Learning Words and Their Meanings from Unsegmented Child-directed Speech

Bevan K. Jones & Mark Johnson
Dept of Cognitive and Linguistic Sciences
Brown University
Providence, RI 02912, USA
{Bevan_Jones,Mark_Johnson}@Brown.edu

Michael C. Frank
Dept of Brain and Cognitive Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
mcfrank@mit.edu

Abstract

Most work on language acquisition treats word segmentation—the identification of linguistic segments from continuous speech—and word learning—the mapping of those segments to meanings—as separate problems. These two abilities develop in parallel, however, raising the question of whether they might interact. To explore the question, we present a new Bayesian segmentation model that incorporates aspects of word learning and compare it to a model that ignores word meanings. The model that learns word meanings proposes more adult-like segmentations for the meaning-bearing words. This result suggests that the non-linguistic context may supply important information for learning word segmentations as well as word meanings.

1 Introduction

Acquiring a language entails mastering many learning tasks simultaneously, including identifying where words begin and end in continuous speech and learning meanings for those words. It is common to treat these tasks as separate, sequential processes, where segmentation is a prerequisite to word learning but otherwise there are few if any dependencies. The earliest evidence of segmentation, however, is for words bordering a child's own name (Bortfeld et al., 2005). In addition, infants begin learning their first words before they achieve adult-level competence in segmentation. These two pieces of evidence raise the question of whether the tasks of meaning learning and segmentation might mutually inform one another.

To explore this question we present a joint model that simultaneously identifies word boundaries and attempts to associate meanings with words. In doing so we make two contributions. First, by modeling the two levels of structure in parallel we simulate a more realistic situation. Second, a joint model allows us to explore possible synergies and interactions. We find evidence that our joint model performs better on a segmentation task than an alternative model that does not learn word meanings.

The picture in Figure 1 depicts a language learning situation from our corpus (originally from Fernald and Morikawa, 1993; recoded in Frank et al., 2009) where a mother talks while playing with various toys. Setting down the dog and picking up the hand puppet of a pig, she asks, "Is that the pig?" Starting out, a young learner not only does not know that the word "pig" refers to the puppet but does not even know that "pig" is a word at all. Our model simulates the learning task, taking as input the unsegmented phonemic representation of the speech along with the set of objects in the non-linguistic context as shown in Figure 1 (a), and infers both a segmentation and a word-object mapping as in Figure 1 (b).

One can formulate the word learning task as that of finding a reasonably small set of reusable word-meaning pairs consistent with the underlying communicative intent. Infant directed speech often refers to objects in the immediate environment, and early word learning seems to involve associating frequently co-occurring word-object pairs (Akhtar and Montague, 1999; Markman, 1990). Several computational models are based on this idea that a word

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 501–509,
Los Angeles, California, June 2010.

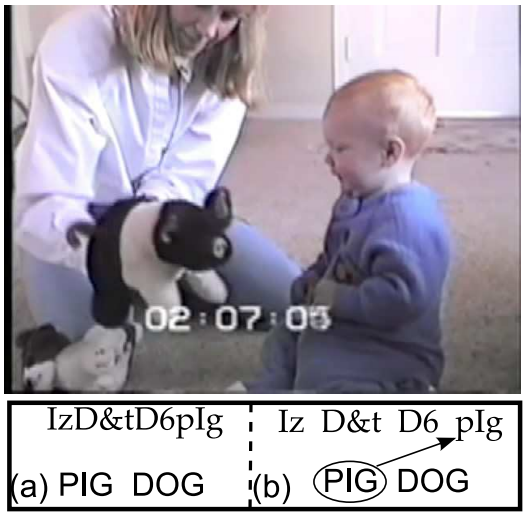

Figure 1: (a) The input to our system for the utterance "Is that the pig?" consists of an unsegmented sequence of phonemes and the set of objects representing the non-linguistic context. These objects were manually identified by inspecting the associated video, a frame from which is shown above. (b) The gold-standard segmentation and word-object assignments of the same utterance, against which the output of our system is evaluated (all words except "pIg" are mapped to a special "null" object, as explained in the text).

that frequently occurs in the presence of an object and not so frequently in its absence is likely to refer to that object (Frank et al., 2009a; Siskind, 1996; Yu and Ballard, 2007). Importantly, all these models assume words are pre-segmented in the input.

