Improving Word Segmentation by Simultaneously Learning Phonotactics

Daniel Blanchard
Computer & Information Sciences
University of Delaware
dsblanch@udel.edu

Jeffrey Heinz
Linguistics & Cognitive Science
University of Delaware
heinz@udel.edu

Abstract

The most accurate unsupervised word segmentation systems that are currently available (Brent, 1999; Venkataraman, 2001; Goldwater, 2007) use a simple unigram model of phonotactics. While this simplifies some of the calculations, it overlooks cues that infant language acquisition researchers have shown to be useful for segmentation (Mattys et al., 1999; Mattys and Jusczyk, 2001). Here we explore the utility of using bigram and trigram phonotactic models by enhancing Brent's (1999) MBDP-1 algorithm. The results show the improved MBDP-Phon model outperforms other unsupervised word segmentation systems (e.g., Brent, 1999; Venkataraman, 2001; Goldwater, 2007).

1 Introduction

How do infants come to identify words in the speech stream? As adults, we break up speech into words with such ease that we often think that there are audible pauses between words in the same sentence. However, unlike some written languages, speech does not have any completely reliable markers for the breaks between words (Cole and Jakimik, 1980). In fact, languages vary on how they signal the ends of words (Cutler and Carter, 1987), which makes the task even more daunting. Adults at least have a lexicon they can use to recognize familiar words, but when an infant is first born, they do not have a pre-existing lexicon to consult. In spite of these challenges, by the age of six months infants can begin to segment words out of speech (Bortfeld et al., 2005). Here we present an efficient word segmentation system aimed to model how infants accomplish the task.

While an algorithm that could reliably extract orthographic representations of both novel and familiar words from acoustic data is something we would like to see developed, following earlier researchers, we simplify the problem by using a text that does not contain any word boundary markers. Hereafter, we use the phrase "word segmentation" to mean some process which adds word boundaries to a text that does not contain them.

This paper's focus is on unsupervised, incremental word segmentation algorithms; i.e., those that do not rely on preexisting knowledge of a particular language, and those that segment the corpus one utterance at a time. This is in contrast to supervised word segmentation algorithms (e.g., Teahan et al., 2000), which are typically used for segmenting text in documents written in languages that do not put spaces between their words like Chinese. (Of course, unsupervised word segmentation algorithms also have this application.) This also differs from batch segmentation algorithms (Goldwater, 2007; Johnson, 2008b; Fleck, 2008), which process the entire corpus at least once before outputting a segmentation of the corpus. Unsupervised incremental algorithms are of interest to some psycholinguists and acquisitionists interested in the problem of language learning, as well as theoretical computer scientists who are interested in what unsupervised, incremental models are capable of achieving.

Phonotactic patterns are the rules that determine what sequences of phonemes or allophones are allowable within words. Learning the phonotactic patterns of a language is usually modeled

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (http://creativecommons.org/licenses/by-nc-sa/3.0/). Some rights reserved.

CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning, pages 65–72
Manchester, August 2008

separately from word segmentation; e.g., current phonotactic learners such as Coleman and Pierrehumbert (1997), Heinz (2007), or Hayes and Wilson (2008) are given word-sized units as input.

However, infants appear to simultaneously learn which phoneme combinations are allowable within words and how to extract words from the input. It is reasonable that the two processes feed into one another, and when infants acquire a critical mass of phonotactic knowledge, they use it to make judgements about what phoneme sequences can occur within versus across word boundaries (Mattys and Jusczyk, 2001). We use this insight, also suggested by Venkataraman (2001) and recently utilized by Fleck (2008) in a different manner, to enhance Brent's (1999) model MBDP-1, and significantly increase segmentation accuracy. We call this modified segmentation model MBDP-Phon.

2 Related Work

2.1 Word Segmentation

The problem of unsupervised word segmentation has attracted many earlier researchers over the past fifty years (e.g., Harris, 1954; Olivier, 1968; de Marcken, 1995; Brent, 1999). In this section, we describe the base model MBDP-1, along with two other segmentation approaches, Venkataraman (2001) and Goldwater (2007). In §4, we compare MBDP-Phon to these models in more detail. For a thorough review of word segmentation literature, see Brent (1999) or Goldwater (2007).

2.1.1 MBDP-1

Brent's (1999) MBDP-1 (Model Based Dynamic Programming) algorithm is an implementation of the INCDROP framework (Brent, 1997) that uses a Bayesian model of how to generate an unsegmented text to insert word boundaries. The generative model consists of five steps:

1. Choose a number of word types, n.
2. Pick n distinct strings from $\Sigma^+\#$, which will make up the lexicon, L. Entries in L are labeled $W_1 \ldots W_n$. $W_0 = \$$, where \$ is the utterance boundary marker.
3. Pick a function, f, which maps word types to their frequency in the text.
4. Choose a function, s, to map positions in the text to word types.
5. Concatenate the words in the order specified by s, and remove the word delimiters ($\#$).

