

Minimally Supervised Model of Early Language Acquisition

Michael Connor

Department of Computer Science
University of Illinois
connor2@uiuc.edu

Yael Gertner

Department of Psychology
University of Illinois
ygertner@cyrus.psych.uiuc.edu

Cynthia Fisher

Department of Psychology
University of Illinois
cfisher@cyrus.psych.uiuc.edu

Dan Roth

Department of Computer Science
University of Illinois
danr@uiuc.edu

Abstract

Theories of human language acquisition assume that learning to understand sentences is a partially-supervised task (at best). Instead of using ‘gold-standard’ feedback, we train a simplified “Baby” Semantic Role Labeling system by combining world knowledge and simple grammatical constraints to form a potentially noisy training signal. This combination of knowledge sources is vital for learning; a training signal derived from a single component leads the learner astray. When this largely unsupervised training approach is applied to a corpus of child directed speech, the BabySRL learns shallow structural cues that allow it to mimic striking behaviors found in experiments with children and begin to correctly identify agents in a sentence.

1 Introduction

Sentence comprehension involves assigning semantic roles to sentence constituents, determining who does what to whom. How do young children begin learning to interpret sentences? The structure-mapping view of early verb and syntax acquisition proposes that children treat the number of nouns in the sentence as a cue to its semantic predicate-argument structure (Fisher, 1996), and represent language experience in an abstract format that promotes generalization to new verbs (Gertner et al., 2006).

Theories of human language acquisition assume that learning to understand sentences is naturally a partially-supervised task: the fit of the learner’s predicted meaning with the referential context and

background knowledge provides corrective feedback (e.g., Pinker (1989)). But this feedback must be noisy; referential scenes provide ambiguous information about the semantic roles of sentence participants. For example, the same participant could be construed as an agent who ‘fled’ or as a patient who is ‘chased’.

In this paper, we address this problem by designing a Semantic Role Labeling system (SRL), equipped with shallow representations of sentence structure motivated by the structure-mapping account, that learns with no gold-standard feedback at all. Instead, the SRL provides its own internally-generated feedback based on a combination of world knowledge and linguistic constraints. As a simple stand-in for world knowledge, we assume that the learner has animacy information for some set of nouns, and uses this knowledge to determine their likely roles. In terms of linguistic constraints, the learner uses simple knowledge about the possible arguments verbs can appear with.

This approach has two goals. The first is to inform theories of language learning by investigating the utility of the proposed internally-generated feedback as one component of the human learner’s tools. Second, from an NLP and Machine Learning perspective we propose to inject information into a supervised learning algorithm through a channel other than labeled training data. From both perspectives, our key question is whether the algorithm can use these internally labeled examples to extract general patterns that can be applied to new cases.

By building a model that uses shallow representations of sentences and minimal feedback, but that

mimics features of language development in children, we can explore the nature of initial representations of syntactic structure.

1.1 Background

The structure-mapping account of early verb and syntax acquisition makes strong predictions. First, it predicts early use of simple structural cues to sentence interpretation. As soon as children can identify some nouns, they should assign different interpretations to transitive and intransitive sentences, simply by assuming that each noun in the sentence bears a distinct semantic role. Similarly, language-specific syntactic learning should transfer rapidly to new verbs. Second, however, this account predicts striking errors. In “Fred and Ginger danced”, an intransitive verb occurs with two nouns. If children interpret any two-noun sentence as if it were transitive, they should mistakenly interpret the order of two nouns in such conjoined-subject intransitive sentences as agent-patient. Experiments with young children support these predictions. 21-month-olds use the number of nouns to understand sentences containing new verbs (Yuan et al., 2007), generalize what they have learned about transitive word-order to new verbs (Gertner et al., 2006), and make the predicted error, treating intransitive sentences containing two nouns as if they were transitive (Gertner and Fisher, 2006). By 25 months, children have learned enough about English syntax to interpret conjoined-subject intransitives differently from transitives (Naigles, 1990).