While the word segmentation task relates less clearly to the communicative content, it can be formulated according to a similar objective, that of attempting to explain the sound sequences in the input in terms of some reasonably small set of reusable units, or words. Computational models have successfully addressed the problem in much this way (Johnson and Goldwater, 2009; Goldwater et al., 2009; Brent, 1999), and the general approach is consistent with experimental observations that humans are sensitive to statistics of sound sequences (Saffran et al., 1996; Frank et al., 2007).

The two tasks can be integrated in a relatively seamless way, since, as we have just formulated them, they have a common objective, that of finding a minimal, consistent set of reusable units. However, the two deal with different types of information with different dependencies. The basic idea is that learning a vocabulary that both meets the constraints of the word-learning task and is consistent with the objective of the segmentation task can yield a better segmentation. That is, we hope to find a synergy in the joint inference of meaning and segmentation.

Note that to the best of our knowledge there is very little computational work that combines word form and word meaning learning (Frank et al. 2006 takes a first step but their model is applicable only to small artificial languages). Frank et al. (2009a) and Regier (2003) review pure word learning models and, in addition to the papers we have already cited, Brent (1999) presents a fairly comprehensive review of previous pure segmentation models. However, none of the models reviewed make any attempt to jointly address the two problems. Similarly, in the behavioral literature on development, we are aware of only one segmentation study (Graf-Estes et al., 2007) that involves non-linguistic context, though this study treats the two tasks sequentially rather than jointly.

We now describe our model and inference procedure and follow with evaluation and discussion.

2 Model Definition

Cross-situational meaning learning in our joint word learning and segmenting model is inspired by the model of Frank et al. (2009a). Our model can be viewed as a variant of the Latent Dirichlet Allocation (LDA) topic model of Blei et al. (2003), where topics are drawn from the objects in the non-linguistic context. The model associates each utterance with a single referent object, the topic, and every word in the utterance is either generated from a distribution over words associated with that object or else from a distribution associated with a special "null" object shared by all utterances. Note that in this paper we use "topic" to denote the referent object of an utterance, otherwise we depart from topic modeling convention and use the term "object" instead.

Segmentation is based on the unigram model proposed by Brent (1999) and reformulated by Goldwater et al. (2009) in terms of a Dirichlet process. Since both LDA and the unigram segmenter are based on unigram distributions it is relatively straightforward

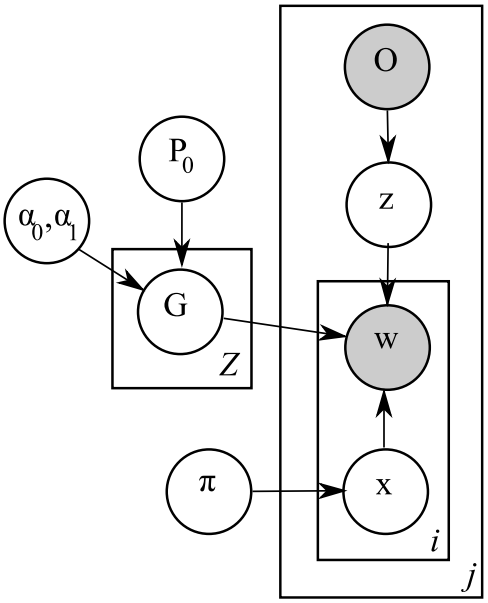

Figure 2: Topical Unigram Model: O_j is the set of objects in the non-linguistic context of the j^{th} utterance, z_j is the utterance topic, w_{ji} is the i^{th} word of the utterance, x_{ji} is the category of the word (referring or non-referring), and the other variables are distribution parameters.

to integrate the two to simultaneously infer word boundaries and word-object associations.

