Representing Intentions in a Cognitive Model of Language Acquisition: Effects of Phrase Structure on Situated Verb Learning

Michael Fleischman (mbf@mit.edu)
Deb Roy (dkroy@media.mit.edu)
The Media Laboratory
Massachusetts Institute of Technology

Abstract

A recent trend in the cognitive sciences is the development of models of language acquisition in which word meaning is grounded in the learner's perceptions and actions. Such physical descriptions of meaning are inadequate for many verbs, however, because of the ambiguous nature of intentional action. We describe a model that addresses such ambiguities by explicitly representing the role of intention recognition in word learning. By augmenting this model with phrase boundary information, we show improvement in learning compared to the original syntax-free model. Greater relative improvement is found in learning verbs than nouns. Evaluations are performed using data collected in a virtual environment. Results highlight the importance of representing intentions in cognitive models and suggest a greater role for the representation of intentions in applied areas of Artificial Intelligence.

Introduction

Computational models of word meaning have historically been rooted in the tradition of structural linguistics. In such models, the meaning of a word is defined strictly by its relations to other words or word-like symbols (e.g., Miller et al., 1990; Landauer et al., 1995, Lenat, 1995). A limitation of these models is their inability to explain how words are used to refer to non-linguistic referents (Harnad, 1990). A recent trend in the cognitive sciences is to address these limitations by modeling word meaning in terms of the non-linguistic context, or *situation,* surrounding language use (for a review see Roy (2005); Roy & Reiter, 2005). The work described here extends these efforts by presenting a situated model of word learning in which the intentions of an agent are explicitly modeled.

Recent efforts to model language acquisition have focused on models that ground the meaning of words in a learner's perceptions and actions. Such models ground the meaning of nouns in directly observable phenomena, such as object color and shape (e.g., Roy & Pentland, 2002). Models that focus on the meaning of verbs have also been introduced that ground meaning in motor control structures (Feldman & Narayanan, 2004) and perceived movements of objects (Siskind, 2001). A limitation of all these models, and a motivation for our current work, is that they are unable to account for the role that intentions play in word meaning.

Many of the most common verbs defy description in purely perceptual terms. For example, two different words, such as the words *chase* and *flee,* can be equally described by the same perceptual characteristics, while a single word, such as *open*, can describe any number of distinct activities that each bare different physical descriptions (e.g., opening with a key vs. opening with a pass code). In both cases, the semantics of the verbs are tied not to physical descriptions of the activity, but to the intentions of the agent who performs them. Although the role that intentions play has long been stressed in the empirical literature on word learning (e.g., Tomasello, 2001), in work on computational modeling, these issues remain largely unexplored.

In this work we describe a computational model that highlights the role of intention recognition in word learning (Fleischman and Roy, 2005). Similar to children, this situated model learns nouns faster than verbs (Gentner, 1982). We then describe an extension of this model that, like humans, exploits phrase structure information in the utterance to lessen noun/verb asymmetry (Gleitman, 1990). The model operates on data collected using a virtual environment; a methodology for computational modeling that allows subjects to interact in complex tasks while facilitating the encoding of situational context. Although by no means exhaustive in its account, these results demonstrate the feasibility and necessity of computationally modeling intentions in word learning

Model Overview

In Fleischman and Roy (2005), a model was developed and tested in a virtual environment based on a multiplayer videogame. In this environment, a game was designed in which a human player must navigate their way through a cavernous world, collecting specific objects, in order to escape. Subjects were paired such that one, the *expert*, would control the virtual character, while the other, the *novice*, guided her through the world via spoken instructions. 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. Both the movements and speech were recorded (Fig. 1a) and input into the model, which operates in two phases: intention recognition and linguistic mapping.

