Starting From Scratch in Semantic Role Labeling

Michael Connor
University of Illinois
connor2@uiuc.edu

Yael Gertner
University of Illinois
ygertner@cyrus.psych.uiuc.edu

Cynthia Fisher
University of Illinois
cfisher@cyrus.psych.uiuc.edu

Dan Roth
University of Illinois
danr@illinois.edu

Abstract

A fundamental step in sentence comprehension involves assigning semantic roles to sentence constituents. To accomplish this, the listener must parse the sentence, find constituents that are candidate arguments, and assign semantic roles to those constituents. Each step depends on prior lexical and syntactic knowledge. Where do children learning their first languages begin in solving this problem? In this paper we focus on the parsing and argument-identification steps that precede Semantic Role Labeling (SRL) training. We combine a simplified SRL with an unsupervised HMM part of speech tagger, and experiment with psycholinguistically-motivated ways to label clusters resulting from the HMM so that they can be used to parse input for the SRL system. The results show that proposed shallow representations of sentence structure are robust to reductions in parsing accuracy, and that the contribution of alternative representations of sentence structure to successful semantic role labeling varies with the integrity of the parsing and argument-identification stages.

1 Introduction

In this paper we present experiments with an automatic system for semantic role labeling (SRL) that is designed to model aspects of human language acquisition. This simplified SRL system is inspired by the syntactic bootstrapping theory, and by an account of syntactic bootstrapping known as 'structure-mapping' (Fisher, 1996; Gillette et al., 1999; Lidz et al., 2003). Syntactic bootstrapping theory proposes that young children use their very partial knowledge of syntax to guide sentence comprehension. The structure-mapping account makes three key assumptions: First, sentence comprehension is grounded by the acquisition of an initial set of concrete nouns. Nouns are arguably less dependent on prior linguistic knowledge for their acquisition than are verbs; thus children are assumed to be able to identify the referents of some nouns via cross-situational observation (Gillette et al., 1999). Second, these nouns, once identified, yield a skeletal sentence structure. Children treat each noun as a candidate argument, and thus interpret the number of nouns in the sentence as a cue to its semantic predicate-argument structure (Fisher, 1996). Third, children represent sentences in an abstract format that permits generalization to new verbs (Gertner et al., 2006).

The structure-mapping account of early syntactic bootstrapping makes strong predictions, including predictions of tell-tale errors. In the sentence “Ellen and John laughed”, an intransitive verb appears with two nouns. If young children rely on representations of sentences as simple as an ordered set of nouns, then they should have trouble distinguishing such sentences from transitive sentences. Experimental evidence suggests that they do: 21-month-olds mistakenly interpreted word order in sentences such as “The girl and the boy kradded” as conveying agent-patient roles (Gertner and Fisher, 2006).

Previous computational experiments with a system for automatic semantic role labeling (BabySRL: (Connor et al., 2008)) showed that it is possible to learn to assign basic semantic roles based on the shallow sentence representations proposed by the structure-mapping view. Furthermore, these simple structural features were robust to drastic reductions in the integrity of the semantic-role feedback (Connor et al., 2009). These experiments showed that representations of sentence structure as simple as ‘first of two nouns’ are useful, but the experiments relied on perfect

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 989–998,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

knowledge of arguments and predicates as a start to classification.

Perfect built-in parsing finesses two problems facing the human learner. The first problem involves classifying words by part-of-speech. Proposed solutions to this problem in the NLP and human language acquisition literatures focus on distributional learning as a key data source (e.g., (Mintz, 2003; Johnson, 2007)). Importantly, infants are good at learning distributional patterns (Gomez and Gerken, 1999; Saffran et al., 1996). Here we use a fairly standard Hidden Markov Model (HMM) to generate clusters of words that occur in similar distributional contexts in a corpus of input sentences.

The second problem facing the learner is more contentious: Having identified clusters of distributionally-similar words, how do children figure out what role these clusters of words should play in a sentence interpretation system? Some clusters contain nouns, which are candidate arguments; others contain verbs, which take arguments. How is the child to know which are which? In order to use the output of the HMM tagger to process sentences for input to an SRL model, we must find a way to automatically label the clusters.

Our strategies for automatic argument and predicate identification, spelled out below, reflect core claims of the structure-mapping theory: (1) The meanings of some concrete nouns can be learned without prior linguistic knowledge; these concrete nouns are assumed based on their meanings to be possible arguments; (2) verbs are identified, not primarily by learning their meanings via observation, but rather by learning about their syntactic argument-taking behavior in sentences.

By using the HMM part-of-speech tagger in this way, we can ask how the simple structural features that we propose children start with stand up to reductions in parsing accuracy. In doing so, we move to a parser derived from a particular theoretical account of how the human learner might classify words, and link them into a system for sentence comprehension.