Figure 2 illustrates a slightly simplified form of the model, and the the relevant distributions are as follows:

$$
\begin{aligned}
z|O &\sim Uniform(O) \\
G_z|z,\alpha_0,\alpha_1,P_0 &\sim \begin{cases} DP(\alpha_1,P_0) & \text{if } z \neq 0 \\ DP(\alpha_0,P_0) & \text{otherwise} \end{cases} \\
\pi &\sim Beta(1,1) \\
x|\pi &\sim Bernoulli(\pi) \\
w|G,z,x &\sim \begin{cases} G_z & \text{if } x = 1 \\ G_0 & \text{if } x = 0 \end{cases}
\end{aligned}
$$

Note that $Uniform(O)$ denotes a discrete uniform distribution over the elements of the set O. P_0 is described later.

Briefly, each utterance has a single topic z_j, drawn from the objects in the non-linguistic context O_j, and then for each word w_{ji} we first flip a coin x_{ji} to determine if it refers to the topic or not. Then, depending on x_{ji} the word is either drawn from a distribution specific to the topic ($x_{ji} = 1$) or from a distribution associated with the "null" object ($x_j i = 0$). In slightly greater detail but still glossing over the details of how the multinomial parameters are generated, the generative story proceeds as follows:

1. For each utterance, indexed by j
2. (a) Pick a single topic z_j uniformly from the set of objects in the environment O_j
   (b) For each word w_{ji} of the utterance
   (c) i. Determine if it refers to z_j or not by setting x_{ji} to 1 (referring) with probability π, and to 0 (non-referring) otherwise.
   ii. if x_{ji} is 1, draw w_{ji} from the topic specific distribution over words G_{z_j}.
   iii. otherwise, draw w_{ji} from G_0, the distribution over words associated with the "null" object.

This generative story is a simplification since it does not describe how we model utterance boundaries. It is important for segmentation purposes to explicitly model utterance boundaries since, unlike utterance-internal word boundaries, we assume utterance boundaries are observed. Thus, the story is complicated by the fact that there is a chance each time we generate a word that we also generate an utterance boundary. The choice of whether to terminate the utterance or not is captured by a $Bernoulli(\gamma)$ random variable $\$_{ji}$ indicating whether the i^{th} word was the last word of the j^{th} utterance.

$$
\begin{aligned}
\gamma &\sim Beta(1,1) \\
\$|\gamma &\sim Bernoulli(\gamma)
\end{aligned}
$$

The G_z multinomial parameters are generated from a Dirichlet process with base distribution over words, P_0, which describes how new word types are generated from their constituent phonemes. Phonemes are generated sequentially, i.i.d. uniformly from m phonemic types. In addition, there is a probability $p_\#$ of generating a word boundary.

$$P_0(w) = (1-p_\#)^{|w|-1} p_\# \frac{1}{m^{|w|}}$$

The concentration parameters α_0 and α_1 also play a critical role in the generation of words and word types. Any given word has a certain probability of either being produced from the set of previously seen word types, or from an entirely new one. The

greater the concentration parameter, the more likely the model is to appeal to the base distribution P_0 to introduce a new word type.

Like Frank et al. (2009a), we distinguish between two coarse grammatical categories, referring and non-referring. Referring words are generated by the topic, while non-referring words are drawn from G_0, a distribution associated with the "null" object. The distinction ensures sparse word-object maps that obey the principle of mutual exclusion. Otherwise all words in the utterance would be associated with the topic object, resulting in a very large set of words for each object that is very likely to overlap with the words for other objects. As a further bias toward a small lexicon, we employ different concentration parameters (α_0 and α_1) for the non-referring and referring words, using a much smaller value for the referring words. Intuitively, there should be a relatively small prior probability of introducing a new word-object pair, corresponding to a small α_1 value. On the other hand, most other words don't refer to the topic object (or any other object for that matter), corresponding to a much larger α_0 value.

Note that this topical unigram model is a straightforward generalization of the unigram segmentation model (Goldwater et al., 2009) to the case of multiple topics. In fact, if all words were assumed to refer to the same object (or to no object at all) the models would be identical.

