain.

Each object must have a `type` property and may have additional properties (e.g., color, size, shape, material). Each object property is annotated with a concept (WordNet synset in the format of "`<word>#<pos-tag>#<sense-id>`". The `type` property has a list of singular and plural *gold standard* nouns that can be used to refer to the object. Other properties each have a list of *gold standard* attributes that correspond to the object. The *gold standard* terminology is manually compiled from all users' speech transcripts. Additional words are added for terms that can be spelled multiple ways (e.g. bathtub and bath_tub), with tokens in compound terms separated by an underscore. An example object is shown in Figure D.1, where:

- Object `apple` has two properties: `type` and `color`.
- The concept of property `type` is "`apple#n#1`"; the concept of property `color` is "`color#n#1`".
- The *gold standard* singular nouns are "apple" and "fruit".
- The *gold standard* plural nouns are "apples", "fruit", and "fruits".
- The *gold standard* word for property `color` is "red".

```
<object name="apple">
  <properties>
    <type text="apple#n#1">
      <noun_singular>apple fruit</noun_singular>
      <noun_plural>apples fruit fruits</noun_plural>
    </type>
    <color text="color#n#1">red</color>
  </properties>
</object>
```

Figure D.1: Example `object` in `semanticDomainKnowledge.xml` domain model file

The object-independent terminology includes adjectives, prepositions, definite and indefinite articles, demonstratives, locatives, pronouns, and numeric expressions. Each term may include zero or more properties with each property having a single value. Figure D.2 shows an example demonstrative with three properties: `category_anaphora`, `anaphora_manner`, and `topic` having the respective values "yes", "demonstrative", and "collection".

objects.csv:

The comma-separated values file is composed of a list of objects in the Treasure Hunting virtual world along with their properties. Very similar objects are grouped into

```
<dem name="these">
  <properties>
    <category_anaphora value="yes"/>
    <anaphora_manner value="demonstrative"/>
    <topic value="collection"/>
  </properties>
</dem>
```

Figure D.2: Example demonstrative (`dem`) in `semanticDomainKnowledge.xml` domain model file

one semantic name. For example, each unique door object in the virtual world is semantically named a `door` object. Each row in the file contains an object ID, the room in the virtual world that contains this object, a list of constituent objects (delimited by a space character), and the semantic name. Three sample rows are shown in Table D.1. Here, two specific door objects (`door_bathroom1` and `door_bathroom2`) compose the `door_bathroom` object. Each of these objects are in the bathroom and each of them have the semantic name `door`.

ID	ROOM	CONSTITUENTS	NAME
door_bathroom	bathroom	door_bathroom1 door_bathroom2	door
door_bathroom1	bathroom		door
door_bathroom2	bathroom		door

Table D.1: Example row in `objects.csv` file

cnfParserules.csv:

The comma-separated values file defines the parsing grammar, which is shown in its entirety in Appendix B. The file is composed of a list of production rules, with one rule per row. In each row, the first column denotes a list of non-terminal symbols that can be generated from the non-terminal symbol in the second column.

D.2 User Data Files

The following are data files that were constructed via user studies in the Treasure Hunting domain:

*_annotated.xml

This log file contains a list of dialogue turns with zero or more `<user_input/>` tags and exactly one `<system_response/>` tag, denoting user and system dialogue turns, respectively. Each user input consists of recognized, transcribed, and annotated speech (with reference resolution tags) along with gaze fixation information. A sample user input is shown in Figure D.3, where:

- The user's speech starts at "12868124101485" (system time in ms) and lasts for 8380 ms.
- The `<transcript/>` tag contains manually transcribed (gold standard) speech.
- The `<rr_annotation/>` tag contains reference resolution annotations of the speech transcript using the following XML-like markup—with angle brackets (<) replaced by square brackets ([):
  - The text in each `[ref/]` tag denotes a referring expression.
  - The `object` attribute denotes a set of objects referred to by the referring expression.
- The `<wavefile/>` tag contains the relative path to the audio file that was used to generate the recognized speech. Additionally, the wavefile name is used to uniquely identify the utterance (e.g. 20081010-105510-601).
- Each `<phrase/>` tag contains a recognition hypothesis (Microsoft Speech Recognizer).
- The `<gaze/>` tag contains all of the gaze fixations occurring concurrently with the speech input.
- Each gaze fixation (specified by `<gaze_fixation/>` tags) starts at a particular time (system time in ms) and is positioned on particular screen coordinates (e.g. `pos="568,364"`).

