Online Learning Mechanisms for Bayesian Models of Word Segmentation

Sharon Goldwater, Lisa Pearl, & Mark Steyvers

In recent years, there has been growing interest in computational-level models of human cognition, which often analyze the acquisition problem in terms of optimal Bayesian behavior. Researchers have shown that human behavior is consistent with the predictions of Bayesian ideal learners in a range of domains, including language (e.g., Xu & Tenenbaum, 2007; Griffiths & Tenenbaum, 2005). These models aim to explain why humans behave as they do given the task and data they encounter, but typically avoid some questions addressed by more traditional psychological models, such as *how* the observed behavior is produced given constraints on memory, processing, and so forth. Here, we use the task of word segmentation as a case study for investigating these questions within a Bayesian framework. We consider some limitations of the infant learner, and develop several learning algorithms that take these limitations into account. Each algorithm can be viewed as a different method of approximating the same ideal learner. When tested on a corpus of English child-directed speech (Bernstein-Ratner, 1984), we find that the learner's behavior depends non-trivially on how the learner's limitations are implemented. While these algorithms do not segment realistic speech as well as the most successful model of the ideal learner, they are more successful than other purely statistical language-independent learning strategies, such as using syllable transitional probability (Saffran et al., 1996, see Gambell & Yang (2006)).

The starting point of our research is the work of Goldwater, Griffiths, and Johnson (2006) (GGJ), which provides an ideal learning analysis of how statistical information, a language-independent cue for word segmentation preferred over language-dependent cues very early in development (Thiessen & Saffran, 2003), could be used by infants to begin to segment words from continuous speech. GGJ develop two models within a Bayesian framework where the learner is presented with some data d (an unsegmented corpus of phonological transcriptions) and seeks a hypothesis h (a sequence of words) that both explains the data (i.e., concatenating together the words in h forms d) and has high prior probability. The prior encodes the intuitions that words should be relatively short, and the lexicon should be relatively small. In addition, each of the two models encodes a different expectation about word behavior: in the *unigram* model, the learner assumes that words are statistically independent (i.e. context is not predictive); in the *bigram* model, words are assumed to be predictive units. GGJ show that an ideal learner biased to heed context (as in the bigram model) achieves far more successful segmentation, while a unigram ideal learner will severely undersegment the corpus, identifying common collocations as single words.

Notably, the results above are achieved using a learning algorithm where the entire corpus is available simultaneously to the learner, which is equivalent to an infant remembering all the utterances they are exposed to. However, it is possible to pair the probabilistic model of GGJ with alternative learning algorithms that make more realistic assumptions about memory and processing. We investigate three such algorithms here, asking how memory and processing limitations might affect the learner's ability to achieve the optimal solution to the segmentation task (i.e., the solution found by the ideal learners in GGJ). To simulate limited memory resources, all the learning algorithms we present operate in an online fashion, so that each utterance is processed (segmented) and then discarded. Under the GGJ model, the only information that is necessary to compute the probability of any particular segmentation of an utterance is the number of times each word (or bigram, in the case of the bigram model) has occurred previously. Thus, in each of our online learners, the lexicon counts are updated after segmenting each utterance. The primary differences between our algorithms lie in the additional details of how memory limitations are implemented, and whether the learner is assumed to sample segmentations from the posterior distribution or choose the most probable segmentation.

Our first learning algorithm, which we call Dynamic Programming Maximization (DPM), processes each utterance as a whole, using dynamic programming (specifically the Viterbi algorithm) to efficiently compute the probability of every possible segmentation given the current lexicon. It then chooses the segmentation with the highest probability, adds the words from that segmentation to the

lexicon, and moves to the next utterance. This algorithm is the only one of our three that has been previously applied to word segmentation (Brent, 1999).

Our second learning algorithm, Dynamic Programming Sampling (DPS), is similar to DPM except that instead of choosing the segmentation with the highest probability, the learner samples a segmentation from the conditional distribution of segmentations given the utterance and the current lexicon. This is equivalent to an algorithm known as *particle filtering*, with only a single particle. Using additional particles requires more memory but should allow closer approximations to the ideal learner. We hope to report results on the trade-off between memory and accuracy at the workshop, although results reported here are only for the single-particle filter.