Unlike LDA, each "document" has only one topic, which is necessitated by the fact that in our model documents correspond single utterances. The utterances in our corpus of child directed speech are often only four or five words long, whereas the general LDA model assumes documents are much larger. Thus, there may not be enough words to infer a useful utterance specific distribution over topics. Consequently, rather than inferring a separate topic distribution for each utterance, we simply assume a uniform distribution over objects in the non-linguistic context. In effect, we rely entirely on the non-linguistic context and word-object associations to infer topics. Though necessitated by data sparsity issues, we also note that it is very rare in our corpus for utterances to refer to more than one object in the non-linguistic context, so the choice of a single topic may also be a more accurate model. In fact, even with multi-sentence documents, LDA may perform better if only one topic is assumed per sentence (Gruber et al., 2007).

3 Inference

We use a collapsed Gibbs sampling procedure, integrating over all possible G_z, π, and γ values and then iteratively sample values for each variable conditioned on the current state of all other variables. We visit each utterance once per iteration, sample a topic, and then visit each possible word boundary location to sample the boundary and word categories simultaneously according to their joint probability.

A single topic is sampled for each utterance, conditioned on the words and their current determinations as referring or non-referring. Since z_j is drawn from a uniform distribution, this probability is simply proportionate to the conditional probability of the words given z_j and the x_{ji} variables.

$$P(z_j|\mathbf{w_j}, \mathbf{x_j}, \mathbf{h}^{-j}) \propto \frac{\Gamma(\sum_w^{W_j} n_{w,z_j}^{(\mathbf{h}^-)} + \alpha_1 P_0(w))}{\Gamma(\sum_w^{W_j} n_{w,z_j}^{(\mathbf{h})} + \alpha_1 P_0(w))} \cdot \prod_w^{W_j} \frac{\Gamma(n_{w,z_j}^{(\mathbf{h})} + \alpha_1 P_0(w))}{\Gamma(n_{w,z_j}^{(\mathbf{h}^-)} + \alpha_1 P_0(w))}$$

Here, $P(z_j|\mathbf{w_j}, \mathbf{x_j}, \mathbf{h^{-j}})$ is the probability of topic z_j given the current hypothesis $\mathbf{h}$ for all variables excluding those for the current utterance. Also, $n_{w,z_j}^{(\mathbf{h}^{-j})}$ is the count of occurrences of word type w that refer to topic z_j among the current variable assignments, and W_j is the set of word types appearing in utterance j. The vectors of word and category variables in utterance j are represented as $\mathbf{w_j}$ and $\mathbf{x_j}$, respectively. Note that only referring words have any bearing on the appropriate selection of z_j and so all factors involving only non-referring words are absorbed by the constant of proportionality.

The word categories can be sampled conditioned on the current word boundary states according to the following conditional probability, where $n_{x_{ji}}^{(\mathbf{h}^{-ji})}$ is the number of words categorized according to label

x_{ji} over the entire corpus excluding word w_{ji}.

$$\begin{aligned} P(x_{ji}|w_{ji}, z_j, \mathbf{h}^{-ji}) &\propto P(w_{ji}|z_j, x_{ji}, \mathbf{h}^{-ji}) \\ &\quad \cdot P(x_{ji}|\mathbf{h}^{-ji}) \\ &= \frac{n^{(\mathbf{h}^{-ji})}_{w_{ji},x_{ji}z_j} + \alpha_{x_{ji}} P_0(w_{ji})}{n^{(\mathbf{h}^{-ji})}_{\bullet,x_{ji}z_j} + \alpha_{x_{ji}}} \cdot \frac{n^{(\mathbf{h}^{-ji})}_{x_{ji}} + 1}{n^{(\mathbf{h}^{-ji})}_{\bullet} + 2} \end{aligned} \quad (1)$$

In practice, however, we actually sample the word category variables jointly with the boundary states, using a scheme similar to that outlined in Goldwater et al. (2009). We visit each possible word boundary location (any point between two consecutive phonemes) and compute probabilities for the hypotheses for which the phonemic environment makes up either one word or two. As illustrated below there are two sets of cases: those where we treat the segment as a single word, and those where we treat it as two words.

$$\underset{\uparrow}{\overset{x_1}{\ldots\#w_1\#\ldots}} \quad \text{vs.} \quad \underset{\uparrow}{\overset{x_2 \quad x_3}{\ldots\#w_2\#w_3\#\ldots}}$$

The probabilities of the hypotheses can be derived by application of equation 1. Since the $\mathbf{x}$ variables can each describe two possible events, there are a total of six different cases to consider for each boundary assignment: two cases without and four with a word boundary.

