Concurrent Acquisition of Word Meaning and Lexical Categories

Afra Alishahi
a.alishahi@uvt.nl
Communication and Information Sciences
Tilburg University, The Netherlands

Grzegorz Chrupała
gchrupala@lsv.uni-saarland.de
Spoken Language Systems
Saarland University, Germany

Abstract

Learning the meaning of words from ambiguous and noisy context is a challenging task for language learners. It has been suggested that children draw on syntactic cues such as lexical categories of words to constrain potential referents of words in a complex scene. Although the acquisition of lexical categories should be interleaved with learning word meanings, it has not previously been modeled in that fashion. In this paper, we investigate the interplay of word learning and category induction by integrating an LDA-based word class learning module with a probabilistic word learning model. Our results show that the incrementally induced word classes significantly improve word learning, and their contribution is comparable to that of manually assigned part of speech categories.

1 Learning the Meaning of Words

For young learners of a natural language, mapping each word to its correct meaning is a challenging task. Words are often used as part of an utterance rather than in isolation. The meaning of an utterance must be inferred from among numerous possible interpretations that the (usually complex) surrounding scene offers. In addition, the linguistic and visual context in which words are heard and used is often noisy and highly ambiguous. Particularly, many words in a language are polysemous and have different meanings.

Various learning mechanisms have been proposed for word learning. One well-studied mechanism is *cross-situational learning*, a bottom-up strategy based on statistical co-occurrence of words and referents across situations (Quine 1960, Pinker 1989). Several experimental studies have shown that adults and children are sensitive to cross-situational evidence and use this information for mapping words to objects, actions and properties (Smith and Yu 2007, Monaghan and Mattock 2009). A number of computational models have been developed based on this principle, demonstrating that cross-situational learning is a powerful and efficient mechanism for learning the correct mappings between words and meanings from noisy input (e.g. Siskind 1996, Yu 2005, Fazly et al. 2010).

Another potential source of information that can help the learner to constrain the relevant aspects of a scene is the sentential context of a word. It has been suggested that children draw on syntactic cues provided by the linguistic context in order to guide word learning, a hypothesis known as *syntactic bootstrapping* (Gleitman 1990). There is substantial evidence that children are sensitive to the structural regularities of language from a very young age, and that they use these structural cues to find the referent of a novel word (e.g. Naigles and Hoff-Ginsberg 1995, Gertner et al. 2006). In particular, young children have robust knowledge of some of the abstract lexical categories such as nouns and verbs (e.g. Gelman and Taylor 1984, Kemp et al. 2005).

Recent studies have examined the interplay of cross-situational learning and sentence-level learning mechanisms, showing that adult learners of an artificial language can successfully and simultaneously apply cues and constraints from both sources of information when mapping words to their referents (Gillette et al. 1999, Lidz et al. 2010, Koehne and Crocker 2010; 2011). Several computational models have also investigated this interaction by adding manually annotated part-of-speech tags as

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 643–654, Jeju Island, Korea, 12–14 July 2012.

input to word learning algorithms, and suggesting that integration of lexical categories can boost the performance of a cross-situational model (Yu 2006, Alishahi and Fazly 2010).

However, none of the existing experimental or computational studies have examined the acquisition of word meanings and lexical categories in parallel. They all make the simplifying assumption that *prior* to the onset of word learning, the categorization module has already formed a relatively robust set of lexical categories. This assumption can be justified in the case of adult learners of a second or artificial language. But children's acquisition of categories is most probably interleaved with the acquisition of word meaning, and these two processes must ultimately be studied simultaneously.

In this paper, we investigate concurrent acquisition of word meanings and lexical categories. We use an online version of the LDA algorithm to induce a set of word classes from child-directed speech, and integrate them into an existing probabilistic model of word learning which combines cross-situational evidence with cues from lexical categories. Through a number of simulations of a word learning scenario, we show that our automatically and incrementally induced categories significantly improve the performance of the word learning model, and are closely comparable to a set of gold-standard, manually-annotated part of speech tags.

2 A Word Learning Model

We want to investigate whether lexical categories (i.e. word classes) that are incrementally induced from child-directed speech can improve the performance of a cross-situational word learning model. For this purpose, we use the model of Alishahi and Fazly (2010). This model uses a probabilistic learning algorithm for combining evidence from word–referent co-occurrence statistics and the meanings associated with a set of pre-defined categories. They use child-directed utterances, manually annotated with a small set of part of speech tags, from the Manchester corpus (Theakston et al. 2001) in the CHILDES database (MacWhinney 1995). Their experimental results show that integrating these gold-standard categories into the algorithm boosts its performance over a pure cross-situational version.