2 Model

We model language learning as a Semantic Role Labeling (SRL) task (Carreras and Màrquez, 2004). This allows us to ask whether a learner, equipped with particular theoretically-motivated representations of the input, can learn to understand sentences at the level of who did what to whom. The architecture of our system is similar to a previous approach to modeling early language acquisition (Connor et al., 2009), which is itself based on the standard architecture of a full SRL system (e.g. (Punyakanok et al., 2008)).

This basic approach follows a multi-stage pipeline, with each stage feeding in to the next. The stages are: (1) Parsing the sentence, (2) Identifying potential predicates and arguments based on the parse, (3) Classifying role labels for each potential argument relative to a predicate, (4) Applying constraints to find the best labeling of arguments for a sentence. In this work we attempt to limit the knowledge available at each stage to the automatic output of the previous stage, constrained by knowledge that we argue is available to children in the early stages of language learning.

In the parsing stage we use an unsupervised parser based on Hidden Markov Models (HMM), modeling a simple 'predict the next word' parser. Next the argument identification stage identifies HMM states that correspond to possible arguments and predicates. The candidate arguments and predicates identified in each input sentence are passed to an SRL classifier that uses simple abstract features based on the number and order of arguments to learn to assign semantic roles.

As input to our learner we use samples of natural child directed speech (CDS) from the CHILDES corpora (MacWhinney, 2000). During initial unsupervised parsing we experiment with incorporating knowledge through a combination of statistical priors favoring a skewed distribution of words into classes, and an initial hard clustering of the vocabulary into function and content words. The argument identifier uses a small set of frequent nouns to seed argument states, relying on the assumptions that some concrete nouns can be learned as a prerequisite to sentence interpretation, and are interpreted as candidate arguments.

The SRL classifier starts with noisy largely unsupervised argument identification, and receives feedback based on annotation in the PropBank style; in training, each word identified as an argument receives the true role label of the phrase that word is part of. This represents the assumption that learning to interpret sentences is naturally supervised by the fit of the learner's predicted meaning with the referential context. The provision of perfect 'gold-standard' feedback over-estimates the real child's access to this supervision, but allows us to investigate the consequences of noisy argument identification for SRL performance. We show that even with imperfect parsing, a learner can identify useful abstract patterns for sentence interpretation. Our ultimate goal is to 'close the loop' of this system, by using learning in the SRL system to improve the initial unsupervised parse and argument identification.

The training data were samples of parental speech to three children (Adam, Eve, and Sarah; (Brown, 1973)), available via CHILDES. The SRL training corpus consists of parental utterances in samples Adam 01-20 (child age 2;3 - 3;1), Eve 01-18 (1;6 - 2;2), and Sarah 01-83 (2;3 - 3;11). All verb-containing utterances without symbols indicating disfluencies were automatically parsed with the Charniak parser (Charniak, 1997), annotated using an existing SRL system (Punyakanok et al., 2008) and then errors were hand-corrected. The final annotated sample contains about 16,730 propositions, with 32,205 arguments.

3 Unsupervised Parsing

As a first step of processing, we feed the learner large amounts of unlabeled text and expect it to learn some structure over this data that will facilitate future processing. The source of this text is child directed speech collected from various projects in the CHILDES repository[1]. We removed sentences with fewer than three words or markers of disfluency. In the end we used 160 thousand sentences from this set, totaling over 1 million tokens and 10 thousand unique words.

The goal of the parsing stage is to give the learner a representation permitting it to generalize over word forms. The exact parse we are after is a distributional and context-sensitive clustering of words based on sequential processing. We chose an HMM based parser for this since, in essence the HMM yields an unsupervised POS classifier, but without names for states. An HMM trained with expectation maximization (EM) is analogous to a simple process of predicting the next word in a stream and correcting connections accordingly for each sentence.

With HMM we can also easily incorporate additional knowledge during parameter estimation. The first (and simplest) parser we used was an HMM trained using EM with 80 hidden states. The number of hidden states was made relatively large to increase the likelihood of clusters corresponding to a single part of speech, while preserving some degree of generalization.

Johnson (2007) observed that EM tends to create word clusters of uniform size, which does not reflect the way words cluster into parts of speech in natural languages. The addition of priors biasing the system toward a skewed allocation of words to classes can help. The second parser was an 80 state HMM trained with Variational Bayes EM (VB) incorporating Dirichlet priors (Beal, 2003).[2]

In the third and fourth parsers we experiment with enriching the HMM POS-tagger with other psycholinguistically plausible knowledge. Words of different grammatical categories differ in their phonological as well as in their distributional properties (e.g., (Kelly, 1992; Monaghan et al., 2005; Shi et al., 1998)); combining phonological and distributional information improves the clustering of words into grammatical categories. The phonological difference between content and function words is particularly striking (Shi et al., 1998). Even newborns can categorically distinguish content and function words, based on the phonological difference between the two classes (Shi et al., 1999). Human learners may treat content and function words as distinct classes from the start.