The probability of each of the two cases without a word boundary can be computed as follows:

$$\begin{aligned} P(w_1, x_1|z, \mathbf{h}^-) = &\frac{n^{(\mathbf{h}^-)}_{w_1,x_1z} + \alpha_{x_1} P_0(w_1)}{n^{(\mathbf{h}^-)}_{\bullet,x_1z} + \alpha_{x_1}} \\ &\cdot \frac{n^{(\mathbf{h}^-)}_{x_1} + 1}{n^{(\mathbf{h}^-)}_{\bullet} + 2} \cdot \frac{n^{(\mathbf{h}^-)}_{\$_1} + 1}{n^{(\mathbf{h}^-)}_{\bullet} + 2} \end{aligned}$$

Here $\mathbf{h}^-$ signifies the current hypothesis for all variables excluding those for the current segment and $n^{(\mathbf{h}^-)}_{\$_1}$ is the count for $\mathbf{h}^-$ of either utterance final words if w_1 is utterance final or non-utterance final words if w_1 is also not utterance final.

In the four cases with a word boundary, we have two words and two categories to sample.

$$\begin{aligned} P(w_2, x_2, w_3, x_3|z, \mathbf{h}^-) = &\frac{n^{(\mathbf{h}^-)}_{w_2,x_2z} + \alpha_{x_2} P_0(w_2)}{n^{(\mathbf{h}^-)}_{\bullet,x_2z} + \alpha_{x_2}} \\ &\cdot \frac{n^{(\mathbf{h}^-)}_{x_2} + 1}{n^{(\mathbf{h}^-)}_{\bullet} + 2} \cdot \frac{n^{(\mathbf{h}^-)}_{\$_2=0} + 1}{n^{(\mathbf{h}^-)}_{\bullet} + 2} \\ &\cdot \frac{n^{(\mathbf{h}^-)}_{w_3,x_3z} + \delta_{x_2}(x_3)\delta_{w_2}(w_3) + \alpha_{x_3} P_0(w_3)}{n^{(\mathbf{h}^-)}_{\bullet,x_3z} + \delta_{x_2}(x_3) + \alpha_{x_3}} \\ &\cdot \frac{n^{(\mathbf{h}^-)}_{x_3} + \delta_{x_2}(x_3) + 1}{n^{(\mathbf{h}^-)}_{\bullet} + 3} \cdot \frac{n^{(\mathbf{h}^-)}_{\$_3} + \delta_{\$_2}(\$_3) + 1}{n^{(\mathbf{h}^-)}_{\bullet} + 3} \end{aligned}$$

Here $\delta_x(y)$ is 1 if $x = y$ and 0 otherwise.

4 Results & Model Comparisons

4.1 Corpus

Our training corpus (Fernald and Morikawa, 1993; Frank et al., 2009b) consists of about 22,000 words and 5,600 utterances. Video recordings consisting of mother-child play over pairs of toys were orthographically transcribed, and each utterance was annotated with the set of objects present in the non-linguistic context. The object referred to by the utterance, if any, was noted, as described in Frank et al. (2009b). We used the VoxForge dictionary to map orthographic words to phoneme sequences in a process similar to that described in Brent (1999).

Figure 1 (a) presents an example of the coding of phonemic transcription and non-linguistic context for a single utterance. The input to the system consists solely of the phonemic transcription and the objects in the non-linguistic context.

4.2 Evaluation

We ran the sampler ten times for 100,000 iterations with parameter settings of $\alpha_1 = 0.01$, $\alpha_0 = 20$, and $p_\# = 0.5$, keeping only the final sample for evaluation. We defined the word-object pairs for a sample as the words in the referring category that were paired at least once with a particular topic. These pairs were then compared against a gold standard set of word-object pairs, while segmentation performance was evaluated by comparing the final boundary assignments against the gold standard segmentation.