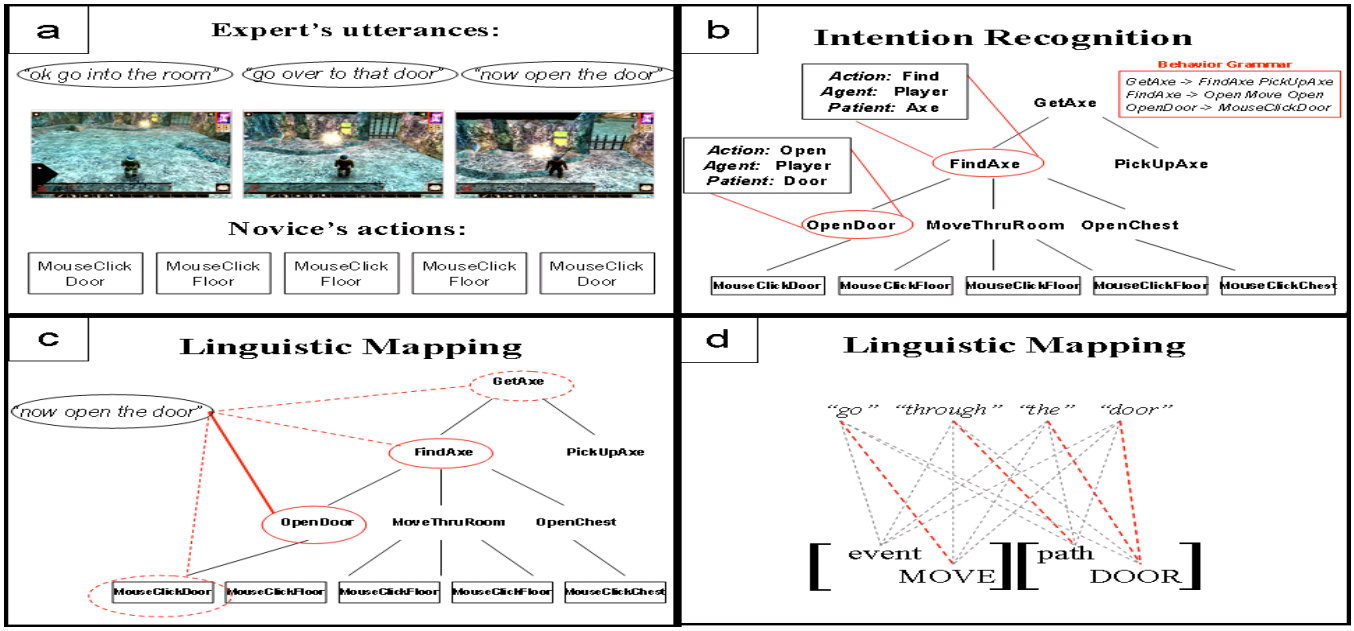

Figure 1. a) Parallel sequences of speech and actions are recorded from subjects as the expert guides the novice through a virtual environment. b) An intentional tree is inferred over the novice's sequence of observed actions using a probabilistic context free grammar of behaviors. Each node in the tree is a different level of intentional action and is encoded by a semantic frame. c) The vertical path from a leaf node in the tree (i.e. observed action) to the root (i.e. highest order intentional action) contains multiple possible levels of intention to which an utterance may refer. Linguistic mapping uses d) Expectation Maximization to estimate the conditional probabilities of words given roles to resolve this ambiguity.

Intention Recognition

Intention recognition is the ability to infer the reasons underlying an agent's behavior based on a sequence of their observed actions. A great deal of work has focused on the role of intentions in dialogue systems (e.g., Grosz & Sidner, 1986; Ferguson & Allen, 1998; Carbury, 2000). Unlike this previous work, we follow work in plan recognition (Pynadath, 1999) and event recognition (Ivanov & Bobick, 2000) and represent intentions using a probabilistic context free grammar (PCFG) of behaviors. Representing behaviors as a grammar enables us to treat intention recognition as a parsing problem over observed sequences of movements, in much the same way that a PCFG of syntax enables parsing of words in a sentence (e.g., Stolcke, 1994).

The idea of a grammar of behavior goes back at least to Miller et al. (1960). In our formalization, a grammar consists of *intention rules* that describe how an agent's high level intentional actions (e.g., *find axe*) can lead to sequences of lower level intentional actions (e.g. *open door*, *go through door*, *open chest*) (Fig. 1b inset). Analogous to syntactic parse trees, a behavior grammar produces *intention trees* by parsing observed movements. Each element in an intention rule (and thus, each node in an intention tree) encodes a semantic frame that contains the participants of the action and their thematic roles (actor, patient, object, etc.) (Fig. 1b inset). In this initial work, the intention rules are created by hand – currently we are exploring automatic learning of such rules.