To implement this division into function and content words[3], we start with a list of function word POS tags[4] and then find words that appear predominantly with these POS tags, using tagged WSJ data (Marcus et al., 1993). We allocated a fixed number of states for these function words, and left the rest of the states for the rest of the words. This amounts to initializing the emission matrix for the HMM with a block structure; words from one class cannot be emitted by states allocated to the other class. This trick has been used before in speech recognition work (Rabiner,

[1] We used parts of the Bloom (Bloom, 1970; Bloom, 1973), Brent (Brent and Siskind, 2001), Brown (Brown, 1973), Clark (Clark, 1978), Cornell, MacWhinney (MacWhinney, 2000), Post (Demetras et al., 1986) and Providence (Demuth et al., 2006) collections.

[2] We tuned the prior using the same set of 8 value pairs suggested by Gao and Johnson (2008), using a held out set of POS-tagged CDS to evaluate final performance.

[3] We also include a small third class for punctuation, which is discarded.

[4] TO,IN,EX,POS,WDT,PDT,WRB,MD,CC,DT,RP,UH

1989), and requires far fewer resources than the full tagging dictionary that is often used to intelligently initialize an unsupervised POS classifier (e.g. (Brill, 1997; Toutanova and Johnson, 2007; Ravi and Knight, 2009)).

Because the function and content word preclustering preceded parameter estimation, it can be combined with either EM or VB learning. Although this initial split forces sparsity on the emission matrix and allows more uniform sized clusters, Dirichlet priors may still help, if word clusters within the function or content word subsets vary in size and frequency. The third parser was an 80 state HMM trained with EM estimation, with 30 states pre-allocated to function words; the fourth parser was the same except that it was trained with VB EM.

3.1 Parser Evaluation

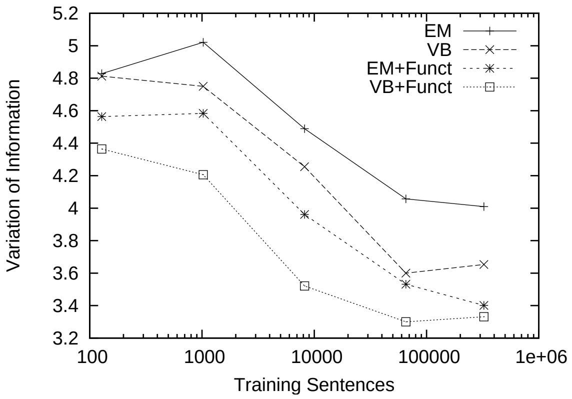


Figure 1: Unsupervised Part of Speech results, matching states to gold POS labels. All systems use 80 states, and comparison is to gold labeled CDS text, which makes up a subset of the HMM training data. Variation of Information is an information-theoretic measure summing mutual information between tags and states, proposed by (Meilă, 2002), and first used for Unsupervised Part of Speech in (Goldwater and Griffiths, 2007). Smaller numbers are better, indicating less information lost in moving from the HMM states to the gold POS tags. Note that incorporating function word preclustering allows both EM and VB algorithms to achieve the same performance with an order of magnitude fewer sentences.

We first evaluate these parsers (the first stage of our SRL system) on unsupervised POS tagging. Figure 1 shows the performance of the four systems using Variation of Information to measure match between gold states and unsupervised parsers as we vary the amount of text they receive. Each point on the graph represents the average result over 10 runs of the HMM with different samples of the unlabeled CDS. Another common measure for unsupervised POS (when there are more states than tags) is a many to one greedy mapping of states to tags. It is known that EM gives a better many to one score than VB trained HMM (Johnson, 2007), and likewise we see that here: with all data EM gives 0.75 matching, VB gives 0.74, while both EM+Funct and VB+Funct reach 0.80.

Adding the function/content word split to the HMM structure improves both EM and VB estimation in terms of both tag matching accuracy and information. However, these measures look at the parser only in isolation. What is more important to us is how useful the provided word clusters are for future semantic processing. In the next sections we use the outputs of our four parsers to identify arguments and predicates.

4 Argument Identification

The unsupervised parser provides a state label for each word in each sentence; the goal of the argument identification stage is to use these states to label words as potential arguments, predicates or neither. As described in the introduction, core premises of the structure-mapping account offer routes whereby we could label some HMM states as argument or predicate states.

The structure-mapping account holds that sentence comprehension is grounded in the learning of an initial set of nouns. Children are assumed to identify the referents of some concrete nouns via cross-situational learning (Gillette et al., 1999; Smith and Yu, 2008). Children then assume, by virtue of the meanings of these nouns, that they are candidate arguments. This is a simple form of semantic bootstrapping, requiring the use of built-in links between semantics and syntax to identify the grammatical type of known words (Pinker, 1984). We use a small set of known nouns to transform unlabeled word clusters into candidate arguments for the SRL: HMM states that are dominated by known names for animate or inanimate objects are assumed to be argument states.

