Speech, Space and Purpose: Situated Language Understanding in Computer Games

Peter Gorniak, Jeff Orkin, Deb Roy
{pgorniak,jorkin,dkroy}@media.mit.edu
MIT Media Laboratory, 20 Ames Street
Cambridge, MA 02139 USA

Abstract

Many common types of language understanding depend on situational context. In the extreme, utterances like "the one to the left of the green ones", "let's do that again" or "can you help me?" provide little content or restrictions through their words, but can be readily understood and acted upon by a human listener embedded in the same situation as the speaker. We describe a series of computational models of situated language understanding that take into account the context provided by a game the language users are playing. Starting with a game focusing on spatial disambiguation, we proceed to a model taking into account player's recognized intentions to perform referent disambiguation and end with a system that understands highly situated commands directly in terms of recognized plan fragments. Finally, we discuss our use of these models in building artificial agents that plan alongside the player in the game world and co-operate through language and their own initiative.

Introduction

The meaning of language often depends on its embedding situation. Especially interactive spoken language, such as occurs when giving directions, eating out or playing basketball, can rarely be fully understood without detailed knowledge of where, when and why it occurred. It is thus necessary to access and model the embedding situation in order to design computational models of situated language understanding, and to build systems that interact with human beings via situated language. Perceiving and acting in the real world, however, is a difficult problem, limited by the current sensing and manipulation abilities of machines. Only constrained and controlled situations can today be modeled by natural language understanding systems embedded in the real world, such as interactive robots. Computer games, on the other hand, have been growing in the complexity of the worlds they let players explore, as well as in popularity. These games create immersive experiences that often demand and encourage social and collaborative language use. While they are not a faithful model of real world physical interaction, they let us explore human language use in spatially rich and purposeful, yet easily sensed and controllable settings.

In this paper we describe a sequence of four games used as platforms for situated computational language understanding. Except for the last, our research using these games follows the same experimental setup: Two players, a speaker and a listener, play the game collaboratively using in-game actions and language (either speech or typed text). We record the actions together with the audio or the text message events, and in each case build a computer model of situated understanding that replaces the human listener. The system attempts to analyse the recorded language and situation, and predicts the listener's next action.

The first game is designed as a purely spatial game: the players are given no instructions but to describe objects to each other that are indistinguishable except for their spatial locations. While this study yields insight into parsing spatial language in terms of visual features, it creates an overly simplistic situation with respect to the purpose of the interaction. Much of the context in reference resolution, however, is provided by understanding purpose: if the speaker knows what the listener is trying to achieve, this knowledge limits the set of possible referents. In the next game, therefore, the players are not

only describing objects, but they do so to solve a puzzle. In that study we show that combining plan recognition with object reference substantially improves language understanding performance over using only one or the other. The third study does away with the distinction between reference resolution and general language understanding and demonstrates that highly situated language understanding can be achieved by grounding language directly in the recognized plans of players. Lastly, we describe a preliminary implementation of an artificial character in a puzzle solving game whose understanding and actions are driven by insights gleaned from the first three game scenarios.

Related Work

Much work in language understanding assumes a static, propositional knowledge base for language understanding [Montague, 1974], though some work explicitly acknowledges the need to model situations [Barwise and Perry, 1983]. Discourse and plans recognized from discourse have played a role in prior computational language understanding work [Grosz and Sidner, 1986, Allen and Perrault, 1980], but such work focuses mostly on the words said, not on the spatial, social and goal-dictated context of language. The state of a shared environment with the user does play a role in discourse understanding for some work in human-computer collaboration [Lesh et al., 1999, Rich and Sidner, 1998], but in this case the similarity to real world physical scenarios is very limited.

From the embodiment literature comes support for taking the physical situation as well as possible interactions into account during language understanding [Glenberg and Kaschak, 2002, Barsalou, 1999, Zwaan, 2003]. Real-world efforts at such types of language understanding using robots include work on grounding verbs in video data [Siskind, 2001], understanding newspaper articles in terms of action representations [Narayanan, 1997] and our own work on word learning in conjunction with perceptual input [Roy and Pentland, 2002, Roy, 2002, Roy et al., 2002]. Finally, language understanding studies and systems embodied in computer games and their characters are starting to appear [Byron and Stoia, 2005, Fleischman and Roy, 2005].

