Intentional Context in Situated Natural Language Learning

Michael Fleischman and Deb Roy
Cognitive Machines
The Media Laboratory
Massachusetts Institute of Technology
mbf@mit.edu, dkroy@media.mit.edu

Abstract

Natural language interfaces designed for situationally embedded domains (e.g. cars, videogames) must incorporate knowledge about the users' context to address the many ambiguities of situated language use. We introduce a model of situated language acquisition that operates in two phases. First, intentional context is represented and inferred from user actions using probabilistic context free grammars. Then, utterances are mapped onto this representation in a noisy channel framework. The acquisition model is trained on unconstrained speech collected from subjects playing an interactive game, and tested on an understanding task.

1 Introduction

As information technologies move off of our desktops and into the world, the need for Natural Language Processing (NLP) systems that exploit information about the environment becomes increasingly apparent. Whether in physical environments (for cars and cell phones) or in virtual ones (for videogames and training simulators), applications are beginning to demand language interfaces that can understand unconstrained speech about constrained domains. Unlike most text-based NLP research, which focuses on open-domain problems, work we refer to as *situated* NLP focuses on improving language processing by exploiting domain-specific information about the non-linguistic situational context of users' interactions. For applications where agents interact in shared environments, such information is critical for successful communication.

Previous work in situated NLP has focused on methods for grounding the meaning of words in physical and virtual environments. The motivation for this work comes from the inability of text-based NLP technologies to offer viable models of semantics for human computer interaction in shared environments. For example, imagine a situation in which a human user is interacting with a robotic arm around a table of different colored objects. If the human were to issue the command "give me the blue one," both the manually-coded (Lenat, 1995; Fellbaum, 1998) and statistical models (Manning and Schutze, 2000) of meaning employed in text-based NLP are inadequate; for, in both models, the meaning of a word is based only on its relations to other words. However, in order for the robot to successfully "give me the blue one," it must be able to link the meaning of the words in the utterance to its perception of the environment (Roy, Hsiao, & Mavridis, 2004). Thus, recent work on grounding meaning has focused on how words and utterances map onto physical descriptions of the environment: either in the form of perceptual representations (Roy, in press, Siskind, 2001, Regier, 1996) or control schemas (Bailey, 1997 Narayanan, 1999).[1]

While such physical descriptions are useful representations for some classes of words (e.g., colors, shapes, physical movements), they are insufficient for more abstract language, such as that which denotes intentional action. This insufficiency stems from the fact that intentional actions (i.e. actions performed with the purpose of achieving a goal) are highly ambiguous when described only in terms of their physically observable characteristics. For example, imagine a situation in which one person moves a cup towards another person and utters the unknown word

[1] Note that Narayanan's work moves away from purely physical to metaphorical levels of description.

"*blicket*." Now, based only on the physical description of this action, one might come to think of "blicket" as meaning anything from "give cup", to "offer drink", to "ask for change." This ambiguity stems from the lack of contextual information that strictly perceptual descriptions of action provide.

This research presents a methodology for modeling the intentional context of utterances and describes how such representations can be used in a language learning task. We decompose language learning into two phases: intention recognition and linguistic mapping. In the first phase, we model intentional action using a probabilistic context free grammar. We use this model to parse sequences of observed physical actions, thereby inferring a hierarchical tree representation of a user's intentions. In the second phase, we use a noisy channel model to learn a mapping between utterances and nodes in that tree representation. We present pilot situated language acquisition experiments using a dataset of paired spontaneous speech and action collected from human subjects interacting in a shared virtual environment. We evaluate the acquired model on a situated language understanding task.

2 Intention Recognition

The ability to infer the purpose of others' actions has been proposed in the psychological literature as essential for language learning in children (Tommasello, 2003, Regier, 2003). In order to understand how such intention recognition might be modeled in a computational framework, it is useful to specify the types of ambiguities that make intentional actions difficult to model. Using as an example the situation involving the cup described above, we propose that this interaction demonstrates two distinct types of ambiguity. The first type, which we refer to as a *vertical ambiguity* describes the ambiguity between the "move cup" vs. "offer drink" meanings of "blicket." Here the ambiguity is based on the level of description that the speaker intended to convey. Thus, while both meanings are correct (i.e., both meanings accurately describe the action), only one corresponds to the word "blicket."