- Each <gaze_fixation/> tag contains a list of potentially fixated objects (specified by <mesh/> tags) along with their fixation probabilities. Multiple visually overlapping objects may be fixated simultaneously. The fixation probabilities are calculated based on the distance between the user's eye camera and an object: the i-th closest object is given the probability
  $$p = \frac{1}{i}\left(\sum_{i=1}^{n}\frac{1}{i}\right)^{-1}$$

*_gsTurns.xml:

This log file contains a list of dialogue turns with zero or more <user_input/> tags and exactly one <system_response/> tag, denoting user and system dialogue turns, respectively. Each user input consists of speech (gold standard, i.e. transcribed and timestamped) and gaze fixation information, exemplified in Figure D.4. In this example:

- The <speech/> tag contains two attributes: (1) a unique utterance id and (2) and a flag specifying whether the user's speech matches its accompanying gaze fixation. In this example, matched="1" because the user mentions an object (desk_75) that has been captured by a gaze fixation.
- The user's speech starts at "12868124101485" (system time in ms) and lasts for 8380 ms.
- Each <phrase/> tag contains the timestamped (in ms) tokens of the speech transcript.
- Each <gaze/> tag contains all of the gaze fixations occurring concurrently with the speech input.
- Each <gaze_fixation/> tag contains a list of potentially fixated objects (specified by <mesh/> tags) along with their fixation probabilities.

*_scene.xml:

The scene log file contains a record of visible entities on the screen each time the user's gaze is captured. An example record is shown in Figure D.5, where:

```
<user_input>
  <speech>
    <transcript>it’s a wooden desk with three drawers on the
        right side</transcript>
    <rr_annotation>[ref object="{desk_75}"]it’s[/ref]
        [ref object="{desk_75}"]a wooden desk[/ref] with
        [ref object="{desk_drawer1, desk_drawer2, desk_drawer3}"]
        three drawers[/ref] on the right side</rr_annotation>
    <wavefile>log\user2001\audio\20081010-105510-601.wav</wavefile>
    <phrase start="12868124101485" length="8380" rank="0">
      <token start="4290" length="390">and</token>
      <token start="4680" length="190">say</token>
      <token start="4870" length="360">wooden</token>
      <token start="5230" length="370">desk</token>
      <token start="5600" length="440">with</token>
      <token start="6070" length="380">three</token>
      <token start="6450" length="630">drawers</token>
      <token start="7170" length="260">from</token>
      <token start="7430" length="130">the</token>
      <token start="7560" length="300">right</token>
      <token start="7860" length="470">side</token>
    </phrase>
    <phrase start="12868124101485" length="8380" rank="1">
      ...
    </phrase>
    ...
  </speech>
  <gaze>
    <gaze_fixation start="12868124100940" length="40" pos="568,364">
      <mesh prob="0.55">computer_monitor</mesh>
      <mesh prob="0.27">desk_75</mesh>
      <mesh prob="0.18">castle</mesh>
    </gaze_fixation>
    <gaze_fixation start="12868124101080" length="20" pos="771,375">
      <mesh prob="0.67">computer_body</mesh>
      <mesh prob="0.33">castle</mesh>
    </gaze_fixation>
    ...
  </gaze>
<user_input>
```

Figure D.3: Example user input in *_annotated.xml file

```
<user_input>
  <speech utt_id="20081010-105510-601" matched="1">
    <transcript>it's a wooden desk with three drawers on the right
    side</transcript>
    <phrase start="12868124101485" length="8380">
      <token start="4290" length="390">it's</token>
      <token start="4680" length="190">a</token>
      <token start="4870" length="360">wooden</token>
      <token start="5230" length="370">desk</token>
      <token start="5600" length="440">with</token>
      <token start="6070" length="380">three</token>
      <token start="6450" length="630">drawers</token>
      <token start="7190" length="240">on</token>
      <token start="7430" length="130">the</token>
      <token start="7560" length="300">right</token>
    <token start="7860" length="390">side</token>
    </phrase>
  </speech>
  <gaze>
    <gaze_fixation start="12868124100940" length="40">
      <mesh prob="0.550000">computer_monitor</mesh>
      <mesh prob="0.270000">desk_75</mesh>
      <mesh prob="0.180000">castle</mesh>
    </gaze_fixation>
    <gaze_fixation start="12868124101080" length="20">
      <mesh prob="0.670000">computer_body</mesh>
      <mesh prob="0.330000">castle</mesh>
    </gaze_fixation>
    ...
  </gaze>
</user_input>
```

Figure D.4: Example user input in `*_gsTurns.xml` file

- A user's gaze fixation lasting 80 ms is captured at time `"12868123694708"` (system time in ms) at position `"605,558"` (screen coordinates in pixels). This gaze fixation falls on two objects, specified by `<mesh/>` tags.