As the model observes sequences of a subject's movements in the virtual environment, an intention tree is inferred by the system. This tree acts as the conceptual scaffolding in which natural language utterances are grounded. In these experiments, the temporal alignment between a spoken utterance and the observed movement to which it corresponds is hand annotated (a focus of future work is the relaxation of this assumption). Given this annotation, there remains an ambiguity for any given observation as to which level within the tree an associated utterance refers. This ambiguity regarding the level of description (Gleitman, 1990) is represented by the multiple nodes that exist along the *vertical path* from the root of the intention tree to the leaf node temporally aligned to the target utterance (Fig. 1c). This ambiguity is resolved in the linguistic mapping procedure (described below) by determining which node along the vertical path a given utterance is most likely to refer.

Linguistic Mapping

Having observed a sequence of movements, the output of intention recognition is an intention tree that represents the model's best guess of the higher order intentions that generated that sequence. The goal of the linguistic mapping phase is to find the links between the words an agent says and the tree that describes what an agent does.

As described above, each node in an inferred intention tree consists of a semantic frame. In the linguistic mapping phase, associations are learned between words in utterances and the elements in these frames (i.e. roles, such as AGENT, *or* role fillers, such as DOOR). These mappings are represented by the conditional probabilities of words

given frame elements [i.e. $p(word|element)$]. By formalizing mappings in this way, we can equate the problem of learning word meanings to one of finding the maximum likelihood estimate of a conditional probability distribution.

Similar to statistical approaches to language translation (Brown et al., 1993), we apply the Expectation Maximization (EM) algorithm to estimate these mappings. EM is a well studied algorithm that attempts to find a locally optimal conditional probability distribution for a dataset by iterating between an Estimation (E) step and a Maximization (M) step.

To understand the use of EM, let us first assume that we know which node in the vertical path is associated with an utterance (i.e., no ambiguity of descriptive level). In the E step, an initial conditional probability distribution is used to collect expected counts of how often a word in an utterance appears with a frame element in its paired semantic frame (Figure 1d). In the M step, these expected counts are used to calculate a new conditional probability distribution. By making a one-to-many assumption -- that each word in an utterance is generated by only one frame element in the parallel frame (but that each frame element can generate multiple words) -- the iterative algorithm is guaranteed to converge to the maximum likelihood estimation of the conditional distribution. Following Brown et al. (1993), we add a NULL role to each semantic frame which acts as a "garbage collector," accounting for common words that don't conceptually map to objects or actions (e.g., "the," "now," "ok," etc.).

The above procedure describes an ideal situation in which one knows which semantic frame from the associated vertical path should be paired with a given utterance. As described above, this is not the case for language learners who, even knowing the intention behind an action, are faced with an ambiguity as to what level of description an utterance was meant to refer (Figure 1c). To address this ambiguity, an outer processing loop is introduced that iterates over all possible pairings of utterances and semantic frames along the vertical path. For each pairing, a conditional probability distribution is estimated using EM. After all pairings have been examined, their estimated distributions are merged, each weighted by their likelihood. This procedure (Figure. 2) continues until a cross-validation stopping criterion is reached. The utterance/frame pair with the highest likelihood yields the most probable resolution of the ambiguity.

Representing linguistic mappings as conditional probabilities not only allows us to apply efficient algorithms to the task of word learning, but also leads to a Bayesian formulation of language understanding. In this formulation, understanding an utterance is equivalent to finding the most likely meaning (i.e. semantic frame) given that utterance:

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

This equation makes understanding utterances particularly easy to model using the two phase model of word learning presented here because of the natural analogues that exist between calculating the posterior probability and the linguistic mapping phase, and between calculating the prior probability and the intention recognition phase. Specifically, the posterior $p(utterance|meaning)$ can be approximated by the probability of the most likely alignment of words in an utterance to elements in a frame (using the probability distribution estimated by EM). Further, the prior $p(meaning)$ can be approximated by the probability of the most likely inferred intentional tree (i.e. the probability given by the by the PCFG parser).