Our previous computational experiments with a system for automatic semantic role labeling (Connor et al., 2008) suggest that it is possible to learn to assign basic semantic roles based on the simple representations proposed by the structure-mapping view. The classifier’s features were limited to lexical information (nouns and verbs only) and the number and order of nouns in the sentence, and trained on a sample of child-directed speech annotated in Prop-Bank (Kingsbury and Palmer, 2002) style. Given this training, our classifier learned to label the first of two nouns as an agent and the second as a patient. Even amid the variability of casual speech, simply representing the target word as the first or the second of two nouns significantly boosts SRL performance (relative to a lexical baseline) on transitive sentences

containing novel verbs. This result depends on key assumptions of the structure-mapping view, including abstract representations of semantic roles, and abstract but simple representations of sentence structure. Another approach was taken by (Alishahi and Stevenson, 2007). Their model learned to assign semantic roles without prior knowledge of abstract semantic roles. Instead, it relied on built-in syntactic knowledge and a rich hierarchical representation of semantic knowledge to learn links between sentence structure and meaning.

However, our previous experimental design has a serious drawback that limits its relevance to the study of how children learn their first language. In training, our SRL received gold standard feedback consisting of correctly labeled sentences. Thus when the SRL made a mistake in identifying the semantic role of any noun in a sentence, it received feedback about the ‘true’ semantic role of this noun. As noted above, this is an unrealistic assumption for the input to human learners.

Here we ask whether an SRL could learn to interpret simple sentences even without gold-standard feedback by relying on world knowledge to generate its own feedback. This internally-generated feedback was based on the following assumptions. First, nouns referring to animate entities are likely to be agents, and nouns referring to inanimate entities are not. Second, each predicate takes at most one agent. Such role uniqueness constraints are typically included in linguistic discussions of thematic roles (Bresnan, 1982; Carlson, 1998). The animacy heuristic is not always correct, of course. For example, in “The door hit you”, an inanimate object is the agent of action, and an animate being is the patient. Nevertheless, it is useful for two reasons. First, there is a strong cross-linguistic association between agency and animacy (Aissen, 1999; Dowty, 1991). Second, from the first year of life, children have strong expectations about the capacities of animate and inanimate entities (Baillargeon et al., in press). Given the universal tendency for speakers to talk about animate action on less animate objects, many sentences will present useful training data to the SRL: In ordinary sentences such as “You broke it,” feedback generated based on animacy will resemble gold-standard feedback.

2 Learning Model

Our learning task is similar to the full SRL task (Carreras and Màrquez, 2004), except that we classify the roles of individual words rather than full phrases. A full automatic SRL system (e.g. (Punyakanok et al., 2005a)) typically involves multiple stages to 1) parse the input, 2) identify arguments, 3) classify those arguments, and then 4) run inference to make sure the final labeling for the full sentence does not violate any linguistic constraints. Our simplified BabySRL architecture essentially replaces the first two steps with developmentally plausible heuristics. Rather than identifying arguments via a learned classifier with access to a full syntactic parse, the BabySRL treats each noun in the sentence as a candidate argument and assigns a semantic role to it. A simple heuristic collapsed compound or sequential nouns to their final noun, an approximation of the head noun of the noun phrase. For example, ‘Mr. Smith’ was treated as the single noun ‘Smith’. Other complex noun phrases were not simplified in this way. Thus, a phrase such as ‘the toy on the floor’ would be treated as two separate nouns, ‘toy’ and ‘floor’. This represents the assumption that young children know ‘Mr. Smith’ is a single name, but they do not know all the predicating terms that may link multiple nouns into a single noun phrase. The simplified learning task of the BabySRL implements a key assumption of the structure-mapping account: that at the start of multiword sentence comprehension children can tell which words in a sentence are nouns (Waxman and Booth, 2001), and treat each noun as a candidate argument.

We further simplify the SRL task such that classification is between two macro-roles: A0 (agent) and A1 (non-agent; all non-A0 arguments). We did so because we reason that this simplified feedback scheme can be primarily informative for a first stage of learning in which learners identify how their language identifies agents vs. non-agents in sentences. In addition, this level of role granularity is more consistent across verbs (Palmer et al., 2005).

For argument classification we use a linear classifier trained with a regularized perceptron update rule (Grove and Roth, 2001). This learning algorithm provides a simple and general linear classifier that works well in other language tasks, and allows

us to inspect the weights of key features to determine their importance for classification.

For the final predictions, the classifier uses predicate-level inference to ensure coherent argument assignments. In our task the only active constraints are that all nouns require a tag, and that they have unique labels, which for this restricted case of A0 vs. not A0 means there will be only one agent.