The model of Alishahi and Fazly (2010) has the suitable architecture for our goal: it provides an integrated learning mechanism which combines evidence from word-referent co-occurrence with cues from the meaning representation associated with word categories. However, the model has two major shortcomings. First, it assumes that lexical categories are formed and finalized prior to the onset of word learning and that a correct and unique category for a target word can be identified at each point in time, assumptions that are highly unlikely. Second, it does not handle any ambiguity in the meaning of a word. Instead, each word is assumed to have only one correct meaning. Considering the high level of lexical ambiguity in most natural languages, this assumption unreasonably simplifies the word learning problem.

To investigate the plausibility of integrating word and category learning, we use an online algorithm for automatically and incrementally inducing a set of lexical categories. Moreover, we use each word in its original form instead of lemmatizing them, which implies that categories contain different morphological forms of the same word. By applying these changes, we are able to study the contribution of lexical categories to word learning in a more realistic scenario.

Representation of input. The input to the model consists of a sequence of utterances, each paired with a representation of an observed scene. We represent an utterance as a set of words, $U = \{w\}$ (e.g. {*she, went, home, ...*}), and the corresponding scene as a set of semantic features, $S = \{f\}$ (e.g. {ANIMATE, HUMAN, FEMALE, ...}).

Word and category meaning. We represent the meaning of a word as a time-dependent probability distribution $p^{(t)}(\cdot|w)$ over all the semantic features, where $p^{(t)}(f|w)$ is the probability of feature f being associated with word w at time t. In the absence of any prior knowledge, the model assumes a uniform distribution over all features as the meaning of a novel word. Also, a function $\mathrm{cat}^{(t)}(w)$ gives us the category to which a word w in utterance $U^{(t)}$ belongs.

At each point in time, a category c contains a set of word tokens. We assign a meaning to each cat-

egory as a weighted sum of the meaning learned so far for each of its members, or $p^{(t)}(f|c) = (1/|c|)\sum_{w \in c} p^{(t)}(f|w)$, where $|c|$ is the number of word tokens in c at the current moment.

Learning algorithm. Given an utterance-scene pair $(U^{(t)}, S^{(t)})$ received at time t, the model first calculates an alignment score a for each word $w \in U^{(t)}$ and each semantic feature $f \in S^{(t)}$. A semantic feature can be aligned to a word according to the meaning acquired for that word from previous observations (word-based alignment, or a_w). Alternatively, distributional clues of the word can be used to determine its category, and the semantic features can be aligned to the word according to the meaning associated to its category (category-based alignment, or a_c). We combine these two sources of evidence when estimating an alignment score:

$$a(w|f, U^{(t)}, S^{(t)}) = \lambda(w) \times a_w(w|f, U^{(t)}, S^{(t)}) + (1 - \lambda(w)) \times a_c(w|f, U^{(t)}, S^{(t)}) \quad (1)$$

where the word-based and category-based alignment scores are estimated based on the acquired meanings of the word and its category, respectively:

$$a_w(w|f, U^{(t)}, S^{(t)}) = \frac{p^{(t-1)}(f|w)}{\sum_{w_k \in \mathrm{U}^{(t)}} p^{(t-1)}(f|w_k)}$$

$$a_c(w|f, U^{(t)}, S^{(t)}) = \frac{p^{(t-1)}(f|\mathrm{cat}(w))}{\sum_{w_k \in \mathrm{U}^{(t)}} p^{(t-1)}(f|\mathrm{cat}(w_k))}$$

The relative contribution of the word-based versus the category-based alignment is determined by the weight function $\lambda(w)$. Cross-situational evidence is a reliable cue for frequent words; on the other hand, the category-based score is most informative when the model encounters a low-frequency word (See Alishahi and Fazly (2010) for a full analysis of the frequency effect). Therefore, we define $\lambda(w)$ as a function of the frequency of the word $n(w)$:

$$\lambda(w) = n(w)/(n(w) + 1)$$

Once an alignment score is calculated for each word $w \in U^{(t)}$ and each feature $f \in S^{(t)}$, the model revises the meanings of all the words in $U^{(t)}$ and their corresponding categories as follows:

$$\mathrm{assoc}^{(t)}(w, f) = \mathrm{assoc}^{(t-1)}(w, f) + a(w|f, \mathrm{U}^{(t)}, \mathrm{S}^{(t)})$$

where $\mathrm{assoc}^{(t-1)}(w,\ f)$ is zero if w and f have not co-occurred before. These association scores are then used to update the meaning of the words in the current input:

$$p^{(t)}(f|w) = \frac{\mathrm{assoc}^{(t)}(f,\ w)}{\sum_{f_j \in \mathcal{F}} \mathrm{assoc}^{(t)}(f_j,\ w)} \quad (2)$$

where $\mathcal{F}$ is the set of all features seen so far. We use a smoothed version of this formula to accommodate noisy or rare input. This process is repeated for all the input pairs, one at a time.