The second type of ambiguity, referred to as *horizontal ambiguity* describes the ambiguity between the "offer drink" vs. "ask for change" interpretations of "blicket." Here there is an ambiguity based on what actually is the intention behind the physical action. Thus, it is the case that only one of these meaning corresponds to "blicket" *and* the other meaning is not an accurate description of the intended action.

Figure 1 shows a graphical representation of these ambiguities. Here the leaf nodes represent a basic physical description of the action, while the root nodes represent the highest-level actions for which the leaf actions were performed[2]. Such a tree representation is useful in that it shows both the horizontal ambiguity that exists between the nodes labeled "ask for change" and "offer drink," as well as the vertical ambiguity that exits between the nodes labeled "offer drink" and "move cup."

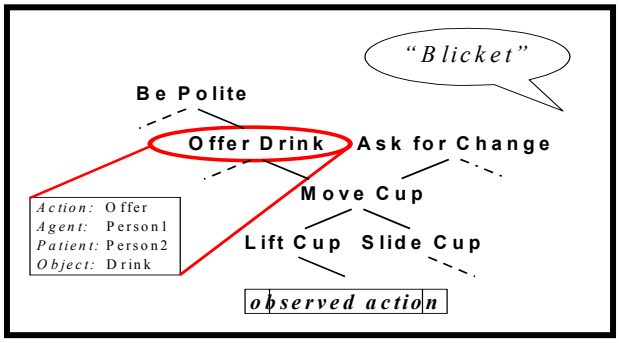

Figure 1: Graphical representation of vertical and horizontal ambiguities for actions.

In order to exploit the intuitive value of such a tree representation, we model intention recognition using probabilistic context free grammars (PCFG)[3]. We develop a small set of production rules in which the left hand side represents a higher order intentional action (e.g., "offer drink"), and the right hand side represents a sequence of lower level actions that accomplish it (e.g. "grasp cup", "move cup", "release cup"). Each individual action (i.e. letter in the alphabet of the PCFG) is further modeled as a simple semantic frame that contains roles for an agent, an object, an action, and multiple optional modifier roles (see inset figure 1). While in this initial work productions are created by hand (a task made feasible by the

[2] In other words, high-level actions (e.g. "be polite) are preformed *by means of* the performance of lower-level actions (e.g. "offer drink").

[3] The idea of a "grammar of behavior" has a rich history in the cognitive sciences dating back at least to Miller et al., 1960

constrained nature of situated domains) learning such rules automatically is discussed in section 4.2.

Just as in the plan recognition work of Pynadath, (1999), we cast the problem of intention recognition as a probabilistic parsing problem in which sequences of physical actions are used to infer an abstract tree representation. Resolving horizontal ambiguities thus becomes equivalent to determining which parse tree is most likely given a sequence of events. Further, resolving vertical ambiguities corresponds to determining which level node in the inferred tree is the correct level of description that the speaker had in mind.

3 Linguistic Mapping

Given a model of intention recognition, the problem for a language learner becomes one of mapping spoken utterances onto appropriate constituents of their inferred intentional representations. Given the intention representation above, this is equivalent to mapping all of the words in an utterance to the role fillers of the appropriate semantic frame in the induced intention tree. To model this mapping procedure, we employ a noisy channel model in which the probability of inferring the correct meaning given an utterance is approximated by the (channel) probability of generating that utterance given that meaning, times the (source) prior probability of the meaning itself (see Equation 1).

$$p(meaning \mid utterance) \approx p(utterance \mid meaning)^{\alpha} \bullet p(meaning)^{(1-\alpha)} \quad \textbf{(1)}$$

Here *utterance* refers to some linguistic unit (usually a sentence) and *meaning* refers to some node in the tree (represented as a semantic frame) inferred during intention recognition[4]. We can use the probability associated with the inferred tree (as given by the PCFG parser) as the source probability. Further, we can learn the channel probabilities in an unsupervised manner using a variant of the EM algorithm similar to machine translation (Brown et al., 1993), and statistical language understanding (Epstein, 1996).

4 Pilot Experiments

4.1 Data Collection

In order to avoid the many physical and perceptual problems that complicate work with robots and sensor-grounded data, this work focuses on language learning in virtual environments. We focus on multiplayer videogames , which support rich types of social interactions. The complexities of these environments highlight the problems of ambiguous speech described above, and distinguish this work from projects characterized by more simplified worlds and linguistic interactions, such as SHRDLU (Winograd, 1972). Further, the proliferation of both commercial and military applications (e.g., Rickel et al., 2002) involving such virtual worlds suggests that they will continue to become an increasingly important area for natural language research in the future.