- The visible objects on the screen are specified in the `<scene/>` tag. Here, each `<mesh/>` tag contains the center position (screen coordinates: `pos="x,y"`) and the surrounding rectangle (screen coordinates: `rect="left,top,bottom,right"`) of the object's visible region.

```
<record>
  <gaze_fixation start="12868123694708" length="80" pos="605,558">
    <mesh>door_dining</mesh>
    <mesh>castle</mesh>
  </gaze_fixation>
  <scene>
    <mesh pos="565,529" rect="487,552,573,578">sword_long</mesh>
    <mesh pos="599,502" rect="496,595,545,604">sword_short</mesh>
    <mesh pos="564,788" rect="244,0,968,1135">door_dining</mesh>
  </scene>
</record>
```

Figure D.5: Example `record` in `*_scene.xml` file

nbest*.nbest:

This directory contains files generated from the `*_annotated.xml` log specifying an n-best list for each user utterance, identified by the utterance ID. The files use the SRI Decipher(TM) `NBestList2.0` format. Each recognition hypothesis in the file has the following format:

```
(score) w1 ( st:  st1 et:  et1 g:  g1 a:  a1 ) w2 ...
```

Here, a word is followed by a start and end time, language model and acoustic score. The scores are in bytelog scale; a bytelog is a logarithm to base 1.0001, divided by 1024 and rounded to an integer. More information about this file format can be found in the SRILM manual:

`http://www-speech.sri.com/projects/srilm/manpages/nbest-format.5.html`

confusionNetwork*.cn:

This directory contains files specifying a confusion network (also called word mesh) for each user utterance, identified by the utterance ID. These files use the `wlat-format` (a file format for SRILM word posterior lattice) and are generated from the `nbest\*.nbest` files with the SRILM toolkit. The file is formatted as follows:

name s

numaligns N

posterior P

align $a\ w_1\ p_1\ w_2\ p_2$ ...

info $a\ w\ start\ dur\ ascore\ gscore\ phones\ phonedurs$

...

The file format specifies the name of the confusion network s, the number of alignment positions A and the total posterior probability mass P contained in the confusion network, followed by one or more confusion set specifications. For each alignment position a, the hypothesized words w_i and their posterior probabilities p_i are listed in alternation. The ***DELETE*** tag represents an empty hypothesis word. Following the **info** tag, word-level information is specified for alignment position a and hypothesized word w. For the purpose of this work, the word start time $start$ and duration dur (in ms) is considered. The remaining acoustic information is ignored. More information about this file format can be found in the SRILM manual:

`http://www-speech.sri.com/projects/srilm/manpages/wlat-format.5.html`

REFERENCES

REFERENCES

[1] A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. In *Biometrika*, volume 83, pages 715–726, 1996.

[2] N. Bee, E. André, and S. Tober. 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, 2009.

[3] T. Bickmore and J. Cassell. *Social Dialogue with Embodied Conversational Agents*, chapter Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems. Kluwer Academic, 2004.

[4] R. A. Bolt. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, volume 14, pages 262–270, New York, NY, USA, 1980. ACM.

[5] D. K. Byron. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, 2002.

[6] D. K. Byron. Understanding referring expressions in situated language: Some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World*, 2003.

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

[8] E. Campana, J. Baldridge, J. Dowding, B. A. Hockey, R. W. Remington, and L. S. Stone. Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of Perceptive User Interfaces*, 2001.

[9] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea.

In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 520–527, New York, NY, USA, 1999. ACM.

[10] J. Y. Chai, P. Hong, M. Zhou, and Z Prasov. Optimization in multimodal interpretation. In *Proceedings of 42nd Annual Meeting of Association for Computational Linguistics (ACL)*, pages 1–8, 2004.

[11] J. Y. Chai, Z. Prasov, J. Blaim, and R. Jin. Linguistic theories in efficient multimodal reference resolution: An empirical investigation. In *ACM International Conference of Intelligent User Interfaces (IUI05)*. ACM Press, 2005.

[12] J. Y. Chai, Z. Prasov, and P. Hong. Performance evaluation and error analysis for multimodal reference resolution in a conversational system. In *Proceedings of HLT-NAACL (Companion Volumn)*, pages 41–44, 2004.