Uniform categories. Adding the category-based alignment as a new factor to Eqn. (1) might imply that the role of categories in this model is nothing more than smoothing the cross-situational-based alignment of words and referents. In order to investigate this issue, we use the following alignment formula as an informed baseline in our experiments, where we replace $a_c(\cdot|f, U^{(t)}, S^{(t)})$ with a uniform distribution:[1]

$$a(w|f, U^{(t)}, S^{(t)}) = \lambda(w) \times a_w(w|f, U^{(t)}, S^{(t)}) + (1 - \lambda(w)) \times \frac{1}{|U^{(t)}|} \quad (3)$$

where $a_w(w|f, U^{(t)}, S^{(t)})$ and $\lambda(w)$ are estimated as before. In our experiments in Section 4, we refer to this baseline as the 'uniform' condition.

3 Online induction of word classes with LDA

Empirical findings suggest that young children form their knowledge of abstract categories, such as verbs, nouns, and adjectives, gradually (e.g. Gelman and Taylor 1984, Kemp et al. 2005). In addition, several unsupervised computational models have been proposed for inducing categories of words which resemble part-of-speech categories, by

[1] We thank an anonymous reviewers for suggesting this condition as an informed baseline.

drawing on distributional properties of their context (see for example Redington et al. 1998, Clark 2000, Mintz 2003, Parisien et al. 2008, Chrupała and Alishahi 2010). However, explicit accounts of how such categories can be integrated in a cross-situational model of word learning have been rare. Here we adopt an online version of the model proposed in Chrupała (2011), a method of soft word class learning using Latent Dirichlet Allocation. The approach is much more efficient than the commonly used alternative (Brown clustering, (Brown et al. 1992)) while at the same time matching or outperforming it when the word classes are used as automatically learned features for supervised learning of various language understanding tasks. Here we adopt this model as our approach to learning lexical categories.

In Section 3.1 we describe the LDA model for word classes; in Section 3.2 we discuss the online Gibbs sampler we use for inference.

3.1 Word class learning with LDA

Latent Dirichlet Allocation (LDA) was introduced by Blei et al. (2003) and is most commonly used for modeling the topic structure in document collections. It is a generative, probabilistic hierarchical Bayesian model that induces a set of latent variables, which correspond to the topics. The topics themselves are multinomial distributions over words.

The generative structure of the LDA model is the following:

$$\begin{aligned} \phi_k &\sim \text{Dirichlet}(\beta), & k &\in [1, K] \\ \theta_d &\sim \text{Dirichlet}(\alpha), & d &\in [1, D] \\ z_{n_d} &\sim \text{Categorical}(\theta_d), & n_d &\in [1, N_d] \\ w_{n_d} &\sim \text{Categorical}(\phi_{z_{n_d}}), & n_d &\in [1, N_d] \end{aligned} \tag{4}$$

Chrupała (2011) reinterprets the LDA model in terms of word classes as follows: K is the number of classes, D is the number of unique word types, N_d is the number of context features (such as right or left neighbor) associated with word type d, z_{n_d} is the class of word type d in the n_d^{th} context, and w_{n_d} is the n_d^{th} context feature of word type d. Hyperparameters α and β control the sparseness of the vectors θ_d and ϕ_k.

Wordtype	Features		
How	do_R		
do	How_L	you_R	you_L
you	do_L	do_R	

Table 1: Matrix of context features

1.8M words (CHILDES)		100M words (BNC)	
train	car	can	will
give	bring	June	March
shoes	clothes	man	woman
book	hole	black	white
monkey	rabbit	business	language

Table 2: Most similar word pairs

As an example consider the small corpus consisting of the single sentence *How do you do*. The rows in Table 1 show the features $w_1 \dots w_{N_d}$ for each word type d if we use each word's left and right neighbors as features, and subscript words with $_L$ and $_R$ to indicate left and right.

After inference, the θ_d parameters correspond to word class probability distributions given a word type while the ϕ_k correspond to feature distributions given a word class: the model provides a probabilistic representation for word types independently of their context, and also for contexts independently of the word type.

Probabilistic, *soft* word classes are more expressive than hard categories. First, they make it easy and efficient to express shared ambiguities: Chrupała (2011) gives an example of words used as either first names or surnames, and this shared ambiguity is reflected in the similarity of their word class distributions. Second, with soft word classes it becomes easy to express graded similarity between words: as an example, Table 2 shows a random selection out of the 100 most similar word pairs according to the Jensen-Shannon divergence between their word class distributions, according to a word class model with 25 classes induced from (i) 1.8 million words of the CHILDES corpus or (ii) 100 million word of the BNC corpus. The similarities were measured between each of the 1000 most frequent CHILDES or BNC words.