Given text parsed by the HMM parser and a list of known nouns, the argument identifier proceeds in multiple steps as illustrated in figure 2. The first stage identifies as argument states those states that appear at least half the time in the training data with known nouns. This use of a seed list and distributional clustering is similar to Prototype Driven Learning (Haghighi and Klein, 2006), except we are only providing information on one specific class.

```
Algorithm ARGUMENT STATE IDENTIFICATION
    INPUT: Parsed Text $T$ = list of (word, state) pairs
           Set of concrete nouns $N$
    OUTPUT: Set of argument states $A$
            Argument count likelihood $ArgLike(s, c)$

    Identify Argument States
    Let $freq(s) = |\{(*, s) \in T\}|$
    Let $freq_N(s) = |\{(w, s) \in T | w \in N\}|$

    For each $s$:
      If $freq_N(s) \geq freq(s)/2$
        Add $s$ to $A$

    Collect Per Sentence Argument Count statistics
    For each Sentence $S \in T$:
      Let $Arg(S) = |\{(w, s) \in S | s \in A\}|$
      For $(w, s) \in S$ s.t. $s \notin A$
        Increment $ArgCount(s, Arg(S))$

    For each $s \notin A$, and argument count $c$:
      $ArgLike(s, c) = ArgCount(s, c)/freq(s)$
```

(a) Argument Identification

```
Algorithm PREDICATE STATE IDENTIFICATION
    INPUT: Parsed Sentence $S$ = list of (word, state) pairs
           Set of argument states $A$
           Sentence Argument Count $ArgLike(s, c)$
    OUTPUT: Most likely predicate $(v, s_v)$

    Find Number of arguments in sentence
    Let $Arg(S) = |\{(w, s) \in S | s \in A\}|$

    Find Non-argument state in sentence most likely
     to appear with this number of arguments
    $(v, s_v) = argmax_{(w,s) \in S} ArgLike(s, Arg(S))$
```

(b) Predicate Identification

Figure 2: Argument identification algorithm. This is a two stage process: argument state identification based on statistics collected over entire text and per sentence predicate identification.

As a list of known nouns we collected all those nouns that appear three times or more in the child directed speech training data and judged to be either animate or inanimate nouns. The full set of 365 nouns covers over 93% of noun occurences in our data. In upcoming sections we experiment with varying the number of seed nouns used from this set, selecting the most frequent set of nouns. Reflecting the spoken nature of the child directed speech, the most frequent nouns are pronouns, but beyond the top 10 we see nouns naming people ('daddy', 'ursula') and object nouns ('chair', 'lunch').

What about verbs? A typical SRL model identifies candidate arguments and tries to assign roles to them relative to each verb in the sentence. In principle one might suppose that children learn the meanings of verbs via cross-situational observation just as they learn the meanings of concrete nouns. But identifying the meanings of verbs is much more troublesome. Verbs' meanings are abstract, therefore harder to identify based on scene information alone (Gillette et al., 1999). As a result, early vocabularies are dominated by nouns (Gentner, 2006). On the structure-mapping account, learners identify verbs, and begin to determine their meanings, based on sentence structure cues. Verbs take noun arguments; thus, learners could learn which words are verbs by detecting each verb's syntactic argument-taking behavior. Experimental evidence provides some support for this procedure: 2-year-olds keep track of the syntactic structures in which a new verb appears, even without a concurrent scene that provides cues to the verb's semantic content (Yuan and Fisher, 2009).

We implement this behavior by identifying as predicate states the HMM states that appear commonly with a particular number of previously identified arguments. First, we collect statistics over the entire HMM training corpus regarding how many arguments are identified per sentence, and which states that are not identified as argument states appear with each number of arguments. Next, for each parsed sentence that serves as SRL input, the algorithm chooses as the most likely predicate the word whose state is most likely to appear with the number of arguments found in the current input sentence. Note that this algorithm assumes exactly one predicate per sentence. Implicitly, the argument count likelihood divides predicate states up into transitive and intransitive predicates based on appearances in the simple sentences of CDS.