Our final learning algorithm, Decayed Markov Chain Monte Carlo (DMCMC) (Marthi et al., 2002) processes an utterance by probabilistically sampling s word boundaries from all the utterances encountered so far. The probability that a particular potential boundary b is sampled is given by the exponentially decaying function b_a^{-d}, where b_a is the number of potential boundary locations between b and the end of the current utterance, and d is the decay rate. Thus, the further b is from the end of the current utterance, the less likely it is to be sampled. This induces a recency effect in the learner, focusing processing on more recent potential boundaries. Intuitively, this corresponds to the notion that the learner's memory is weaker for less recent potential boundaries.

Detailed results for all the learners are shown in Table 1 and Figure 1. We find that the performance of all the learners stabilizes relatively rapidly after an initial learning curve over the beginning of the corpus. For the unigram models, DPM and DPS have qualitatively similar performance (probably due to the similar nature of the algorithms), although DPS is slightly lower across the board. Unlike the ideal learner model, these models have a tendency to oversegment rather than undersegment (oversegmentation is characterized by high boundary recall and a low boundary precision, undersegmentation by the reverse). The DMCMC unigram model is more similar to the ideal learner in exhibiting undersegmentation, though not to the same extent. Taken together, these results suggest that undersegmentation may be less of a problem for learners with limited memory than for statistically optimal learners.

Turning to the bigram learners, we find that DPM performs better than its unigram counterpart, particularly on rare words (shown by greater improvements in lexicon accuracy than token accuracy). DPS and DMCMC, in contrast (and perhaps surprisingly given the ideal learner model), perform significantly worse than their unigram counterparts: DPS increasingly oversegments while DMCMC increasingly undersegments. Thus, the utility of context in word segmentation depends strongly on the learner's implementation.

Though the models vary in their performance, all models outperform the syllable transitional probability strategy, which has a token precision of 41.6 and recall of 23.3 on English child-directed speech (cf. Gambell & Yang (2006)). More practically for language acquisition, while the general word segmentation performance of these online algorithms is not as good as that of the ideal learning model, they still surpass the performance of other purely statistical online learning strategies that children seem able to use. As such, if infants require a seed pool of words to identify language-dependent strategies, these online language-independent strategies may provide a pool reliable enough to do so.

One moral of this investigation is that simple intuitions about human cognition, such as having memory limitations, can be cashed out multiple ways in online learning algorithms – processing utterances incrementally, keeping only a single lexicon hypothesis in memory, implementing recency effects with exponential decay functions, and so on. Having explored several algorithm instantiations incorporating this intuition, we find that the learning assumptions or biases that work best depend on how memory limitations are implemented. And in fact, some biases that are helpful for an ideal learner, such as using context to guide hypotheses, may hinder a learner with limited memory, as in the DMCMC and DPS models. The transition from a computational-level solution for an acquisition problem to the algorithmic-level approximation is not necessarily straightforward.