3.2 Online Gibbs sampling for LDA

There have been a number of attempts to develop online inference algorithms for topic modeling with LDA. A simple modification of the standard Gibbs sampler (**o-LDA**) was proposed by Song et al. (2005) and Banerjee and Basu (2007).

Canini et al. (2009) experiment with three sampling algorithms for online topic inference: (i) **o-LDA**, (ii) incremental Gibbs sampler, and (iii) a particle filter. Only **o-LDA** is truly online in the sense that it does not revisit previously seen documents. The other two, the incremental Gibbs sampler and the particle filter, keep seen documents and periodically resample them. In Canini et al.'s experiments all of the online algorithms perform worse than the standard batch Gibbs sampler on a document clustering task.

Hoffman et al. (2010) develop an online version of the variational Bayes (VB) optimization method for inference for topic modeling with LDA. Their method achieves good empirical results compared to batch VB as measured by perplexity on held-out data, especially when used with large minibatch sizes.

Online VB for LDA is appropriate when streaming documents: with online VB documents are represented as word count tables. In our scenario where we apply LDA to modeling word classes we need to process context features from sentences arriving in a stream: i.e. we need to sample entries from a table like Table 1 in order of arrival rather than row by row. This means that online VB is not directly applicable to online word-class induction.

However it also means that one issue with **o-LDA** identified by Canini et al. (2009) is ameliorated. When sampling in a topic modeling setting, documents are unique and are never seen again. Thus, the topics associated with old documents get *stale* and need to be periodically *rejuvenated* (i.e. resampled). This is the reason why the incremental Gibbs sampler and the particle filter algorithms in Canini et al. (2009) need to keep old documents around and cannot run in a true online fashion. Since for word class modeling we stream context features as they arrive, we will continue to see features associated with the seen word types, and will automatically resample their class assignments. In exploratory experiments we have seen that this narrows the performance gap between the **o-LDA** sampler and the batch collapsed Gibbs sampler.

We present our version of the **o-LDA** sampler in Algorithm 1. For each incoming sentence t we run J passes of sampling, updating the counts tables after each sampling step. We sample the class assignment z_{t_i} for feature w_{t_i} according to:

$$P(z_t|\mathbf{z}_{t-1}, \mathbf{w}_t, \mathbf{d}_t) \propto \frac{(n_{t-1}^{z_t,d_t} + \alpha) \times (n_{t-1}^{z_t,w_t} + \beta)}{\sum_{j=1}^{V_{t-1}} n_{t-1}^{z_t,w_j} + \beta}, \quad (5)$$

where $n_t^{z,d}$ stands for the number of times class z co-occurred with word type d up to step t, and similarly $n_t^{z,w}$ is the number of times feature w was assigned to class z. V_t is the number of unique features seen up to step t, while α and β are the LDA hyperparameters. There are two differences between the original **o-LDA** and our version: we do not initialize the algorithm with a batch run over a prefix of the data, and we allow more than one sampling pass per sentence.[2] Exploratory experiments have shown that batch initialization is unnecessary, and that multiple passes typically improve the quality of the induced word classes.

Algorithm 1 Online Gibbs sampler for word class induction with LDA

```
for t = 1 → ∞ do
    for j = 1 → J do
        for i = 1 → I_t do
            sample z_{t_i} ~ P(z_{t_i}|z_{t_i-1}, w_{t_i}, d_{t_i})
            increment n_t^{z_{t_i},w_{t_i}} and n_t^{z_{t_i},d_{t_i}}
```

Figure 1 shows the top 10 words for each of the 10 word classes induced with our online Gibbs sampler from 1.8 million words of CHILDES. Similarly, Figure 2 shows the top 10 words for 5 randomly chosen topics out of 50, learned online from 100 million words of the BNC.

The topics are relatively coherent and at these levels of granularity express mostly part of speech and subcategorization frame information.

Note that for each word class we show the words most frequently assigned to it while Gibbs sampling.

[2] Note that we do not allow multiple passes over the stream of sentences. Rather, while processing the current sentence, we allow the words in this sentence to be sampled more than once.

do are have can not go put did get play
is that it what not there he was where put
you not I the we what it they your a
to you we and I will not can it on
it a that the not he this right got she
are do is have on in can want did going
one I not shall there then you are we it
is in are on oh with and of have do
the a your of that it this some not very
going want bit go have look got will at little

Figure 1: Top 10 words for 10 classes learned from CHILDES

I you he it they we she , You He
a the more some all no The other I two
as if when that where how because If before what
was is 's had , has are would did said
the his her their this an that its your my

Figure 2: Top 10 words of 5 randomly chosen classes learned from BNC

Since we are dealing with soft classes, most word-types have non-zero assignment probabilities for many classes. Thus frequently occurring words such as *not* will typically be listed for several classes.