4.2.1 Word Learning

To explore the contribution of word boundaries to the joint word learning and segmenting task, we compare our full joint model against a variant that only infers topics, using the gold standard segmentation as input. In this way we also reproduce the usual assumption of a sequential relationship between segmentation and word learning and test the necessity of the simplifying assumption. The results are shown in Table 2. We compare them with three different metric types: topic accuracy; precision, recall, and F-score of the word-object pairs; and Kullback-Liebler (KL) divergence.

First, treating utterances with no referring words as though they have no topic, we compute the accuracy of the inferred topics. Note that we don't report accuracy for the the variant with no non-linguistic context, since in this case the objects are interchangeable, and we have a problem identifying which cluster corresponds to which topic. Table 2 shows that the joint segmentation and word learning model gets the topic right for 81% of the utterances. The variant that assumes pre-segmented input does comparably well with an accuracy of 79%. Surprisingly, it seems that knowing the gold segmentation doesn't add very much, at least for the topic inference task.

To evaluate how well we discovered the word-object map, we manually compiled a list of all the nouns in the corpus that named one of the 30 objects. We used this set of nouns, cross-referenced with their topic objects, as a gold standard set of word-object pairs. By counting the co-occurrences, we also compute a gold standard probability distribution for the words given the topic, $P(w|z, x = 1)$.

Precision, recall and F-score are computed as per Frank et al. (2009a). In particular, precision is the fraction of gold pairs among the sampled set and recall is the fraction of sampled pairs among the gold standard pairs.

$$\text{p} = \frac{|\text{Sampled} \cap \text{Gold}|}{|\text{Sampled}|}, \qquad \text{r} = \frac{|\text{Sampled} \cap \text{Gold}|}{|\text{Gold}|}$$

KL divergence is a way of measuring the difference between distributions. Small numbers generally indicate a close match and is zero only when the two are equal. Using the empirical distribution over gold topics $P(z)$, we use the standard formula for KL divergence to compare the gold standard distribution P against the inferred distribution Q. I.e., we compute $D_{KL}(P(w, z)||Q(w, z))$.

Object	Words			
BOX	thebox	box		
BRUSH	brush			
BUNNY	rabbit	Rosie		
BUS	bus			
CAR	car	thecar		
CHEESE	cheese			
DOG	thedoggy	doggy		
DOLL	doll	thedoll	yeah	benice
DOUGH	dough			
ERNIE	Ernie			

Table 1: Subset of an inferred word-object mapping. For clarity, the proposed words have been converted to standard English orthography.

	p	r	f	KL	acc
Joint	**0.21**	0.45	0.28	2.78	**0.81**
Gold Seg	**0.21**	**0.60**	**0.31**	**1.82**	0.79

Table 2: Word Learning Performance. Comparing precision, recall, and F-score of word-object pairs, $D_{KL}(P(w, z)||Q(w, z))$, and accuracy of utterance topics for the full joint model and a variant that only infers meanings given a gold standard segmentation.

The model learns fairly meaningful word-object associations; results are shown in Table 2. As in the case of topic accuracy, the joint and word learning only variants perform similarly, this time with somewhat better performance for the easier task with an F-score and KL divergence of 0.31 and 1.82 vs. 0.28 and 2.78 for the joint task.

Table 1 illustrates the sort of word-object pairs the model discovers. As can be seen, many of the errors are due to the segmentation, usually under-segmentation errors where it segments two words as one. This is a general problem with the unigram segmenter on which our model is based (Goldwater et al., 2009). Yet, even though these segmentation errors are also counted as word learning errors, they are often still meaningful in the sense that the true referring word is a subsequence.

So, word segmentation has an impact on word learning. Yet, the joint model still tends to uncover reasonable meanings. The next question is whether these meanings have an impact on the segmentation.

	NoCon	Random	Joint
Referring Nouns	0.36	0.35	**0.50**
Neighbors	0.33	0.33	**0.37**
Utt Final Nouns	0.36	0.36	**0.52**
Entire Corpus	0.53	0.53	**0.54**

Table 3: Segmentation performance. F-score for three subsets and the full corpus for three variants: the model without non-linguistic context, the model with random topics, and the full joint model.