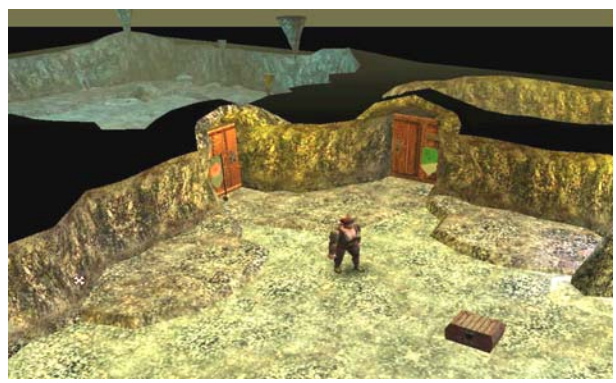

Figure 2: Screen shot of Neverwinter Nights game used in experimentation.

In order to test our model, we developed a virtual environment based on the multi-user videogame Neverwinter Nights.[5] The game, shown in Figure 2, provides useful tools for generating modules in which players can interact. The game was instrumented such that all players' speech/text language and actions are recorded during game play. For data collection, a game was designed in which a single player must navigate their way through a cavernous world, collecting specific objects, in order to escape. Subjects were paired such that one, the *novice*, would control the virtual character, while the other, the *expert,* guided her through the world. While the expert could say anything in order to tell the novice where to go and what to do, the novice was instructed not to speak, but only to follow the commands of the expert.

[4] α refers to a weighting coefficient.

[5] http://nwn.bioware.com/

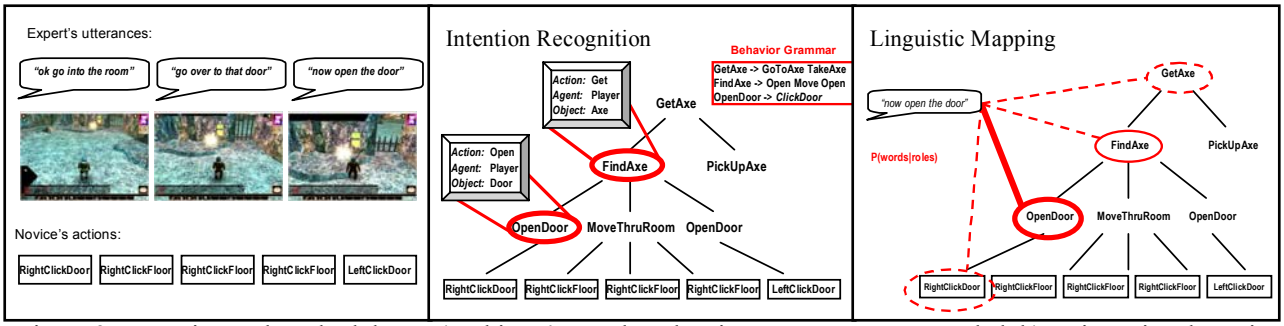

Figure 3. Experimental methodology: a) subjects' speech and action sequences are recorded; b) an intentional tree is inferred over the sequence of observed actions using a PCFG parser; c) the linguistic mapping algorithm examines the mappings between the utterance and all possible nodes to learn the best mapping of words given semantic roles.

The purpose behind these restrictions was to elicit free and spontaneous speech that is only constrained by the nature of the task. This environment seeks to emulate the type of speech that a real situated language system might encounter: i.e., natural in its characteristics, but limited in its domain of discourse.

The subjects in the data collection were university graduate and undergraduate students. Subjects (8 male, 4 female) were staggered such that the novice in one trial became the expert in the next. Each pair played the game at least five times, and for each of those trials, all speech from the expert and all actions from the novice were recorded. Table 1 shows examples of utterances recorded from game play, the observed actions associated with them, and the actions' inferred semantic frame.