4 Evaluation

4.1 Experimental setup

As training data, we extract utterances from the Manchester corpus (Theakston et al. 2001) in the CHILDES database (MacWhinney 1995), a corpus that contains transcripts of conversations with children between the ages of 1 year, 8 months and 3 years. We use the mother's speech from transcripts of 12 children (henceforth referred to by children's names).

We run word class induction while simultaneously outputting the highest scoring word-class label for each word: for a new sentence, we sample class assignments for each feature (doing J passes), update the counts, and then for each word d_{t_i} output the highest scoring class label according to $\text{argmax}_z\, n_t^{z,d_{t_i}}$ (where $n_t^{z,d_{t_i}}$ stands for the number of times class z co-occurred with word type d_{t_i} up to step t).

During development we ran the online word class induction module on data for Aran, Becky, Carl and Anne and then started the word learning module for the Anne portion while continuing inducing categories. We then evaluated word learning on Anne. We chose the parameters of the word class induction module based on those development results: $\sum_{1=1}^{K} \alpha = 10$, $\beta = 0.1$, $K = 10$ and $J = 20$.

We used cross-validation for the final evaluation. For each of six data files (Anne, Aran, Becky, Carl, Dominic and Gail), we ran word-class induction on the whole corpus with the chosen file last, and then started applying the word-learning algorithm on this last chosen file (while continuing with category induction). We evaluated how well word meanings were learned in those six cases.

We follow Alishahi and Fazly (2010) in the construction of the input. We need a semantic representation paired with each utterance. Such a representation is not available from the corpus and has to be constructed. We automatically construct a gold lexicon for all nouns and verbs in this corpus as follows. For each word, we extract all hypernyms for its first sense in the appropriate (verb or noun) hierarchy in WordNet (Fellbaum 1998), and add the first word in the synset of each hypernym to the set of semantic features for the target word. For verbs, we also extract features from VerbNet (Kipper et al. 2006). A small subset of words (pronouns and frequent quantifiers) are also manually added. This lexicon represents the *true* meaning of each word, and is used in generating the scene representations in the input and in evaluation.

For each utterance in the input corpus, we form the union of the feature representations of all its words. Words not found in the lexicon (i.e. for which we could not extract a semantic representation from WordNet and VerbNet) are removed from the utterance (only for the word learning module).

In order to simulate the high level of noise that children receive from their environment, we follow Alishahi and Fazly (2010) and pair each utterance with a combination of its own scene representation and the scene representation for the following utterance. This decision was based on the intuition that consequent utterances are more likely to be about re-

Utterance:	{	*mommy, ate, broccoli* }
Scene:	{	ANIMATE, HUMAN, ...,
		CONSUMPTION, ACTION, ...
		BROCCOLI, VEGETABLE, ...
		PLATE, OBJECT, ... }

Figure 3: A sample input item to the word learning model

Category	Avg. MAP	Std. Dev.
None	0.626	0.032
Uniform	0.633	0.032
LDA	0.659	0.029
POS	0.672	0.030

Table 3: Final Mean Average Precision scores

lated topics and scenes. This results in a (roughly) 200% ambiguity. In addition, we remove the meaning of one random word from the scene representation of every second utterance in an attempt to simulate cases where the referent of an uttered word is not within the perception field (such as 'daddy is not home yet'). A sample utterance and its corresponding scene are shown in Figure 3.

As mentioned before, many words in our input corpus are polysemous. For such words, we extract different sets of features depending on their manually tagged part of speech and keep them in the lexicon (e.g. the lexicon contains two different entries for *set:N* and *set:V*). When constructing a scene representation for an utterance which contains an ambiguous word, we choose the correct sense from our lexicon according to the word's part of speech tag in Manchester corpus.

In the experiments reported in the next section, we assess the performance of our model on learning words at each point in time: for each target word, we compare its set of features in the lexicon with its probability distribution over the semantic features that the model has learned. We use mean average precision (MAP) to measure how well $p^{(t)}(\cdot|w)$ ranks the features of w.

4.2 Learning curves

To understand whether our categories contribute to learning of word–meaning mappings, we compare the pattern of word learning over time in four conditions. The first condition represents our baseline, in which we do not use category-based alignment in the word learning model by setting $\lambda(w) = 1$ in Eqn. (1). In the second condition we use a set of uniformly distributed categories for alignment, as estimated by Eqn. (3) on page 3 (this condition is introduced to examine whether categories act as more than a simple smoothing factor in the alignment process.) In the third condition we use the categories induced by online LDA in the word learning model. The fourth condition represents the performance ceiling, in which we use the pre-defined and manually annotated part of speech categories from the Manchester corpus.