1. Set uniform likelihoods for all utterance/frame pairings
2. For each pair, run standard EM
3. Merge output distributions of EM (weighting each by the likelihood of the pairing)
4. Use merged distribution to recalculate likelihoods of all utterance/frame pairings
5. Go to Step 2

Figure 2. Intentional Expectation Maximization algorithm

Incorporating Syntactic Information

The linguistic mapping phase as described thus far treats utterances as unstructured bags of words. Findings in development psychology suggest that children are able to take advantage of structural cues in utterances in order to aid in early word learning (e.g., Snedeker and Gleitman, 2004). We now describe an extension of Fleischman and Roy (2005), in which the linguistic mapping phase is extended to leverage knowledge of syntactic phrases boundaries.

The first step in exploiting phrase boundary information is to be able to find phrase boundaries in the input. Phrase boundaries within utterances are found using a highly accurate automatic phrase chunker (Daume and Marcu, 2005) that uses local lexical features and is trained on a large corpus of annotated text. We make no claims as to the appropriateness of the phrase chunker as a model for linguistic development. Rather, the chunker is treated only as a tool by which the effects of phrase boundary information on word learning may be studied. Since we seek to examine the effect of minimal syntactic information on language learning, only phrase boundary locations are used by the model. Thus, although they may be also useful for word learning, no detailed parse trees, embedded clause structures, or other syntactic information (e.g. noun phrase vs. prepositional phrase categorization) is provided to our model. Figure 3 gives an example of the phrase boundaries found by the phrase chunker for a sample utterance.

In addition to finding phrase boundaries in the input utterances, the form of the semantic frames generated during intention recognition have been slightly modified. We introduce the notion of a *semantic chunk* and define it as the set which contains both a semantic role *and* its corresponding role filler (see Figure 3). To leverage the boundary information provided by this chunking, the original linguistic mapping algorithm is modified by nesting another layer into the original two-layer EM learning procedure. This new nested layer aligns phrases to semantic

chunks and replaces the use of standard EM in Figure 2 Step 2, with a new phrasal EM procedure described in Figure 4.

The new model comprises three nested layers of EM that operate as follows: (1) Utterances are aligned to frames (to account for ambiguities of descriptive level); (2) Phrases from within the utterances are aligned to semantic chunks from within the frames, and (3) Words from within the phrases are aligned to frame elements from within the semantic chunks. Probability distributions estimated at the lowest layers [*p(word|element)*] are propagated up to the higher layers where they are merged and used to calculate the likelihoods of the proposed alignments between both phrases and semantic chunks, and finally between utterances and frames. Although adding this phrasal layer adds algorithmic complexity, because the number of phrase to chunk alignments is relatively small, the overall number of expected counts that the algorithm must examine in estimating the conditional probability distributions is dramatically *reduced* (see Discussion for more details).

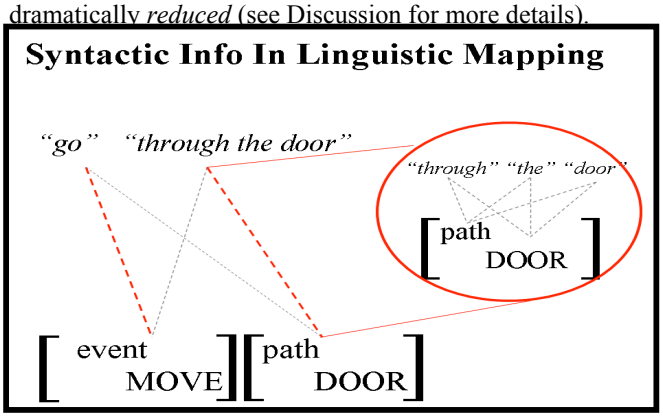

Figure 3. Syntactic phrase boundaries are used in the Phrasal Expectation Maximization Algorithm to reduce the hypothesis space of possible alignments between words and semantic frame elements.

Model Evaluation

Data Collection

In order to evaluate the model, we developed a virtual environment based on the multi-user videogame Neverwinter Nights. [1] The subjects in the data collection were university students (8 male, 4 female). Subjects were staggered such that the novice controlling the virtual character in one trial became the expert issuing commands in the next. The game was instrumented so that all the experts' speech and all of the novices' actions were recorded during play. Figure 1a shows screen shots of a game with the associated sequences of data: the expert's speech and novice's actions.