Utterance	Action	Frame
ok this time you are gonna get the axe first	MOVE ROOM1	*act*: GET *obj*: AXE
through the red archway on your right	MOVE ROOM2	act: MOVE *goal*: ARCH *manner*: THRU
now open that door	CLICK_ON LEVER	*act*: OPEN *obj*: DOOR
ok now take the axe	CLICK_ON CHEST	*act*: TAKE *obj*: AXE *source*: CHEST

Table 1: Representative test utterances collected from subjects with associated game actions and frames

Data collection produces two parallel streams of information: the sequence of actions taken by the novice and the audio stream produced by the expert (figure 3a). The audio streams are automatically segmented into utterances using a speech endpoint detector, which are then transcribed by a human annotator. Each action in the sequence is then automatically parsed, and each node in the tree is replaced with a semantic frame (figure 3b).[6] The data streams are then fed into the linguistic mapping algorithms as a parallel corpus of the expert's transcribed utterances and the inferred semantic roles associated with the novice's actions (figure 3c).

4.2 Algorithms

Intention Recognition

As described in section 2, we represent the task model associated with the game as a set of production rules in which the right hand side consists of an intended action (e.g. "find key") and the left hand side consists of a sequence of sub-actions that are sufficient to complete that action (e.g. "go through door, open chest, pick_up key"). By applying probabilities to the rules, intention recognition can be treated as a probabilistic context free parsing problem, following Pynadath, 1999. For these initial experiments we have hand-annotated the training data in order to generate the grammar used for intention recognition, estimating their maximum likelihood probabilities over the training set. In future work, we intend to examine how such grammars can be learned in conjunction with the language itself; extending research on learning task models (Nicolescu and Mataric, 2003) and work on learning PCFGs (Klein and Manning, 2004) with our own work on unsupervised language learning.

Given the PCFG, we use a probabilistic Earley parser (Stolcke, 1994), modified slightly to output

[6] We use 65 different frames, comprised of 35 unique role fillers.

partial trees (with probabilities) as each action is observed. Figure 4 shows a time slice of an inferred intention tree after a player mouse clicked on a lever in the game. Note that both the vertical and horizontal ambiguities that exist for this action in the game parallel the ambiguities shown in Figure 1. As described above, each node in the tree is represented as a semantic frame (see figure 4 insets), whose roles are aligned to the words in the utterances during the linguistic mapping phase.

Linguistic Mapping

The problem of learning a mapping between linguistic labels and nodes in an inferred intentional tree is recast as one of learning the channel probabilities in Equation 1. Each node in a tree is treated as a simple semantic frame and the role fillers in these frames, along with the words in the utterances, are treated as a parallel corpus. This corpus is used as input to a standard Expectation Maximization algorithm that estimates the probabilities of generating a word given the occurrence of a role filler. We follow IBM Model 1 (Brown et al., 1993) and assume that each word in an utterance is generated by exactly one role in the parallel frame

Using standard EM to learn the role to word mapping is only sufficient if one knows to which level in the tree the utterance should be mapped. However, because of the vertical ambiguity inherent in intentional actions, we do not know in advance which is the correct utterance-to-level mapping. To account for this, we extend the standard EM algorithm as follows (see figure 3c):

1) set uniform likelihoods for all utterance-to-level mappings
2) for each mapping, run standard EM
3) merge output distributions of EM (weighting each by its mapping likelihood)
4) use merged distribution to recalculate likelihoods of all utterance-to-level mappings
5) goto step 2

4.3 Experiments

Methodologies for evaluating language acquisition tasks are not standardized. Given our model, there exists the possibility of employing intrinsic measures of success, such as word alignment accuracy. However, we choose to measure the success of learning by examining the related (and more natural) task of language understanding.

For each subject pair, the linguistic mapping algorithms are trained on the first four trials of game play and tested on the final trial. (This gives on average 130 utterances of training data and 30 utterances of testing data per pair.) For each utterance in the test data, we calculate the likelihood that it was generated by each frame seen in testing. We select the maximum likelihood frame as the system's hypothesized meaning for the test utterance, and examine both how often the maximum likelihood estimate exactly matches the true frame (*frame accuracy*), and how many of the role fillers within the estimated frame match the role fillers of the true frame (*role accuracy*).[7]

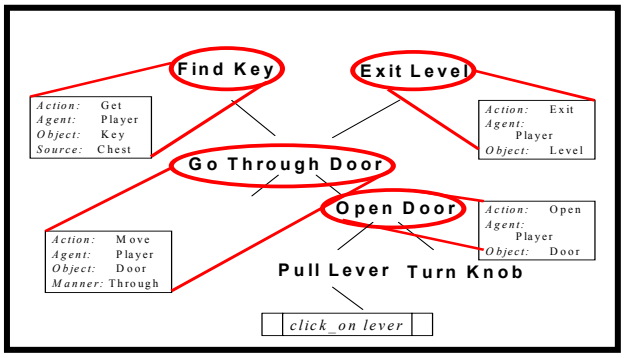

Figure 4: Inferred intention tree (with semantic frames) from human subject game play.