Supplementary Material

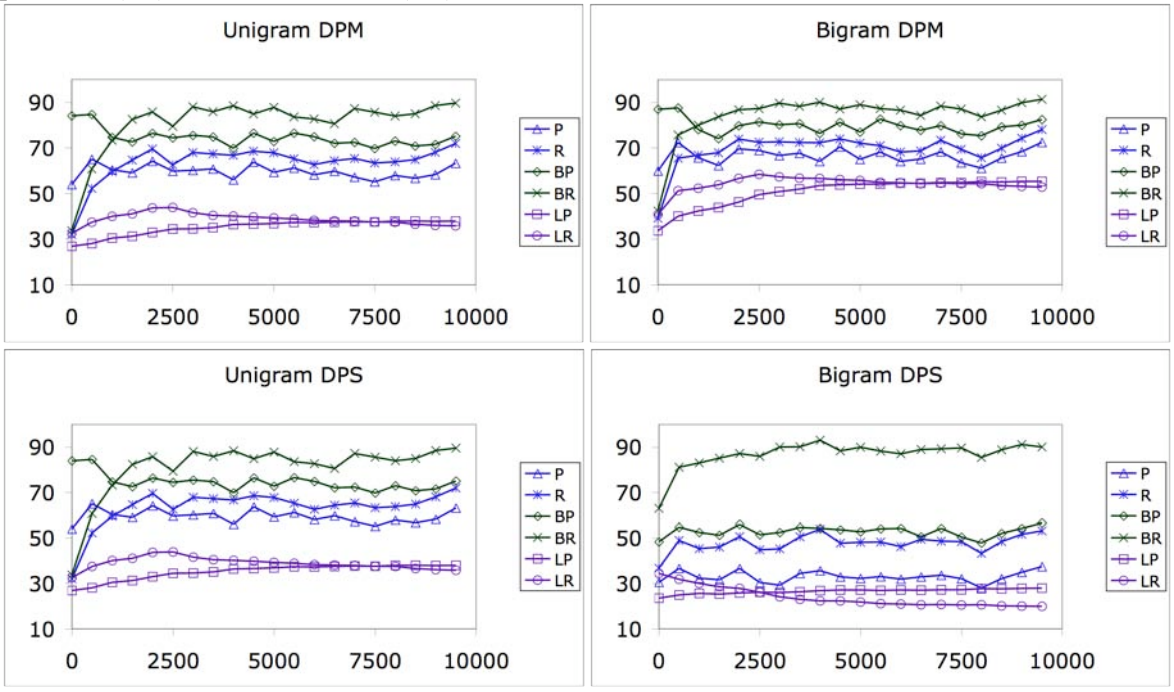
A. Tables

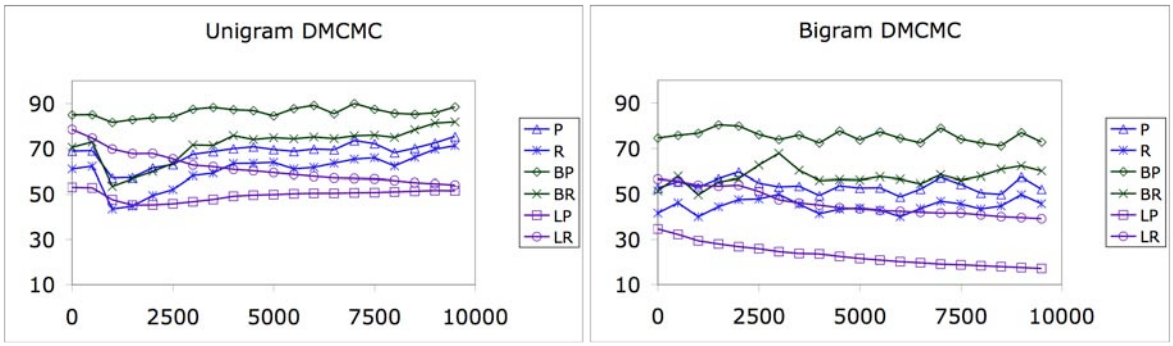
Table 1. Performance of different learning models on the second half of the corpus to factor out convergence time differences. Precision and recall over word tokens, word boundaries, and the lexicon resulting from the word segmentation are shown. All DMCMC learners use decay rate $d = 1$, and sample 10000 boundaries per utterance.

	Token Precision	Token Recall	Boundary Precision	Boundary Recall	Lexicon Precision	Lexicon Recall
Unigram Models (No Context)						
GGJ – Ideal	61.7	47.1	92.7	61.6	55.1	66.0
DPM	64.5	69.3	77.7	85.9	59.5	48.7
DPS	58.6	65.5	72.9	85.3	51.8	39.0
DMCMC	70.7	64.7	86.9	76.4	56.7	61.9
Bigram Models (Context)						
GGJ – Ideal	74.6	68.4	90.4	79.8	63.3	62.6
DPM	66.0	70.8	79.1	87.3	64.4	55.4
DPS	32.7	48.4	52.7	88.8	34.1	23.3
DMCMC	52.7	44.5	74.4	58.0	22.5	45.3

B. Figures

Figure 1. The graphs below display the detailed performance of the learning algorithms over the corpus as a whole, divided into groups of 500 utterances (x-axis = utterance number, y-axis = score percentage). Scores for token precision (P), token recall (R), boundary precision (BP), boundary recall (BR), lexicon precision (LP), and lexicon recall (LR) are shown.





References:

Bernstein-Ratner, N. (1984). Patterns of vowel Modification in motherese. *Journal of Child Language. 11,* 557-578.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. Machine Learning, 34:71–105.

Gambell, T. & Yang, C. (2006). Word Segmentation: Quick but not Dirty. Manuscript. Yale University.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 354-384.

Goldwater, S., Griffiths, T., and Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of COLING/ACL*, Sydney.

Marthi, B., Pasula, H., Russell, S, & Peres, Y. et al. 2002. Decayed MCMC Filtering. In *Proceedings of 18th UAI*, 319-326.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-olds. *Science*, *274*, 1926-928.

Thiessen, E., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706–716.

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114:245-272.