Games, Systems and Studies

As mentioned earlier, our language understanding studies have evolved from initially using systems that focus on modelling spatial language and phenomena, thus limiting the need for plans or purpose, to models that fully represent the situation in terms of the speaker's and listener's plans. We now sketch several studies and systems to illustrate this path and its insights.

Bishop: Space is Everything

Like all of our studies presented here, the Bishop task [Gorniak and Roy, 2004] allowed a pair of participants to use unconstrained language to accomplish a given task. In a single session, one participant acted as the speaker and one as the listener. Both could view a current scene of up to 30 objects, such as the one in Figure 1 on computer monitors arranged such that participants could not see each other. Both monitors always showed identical scenes. It was the speaker's task to pick one object on the screen, use the computer's mouse to select it, and then verbally describe the selected object to the listener. It was the listener's task to identify the object being described and select it on his or her own screen. If speaker and listener picked the same object, this object disappeared and the speaker moved on to select another object. If the selections did not match, the speaker got another chance to describe the object. The task is not hard, and listeners select the correct object 96.5% on first attempt.

As the objects are indistinguishable except for appearing in two colours, and are randomly arranged in the scene, speakers are forced to use the spatial configuration of the scene to distin-

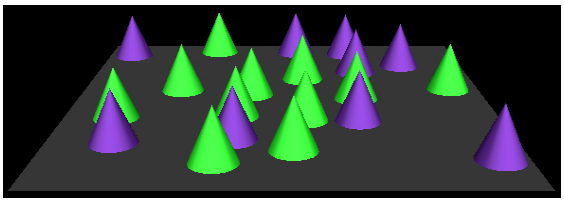

Figure 1: Sample Spatial Arrangement Used in the Bishop Task

guish objects. In analysing data collected in this task, we distinguish a set of *descriptive strategies*: a visual feature measured on the current scene, together with their linguistic realization. The descriptive strategies that cover most of the data are:

colour almost every utterance employs colour names ("green" or "purple")

spatial regions and extrema as in "the green one that's closest to us in the front" or "the purple one on the left side"

grouping participants used groups of objects ("the green cones") both to select within them, and to select relative to the group ("the one behind the three purple"

spatial relations phrases like "behind" or "to the left of" occur both in relation to individual object and using groups as landmarks

anaphora as the successfully identified object disappeared, participants would sometimes use its prior location as a reference point ("the one behind that one")

It is the goal of the Bishop system to model these observed descriptive strategies. The system, like all systems discussed here, employs a grounded semantic composition strategy to understand language in terms of its semantic embedding in the world. A robust chart parser finds islands of grammaticality in the transcribed and often ungrammatical speech. Whenever the parser completes a grammatical constituent, it uses entries from a visually grounded lexicon designed from the data to interpret the constituent in terms of the visual scene, producing what we call a *concept*. In the Bishop task, concepts are ranked subsets of the objects present in the current scene. At the word level, the parsing process starts out with concepts referring to all objects in the scene. The lexical entries then apply to these concepts, or the concepts produced by other constituents, to produce new concepts - in essence filtering the set of possible referents as parsing progresses. For example, the lexical entry for "left" would produce a concept ranking objects by their horizontal position in the scene, whereas in the constituent "the green one on the left" it would be applied to the set of all green objects (a concept produced by "green one"), ranking only these objects in terms of their horizontal position. We will encounter this method of incremental situated semantic interpretation throughout the systems presented in this paper.

Lexical entries have to cover many linguistic and semantic subtleties, and they do cover all of them in the Bishop system. For example, the Bishop system distinguishes between relative and absolute position to distinguish otherwise similar terms like "left" and "middle". As in the example above, participants usually understand "the green one on the left" to mean the the green object furthest to the left, independent of its absolute horizontal position, whereas "middle" often picks out an object in the absolute centre of the scene. However, "middle" is also used in other ways, without any linguistic phenomena indicating a difference: it can indicate the object in the absolute middle of the scene, the object in the middle of a visually salient group of objects, the object between two salient groups of objects (or between two objects), among other uses. Despite not covering all such subtle influences on visual word meanings, the Bishop system selects the correct target object over 80% of the time on data the system was built from and over 70% on an independent test set of utterances with descriptive strategies it has implemented (versus a random

baseline of about 13%).