4.2.2 Word Segmentation

To measure the impact of word learning on segmentation, we again compare the model on the full joint task against two other variants: one where topics are randomly selected, and one that ignores the non-linguistic context. For the random topics variant, we choose each topic during initialization according to the empirical distribution over gold topics and treat these topic assignments as observed variables for subsequent iterations. The variant that ignores non-linguistic context draws topics uniformly from the entire set of objects ever discussed in the corpus, another test of the contribution of the non-linguistic context to segmentation. We report token F-score, computed as per Goldwater et al. (2009), where any segment proposed by the model is a true positive only if it matches the gold segmentation and is a false positive otherwise. Any segment in the gold data not found by the model is a false negative.

Table 3 shows the segmentation performance for various subsets as well as for the entire corpus. Because we are primarily interested in synergies between word learning and segmentation, we focus on the words most directly impacted by the meanings: gold standard referring nouns and their neighboring words.

The model behaves the same with randomized topics as without context; it learns none of the gold standard pairs (no matter how we identify clusters with topics for the contextless case). On all subsets, the full joint model outperforms the other two variants. In particular, the greatest gain is for the referring nouns, with a 21% reduction in error. Also, similar to the findings of Bortfeld et al. (2005) regarding 6 month olds' abilities to segment words adjoining a familiar name, we also find that neighboring words benefit from sharing a word boundary with a learned word.

The model performs exceptionally well on utterance final referring nouns, with a 24% reduction in error. This may explain certain psycholinguistic observations. Frank et al. (2006) performed an artificial language experiment with humans subjects demonstrating that adults were able to learn words at the same time as they learned to segment the language. However, subjects did much better on a word learning task when the meaning bearing words were consistently placed at the end of utterances. There are several possible reasons why this might have been the case. For instance, it is common in English for the object noun to occur at the end of the sentence, and since the subjects were all English speakers, they may have found it easier to learn an artificial language with a similar pattern. However, our model predicts another simple possibility: the segmentation task is easier at the end because one of the two word boundaries is already known (the utterance boundary itself).

4.3 Discussion

The word learning model generally prefers a very sparse word-to-object map. This is enforced by using a concentration parameter α_1 that is quite small relative to the α_0 parameter, and it biases the model so that the distributions over referring words are very different from that over non-referring words. A small concentration parameter biases the estimator to prefer a small set of word types. In contrast, the relatively large concentration parameter for the non-referring words tends to result in most of the words receiving highest probability as non-referring words. The model thus categorizes words accordingly. It is in part due to this tendency towards sparse word-object maps that the model enforces mutual exclusivity, a phenomenon well documented among natural word learners (Markman, 1990).

Aside from contributing to mutual exclusivity and specialization among the topical word distributions, the small concentration parameter also has important implications for the segmentation task. A very small value for α_1 discourages the learner from acquiring more word types for each meaning than absolutely necessary, thereby forcing the segmenter to use fewer types to explain the sequence of phonemes. A model without any notion

of meaning cannot maintain separate distributions for different topics, and must in some sense treat all words as non-referring. A segmenting model without meanings cannot share the word learner's reluctance to propose new meaning-bearing word types and might propose three separate types for "your book", "a book", and "the book". However, with a small enough prior on new referring word types, the word learner that discovers a common referent for all three sequences and, preferring fewer referring word types, is more likely to discover the common subsequence "book". With a single word-object pair ("book", BOOK), the word learner could explain reference for all three sequences instead of using the three separate pairs ("yourbook", BOOK), ("abook", BOOK), and ("thebook", BOOK).

While relying on non-linguistic context helps segment the meaning-bearing words, the overall improvement is small in our current corpus. One reason for this small improvement was that only 9% of the tokens in the corpus were referring words. In corpora containing a larger variety of objects – and in cases where sub- and super-ordinate labels like "eyes" and "ears" are coded – this percentage is likely to be much higher, leading to a greater boost in overall segmentation performance.