4.1 Argument Identification Evaluation

Figure 3 shows argument and predicate identification accuracy for each of the four parsers when provided with different numbers of known nouns. The known word list is very skewed with its most frequent members dominating the total noun occurrences in the data. The ten most frequent words[5] account for 60% of the total noun occurrences. We achieve the different occurrence coverage numbers of figure 3 by using the most frequent N words from the list that give the specific coverage[6]. Pronouns refer to people or objects, but are abstract in that they can refer to any person or object. The inclusion of pronouns in our list of

[5] you, it, I, what, he, me, ya, she, we, her

[6] N of 5, 10, 30, 83, 227 cover 50%, 60%, 70%, 80%, 90% of all noun occurrences

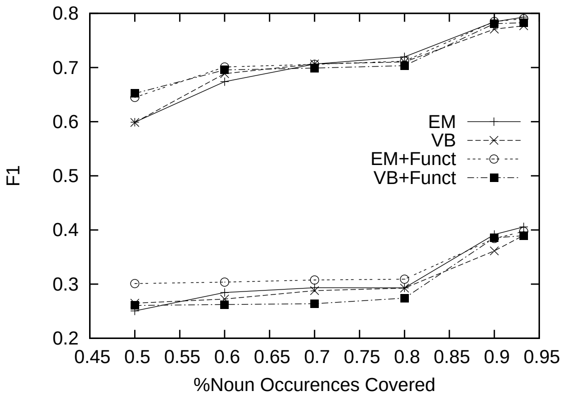


Figure 3: Effect of number of concrete nouns for seeding argument identification with various unsupervised parsers. Argument identification accuracy is computed against true argument boundaries from hand labeled data. The upper set of results show primary argument (A0-4) identification F1, and bottom lines show predicate identification F1.

known nouns represents the assumption that toddlers have already identified pronouns as referential terms. Even 19-month-olds assign appropriately different interpretations to novel verbs presented in simple transitive versus intransitive sentences with pronoun arguments ("He's kradding him!" vs. "He's kradding!"; (Yuan et al., 2007)). In ongoing work we experiment with other methods of identifying seed nouns.

Two groups of curves appear in figure 3: the upper group shows the primary argument identification accuracy and the bottom group shows the predicate identification accuracy. We evaluate compared to gold tagged data with true argument and predicate boundaries. The primary argument (A0-4) identification accuracy is the F1 value, with precision calculated as the proportion of identified arguments that appear as part of a true argument, and recall as the proportion of true arguments that have some state identified as an argument. F1 is calculated similarly for predicate identification, as one state per sentence is identified as the predicate.

As shown in figure 3, argument identification F1 is higher than predicate identification (which is to be expected, given that predicate identification depends on accurate arguments), and as we add more seed nouns the argument identification improves. Surprisingly, despite the clear differences in unsupervised POS performance seen in figure 1, the different parsers do not yield very different argument and predicate identification. As we will see in the next section, however, when the arguments identified in this step are used to train SRL classifier, distinctions between parsers reappear, suggesting that argument identification F1 masks systematic patterns in the errors.

5 Testing SRL Performance

Finally, we used the results of the previous parsing and argument-identification stages in training a simplified SRL classifier (BabySRL), equipped with sets of features derived from the structure-mapping account. For argument classification we used a linear classifier trained with a regularized perceptron update rule (Grove and Roth, 2001). In the results reported below the BabySRL did not use sentence-level inference for the final classification, every identified argument is classified independently; thus multiple nouns can have the same role. In what follows, we compare the performance of the BabySRL across the four parsers. We evaluated SRL performance by testing the BabySRL with constructed sentences like those used for the experiments with children described in the Introduction. All test sentences contained a novel verb, to test the model's ability to generalize.

We examine the performance of four versions of the BabySRL, varying in the features used to represent sentences. All four versions include lexical features consisting of the target argument and predicate (as identified in the previous steps). The baseline model has only these lexical features (Lexical). Following Connor et al. (2008; 2009), the key feature type we propose is noun pattern features (NounPat). Noun pattern features indicate how many nouns there are in the sentence and which noun the target is. For example, in "You dropped it!", 'you' has a feature active indicating that it is the first of two nouns, while 'it' has a feature active indicating that it is the second of two nouns. We compared the behavior of noun pattern features to another simple representation of word order, position relative to the verb (VerbPos). In the same example sentence, 'you' has a feature active indicating that it is pre-verbal; for 'it' a feature is active indicating that it is post-verbal. A fourth version of the BabySRL (Combined) used both NounPat and VerbPos features.

We structured our tests of the BabySRL to test the predictions of the structure-mapping account. (1) NounPat features will improve the SRL's ability to interpret simple transitive test sentences containing two nouns and a novel verb, relative to a lexical baseline. Like 21-month-old children (Gertner et al., 2006), the SRL should interpret the first noun as an agent and the second as a patient. (2) Because NounPat features represent word order solely in terms of a sequence of nouns, an SRL equipped with these features will make the errors predicted by the structure-mapping account and documented in children (Gertner and Fisher, 2006). (3) NounPat features permit the SRL to assign different roles to the subjects of transitive and intransitive sentences that differ in their number of nouns. This effect follows from the nature of the NounPat features: These features partition the training data based on the number of nouns, and therefore learn separately the likely roles of the '1st of 1 noun' and the '1st of 2 nouns'.