Table 3 shows the average and the standard deviation of the *final* MAP scores across the six datasets, for the four conditions (no categories, uniform categories, LDA categories and gold part-of-speech tags). The differences between LDA and None, and between LDA and Uniform are statistically significant according to the paired t test ($p < 0.01$), while the difference between LDA and POS is not ($p = 0.16$).

Figure 4 shows the learning curves in each condition, averaged over the six splits explained in the previous section. The top panel shows the average learning curve over the minimum number of sentences across the six sub-corpora (8800 sentences). The curves show that our LDA categories significantly improve the performance of the model over both baselines. That means that using these categories can improve word learning compared to not using them and relying on cross-situational evidence alone. Moreover, LDA-induced categories are not merely acting as a smoothing function the way the 'uniform' categories are. Our results show that they are bringing relevant information to the task at hand, that is, improving word learning by using the sentential context. In fact, this improvement is comparable to the improvement achieved by integrating the 'gold-standard' POS categories.

The middle and bottom panels of Figure 4 zoom in on shorter time spans (5000 and 1000 sentences, respectively). These diagrams suggest that the pattern of improvement over baseline is relatively constant, even at very early stages of learning. In fact, once the model receives enough input data, cross-situational evidence becomes stronger (since fewer

words in the input are encountered for the first time) and the contribution of the categories becomes less significant.

4.3 Class granularity

In Figure 5 we show the influence of the number of word classes used on the performance in word learning. It is evident that in the range between 5 to 20 classes the performance of the word learning module is quite stable and insensitive to the exact class granularity. Even with only 5 classes the model can still roughly distinguish noun-like words from verb-like words from pronoun-like words, and this will help learn the meaning elements derived from the higher levels of WordNet hierarchy. Notwithstanding that, ideally we would like to avoid having to pre-specify the number of classes for the word class induction module: we thus plan to investigate non-parametric models such as Hierarchical Dirichlet Process for this purpose.

5 Related Work

This paper investigates the interplay between two language learning tasks which have so far been studied in isolation: the acquisition of lexical categories from distributional clues, and learning the mapping between words and meanings. Previous models have shown that lexical categories can be learned from unannotated text, mainly drawing on distributional properties of words (e.g. Redington et al. 1998, Clark 2000, Mintz 2003, Parisien et al. 2008, Chrupała and Alishahi 2010).

Independently, several computational models have exploited cross-situational evidence in learning the correct mappings between words and meanings, using rule-based inference (Siskind 1996), neural networks (Li et al. 2004, Regier 2005), hierarchical Bayesian models (Frank et al. 2007) and probabilistic alignment inspired by machine translation models (Yu 2005, Fazly et al. 2010).

There are only a few existing computational models that explore the role of syntax in word learning. Maurits et al. (2009) investigates the joint acquisition of word meaning and word order using a batch model. This model is tested on an artificial language with a simple first order predicate representation of meaning, and limited built-in possibilities for word