For each subject, the algorithm's parameters are optimized using data from all *other* subjects. We assume correct knowledge of the temporal alignment between utterances and actions. In future work, we will relax this assumption to explore the effects of not knowing which actions correspond to which utterances in time.

To examine the performance of the model, three experiments are presented. Experiment 1 examines the basic performance of the algorithms on the language understanding task described above given uniform priors. The system is tested under two conditions: 1) using the extended EM algorithm given an *unknown* utterance-to-level alignment, and 2) using the standard EM algorithm given the *correct* utterance-to-level alignment.

Experiment 2 tests the benefit of incorporating intentional context directly into language understanding. This is done by using the parse probability of each hypothesized intention as the

[7] See Fleischman and Roy (2005) for experiments detailing performance on specific word categories.

source probability in Equation 1. Thus, given an utterance to understand, we cycle through all possible actions in the grammar, parse each one as if it were observed, and use the probability generated by the parser as its prior probability. By changing the weighting coefficient (α) between the source and channel probabilities, we show the range of performances of the system from using no context at all (α=1) to using only context itself (α=0) in understanding.

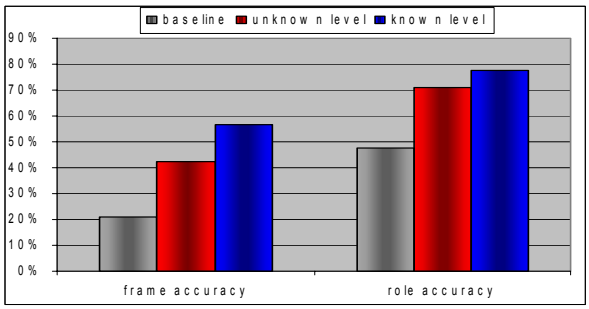

Figure 5: Comparison of models trained with utterance-to-level alignment both known and unknown. Performance is on a language understanding task (baseline equivalent to choosing most frequent frame)

Experiment 3 studies to what extent inferred tree structures are necessary when modeling language acquisition. Although, in section 1, we have presented intuitive reasons why such structures are required, one might argue that inferring trees over sequences of observed actions might not actually improve understanding performance when compared to a model trained only on the observed actions themselves. This hypothesis is tested by comparing a model trained given the correct utterance-to-level alignment (described in experiment 1) with a model in which each utterance is aligned to the leaf node (i.e. observed action) below the correct level of alignment. For example, in figure 4, this would correspond to mapping the utterance "*go through the door*", not to "GO THROUGH DOOR", but rather to "CLICK_ON LEVER."

4.4 Results

Experiment 1: We present the average performance over all subject pairs, trained with the correct utterance-to-level alignment both known and unknown, and compare it to a baseline of choosing the most frequent frame from the training data. Figure 5 shows the percentage of maximum likelihood frames chosen by the system that exactly match the intended frame (frame accuracy), as well as, the percentage of roles from the maximum likelihood frame that overlap with roles in the intended frame (role accuracy).

As expected, the understanding performance goes down for both frames and roles when the correct utterance-to-level alignment is unknown. Interestingly, while the frame performance declines by 14.3%, the performance on roles only declines 6.4%. This difference is due primarily to the fact that, while the mapping from words to action role fillers is hindered by the need to examine all alignments, the mapping from words to object role fillers remains relatively robust. This is due to the fact that while each level of intention carries a different action term, often the objects described at different levels remain the same. For example, in figure 4, the action fillers "TAKE", "MOVE", "OPEN", and "PULL" occur only once along the path. However, the object filler "DOOR" occurs multiple times. Thus, the chance that the role filler "DOOR" correctly maps to the word "*door*" is relatively high compared to the role filler "OPEN" mapping to the word "*open.*"[8]

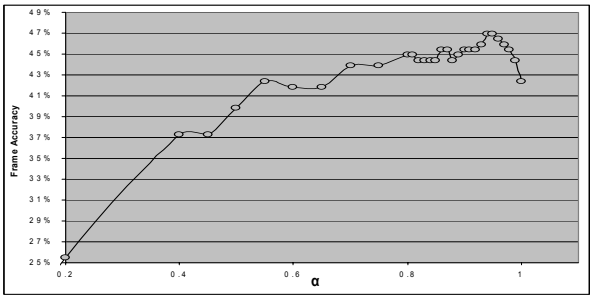

Figure 6: Frame accuracy as a function of α value (Eq. 1) trained on unknown utterance-to-level alignments.