These patterns contrast with the behavior of the VerbPos features: When the BabySRL was trained with perfect parsing, VerbPos promoted agent-patient interpretations of transitive test sentences, and did so even more successfully than NounPat features did, reflecting the usefulness of position relative to the verb in understanding English sentences. In addition, VerbPos features eliminated the errors with two-noun intransitive sentences. Given test sentences such as 'You and Mommy krad', VerbPos features represented both nouns as pre-verbal, and therefore identified both as likely agents. However, VerbPos features did not help the SRL assign different roles to the subjects of simple transitive and intransitive sentences: 'Mommy' in 'Mommy krads you' and 'Mommy krads' are both represented simply as pre-verbal.

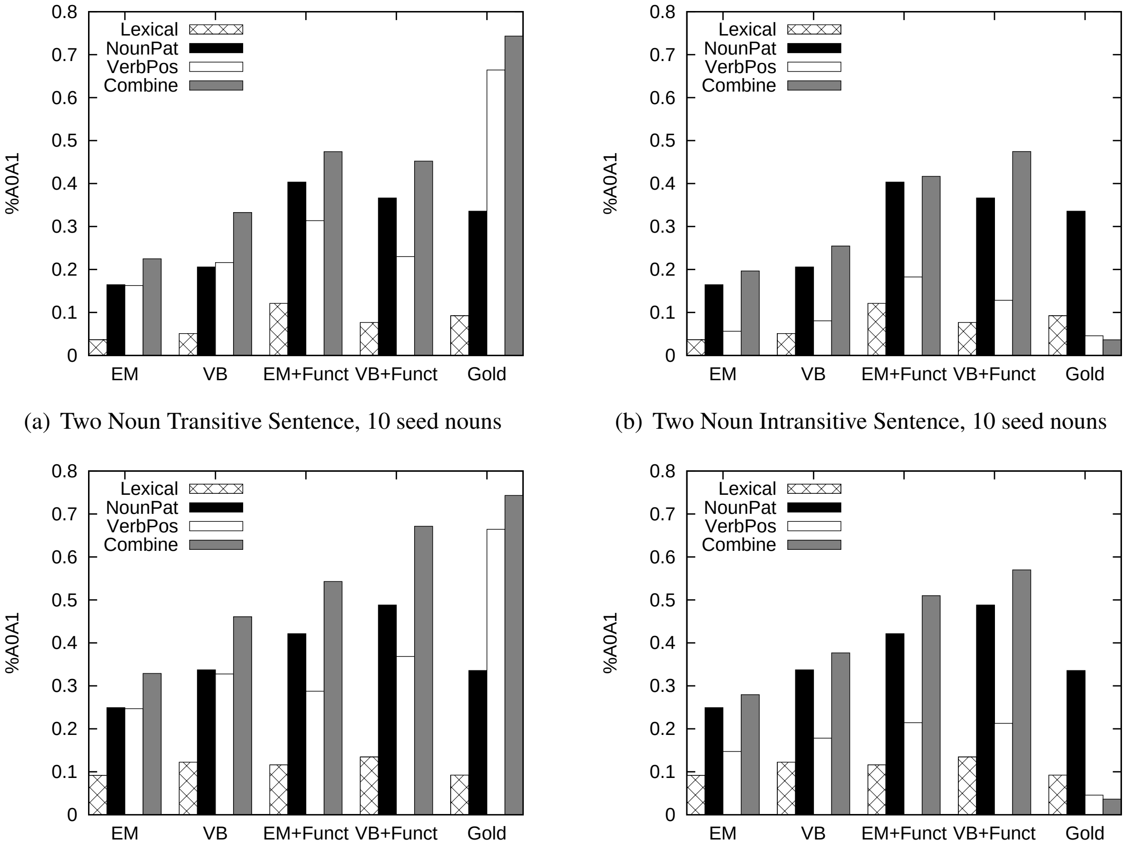
To test the system's predictions on transitive and intransitive two noun sentences, we constructed two test sentence templates: 'A krads B' and 'A and B krad', where A and B were replaced with familiar animate nouns. The animate nouns were selected from all three children's data in the training set and paired together in the templates such that all pairs are represented.

Figure 4 shows SRL performance on test sentences containing a novel verb and two animate nouns. Each plot shows the proportion of test sentences that were assigned an agent-patient (A0-A1) role sequence; this sequence is correct for transitive sentences but is an error for two-noun intransitive sentences. Each group of bars shows the performance of the BabySRL trained using one of the four parsers, equipped with each of our four feature sets. The top and bottom panels in Figure 4 differ in the number of nouns provided to seed the argument identification stage. The top row shows performance with 10 seed nouns (the 10 most frequent nouns, mostly animate pronouns), and the bottom row shows performance with 365 concrete (animate or inanimate) nouns treated as known. Relative to the lexical baseline, NounPat features fared well: they promoted the assignment of A0-A1 interpretations to transitive sentences, across all parser versions and both sets of known nouns. Both VB estimation and the content-function word split increased the ability of NounPat features to learn that the first of two nouns was an agent, and the second a patient. The NounPat features also promote the predicted error with two-noun intransitive sentences (Figures 4(b), 4(d)). Despite the relatively low accuracy of predicate identification noted in section 4.1, the VerbPos features did succeed in promoting an A0A1 interpretation for transitive sentences containing novel verbs relative to the lexical baseline. In every case the performance of the Combined model that includes both NounPat and VerbPos features exceeds the performance of either NounPat or VerbPos alone, suggesting both contribute to correct predictions for transitive sentences. However, the performance of VerbPos features did not improve with parsing accuracy as did the performance of the NounPat features. Most strikingly, the VerbPos features did not eliminate the predicted error with two-noun intransitive sentences, as shown in panels 4(b) and 4(d). The Combined model predicted an A0A1 sequence for these sentences, showing no reduction in this error due to the participation of VerbPos features.