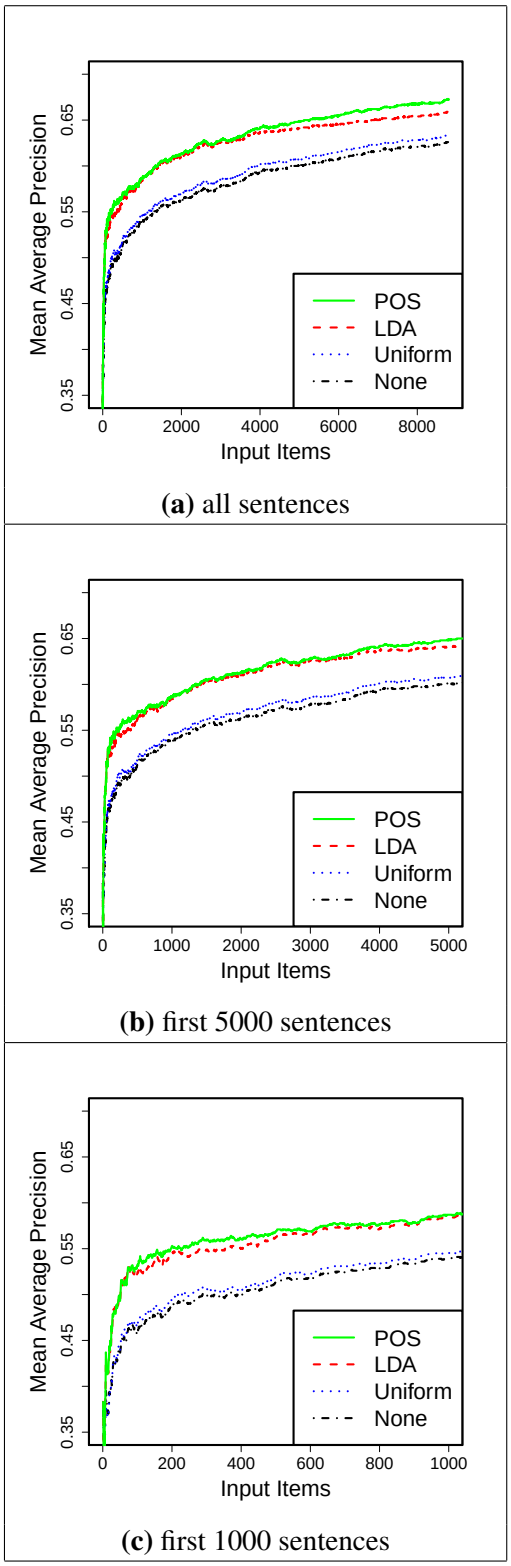

Figure 4: Mean average precision for all observed words at each point in time for four conditions: with gold POS categories, with LDA categories, with uniform categories, and without using categories. Each panel displays a different time span.

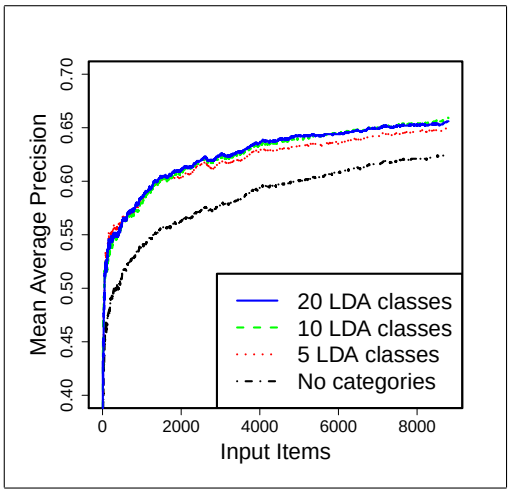

Figure 5: Mean average precision for all observed words at each point in time in four conditions: using online LDA categories of varying numbers of 20, 10 and 5, and without using categories.

order. The model of Niyogi (2002) simulates the mutual bootstrapping effects of syntactic and semantic knowledge in verb learning, that is the use of syntax to aid in inducing the semantics of a verb, and the use of semantics to narrow down possible syntactic frames in which a verb can participate. However, this model relies on manually assigned priors for associations between syntactic and semantic features, and is tested on a toy language with very limited vocabulary and a constrained syntax.

Yu (2006) integrates automatically induced syntactic word categories into his model of cross-situational word learning, showing that they can improve the model's performance. Yu's model also processes input utterances in a batch mode, and its evaluation is limited to situations in which only a coarse distinction between referring words (words that could potentially refer to objects in a scene, e.g. concrete nouns) and non-referring words (words that cannot possibly refer to objects, e.g. function words) is sufficient. It is thus not clear whether information about finer-grained categories (e.g. verbs and nouns) can indeed help word learning in a more naturalistic incremental setting.

On the other hand, the model of Alishahi and Fazly (2010) integrates manually annotated part-of-speech tags into an incremental word learning algorithm, and shows that these tags boost the overall word learning performance, especially for infrequent words.

In a different line of research, a number of models have been proposed which study the acquisition of the link between syntax and semantics within the Combinatory Categorial Grammar (CCG) framework (Briscoe 1997, Villavicencio 2002, Buttery 2006, Kwiatkowski et al. 2012). These approaches set the parameters of a semantic parser on a corpus of utterances paired with a logical form as their meaning.

These models bring in extensive and detailed prior assumptions about the nature of the syntactic representation (i.e. atomic categories such as S and NP, and built-in rules which govern their combination), as well as about the representation of meaning via the formalism of lambda calculus.

This is fundamentally different than the approach taken in this paper, which in comparison only assumes very simple syntactic and semantic representations of syntax. We view word and category learning as stand-alone cognitive tasks with independent representations (word meanings as probabilistic collections of properties or features as opposed to single symbols; categories as sets of word tokens with similar context distribution) and we do not bring in any prior knowledge of specific atomic categories.

6 Conclusion

In this paper, we show the plausibility of using automatically and incrementally induced categories while learning word meanings. Our results suggest that the sentential context that a word appears in across its different uses can be used as a complementary source of guidance for mapping it to its featural meaning representation.

In Section 4 we show that the improvement achieved by our categories is comparable to that gained by integrating gold POS categories. This result is very encouraging, since manually assigned POS tags are typically believed to set the upper bound on the usefulness of category information.

We believe that it automatically induced categories have the potential to do even better: Chrupała and Alishahi (2010) have shown that categories induced from usage data in an unsupervised fashion can be used more effectively than POS categories in

a number of tasks. In our experiments here on the development data we observed some improvements over POS categories. This advantage can result from the fact that our categories are more fine-grained (if also more noisy) than POS categories, which sometimes yields more accurate predictions.