Experiment 2: Figure 6 shows the average frame accuracy of the system trained without knowing the correct utterance-to-level alignment, as a function of varying the α values from Equation 1. The graph shows that including intentional context does improve system performance when it is not given too much weight (i.e., at relatively high alpha values). This suggests that the benefit of intentional context is somewhat outweighed by the power of the learned role to word mappings.

[8] This asymmetry for learning words about actions vs. objects is well known in psychology (Gleitman, 1990) and is addressed directly in Fleischman and Roy, 2005.

Looking closer, we find a strong negative correlation (r=-0.81) between the understanding performance using only channel probabilities (α=1) and the improvement obtained by including the intentional context. In other words, the better one does without context, the less context improves performance. Thus, we expect that in noisier environments (such as when speech recognition is employed) where channel probabilities are less reliable, employing intentional context will be even more advantageous.

Experiment 3: Figure 7 shows the average performance on both frame and role accuracy for systems trained without using the inferred tree structure (on leaf nodes only) and on the full tree structure (given the correct utterance-to-level alignment). Baselines are calculated by choosing the most frequent frame from training.[9]

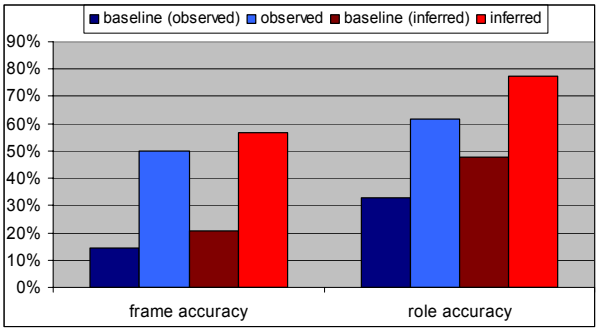

Figure 7: Comparison of models trained on inferred intentional tree vs. directly on observed actions

It is clear from the figure that understanding performance is higher when the intentional tree is used in training. This is a direct result of the fact that speakers often speak about high-level intentions with words that do not directly refer to the observed actions. For example, after opening a door, experts often say: "*go through the door*," for which the observed action is a simple movement (e.g., "MOVE ROOMx"). Also, by referring to high-level intentions, experts can describe sequences of actions that are not immediately referred to. For example, an expert might say: "*get the key*" to describe a sequence of actions that begins with "CLICK_ON CHEST." Thus, the result of not learning over a parsed hierarchical representation of intentions is increased noise, and subsequently, poorer understanding performance.

5 Discussion

The results from these experiments, although preliminary, indicate that this model of language acquisition performs well above baseline on a language understanding task. This is particularly encouraging given the unconstrained nature of the speech on which it was trained. Thus, even free and spontaneous speech can be handled when modeling a constrained domain of discourse.[10]

In addition to performing well given difficult data, the experiments demonstrate the advantages of using an inferred intentional representation both as a contextual aid to understanding and as a representational scaffolding for language learning. More important than these preliminary results, however, is the general lesson that this work suggests about the importance of knowledge representations for situated language acquisition.

As discussed in section 2, learning language about intentional action requires dealing with two distinct types of ambiguity. These difficulties cannot be handled by merely increasing the amount of data used, or switching to a more sophisticated learning algorithm. Rather, dealing with language use for situated applications requires building appropriate knowledge representations that are powerful enough for unconstrained language, yet scalable enough for practical applications. The work presented here is an initial demonstration of how the semantics of unconstrained speech can be modeled by focusing on constrained domains.

As for scalability, it is our contention that for situated NLP, it is not a question of being able to scale up a single model to handle open-domain speech. The complexity of situated communication requires the use of domain-specific knowledge for modeling language use in different contexts. Thus, with situated NLP systems, it is less productive to focus on how to scale up single models to operate beyond their original domains. Rather, as more individual applications are tackled (e.g. cars,

[9] Note that baselines are different for the two conditions, because there are a differing number of frames used in the leaf node only condition.