Table 1 shows SRL performance on the same transitive test sentences ('A krads B'), compared to simple one-noun intransitive sentences ('A krads'). To permit a direct comparison, the table reports the proportion of transitive test sentences for which the first noun was assigned an agent (A0) interpretation, and the proportion of intransitive test sentences with the agent (A0) role assigned to the single noun in the sentence. Here we report only the results from the best-performing parser (trained with VB EM, and content/function word pre-clustering), compared to the same classifiers trained with gold standard argument identification. When trained on arguments identified via the unsupervised POS tagger, noun pattern features promoted agent interpretations of tran-

	Two Noun Transitive, % Agent First				One Noun Intransitive, % Agent Prediction			
	Lexical	NounPat	VerbPos	Combine	Lexical	NounPat	VerbPos	Combine
VB+Funct 10 seed	0.48	0.61	0.55	0.71	0.48	0.57	0.56	0.59
VB+Funct 365 seed	0.22	0.64	0.41	0.74	0.23	0.33	0.43	0.41
Gold Arguments	0.16	0.41	0.69	0.77	0.17	0.18	0.70	0.58

Table 1: SRL result comparison when trained with best unsupervised argument identifier versus trained with gold arguments. Comparison is between agent first prediction of two noun transitive sentences vs. one noun intransitive sentences. The unsupervised arguments lead the classifier to rely more on noun pattern features; when the true arguments and predicate are known the verb position feature leads the classifier to strongly indicate agent first in both settings.



(a) Two Noun Transitive Sentence, 10 seed nouns

(b) Two Noun Intransitive Sentence, 10 seed nouns

(c) Two Noun Transitive Sentence, 365 seed nouns

(d) Two Noun Intransitive Sentence, 365 seed nouns

Figure 4: SRL classification performance on transitive and intransitive test sentences containing two nouns and a novel verb. Performance with gold-standard argument identification is included for comparison. Across parses, noun pattern features promote agent-patient (A0A1) interpretations of both transitive ("You krad Mommy") and two-noun intransitive sentences ("You and Mommy krad"); the latter is an error found in young children. Unsupervised parsing is less accurate in identifying the verb, so verb position features fail to eliminate errors with two-noun intransitive sentences.

sitive subjects, but not for intransitive subjects. This differentiation between transitive and intransitive sentences was clearer when more known nouns were provided. Verb position features, in contrast, promote agent interpretations of subjects weakly with unsupervised argument identification, but equally for transitive and intransitive.

Noun pattern features were robust to increases in parsing noise. The behavior of verb position features suggests that variations in the identifiability of different parts of speech can affect the usefulness of alternative representations of sentence structure. Representations that reflect the position of the verb may be powerful guides for understanding simple English sentences, but representations reflecting only the number and order of nouns can dominate early in acquisition, depending on the integrity of parsing decisions.

6 Conclusion and Future Work

The key innovation in the present work is the combination of unsupervised part-of-speech tagging and argument identification to permit learning in a simplified SRL system. Children do not have the luxury of treating part-of-speech tagging and semantic role labeling as separable tasks. Instead, they must learn to understand sentences starting from scratch, learning the meanings of some words, and using those words and their patterns of arrangement into sentences to bootstrap their way into more mature knowledge.

We have created a first step toward modeling this incremental process. We combined unsupervised parsing with minimal supervision to begin to identify arguments and predicates. An SRL classifier used simple representations built from these identified arguments to extract useful abstract patterns for classifying semantic roles. Our results suggest that multiple simple representations of sentence structure could co-exist in the child's system for sentence comprehension; representations that will ultimately turn out to be powerful guides to role identification may be less powerful early in acquisition because of the noise introduced by the unsupervised parsing.

The next step is to 'close the loop', using higher level semantic feedback to improve the earlier argument identification and parsing stages. Perhaps with the help of semantic feedback the system can automatically improve predicate identification, which in turn allows it to correct the observed intransitive sentence error. This approach will move us closer to the goal of using initial simple structural patterns and natural observation of the world (semantic feedback) to bootstrap more and more sophisticated representations of linguistic structure.