We should acknowledge that the decisions entailed in enriching the annotations are neither trivial nor without theoretic implication, however. It is not immediately obvious how to represent the non-linguistic correlates of verbs, for instance. Developmentally, verbs are typically acquired much later than nouns, and it has been argued that this may be due to the difficulty of producing a cognitive representation of the associated meaning (Gentner and Boroditsky, 2001). Even among concrete nouns, not all are equal. Children tend to have a bias toward whole objects when mapping novel words to their non-linguistic counterparts (Markman, 1990). Decisions about more sophisticated encoding of non-linguistic information may thus require more knowledge about children's representations of the world around them

5 Conclusion and Future Work

We find (1) that it is possible to jointly infer both meanings and a segmentation in a fully unsupervised way and (2) that doing so improves the segmentation performance of our model. In particular, we found that although the word learning side suffered from segmentation errors, and performed worse than a model that learned from a gold standard segmentation, the loss was only slight. On the other hand, segmentation performance for the meaning bearing words improved a great deal. The first result suggests that is not necessary to assume fully segmented input in order to learn word meanings, and that the segmentation and word learning tasks can be effectively modeled in parallel, allowing us to explore potential developmental interactions. The second result suggests that synergies do actually exist and argue not only that we can model the two as parallel processes, but that doing so could prove fruitful.

Our model is relatively simple both in terms of word learning and in terms of word segmentation. For instance, social cues and shared attention, or discourse effects, might all play a role (Frank et al., 2009b). Shared features or other relationships can also potentially impact how quickly one might generalize a label to multiple instances (Tenenbaum and Xu, 2000). There are many ways to elaborate on the word learning task, with additional potential synergistic implications.

We might also elaborate the linguistic structures we incorporate into the word learning model. For instance, Johnson (2008) explores synergies in syllable and morphological structures in word segmentation. Aspects of linguistic structure, such as morphology, may contribute to word meaning learning beyond its contribution to word segmentation performance.

Acknowledgments

This research was funded by NSF awards 0544127 and 0631667 to Mark Johnson and by NSF DDRIG 0746251 to Michael C. Frank. We would also like to thank Anne Fernald for providing the corpus and Maeve Cullinane for help in coding it.

References

Nameera Akhtar and Lisa Montague. 1999. Early lexical acquisition: The role of cross-situational learning. *First Language*, 19(57 Pt 3):347–358.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Heather Bortfeld, James L. Morgan, Roberta Michnick Golinkoff, and Karen Rathbun. 2005. Mommy and me: Familiar names help launch babies into speechstream segmentation. *Psychological Science*, 16(4):298–304.

Michael R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Anne Fernald and Hiromi Morikawa. 1993. Common themes and cultural variations in japanese and american mothers' speech to infants. In *Child Development*, number 3, pages 637–656, June.

Michael C. Frank, Vikash Mansinghka, Edward Gibson, and Joshua B. Tenenbaum. 2006. Word segmentation as word learning: Integrating stress and meaning with distributional cues. In *Proceedings of the 31st Annual Boston University Conference on Language Development*.

Michael C. Frank, Sharon Goldwater, Vikash Mansinghka, Tom Griffiths, and Joshua Tenenbaum. 2007. Modeling human performance in statistical word segmentation. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pages 281–286.

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. 2009a. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 5:578–585.

Michael C. Frank, Noah D. Goodman, Joshua B. Tenenbaum, and Anne Fernald. 2009b. Continuity of discourse provides information for word learning.

Dedre Gentner and Lera Boroditsky. 2001. Individuation, relativity, and early word learning. *Language, culture, & cognition*, 3:215–56.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Katharine Graf-Estes, Julia L. Evans, Martha W. Alibali, and Jenny R. Saffran. 2007. Can infants map meaning to newly segmented words? statistical segmentation and word learning. *Psychological Science*, 18(3):254–260.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. 2007. Hidden topic markov models. In *Artificial Intelligence and Statistics (AISTATS)*, March.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, Colorado, June. Association for Computational Linguistics.

Mark Johnson. 2008. Using adaptor grammars to identifying synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, Columbus, Ohio. Association for Computational Linguistics.

Ellen M. Markman. 1990. Constraints children place on word learning. *Cognitive Science*, 14:57–77.

Terry Regier. 2003. Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences*, 7:263–268.

Jenny R. Saffran, Elissa L. Newport, and Richard N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of memory and Language*, 35:606–621.

Jeffrey M. Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91.

Joshua B. Tenenbaum and Fei Xu. 2000. Word learning as bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 517–522.

Chen Yu and Dana H. Ballard. 2007. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.