2.1 Training and Feedback

The key feature of our BabySRL lies in the way feedback is provided. Ordinarily, during training, SRL classifiers predict a semantic label for an argument and receive gold-standard feedback about its correct semantic role. Such accurate feedback is not available for the child learner. Children must rely on their own error-prone interpretation of events to supply feedback. This internally-generated feedback signal is presumably derived from multiple information sources, including the plausibility of particular combinations of argument-roles given the current situation (Chapman and Kohn, 1978). Here we model this process by combining background knowledge with linguistic constraints to generate a training signal. The ‘unsupervised’ feedback is based on: 1) nouns referring to animate entities are assumed to be agents, while nouns referring to inanimate entities are non-agents and 2) each predicate can have at most one agent.

This internally-generated feedback bears some similarities to Inference Based Training (Punyakanok et al., 2005b). In both cases the feedback to local supervised classifiers depends on global constraints. With IBT, feedback for mistakes is only considered after global inference, but for BabySRL the global inference is applied to the feedback itself. Figure 1 gives an overview of the training and testing procedure, making clear the distinction between training and testing inference.

The training data were samples of parental speech to one child (‘Sarah’; (Brown, 1973), available via Chiles (MacWhinney, 2000)). We trained on parental utterances in samples 1 through 80, recorded at child age 2;3-3;10 years. All verb-containing utterances without symbols indicating long pauses or unintelligible words were automatically parsed with the Charniak parser (Charniak, 1997) and annotated using an existing SRL sys-

tem (Punyakanok et al., 2005a). In this initial pass, sentences with parsing errors that misidentified argument boundaries were excluded. Role labels were hand-corrected using the PropBank annotation scheme. The child-directed speech training set consists of about 8300 tagged arguments over 4700 sentences, of which a majority had a single verb and two labeled nouns¹. The annotator agreement on this data set ranged between 95-97% at the level of arguments. In the current paper these role-tagged examples provide a comparison point for the utility of animacy-based feedback during training.

Our BabySRL did not receive these hand-corrected semantic roles during training. Instead, for each training example it generated its own feedback based in part on an animacy table. To obtain the animacy table we coded the 100 most frequent nouns in our corpus (which constituted less than 15% of the total number of nouns, but 65% of noun occurrences). We considered 84 of these nouns to be unambiguous in animacy: Personal pronouns and nouns referring to people were coded as animate (30). Nouns referring to objects, body parts, locations, and times, were coded as inanimate (54). The remaining 16 nouns were excluded because they were ambiguous in animacy (e.g., dolls, actions).

We test 3 levels of feedback representing increasing amounts of linguistic knowledge used to generate internal interpretations of the sentences. Using the animacy table, Animacy feedback (**Feedback 1**) was generated as follows: for each noun in training, if it was coded as animate it was labeled A0, if it was coded as inanimate it was labeled A1, otherwise no feedback was given. Because of the frequency of animate nouns this gives a skewed distribution of 4091 animate agents and 1337 inanimate non-agents.

(**Feedback 2**) builds on Feedback 1 by adding another linguistic constraint: if a noun was not found in the animacy-table and there is another noun in the sentence that is labeled A0, then the unknown noun is an A1. In the training set this adds non-agent training examples, yielding 4091 A0 and 2627 A1 examples.

Feedback 1 and Feedback 2 allow two nouns in a sentence to be labeled with A0. **Feedback 3** pre-

vents this; it implements a unique agent constraint that incorporates bootstrapping to make an ‘intelligent guess’ about which noun is the correct agent. This decision is made based on the current predictions of the classifier. Given a sentence with multiple animate nouns, the classifier predicts a label for each, and the one with the highest score for A0 is declared the true agent and the rest are classified as non-agent. Note that we cannot apply role uniqueness to the A1 (not A0) role, given that this label encompasses multiple non-agent roles. This feedback scheme, allowing at most one agent per sentence, reduces the number of A0 examples and increases the number of A1 examples to 3019 A0 and 3699 A1.

2.2 Feature Sets

The basic feature we propose is the noun pattern feature (NPattern). We hypothesize that children use the number and order of nouns to represent argument structure. The NPattern feature indicates how many nouns there are in the sentence and which noun the target is. For example, in the two-noun sentence ‘Did you see it?’, ‘you’ has a feature active indicating that it is the first noun of two. Likewise, for ‘it’ a feature is active indicating that it is the second of two nouns. This feature is easy to compute once nouns are identified, and does not require fine-grained part-of-speech distinctions.