The Bishop system thus represents a powerful computational model of human spatial perception and language for the Bishop task. The task, however, is limited in that it was specifically designed to elicit spatial language to the neglect of other influences of the situation at hand on human language use. In most situations, space is not the only meaning-providing context for a situated utterance. The shared intentions of speaker and listener are a far more prevalent context. In the next study, we thus turn from applying the Bishop framework to a purely visual-spatial understanding task to one producing concepts that also take into account the speaker's intentions.

Figure 2: Screenshot from Bioware's Neverwinter Nights

Purposeful Reference

The Bishop framework assumes a relatively clean transcript of human speech, and a single purpose to each utterance: to refer to exactly one object in the current visual scene. Our next two studies and systems use a commercial game, Bioware's Neverwinter Nights (a screenshot from the study is shown in Figure 2). The first study works directly from speaker's acoustically ambiguous speech, and, while setting a high level goal, does not dictate utterance-by-utterance purpose to the speaker. Once more, one speaker and one listener participate in the study. The speaker is given a high level instruction ("light both fires in the first room") and then left to experiment as to how to accomplish this goal. Speaker and listener each control one character in the game, and the design of the puzzle requires the speaker to take advantage of the otherwise inactive listener by issuing commands. For example, to enter a room with a chest containing a needed key, one of the players must pull a level that opens the door to the room, while the other has to walk into the room before the door closes again. The puzzle contains several identical items (for example, three levers and two doors) that serve different purposes during the puzzle solution.

It was the goal of this study to show that taking into account speech and intention at the same time accomplishes understanding of situated utterances that are ambiguous when taking into account only one or the other. To handle the noisy acoustics, we run a probabilistic language parser on the confusion networks [Mangu et al., 1999] produced by a speech recognizer. This parser produces an estimate of the likelihoods of possible words at different times in the speech stream, given their acoustic and grammatical probabilities. Similarly to the Bishop parser, this parser creates a concept whenever it completes a constituent. Possible referents for these concepts are in-game items such as levers and doors, and the groundings of words in the situation are again encoded in the lexicon. As this system employs a probabilistic parser, a concept now is made up of a set of probabilities that represent the likelihood of a fragment of speech referring to a particular object in the game, given the acoustic, grammatical and referential properties of the fragment.

To capture the speaker's intention, we capture high level game events such as characters' movements between rooms, and perform probabilistic hierarchical plan recognition on this sequence of events. The plan recognizer captures higher level events and assigns probabilities to its interpretations of the events stream. For ex-

ample, the three consecutive low level events of one player pulling a lever, a door opening, and the other player changing rooms all together constitute a co-operative room change event. The plan recognizer also computes the probabilities of all possible events that players could engage in next. We can read these probabilities as the recognized intentions of the players. For example, if one player just opened a door by pulling a lever, the plan recognizer might predict that it is the other player's intention to walk through this door. Thus, we can rank the intentions by the probabilities the plan recognizer assigns to them. We then translate these probabilities involving intentions into probabilities involving objects simply by considering which objects a certain action involves - pulling a certain lever involves that lever and the door it opens.

Having arrived at a concept that provides the probabilities of reference of speech segments, and the predicted intentions of players, we integrate these probabilities via Bayes' theorem to yield the probability of reference of a segment of speech given the combined acoustic, grammatical, referential and intentional information gathered from the speech signal and the game. We have evaluated this method by predicting the referents of noun phrases in a data set collected from participants solving the puzzle, and shown that combining intention recognition and reference resolution in this way improves reference resolution dramatically (as much as 100%) over performing only intention recognition or only reference resolution in this task.

This study and the associated computational framework show that recognizing speakers' intentions can be the key to situated language understanding in collaborative scenarios, and that the compositional parsing framework introduced in the Bishop system extends beyond the visual task to cover noisy speech in a game environment. It also shows that through a probabilistic extension this system can coherently integrate external information such as intentional probabilities. It is a shortcoming of extending the Bishop system in this way, however, that intentional information is considered external to the reference resolution process. In fact, one would think it was at the core of reference resolution. Furthermore, it is unclear how to expand this system beyond reference to objects, because the concepts it produces, just like in the Bishop system, contain probabilities over objects. To alleviate these problems we need a new representational basis for concepts that does not have object reference as its core.