The expert's speech is automatically segmented into utterances based on pause structure and then manually transcribed. The novice's action sequences are parsed using a hand built behavior grammar to infer tree representations of the novice's intentions (Fig. 1b). In the current experiments, the entire sequence of actions composing a game trial is parsed at once and linguistic mapping is performed using the most likely tree from that parse.

In hand building the behavior grammar, two sets of rules were created: one to describe agents' possible paths of movement and one to describe non-locomotion actions. The movement rules were built semi-automatically, by enumerating all possible paths between target rooms in the game. The action rules were designed based on the rules of the game in order to match the actions that players must take to win (e.g. opening doors, taking objects, interacting with non-player characters, etc.). Rules were built and refined in an iterative manner, in order to insure that all subject trials could be parsed. Because of limited data, generalization of the rules to held-out data was not examined. Probabilities were set using the frequency of occurrence of the rules on the training data. A major focus of future work will be the automation of this process, which would merge the inter-related problems of language acquisition and task learning.

Having collected the utterances and parsed the actions, the two streams are processed by the learning model such that the semantic roles from the novice's intention tree are mapped to the words in the expert's utterances. By iterating through all possible mappings, the algorithm converges to a probability distribution that maximizes the likelihood of the data (Fig. 1c-d).

1. Set initial distribution using conditional probabilities from intentional EM
2. Generate all possible phrase/chunk pairs
3. For each pair, run standard EM
4. Merge output distributions of standard EM (weighting each by the likelihood of the pairing)
5. Use merged distribution to recalculate likelihoods of all utterance/frame pairings
6. Goto step 2

Figure 4. Phrasal Expectation Maximization algorithm

Experiments

To evaluate the effect of syntactic information on word learning in the model, the linguistic mapping algorithms were trained using utterances both with and without annotated phrase boundary information. For both conditions, the model was trained on the first four trials of game play for all subject pairs and tested on the final trial. This yielded 1040 training, and 240 testing utterances. For each pair, the number of iterations, beam search, and other parameters are optimized using cross-validation.

For each utterance in the test data, the likelihood that it was generated by each possible frame is calculated. We select the maximum likelihood frame as the system's hypothesized meaning for the test utterance, and examine how often the system maps each word of that utterance to the correct semantic role. Word mapping accuracies are separated by word class (nouns and verbs) and compared.

Further, we examine the ability of the system to accurately predict the correct level of description to which test utterances refer. We compare the system trained with

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

syntactic information against the system trained without in two conditions: both when it is and is not trained given the correct utterance/semantic frame (i.e., both with and without descriptive level ambiguities resolved by hand).

Results and Discussion

Figure 5 shows the word accuracy performance on nouns and verbs for the system trained both with and without phrase boundary information. As described in Fleischman and Roy (2005), the model learns nouns better than it learns verbs. Further, the figure indicates that syntactic phrase boundary information improves learning of verbs more than nouns. Figure 6 shows the ability of the system to predict the correct level of description to which a novel test utterance refers. The system performs equally well with or without syntactic information given the correct utterance/frame pairs during training. However, when ambiguities of descriptive level are not resolved by hand, the system's ability to predict the correct level of description becomes dramatically impoverished if access to syntactic information is not provided.

Figure 5 shows that the language learning model takes advantage of phrase chunking. Although word learning improves across the board, the model shows a larger increase in performance for learning verbs than nouns. This result concurs with findings in developmental psychology which suggest that syntactic information, such as the number and order of phrases and the thematic markers they contain, serve as cues to the language learner when acquiring verbs (Snedeker and Gleitman, 2004). Our model shows improved learning even though it is not designed to take advantage of structural cues of this complexity. Rather, the syntactic information is exploited by the model only in its ability to reduce the number of possible mappings that must be considered during training.

As described above, when estimating the conditional probability distribution, the EM algorithm must take expected counts over all possible word to frame element alignments for a given utterance/frame pair (Fig. 1d). The usefulness of the phrase boundary information is in its ability to reduce the number of possible alignments that must be examined when calculating these expected counts. For example, in Figure 1d the EM algorithm applied to the given utterance/frame pair must take expected counts over 4^4=256 different possible word-to-element alignments (the number of elements in the semantic frame raised to the number of words in the utterance). However, using phrasal EM (see Figure 3), only 2^2 phrase-to-chunk alignments are generated (the number of semantic chunks in the frame raised to the number of phrases in the utterance), each one necessitating EM to take expected counts over only 2^1+2^3 word-to-element alignments. Thus, phrase chunking reduces the potential number of alignments from 256 to 40, leading to more effective use of limited data.