One important characteristic of the category induction algorithm we have used in this paper is that it provides a *soft* categorization scheme, where each word is associated with a probability distribution over all categories. In future, we plan to exploit this feature: when estimating the category-based alignment, we can interpolate predictions of multiple categories to which a word belongs, weighted by its probabilities associated with membership in each category.

Acknowledgements

Grzegorz Chrupała was funded by the German Federal Ministry of Education and Research (BMBF) under grant number 01IC10S01O as part of the Software-Cluster project EMERGENT (`www.software-cluster.org`).

References

Alishahi, A. and Fazly, A. (2010). Integrating Syntactic Knowledge into a Model of Cross-situational Word Learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Banerjee, A. and Basu, S. (2007). Topic models over text streams: A study of batch and online unsupervised learning. In *SIAM Data Mining*.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Briscoe, T. (1997). Co-evolution of language and of the language acquisition device. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 418–427. Association for Computational Linguistics.

Brown, P. F., Mercer, R. L., Della Pietra, V. J., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Buttery, P. (2006). Computational models for first language acquisition. *Computer Laboratory, University of Cambridge, Tech. Rep. UCAM-CLTR-675*.

Canini, K., Shi, L., and Griffiths, T. (2009). Online inference of topics with latent dirichlet allocation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Chrupała, G. (2011). Efficient induction of probabilistic word classes with LDA. In *International Joint Conference on Natural Language Processing*.

Chrupała, G. and Alishahi, A. (2010). Online Entropy-based Model of Lexical Category Acquisition. In *CoNLL 2010*.

Clark, A. (2000). Inducing syntactic categories by context distribution clustering. In *Proceedings of the 2nd workshop on Learning Language in Logic and the 4th conference on Computational Natural Language Learning*, pages 91–94. Association for Computational Linguistics Morristown, NJ, USA.

Fazly, A., Alishahi, A., and Stevenson, S. (2010). A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, 34(6):1017–1063.

Fellbaum, C., editor (1998). *WordNet, An Electronic Lexical Database*. MIT Press.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems*, volume 20.

Gelman, S. and Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, pages 1535–1540.

Gertner, Y., Fisher, C., and Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, 17(8):684–691.

Gillette, J., Gleitman, H., Gleitman, L., and Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73(2):135–76.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1:135–176.

Hoffman, M., Blei, D., and Bach, F. (2010). Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*.

Kemp, N., Lieven, E., and Tomasello, M. (2005). Young Children's Knowledge of the" Determiner" and" Adjective" Categories. *Journal of Speech, Language and Hearing Research*, 48(3):592–609.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). Extensive classifications of english verbs. In *Proceedings of the 12th EURALEX International Congress*.

Koehne, J. and Crocker, M. W. (2010). Sentence processing mechanisms influence cross-situational word learning. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Koehne, J. and Crocker, M. W. (2011). The interplay of multiple mechanisms in word learning. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Kwiatkowski, T., Goldwater, S., Zettelmoyer, L., and Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Li, P., Farkas, I., and MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, 17:1345–1362.

Lidz, J., Bunger, A., Leddon, E., Baier, R., and Waxman, S. R. (2010). When one cue is better than two: lexical vs . syntactic cues to verb learning. Unpublished manuscript.

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Lawrence Erlbaum Associates, second edition.

Maurits, L., Perfors, A. F., and Navarro, D. J. (2009). Joint acquisition of word order and word reference. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Monaghan, P. and Mattock, K. (2009). Cross-situational language learning: The effects of grammatical categories as constraints on referential labeling. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Naigles, L. and Hoff-Ginsberg, E. (1995). Input to Verb Learning: Evidence for the Plausibility of Syntactic Bootstrapping. *Developmental Psychology*, 31(5):827–37.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proceedings of the 24th annual conference of the Cognitive Science Society*, pages 697–702.

Parisien, C., Fazly, A., and Stevenson, S. (2008). An incremental bayesian model for learning syntactic categories. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

Quine, W. (1960). *Word and Object*. Cambridge University Press, Cambridge, MA.

Redington, M., Crater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science: A Multidisciplinary Journal*, 22(4):425–469.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29:819–865.

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

Smith, L. and Yu, C. (2007). Infants rapidly learn words from noisy data via cross-situational statistics. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Song, X., Lin, C., Tseng, B., and Sun, M. (2005). Modeling and predicting personal information dissemination behavior. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 479–488. ACM.

Theakston, A. L., Lieven, E. V., Pine, J. M., and Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28:127–152.

Villavicencio, A. (2002). The acquisition of a unification-based generalised categorial grammar. In *Proceedings of the Third CLUK Colloquium*, pages 59–66.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, 17(3–4):381–397.

Yu, C. (2006). Learning syntax–semantics mappings to bootstrap word learning. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.