Referring to Intentions

Even when restricting the situated language understanding task to commands, reference resolution covers only a fraction of the linguistic and semantic phenomena that occur in the game tasks studied here. Players use utterance such as "let's do that again" or even "now!" as commands, both of which do not contain noun phrases that clearly refer to objects in the world. Furthermore, the post-hoc integration of intentions into the reference resolution process in the previous study goes against our intuitions about the central nature of intentions in language understanding.

To handle a wider range of commands and to put intentions at the core of the concepts in our language understanding systems, we turn to the notion of *affordances* [Gibson, 1977] and introduce the notion of *Affordance-Based Concepts* (ABCs) [Gorniak, 2005, Gorniak and Roy, 2006]. The task players engage in is similar to the one in the last study, though the current study analyses players' typed text instead of speech. As before, we perform probabilistic hierarchical plan recognition on high level game events, and parse player's utterances to produce concepts for grammatical constituents. The fundamental difference lies in the fact that the concepts produced by the parser do not contain probabilities over objects anymore, but rather contain probabilities over recognized intentions. Thus each prediction the plan recognizer makes, such as the prediction that a player will walk into

another room or pull a level, is itself a possible component of an ABC. When the parser encounters a word like "door", its lexical entry selects all intentional predictions that the plan recognizer has ever made that involve doors. These might be predictions of players breaking, unlocking, opening and walking through doors. Of course, not all of these predictions came or will come to pass. However, they closely correspond to the notion of affordances: players' possible interactions with the world, the nature of which depends both on the player (for example, his or her physical location, abilities and possessions) and the world (for example, which lever open which doors, which doors can be broken down.) As the parser forms more complex grammatical constituents, it filters concepts as the Bishop parser did. For example, "open the door on the left" might filter for the predicted opening interactions with doors the listener can engage in if she moves towards her left.

The combination of linguistic parser and plan recognizer thus produces concepts that are bundles of ranked affordances. These concepts naturally predict actions, and constitute a rich interpretation of reference (a "door" is the set of all possible interactions the listener could engage in with doors). Beyond reference, this framework covers utterances like "let's do that again" or "now!" - the first is interpreted as a command to repeat the last joint interactions of the two players with the world, whereas the second simply picks the most likely current affordance as its prediction. Once more we have evaluated this approach by predicting the actions of human listeners in response to commands using unseen data. The system achieves 70% accuracy versus a 14% random baseline, showing that the notion of Affordance-Based Concepts serves to capture the relevant intentional aspects of a given situation to interpret otherwise ambiguous language use.

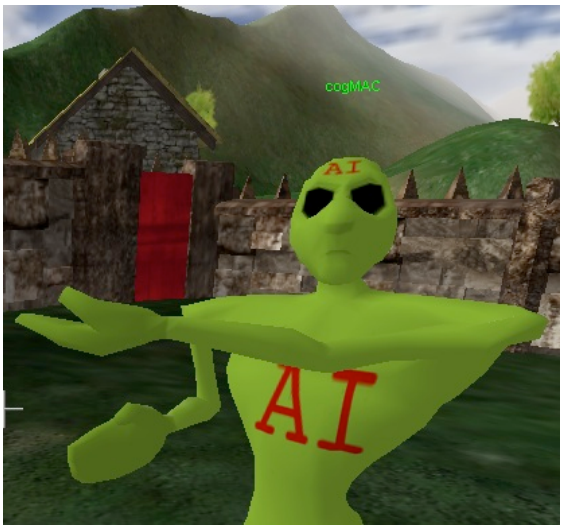

Figure 3: One of the collaborative synthetic game characters

Intention-Recognizing Collaborative Characters

So far our studies have involved two human beings, and have attempted to predict the actions of the human listener in the data collected. One of our goals, however, is to produce artificial agents (like the one in Figure 3) that dynamically take the evolving situation into account when understanding their human collaborator's intentions. To this end, we have transferred the implementation of Affordance-Based Concepts to several interactive game scenarios. In one scenario, a single player has the option of accomplishing a given task in three different ways, each of which involves several steps that require two characters to perform actions simultaneously. The player can ask for help at several points during the game, and the artificial character will attempt to perform the correct collaborative action (for example, opening the door for a player locked in another room or lighting the forge for a player attempting to smelt a key when asked "can you help me with this?"). The character's response is based

on its intention recognition engine. In another scenario, we have added another layer of intention recognition to the character, which lets us loosen the assumption that the player's overall goal is known. The character will now watch the player interact with the world and based on the player's actions guess the player's overall goal (for example: to put one box into every room of the game). While continuously re-evaluating its estimate of the highest-level goal, the character will then collaborate with the player by autonomously planning towards this goal and taking suitable actions to bring the current world state closer to the target state. In this case, it does not require an overall puzzle solution, but rather formulates plan fragments that contribute to the identified goal.