We compare the noun pattern feature to a baseline lexical feature set (Words): the target noun and the root form of the predicate. The NPattern feature set includes lexical features as well as features indicating the number and order of the noun (first of two, second of three, etc.). With gold-standard role feedback, (Connor et al., 2008) found that the NPattern feature allowed the BabySRL to generalize to new verbs: it increased the system’s tendency to predict that the first of two nouns was A0 and the second of two nouns A1 for verbs not seen in training.

To the extent that in child-directed speech the first of two nouns tends to be an agent, and agents tend to be animate, we anticipate that with the NPattern feature the BabySRL will learn the same thing, even when provided with internally-generated feedback based on animacy. In Connor et al. (2008) we showed that, because this NPattern feature set represents only the number and order of nouns, with this feature set the BabySRL reproduced the errors chil-

¹Corpus available at <http://l2r.cs.uiuc.edu/~cogcomp>

```

Algorithm BABYSRL TRAINING
INPUT: Unlabeled Training Sentences
OUTPUT: Trained Argument Classifier

For each training sentence
  Generate Internal Feedback: Find interpreted meaning
  Feedback 1: Apply Animacy Heuristic
  For each argument in the sentence (noun)
    If noun is animate → mark as agent
    If noun is inanimate → mark as non-agent
    else leave unknown
  end

  Feedback 2: Known agent constraint
  Beginning with Feedback 1
  If an agent was found
    Mark all unknown arguments as non-agent

  Feedback 3: Unique agent constraint
  Beginning with Feedback 2
  If multiple agents found
    Find argument with highest agent prediction
    Leave this argument an agent, mark rest as non-agent

  Train Supervised Classifier
  Present each argument to classifier
  Update if interpreted meaning does not match
  classifier prediction
end

```

(a) Training

```

Algorithm BABYSRL TESTING
INPUT: Unlabeled Testing Sentences
OUTPUT: Role labels for each argument

For each test sentence
  Predict roles for each argument
  Test Inference:
  Find assignment to whole sentence with highest sum of
  predictions that doesn't violate uniqueness constraint
end

```

(b) Testing

Figure 1: BabySRL training and testing procedures. Internal feedback is generated using animacy plus optional constraints. This feedback is fed to a supervised learning algorithm to create an agent-identification classifier.

dren make as noted in the Introduction, mistakenly assigning agent- and non-agent roles to the first and second nouns in intransitive test sentences containing two nouns. In the present paper, the linguistic constraints provide an additional cause for this error. In addition, as a first step in examining recovery from the predicted error, Connor et al. (2008) added a verb position feature (VPosition) specifying whether the target noun is before or after the verb. Given these features, the BabySRL’s classification

of transitive and two-noun intransitive test sentences diverged, because the gold-standard training supported the generalization that pre-verbal nouns tend to be agents, and post-verbal nouns tend to be patients. In the present paper we include the VPosition feature for comparison to Connor et al. (2008).

2.3 Testing

To evaluate the BabySRL we tested it with both a held-out sample of child-directed speech, and with constructed sentences containing novel verbs, like those used in the experiments with children described above. These sentences provide a more stringent test of generalization than the customary test on a held-out section of the data. Although the held-out section of data contains unseen sentences, it may contain few unseen verbs. In a held out section of our data, 650 out of 696 test examples contain a verb that was encountered in training. Therefore, the customary test cannot tell us whether the system generalizes what it learned to novel verbs.

All constructed test sentences contained a novel verb (‘gorp’). We constructed two test sentence templates: ‘A gorp B’ and ‘A and B gorp’, where A and B were replaced with nouns that appeared more than twice in training. For each test sentence template we built a test set of 100 sentences by randomly sampling nouns in two different ways described next.

Full distribution: The first nouns in the test sentences (A) are chosen from the set of all first nouns in our corpus, taking their frequency into account when sampling. The second nouns in the sentences (B) are chosen from the set of nouns appearing as second nouns in the sentence of our corpus. This way of sampling the nouns will maximize the SRL’s test performance based on the baseline feature set of lexical information alone (Words). This is so because in our data many sentences have an animate first noun and an inanimate second noun. Based on these words alone the SRL could learn to predict an A0-A1 role sequence for our test sentences. Nevertheless, we expect that when the BabySRL is also given the NPattern feature it should be able to perform better than this high lexical baseline.