[10] Notably, situated applications for which natural language interfaces are required typically have limited domains (e.g., talking to one's car doesn't require open-domain language processing).

phones, videogames, etc.) the interesting question becomes one of how agents can learn to switch between different models of language as they interact in different domains of discourse.

6 Conclusion

We have introduced a model of language acquisition that explicitly incorporates intentional contexts in both learning and understanding. We have described pilot experiments on paired language and action data in order to demonstrate both the model's feasibility as well as the efficacy of using intentional context in understanding. Although we have demonstrated a first step toward an advanced model of language acquisition, there is a great deal that has not been addressed. First, what is perhaps most obviously missing is any mention of syntax in the language learning process and its role in bootstrapping for language acquisition. Future work will focus on moving beyond the IBM Model 1 assumptions, to develop more syntactically-structured models.

Further, although the virtual environment used in this research bears similarity to situated applications that demand NL interfaces, it is not known exactly how well the model will perform "in the real world." Future work will examine installing models in real world applications. In parallel investigations, we will explore our method as a cognitive model of human language learning.

Finally, as was mentioned previously, the task model for this domain was hand annotated and, while the constrained nature of the domain simplified this process, further work is required to learn such models jointly with language.

In summary, we have presented first steps toward tackling problems of ambiguity inherent in grounding the semantics of situated language. We believe this work will lead to practical applications for situated NLP, and provide new tools for modeling human cognitive structures and processes underlying situated language use (Fleischman and Roy, 2005).

Acknowledgments

Peter Gorniak developed the software to capture data from the videogame used in our experiments.

References

D. Bailey, J Feldman, S. Narayanan., & G. Lakoff.. Embodied lexical development. 19th Cognitive Science Society Meeting. Mahwah, NJ, 1997.

P. F. Brown, V. J. Della Pietra, S. A. Della Pietra & R. L. Mercer. "The Mathematics of Sta-tistical Machine Translation: Parameter Estimation," Computational Linguistics 19(2). 1993.

M Epstein Statistical Source Channel Models for Natural Language Understanding Ph. D. thesis, New York University, September, 1996

C. Fellbaum WordNet: An On-line Lexical Database and Some of its Applications. MIT Press, 1998.

M. Fleischman and D.K. Roy. *Why Verbs are Harder to Learn than Nouns*: Initial Insights from a Computational Model of Intention Recognition in Situated Word Learning. CogSci. Italy, 2005.

L. Gleitman. "The structural sources of verb meanings." Language Acquisition, 1(1), 1990.

D. Klein and C. Manning, "Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency", *Proc. of the 42nd ACL,* 2004

D. B Lenat,. CYC: A Large-Scale Investment in Knowledge Infrastructure". Comm. of ACM, 1995.

C. Manning, H. Schutze,. Foundations of Statistical Natural Language Processing. MIT Press, 2001.

G. A. Miller, E. Galanter, and K. Pribram 1960. Plans and the Structure of Behavior. New York: Halt.

S. Narayanan.. Moving right along: A computational model of metaphoric reasoning about events. In Proc. of AAAI. Orlando, FL, 1999.

M. Nicolescu, M. Mataric´, Natural Methods for Robot Task Learning: Instructive Demonstration, Generalization and Practice, AGENTS, Australia, 2003.

D. Pynadath, 1999. Probabilistic Grammars for Plan Recognition. Ph.D. Thesis, University of Michigan.

T. Regier. The human semantic potential. MIT Press, Cambridge, MA, 1996.

T. Regier. Emergent constraints on word-learning: A computational review. TICS, 7, 263-268, 2003.

J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum and W. Swartout, "Towards a New Generation of Virtual Humans for Interactive Experiences," in IEEE Intelligent Systems July/August 2002.

D.Roy, K. Hsiao, and N. Mavridis. Mental imagery for a conversational robot. IEEE Trans. on Systems, Man, and Cybernetics, 34(3) 2004.

D. Roy. (in press). Grounding Language in the World: Schema Theory Meets Semiotics. AI.

J. Siskind. Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. JAIR, 2001.

A. Stolcke. Bayesian Learning of Probabilistic Language Models. Ph.d., UC Berkeley, 1994.

M. Tomasello. Constructing a Language: A Usage-Based Theory of Language Acquisition. Harvard University Press, 2003.

T. Winograd. *Understanding Natural Language.* Academic Press, 1972.