[13] J. Y. Chai, Z. Prasov, and S. Qu. Cognitive principles in robust multimodal interpretation. *Journal of Artificial Intelligence Research*, 27:55–83, 2006.

[14] Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. Quickset: multimodal interaction for distributed applications. In *MULTIMEDIA '97: Proceedings of the fifth ACM international conference on Multimedia*, pages 31–40, New York, NY, USA, 1997. ACM.

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

[16] N. J. Cooke. *Gaze-Contingent Automatic Speech Recognition*. PhD thesis, University of Birmingham, 2006.

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

[18] A. T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

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

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

[21] J.M. Findlay. Eye scanning and visual search. In J. M. Henderson and F. Ferreira, editors, *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press, 2004.

[22] A. Genkin, D. Lewis, and D. Madigan. Large scale bayesian logistic regression for text categorization. In *Journal of Machine Learning, submitted*, 2004.

[23] G. Gillund and R. M. Shiffrin. A retrieval model for both recogintion and recall. *Psychological Review*, 91:1–67, 1984.

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

[25] P. Gorniak and D. Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.

[26] H. P. Grice. Speech acts. In *Logic and conversation*, pages 41–58. New York: Academic Press., 1975.

[27] Z. M. Griffin. Gaze durations during speech reflect word selection and phonological encoding. In *Cognition*, volume 82, pages B1–B14, 2001.

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

[29] J. K. Gundel, N. Hedberg, and R. Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307, 1993.

[30] J.M. Henderson and F. Ferreira. In *The interface of language, vision, and action: Eye movements and the visual world*. Taylor & Francis, 2004.

[31] C. Hennessey, B. Noureddin, and P. Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, pages 87–94, San Diego, California, 2006.

[32] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classiers*, pages 115–132. MIT Press, 2000.

[33] R. J. K. Jacob. The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems*, 9(3):152–169, 1991.

[34] R. J. K. Jacob. Eye tracking in advanced interface design. *In W. Barfield and T. Furness, editors, Advanced Interface Design and Virtual Environments*, pages 258–288, 1995.

[35] R. J. K. Jacob and K. S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (section commentary). Elsevier Science, Amsterdam, 2003.

[36] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 2002.

[37] M. Johnston. Unification-based multimodal parsing. In *Proceedings of ACL/COLING'98*, 1998.

[38] M. Johnston and S Bangalore. Finite-state multimodal parsing and understanding. In *In Proceedings of COLING00*, 2000.

[39] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. Match: an architecture for multimodal dialogue systems. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 376–383, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[40] T. Jording and I. Wachsmuth. An anthropomorphic agent for the use of spatial language. In K. Coventry and P. Olivier, editors, *Spatial Language: Cognitive and Computational Aspects*, pages 69–86. Kluwer Academic, Dordrecht, 2001.

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

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

[43] M. Kaur, M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi. Where is "it"? event synchronization in gaze-speech input systems. In *Proceedings of Fifth International Conference on Multimodal Interfaces*, pages 151–157. ACM Press, 2003.

[44] A. Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI00*, 2000.

[45] J. Kelleher, F. Costello, and J. A. van Genabith. Dynamically updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence Journal*, 67(1–2):62–102, 2005.

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

[47] L. Kievit, P. Piwek, R. Beun, and H. Bunt. Multimodal cooperative resolution of referential expressions in the DenK system. In *Lecture Notes In Computer Science. Revised Papers from the Second International Conference on Cooperative Multimodal Communication*, volume 2155, pages 197–216. 1998.

[48] C. Koch and L. Itti. Computational modelling of visual attention. *Nature Reviews: Neuroscience*, 2(3):194–203, 2001.

[49] D. B. Koons, C. J. Sparrell, and K. R. Thorisson. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent multimedia interfaces*, pages 257–276. American Association for Artificial Intelligence, 1993.

[50] F. Landragin, N. Bellalem, , and L. Romary. Visual salience and perceptual grouping in multimodal interactivity. In *First International Workshop on Information Presentation and Natural Multimodal Dialogue*, pages 151–155, Verona, Italy, 2001.

[51] M. Lange and H. Leiß. To CNF or not to CNF? An efficient yet presentable version of the CYK algorithm. *Informatica Didactica*, 8, 2009.

[52] D. Li, J. Babcock, and D. J. Parkhurst. openeyes: A low-cost head-mounted eye-tracking solution. In *Proceedings of the ACM Eye Tracking Research and Applications Symposium.*, 2006.