Two animate nouns: In these test sentences the A and B nouns are chosen from our list of animate nouns. We chose nouns from this list that were fairly frequent (ranging from 8 to 240 uses in the

corpus), and that occurred roughly equally as the first and second noun. This mimics the sentences used in the experiments with children (e.g., “The girl is kradding the boy!”). The lexical baseline system’s tendency to assign an A0-A1 sequence to these nouns should be much lower for these test sentences. We therefore expect the contribution of the NPattern feature to be more apparent in these test sentences.

The test sentences with novel verbs ask whether the classifier transfers its learning about argument role assignment to unseen verbs. Does it assume the first of two nouns in a simple transitive sentence (‘A gorp B’) is the agent (A0) and the second is not an agent (A1)? In (Connor et al., 2008) we showed that a system with the same feature and representations also over-generalized this rule to two-noun intransitives (‘A and B gorp’), mimicking children’s behavior. In the present paper this error is over-determined, because the classifier learns only an agent/non-agent contrast, and the linguistic constraints forbid duplicate agents in a sentence. However, for comparison to the earlier paper we test our system on the ‘A and B gorp’ sentences as well.

3 Experimental Results

Our experiments use internally-generated feedback to train simple, abstract structural features: the NPattern features that proved useful with gold-standard training in Connor et al. (2008). Section 3.1 tests the system on agent-identification in held-out sentences from the corpus, and demonstrates that the animacy-based feedback is useful, yielding SRL performance comparable to that of a system trained with 1000 sentences of gold-standard feedback. Section 3.2 presents the critical novel-verb test data, demonstrating that this system replicates key findings of (Connor et al., 2008) with no gold standard feedback. Using only noisy internally-generated feedback, the BabySRL learned that the first of two nouns is an agent, and generalized this knowledge to sentences with novel verbs.

3.1 Comparing Self Generated Feedback with Gold Standard Feedback

Table 1 reports for the varying feedback schemes, the A0 F1 performance for a system with either lexical baseline feature (Words) or structural features

Feedback	Words	+NPattern
1. Just Animacy	0.72	0.73
2. + non A0 Inference	0.74	0.75
3. + unique A0 bootstrap	0.70	0.74
10 Gold	0.43	0.47
100 Gold	0.61	0.65
1000 Gold	0.75	0.76

Table 1: Agent identification results (A0 F1) on held-out sections of the Sarah Childe corpus. We compare a classifier trained with various amounts of gold labeled data (averaging over 10 different samples at each level of data). For noun pattern features the internally generated bootstrap feedback provides comparable accuracy to training with between 100-1000 fully labeled examples.

(+NPattern) when tested on a held-out section of the Sarah Childe corpus section 84-90, recorded at child ages 3;11-4;1 years. Agent identification based on lexical features is quite accurate given animacy feedback alone (Feedback 1). As expected, because many agents are animate, the animacy tagging heuristic itself is useful. As linguistic constraints are added via non-A0 inference (Feedback 2), performance increases for both the lexical baseline and NPattern feature-set, because the system experiences more non-A0 training examples.

When the unique A0 constraint is added (Feedback 3), the lexical baseline performance decreases, because for the first time animate nouns are being tagged as non-agents. With this feedback the NPattern feature set yields a larger improvement over lexical baseline, showing that it extracts more general patterns. We discuss the source of these feedback differences in the novel-verb test section below.

We compared the usefulness of the internally-generated feedback to gold-standard feedback by training a classifier equipped with the same features on labeled sentences. We reduced the SRL labeling for the training sentences to the binary agent/non-agent set, and trained the classifier with 10, 100, or 1000 labeled examples. Surprisingly, the simple feedback derived from 84 nouns labeled with animacy information yields performance equivalent to between 100 and 1000 hand-labeled examples.

Feedback	Full Distribution Nouns			Animate Nouns		
	Words	NPattern	VPosition	Words	NPattern	VPosition
‘A gorps B’						
1. Animacy	0.86	0.86	0.87	0.76	0.79	0.70
2. + non A0 Inference	0.87	0.92	0.90	0.63	0.86	0.85
3. + unique A0 bootstrap	0.87	0.95	0.89	0.63	0.82	0.66
‘A and B gorp’						
1. Animacy	0.86	0.86	0.84	0.76	0.79	0.68
2. + non A0 Inference	0.87	0.92	0.85	0.63	0.86	0.66
3. + unique A0 bootstrap	0.87	0.95	0.86	0.63	0.82	0.63

Table 2: Percentage of sentences interpreted as agent first (%A0-A1) by the BabySRL when trained on unlabeled data with the 3 internally-generated feedback schemes described in the text. Two different two-noun sentence structures were used (‘A gorps B’, ‘A and B gorp’), along with two different methods of sampling the nouns (Full Distribution, Animate Nouns) to create test sets with 100 sentences each.