This reduction follows from the fact that phrase boundaries do not allow mappings in which words from a particular phrase are aligned to frame elements from different semantic chunks (e.g., it can never be the case that "through" aligns to an element in [PATH DOOR], while "door" aligns to an element in [EVENT MOVE]). By pruning out such superfluous alignments, the algorithm is able to converge to a less noisy estimate for the conditional probability distribution.

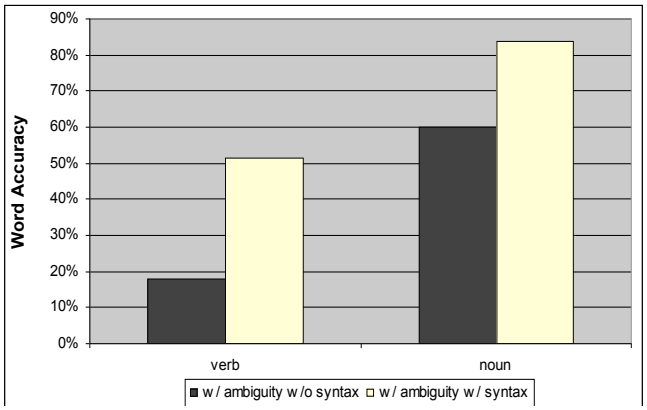

Figure 5. Incorporating syntactic phrase information in the model improves performance on learning verbs more than on nouns.

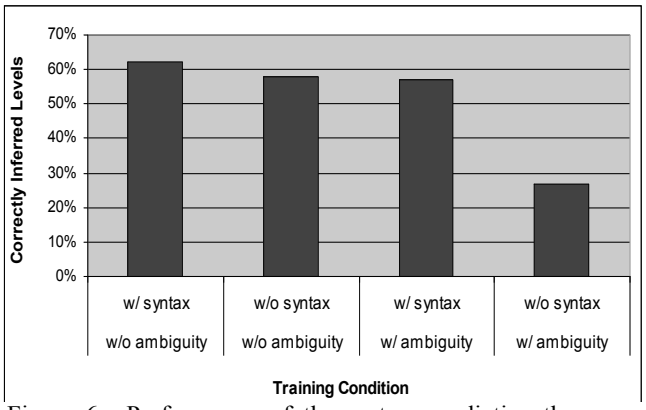

Figure 6. Performance of the system predicting the correct level of description to which a novel test utterance refers.

This reduction in noise explains why word learning increases in general, but does not explain why verbs in particular benefit so much from syntactic information. In the original model without phrase chunking (Fleischman and Roy, 2005) we showed that one cause of the verb/noun learning asymmetry in our model was the fact that, while each node of an intentional tree (i.e. semantic frame) has a different action role, often the object roles in different levels are the same. This same reasoning explains the verb/noun asymmetry in the current model.

For example, in Figure 1c, the actions associated with the nodes (e.g., *finding, opening, getting*) occur only once along the vertical path from root to leaf. However, the objects associated with those nodes (e.g., *axe*) occur multiple times along that same vertical path. This means that even if the model misinterprets what level of intention an utterance describes, because object roles are repeated at multiple levels, the model is still able to map nouns to correct referents. However, because action roles are more specific to their level of description, if the model misinterprets the level, linguistic mapping for the verb may fail.

This explanation for the slower learning of verbs than nouns in the original model can now be used to understand how syntactic information increases performance on verbs more than nouns. The improved conditional probability distributions estimated using the phrasal EM algorithm allow the system to more accurately determine the correct level of description to which novel utterances refer. As shown in Figure 6, training with phrase boundary information enables the system to determine the level of description with nearly the same accuracy as systems that were given the correct level of description during training. Thus, the syntactic phrase information enables the system to perform nearly as well as systems for which no ambiguity was present during training at all. Because the system can determine levels of description more accurately, the ambiguity that caused the slower learning of verbs than nouns in the original model is reduced and verbs are acquired with more ease.