[53] T. Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

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

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

[56] E. Matin. Saccadic suppression: a review and an analysis. In *Psychological Bulletin*, volume 81, pages 899–917, 1974.

[57] P. McKevitt, editor. *Integration of Natural Language and Vision Processing*, volume I–IV. Kluwer Academic, Dordrecht, 1995/1996.

[58] A. Meyer, M. Böhme, T. Martinetz, and E. Barth. A single-camera remote eye tracker. In *Perception and Interactive Technologies*, pages 208–211, 2006.

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

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

[61] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. 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, 2003.

[62] J. G. Neal, C. Y. Thielman, Z. Dobes, S. M. Haller, and S. C. Shapiro. Natural language with integrated deictic and graphic gestures. In *Readings in intelligent user interfaces*, pages 37–52, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[63] U. Neisser. *Cognitive Psychology.* Appleton-Century-Crofts, 1967.

[64] J. Nielsen and K. Pernice. *Eyetracking Web Usability.* New Riders Publishing, Thousand Oaks, CA, USA, 2009.

[65] H. Noser, O. Renault, D. Thalmann, and N. Magnenat-Thalmann. Navigation for digital actors based on synthetic vision, memory and learning. *Computers and Graphics*, 19(1):7–19, 1995.

[66] T. Ohno and N. Mukawa. A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In *Proceedings of Eye Tracking Research and Applications (ETRA2004)*, pages 115–122, 2004.

[67] S. Oviatt and P. Cohen. Perceptual user interfaces: Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43:45–53, 2000.

[68] S.L. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings Of the Conference on Human Factors in Computing Systems.* ACM, 1999.

[69] B. Pedersen and M. Spivey. Offline tracking of eyes and more with a simple webcam. In *Proceedings of the workshop "What have eye movements told us so far, and what is next?", 28th Annual Meeting of the Cognitive Science Society*, 2006.

[70] C. Perlich, F. Provost, and J. S. Simonoff. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research*, 4:211–255, 2003.

[71] E. D. Petajan. *Automatic lipreading to Enhance Speech Recognition.* PhD thesis, University of Illinois at Urbana-Champaign., 1984.

[72] P. Piwek, R. J. Beun, and A. Cremers. 'proximal' and 'distal' in language and cognition: evidence from deictic demonstratives in dutch. *Journal of Pragmatics*, 40(4):694–718, 2008.

[73] I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3d synthetic agent. *Special Issue on Behavior Planning for Life-Like Characters and Avatars, Journal of AI Communications*, 13(3):169–181, 2000.

[74] Z. Prasov, J. Y. Chai, and H. Jeong. Eye gaze for attention prediction in multimodal human-machine conversation. Technical report, Proceedings of the AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants, 2007.

[75] S. Qu and J. Y. Chai. 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)*, 2007.

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

[77] P. Qvarfordt, D. Beymer, and S. Zhai. Realtourist - a study of augmenting human-human and human-computer dialogue with eye-gaze overlay. In *INTERACT 2005, LNCS 3585*, pages 767–780, 2005.

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

[79] S. Shih and J. Liu. A novel approach to 3-d gaze tracking using stereo cameras. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 34(1):234–245, 2004.

[80] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh. 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, 2004.

[81] Y. So. A tutorial on logistic regression. In *Proceedings Eighteenth Annual SAS Users Group International Conference*, 1993.

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

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

[84] M. K. Tanenhaus and J. C. Trueswell. *Approaches to Studying World-Situated Language Use*, chapter Eye Movements as a Tool for Bridging the Language-as-Product and Language-as-Action Traditions, pages 3–38. MIT Press, 2004.

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

[86] C. Ware and H. H. Mikaelian. An evaluation of an eye tracker as a device for computer input. In *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface (CHI'87)*, pages 183–188, New York, NY, USA, 1987. ACM.

[87] T. Winograd. A procedural model of language understanding. In R. C. Schank and K. M. Colby, editors, *Computer Models of Thought and Language*, pages 152–186. W.H. Freeman and Company, New York, 1973. Reprinted in Grosz et al. (1986), Readings in Natural Language Processing. Morgan Kaufman.

[88] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 98:25–51, 2005.

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

[90] S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'99)*, pages 246–253. ACM Press, 1999.

[91] H. Zhang and Jiang Su. Naive bayesian classifiers for ranking. In *Proceedings of the 15th European Conference on Machine Learning (ECML)*, 2004.