3.2 Comparing Structural Features with Lexical Features

The previous section shows that the BabySRL equipped with simple structural features can use internally generated feedback to learn a simple agent/non-agent classification, and apply it to unseen sentences. In this section we probe what the SRL has learned by testing generalization to new verbs in constructed sentences. Table 2 summarizes these experiments. The results are broken down both by what sentence structure is used in test (‘A gorps B’, ‘A and B gorp’) and how the nouns ‘A’ and ‘B’ are sampled (Full Distribution, Animate Nouns). The results are presented in terms of %A0A1: the percentage of test sentences that are assigned an Agent role for ‘A’ and a non-Agent role for ‘B’.

For the transitive ‘A gorps B’ sentences, A0A1 is the correct interpretation; A should be the agent. As predicted, when A and B are sampled from the full distribution of nouns, simply basing classification on the Words feature-set already strongly predicts this A0A1 ordering for the majority of cases. This is because the data (language in general, child directed speech in particular here) are naturally distributed such that particular nouns that refer to animates tend to be agents, and tend to appear as first nouns, and those that refer to inanimates tend to be non-agents and second nouns. Thus, a learner representing sentence information in terms of words only succeeds with full-distribution ‘A gorps B’ test sentences even with the simplest animacy feedback (Feedback 1);

the A and B nouns in these test sentences reproduce the learned distribution. Also as predicted, given this simple feedback, the additional higher-level features (NPattern, VPosition) do not improve much upon the lexical baseline. This is due to the strictly lexical nature of the animacy feedback: each lexical item (e.g., ‘you’ or ‘it’) will always either be animate or inanimate and therefore either A0 or A1. Therefore, in this case lexical features are the best predictors.

Also as expected, higher-level features (NPattern, and VPosition) improve performance with a more sophisticated self-generated feedback scheme. Adding inferred feedback to label unknown nouns as A1 when the sentence contains a known animate noun (Feedback 2) decreases the ratio of A0 to non-A0 arguments. This feedback is less lexically determined: for nouns whose animacy is unknown, feedback will be provided only if there is another animate noun in the sentence. This leaves room for the abstract structural features to play a role.

Next we test a form of the unique-A0 constraint. In (Feedback 3), in addition to the non-A0 inference added in (Feedback 2), the BabySRL intelligently selects one noun as A0 in sentences with multiple animate nouns. With this feedback we see a striking increase in test performance based on the noun pattern features over the lexical baseline. In principle, this feedback mechanism might permit the classifier to start to learn that animate nouns are not always agents. Early in training, the noun pattern feature learns that first nouns tend to be animate (and therefore interpreted as agents), and it feeds this informa-

tion back into subsequent training examples, generating new feedback that continues to interpret as agents those animate nouns that appear first in sentences containing two animates.

For the nouns sampled from the full distribution we see that structural features improve over the lexical baseline despite the high performance of the lexical baseline. This finding tells us that simple representations of sentence structure can be useful in learning to interpret sentences even with no gold-standard training. Provided only with simple internally-generated feedback based on animacy knowledge and linguistic constraints, the BabySRL learned that the first of two nouns tends to be an agent, and the second of two does not.

The results for the ‘A B gorp’ test sentences demonstrate an important way in which predictions based on different simple structure representations can diverge. As expected, the NPattern feature makes the same overgeneralization error seen by children and the system in (Connor et al., 2008). However, when the VPosition feature is added, different results are obtained for the ‘A gorp B’ and ‘A and B gorp’ sentences. The SRL predicts fewer A0A1 for ‘A and B gorp’ (it cannot predict the expected A0A0 because of the uniqueness constraint used in test inference).

Next, we replicate our findings by performing the same experiments with test sentences in which both ‘A’ and ‘B’ are animate. Because lexical features alone cannot determine if ‘A’ or ‘B’ should be the agent, it is a more sensitive test of generalization.