Conclusion

We have described a model of situated word learning in which the use of intention recognition leads to noun/verb acquisition asymmetries analogous to those found in human learners. We showed how augmenting this model with simple phrase structure information dramatically increases performance on verb learning,. The increased performance of the system follows from the use of phrasal information to reduce the number of possible word meanings that the model must examine during learning.

The model that we describe demonstrates the importance of representing intentions in computational models of word learning. The use of formal grammars of behavior can also be beneficial in practical Artificial Intelligence applications. Fleischman and Hovy (2006) describe a Natural Language Interface (NLI) for a virtual training environment in which intention recognition is used to increase robustness to noisy speech recognition. Gorniak and Roy (2005) use plan recognition to resolve co-reference in video games.

Our current work focuses on addressing some of the simplifying assumptions made in the current model. In particular, we are examining how behavior grammar-like representations can be automatically learned from low level features. As a first step in this direction, Fleischman et al. (2006) examines how hierarchical patterns of movement can be learned from large amounts of home video recordings. Currently, we are extending this work, by applying similar techniques in the domain of broadcast television to support applications such as video search and event classification.

References

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

Carberry, S. Plan Recognition in Natural Language Dialogue. MIT Press, Cambridge MA. 1990.

Daume, H. III & D. Marcu, (2005). Approximate Large Margin Learning for Structured Outputs as a Search Optimization Problem, ICML, Bonn Germany.

Feldman, J. and Narayanan, S. (2004). Embodied Meaning in a Neural Theory of Language, *Brain and Language 89 (2004).*

Ferguson, G. and Allen, J., (1998) TRIPS: An Intelligent Integrated Problem-Solving Assistant. *Fifteenth National Conference on Artificial Intelligence (AAAI-98)*.

Fleischman, M. and Roy, D. (2005) *Why Verbs are Harder to Learn than Nouns:*Initial Insights from a Computational Model of Intention Recognition in Situated Word Learning Proceedings of the Annual Meeting of Cognitive Science Society.

Fleischman, M., Decamp, P. Roy, D. (2006). Mining Temporal Patterns of Movement for Video Content Classification. *Workshop on Multi-Media Information Retrieval.*

Fleischman, M. and Hovy, E. (2006). Taking Advantage of the Situation: Non-Linguistic Context for Natural Language Interfaces to Interactive Virtual Environments. *Intelligent User Interfaces*.

Gorniak, P. and Roy, D. (2005) Probabilistic Grounding of Situated Speech using Plan Recognition and Reference Resolution. *International Conference on Multimodal Interfaces*.

Gentner. (1982) Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj, editor, *Language development: Vol. 2. Language, cognition, and culture*. Erlbaum, Hillsdale, NJ.

Gleitman, L. (1990). The structural sources of word meaning. Language Acquisition, 1, 3-55.

Grosz, B. and Sidner, C. (1986) Attention, Intentions, and the Structure of Discourse, Computational Linguistics. 12(3)

Harnad, S. (1990). The symbol grounding problem. Physica D, 42.

Ivanov, Y. and Aaron F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing", IEEE Transactions on Pattern Analisys & Machine Intelligence, 22(8), August, 2000.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes,* 25, 259-284.

Lenat, D.B. (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11).

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

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Five papers on Wordnet. *International Journal of Lexicology*, 3(4).

Pynadath, D, (1999). Probabilistic Grammars for Plan Recognition. Ph.D. U of Michigan.

Roy, D. (2005). "Grounding Words in Perception and Action: Insights from Computational Models". TICS.

Roy. D. and Reiter, E. (2005). Connecting Language to the World. Artificial Intelligence, 167(1-2), 1-12.

Roy, D. and Pentland, A. (2002) Learning Words from Sights and Sounds: A Computational Model. *Cognitive Science*, 26(1).

Siskind, J. (2001). Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *Journal of Artificial Intelligence Research*, 15, 31-90.

Snedeker, J. & Gleitman, L. (2004). Why it is hard to label our concepts. To appear in Hall & Waxman (eds.), Weaving a Lexicon. Cambridge, MA: MIT Press

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

Tomasello, M. (2001). Perceiving intentions and learning words in the second year of life. *In M. Bowerman & S. Levinson (Eds.), Language Acquisition and Conceptual Development.* Cambridge University Press.