Both of these scenarios are examples of transferring the computational models of language understanding and intention recognition introduced in the previous sections to modern day game scenarios. We believe that intention recognition and intention-based situated language understanding are not only feasible, but essential for producing games with entertaining and non-frustrating non-player characters.

Conclusion

We have presented a series of studies that all share a similar design and language interpretation framework. In content, however, the studies and systems progress from a purposeless, purely spatial task and computational model, to taking into account the intentions behind an utterance in an external manner, to making intention recognition and affordances the core of a situated language understanding system. All of the systems perform well on their specific tasks, but incorporating intention recognition lets us coherently cover a large set of linguistic and intentional phenomena that are not addressed by purely spatial approaches. We have also shown the applicability of these approaches to game environments with synthetic characters.

In the future, we intend to integrate work on discourse into our language understanding systems, and to build more flexible agents that can adapt to new game situations without knowing how the game works ahead of time.

References

[Allen and Perrault, 1980] Allen, J. and Perrault, R. (1980). Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178.

[Barsalou, 1999] Barsalou, L. (1999). Perceptual symbol systems. *Behavioural and Brain Sciences*, 22(4):577–609.

[Barwise and Perry, 1983] Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge, MA.

[Byron and Stoia, 2005] Byron, D. K. and Stoia, L. (2005). An analysis of proximity markers in collaborative dialog. In *Proceedings of the 41st annual meeting of the Chicago Linguistic Society*.

[Fleischman and Roy, 2005] Fleischman, M. and Roy, D. (2005). Why are verbs harder to learn than nouns? In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

[Gibson, 1977] Gibson, J. (1977). The theory of affordances. In Shaw, R. and Bransford, J., editors, *Perceiving, Acting and Knowing*, pages 67–82. Wiley, New York.

[Glenberg and Kaschak, 2002] Glenberg, A. M. and Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin and Review*, 9(3):558–565.

[Gorniak, 2005] Gorniak, P. (2005). *The Affordance-Based Concept*. PhD thesis, Massachusetts Institute of Technology.

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

[Gorniak and Roy, 2006] Gorniak, P. J. and Roy, D. (2006). Perceived affordances as a substrate for linguistic concepts. In *to appear in Proceedings of Cognitive Science*.

[Grosz and Sidner, 1986] Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

[Lesh et al., 1999] Lesh, N., Rich, C., and Sidner, C. L. (1999). Using plan recognition in human-computer collaboration. In *User Modeling:Proceedings of the 7th International Conference, UM99*.

[Mangu et al., 1999] Mangu, L., Brill, E., and Stolcke, A. (1999). Finding consensus among words: Lattice-based word error minimization. In *Proceedings of EUROSPEECH'99*, volume 1, pages 495–498, Budapest.

[Montague, 1974] Montague, R. (1974). *Formal Philosophy*. Yale University Press.

[Narayanan, 1997] Narayanan, S. (1997). *KARMA: Knowledge-based Action Representations for Metaphor and Aspect*. PhD thesis, University of California, Berkeley.

[Rich and Sidner, 1998] Rich, C. and Sidner, C. L. (1998). Collagen: A collaboration manager for software agents.

[Roy, 2002] Roy, D. (2002). Learning words and syntax for a visual description task. *Computer Speech and Language*, 16:353–385.

[Roy et al., 2002] Roy, D., Gorniak, P. J., Mukherjee, N., and Juster, J. (2002). A trainable spoken language understanding system. In *Proceedings of the International Conference of Spoken Language Processing*.

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

[Siskind, 2001] Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15:31–90.

[Zwaan, 2003] Zwaan, R. A. (2003). The immersed experiencer: Toward an embodied theory of language comprehension. *The Psychology of Learning and Motivation*, 44.