When we look at the lexical baseline for animate sentences, the agent-first percentage is lower compared to the full distribution results, because the word features indicate nearly evenly that both nouns should be agents, so the Words baseline model must rely on small, chance differences in its experience with particular words. This percentage is still well above chance due to the method used to apply inference during testing. Recall that the classifier uses predicate-level inference at test to ensure that only one argument is labeled A0. This inference is implemented using a beam search that looks at arguments in a fixed order and roles from A0 up. Thus in the case of ties there is a preference for first seen solutions, meaning A0A1 in this case. This bias has a large effect on the SRL’s baseline performance with

the test sentences containing two animate nouns. Despite this high baseline, however, because lexical features alone cannot determine if ‘A’ or ‘B’ should be the agent, we are able to see more clearly the improvement gained by including structural features.

Regardless of our testing scheme, we see that as the feedback incorporates more information, both added linguistic constraints and the SRL’s own prior learning, the noun pattern structural feature is better used to identify agents beyond the lexical baseline. The largest improvement over this lexical baseline is obtained by combining knowledge of animacy with a single-agent constraint and bootstrapping predictions based on prior learning.

4 Conclusion and Future Work

Conventional approaches to supervised learning require creating large amounts of hand-labeled data. This is labor-intensive, and limits the relevance of the work to the study of how children learn languages. Children do not receive perfect feedback about sentence interpretation. Here we found that our simple SRL classifier can, to a surprising degree, attain performance comparable to training with 1000 sentences of labeled data. This suggests that fully labeled training data can be supplemented by a combination of simple world knowledge (animates make good agents) and linguistic constraints (each verb has only one agent). The combination of these sources provides an informative training signal that allows our BabySRL to learn a high-level semantic task and generalize beyond the training data we provided to it. The SRL learned, based on the distribution of animates in sentences of child-directed speech, that the first of two nouns tends to be an agent. It did so based on representations of sentence structure as simple as the ordered set of nouns in the sentence. This demonstrates that it is possible to learn how to correctly assign semantic roles based on these very simple cues. This together with experimental work (e.g. (Fisher, 1996) suggests that such representations might play a role in children’s early sentence comprehension.

Acknowledgments

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

References

- J. Aissen. 1999. Markedness and subject choice in optimality theory. *Natural Language and Linguistic Theory*, 17:673–711.
- A. Alishahi and S. Stevenson. 2007. A computational usage-based model for learning general properties of semantic roles. In *Proceedings of the 2nd European Cognitive Science Conference*.
- R. Baillargeon, D. Wu, S. Yuan, J. Li, and Y. Luo. (in press). Young infants' expectations about self-propelled objects. In B. Hood and L. Santos, editors, *The origins of object knowledge*. Oxford University Press, Oxford.
- J. Bresnan. 1982. *The mental representation of grammatical relations*. MIT Press, Cambridge MA.
- R. Brown. 1973. *A First Language*. Harvard University Press, Cambridge, MA.
- G. Carlson. 1998. Thematic roles and the individuation of events. In S. D. Rothstein, editor, *Events and Grammar*, pages 35–51. Kluwer, Dordrecht.
- 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.
- R. S. Chapman and L. L. Kohn. 1978. Comprehension strategies in two- and three-year-olds: Animate agents or probable events? *Journal of Speech and Hearing Research*, 21:746–761.
- E. Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proc. National Conference on Artificial Intelligence*.
- 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)*, Aug.
- D. Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67:547–619.
- C. Fisher. 1996. Structural limits on verb mapping: The role of analogy in children's interpretation of sentences. *Cognitive Psychology*, 31:41–81.
- 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.
- A. Grove and D. Roth. 2001. Linear concepts and hidden variables. *Machine Learning*, 42(1/2):123–141.
- P. Kingsbury and M. Palmer. 2002. From Treebank to PropBank. In *Proceedings of LREC-2002*, Spain.
- B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates, Mahwah, NJ.
- L. R. Naigles. 1990. Children use syntax to learn verb meanings. *Journal of Child Language*, 17:357–374.
- M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. In *Computational Linguistics 31(1)*.
- S. Pinker. 1989. *Learnability and Cognition*. Cambridge: MIT Press.
- V. Punyakanok, D. Roth, and W. Yih. 2005a. The necessity of syntactic parsing for semantic role labeling. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1117–1123.
- V. Punyakanok, D. Roth, W. Yih, and D. Zimak. 2005b. Learning and inference over constrained output. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1124–1129.
- S. R. Waxman and A. Booth. 2001. Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology*, 43:217–242.
- 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.