Acknowledgments

This research is supported by NSF grant BCS-0620257 and NIH grant R01-HD054448.

References

M.J. Beal. 2003. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.

L. Bloom. 1970. *Language development: Form and function in emerging grammars*. MIT Press, Cambridge, MA.

L. Bloom. 1973. *One word at a time: The use of single-word utterances before syntax*. Mouton, The Hague.

M.R. Brent and J.M. Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81:31–44.

E. Brill. 1997. Unsupervised learning of disambiguation rules for part of speech tagging. In *Natural Language Processing Using Very Large Corpora*. Kluwer Academic Press.

R. Brown. 1973. *A First Language*. Harvard University Press, Cambridge, MA.

X. Carreras and L. Màrquez. 2004. Introduction to the CoNLL-2004 shared tasks: Semantic role labeling. In *Proceedings of CoNLL-2004*, pages 89–97. Boston, MA, USA.

E. Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proc. National Conference on Artificial Intelligence*.

E.V. Clark. 1978. Awwareness of language: Some evidence from what children say and do. In R. J. A. Sinclair and W. Levelt, editors, *The child's conception of language*. Springer Verlag, Berlin.

M. Connor, Y. Gertner, C. Fisher, and D. Roth. 2008. Baby srl: Modeling early language acquisition. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*, pages xx–yy, Aug.

M. Connor, Y. Gertner, C. Fisher, and D. Roth. 2009. Minimally supervised model of early language acquisition. In *Proc. of the Annual Conference on Computational Natural Language Learning (CoNLL)*, Jun.

M. Demetras, K. Post, and C. Snow. 1986. Feedback to first-language learners. *Journal of Child Language*, 13:275–292.

K. Demuth, J. Culbertson, and J. Alter. 2006. Word-minimality, epenthesis, and coda licensing in the acquisition of english. *Language & Speech*, 49:137–174.

C. Fisher. 1996. Structural limits on verb mapping: The role of analogy in children's interpretation of sentences. *Cognitive Psychology*, 31:41–81.

Jianfeng Gao and Mark Johnson. 2008. A comparison of bayesian estimators for unsupervised hidden markov model pos taggers. In *Proceedings of EMNLP-2008*, pages 344–352.

D. Gentner. 2006. Why verbs are hard to learn. In K. Hirsh-Pasek and R. Golinkoff, editors, *Action meets word: How children learn verbs*, pages 544–564. Oxford University Press.

Y. Gertner and C. Fisher. 2006. Predicted errors in early verb learning. In *31st Annual Boston University Conference on Language Development*.

Y. Gertner, C. Fisher, and J. Eisengart. 2006. Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, 17:684–691.

J. Gillette, H. Gleitman, L. R. Gleitman, and A. Lederer. 1999. Human simulations of vocabulary learning. *Cognition*, 73:135–176.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of 45th Annual Meeting of the Association of Computational Linguists*, pages 744–751.

R. Gomez and L. Gerken. 1999. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70:109–135.

A. Haghighi and D. Klein. 2006. Prototype-drive learning for sequence models. In *Proceedings of NAACL-2006*, pages 320–327.

Mark Johnson. 2007. Why doesnt em find good hmm pos-taggers? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 296–305.

M.H. Kelly. 1992. Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, 99:349–364.

J. Lidz, H. Gleitman, and L. R. Gleitman. 2003. Understanding how input matters: verb learning and the footprint of universal grammar. *Cognition*, 87:151–178.

B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk. Third Edition*. Lawrence Elrbaum Associates, Mahwah, NJ.

M. P. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June.

Marina Meilă. 2002. Comparing clusterings. Technical Report 418, University of Washington Statistics Department.

T. Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90:91–117.

P. Monaghan, N. Chater, and M.H. Christiansen. 2005. The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96:143–182.

S. Pinker. 1984. *Language learnability and language development*. Harvard University Press, Cambridge, MA.

V. Punyakanok, D. Roth, and W. Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2).

L. R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conferenceof the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*.

J.R. Saffran, R.N. Aslin, and E.L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Rushen Shi, James L. Morgan, and Paul Allopenna. 1998. Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language*, 25(01):169–201.

Rushen Shi, Janet F. Werker, and James L. Morgan. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2):B11 – B21.

L.B. Smith and C. Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106:1558–1568.

Kiristina Toutanova and Mark Johnson. 2007. A bayesian lda-based model for semi-supervised part-of-speech tagging. In *Proceedings of NIPS*.

S. Yuan and C. Fisher. 2009. "really? she blicked the baby?": Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, 20:619–626.

S. Yuan, C. Fisher, Y. Gertner, and J. Snedeker. 2007. Participants are more than physical bodies: 21-month-olds assign relational meaning to novel transitive verbs. In *Biennial Meeting of the Society for Research in Child Development*, Boston, MA.