It is important to note that this model treats the generation of the text as a single event in the probability space, which allows Brent to make a number of simplifying assumptions. As the values for n, L, f, and s completely determine the segmentation, the probability of a particular segmentation, $\overline{w}_m$, can be calculated as:

$$P(\overline{w}_m) = P(n, L, f, s) \tag{1}$$

To allow the model to operate on one utterance at a time, Brent states the probability of each word in the text as a recursive function, $R(\overline{w}_k)$, where $\overline{w}_k$ is the text up to and including the word at position k, w_k. Furthermore, there are two specific cases for R: familiar words and novel words. If w_k is familiar, the model already has the word in its lexicon, and its score is calculated as in Equation 2.

$$R(\overline{w}_k) = \frac{f(w_k)}{k} \cdot \left(\frac{f(w_k) - 1}{f(w_k)}\right)^2 \tag{2}$$

Otherwise, the word is novel, and its score is calculated using Equation 3[1] (Brent and Tao, 2001),

$$R(\overline{w}_k) = \\ \frac{6}{\pi^2} \cdot \frac{n}{k} \cdot \frac{P_\Sigma(a_1)...P_\Sigma(a_q)}{1-P_\Sigma(\#)} \cdot \left(\frac{n-1}{n}\right)^2 \tag{3}$$

where P_Σ is the probability of a particular phoneme occurring in the text. The third term of the equation for novel words is where the model's unigram phonotactic model comes into play. We detail how to plug a more sophisticated phonotactic learning model into this equation in §3. With the generative model established, MBDP-1 uses a Viterbi-style search algorithm to find the segmentation for each utterance that maximizes the R values for each word in the segmentation.

Venkataraman (2001) notes that considering the generation of the text as a single event is unlikely to be how infants approach the segmentation problem. However, MBDP-1 uses an incremental search algorithm to segment one utterance at a time, which is more plausible as a model of infants' word segmentation.

[1] Brent (1999) originally described the novel word score as $R(\overline{w}_k) = \frac{6}{\pi^2} \cdot \frac{n_k}{k} \cdot \frac{P_\sigma(W_{n_k})}{1-\frac{n_k-1}{n_k}\cdot\sum_{j=1}^{n_k} P_\sigma(W_j)} \cdot \left(\frac{n_k-1}{n_k}\right)^2$, where P_σ is the probability of all the phonemes in the word occurring together, but the denominator of the third term was dropped in Brent and Tao (2001). This change drastically speeds up the model, and only reduces segmentation accuracy by $\sim 0.5\%$.

2.1.2 Venkataraman (2001)

MBDP-1 is not the only incremental unsupervised segmentation model that achieves promising results. Venkataraman's (2001) model tracks MBDP-1's performance so closely that Batchelder (2002) posits that the models are performing the same operations, even though the authors describe them differently.

Venkataraman's model uses a more traditional, smoothed n-gram model to describe the distribution of words in an unsegmented text.[2] The most probable segmentation is retrieved via a dynamic programming algorithm, much like Brent (1999).

We use MBDP-1 rather than Venkataraman's approach as the basis for our model only because it was more transparent how to plug in a phonotactic learning module at the time this project began.

2.1.3 Goldwater (2007)

We also compare our results to a segmenter put forward by Goldwater (2007). Goldwater's segmenter uses an underlying generative model, much like MBDP-1 does, only her language model is described as a Dirichlet process (see also Johnson, 2008b). While this model uses a unigram model of phoneme distribution, as did MBDP-1, it implements a bigram word model like Venkataraman (2001). A bigram word model is useful in that it prevents the segmenter from assuming that frequent word bigrams are not simply one word, which Goldwater observes happen with a unigram version of her model.

Goldwater uses a Gibbs sampler augmented with simulated annealing to sample from the posterior distribution of segmentations and determine the most likely segmentation of each utterance.[3] This approach requires non-incremental learning.[4] We include comparison with Goldwater's segmenter because it outperforms MBDP-1 and Venkataraman (2001) in both precision and recall, and we are interested in whether an incremental algorithm supplemented with phonotactic learning can match its performance.

2.2 Phonotactic Learning

Phonotactic acquisition models have seen a surge in popularity recently (e.g., Coleman and Pierrehumbert, 1997; Heinz, 2007; Hayes and Wilson, 2008). While Hayes and Wilson present a more complex Maximum Entropy phonotactic model in their paper than the one we add to MBDP-1, they also evaluate a simple n-gram phonotactic learner operating over phonemes. The input to the models is a list of English onsets and their frequency in the lexicon, and the basic trigram learner simply keeps track of the trigrams it has seen in the corpus. They test the model on novel words with acceptable rhymes—some well-formed (e.g., [kɪp]), and some less well-formed (e.g., [stwɪk])—so any ill-formedness is attributable to onsets. This basic trigram model explains 87.7% of the variance in the scores that Scholes (1966) reports his 7th grade students gave when subjected to the same test. When Hayes and Wilson run their Maximum Entropy phonotactic learning model with n-grams over phonological features, the r-score increases substantially to 95.6%.

Given the success and simplicity of the basic n-gram phonotactic model, we choose to integrate this with MBDP-1.

3 Extending MBDP-1 with Phonotactics

The main contribution of our work is adding a phonotactic learning component to MBDP-1 (Brent, 1999). As we mention in §2.1.1, the third term of Equation 3 is where MBDP-1's unigram phonotactic assumption surfaces. The original model simply multiplies the probabilities of all the phonemes in the word together and divides by one minus the probability of a particular phoneme being the word boundary to come up with probability of the phoneme combination. The order of the phonemes in the word has no effect on its score. The only change we make to MBDP-1 is to the third term of Equation 3. In MBDP-Phon this becomes

$$\prod_{i=0}^{q} P_{\mathrm{MLE}}(a_i \ldots a_j) \tag{4}$$

where $a_i \ldots a_j$ is an n-gram inside a proposed word, and a_0 and a_q are both the word boundary symbol, $\#$[5].

It is important to note that probabilities calculated in Equation 4 are maximum likelihood estimates of the joint probability of each n-gram in the word. The maximum likelihood estimate (MLE)

[2] We refer the reader to Venkataraman (2001) for the details of this approach.

[3] We direct the reader to Goldwater (2007) for details.

[4] In our experiments and those in Goldwater (2007), the segmenter runs through the corpus 1000 times before outputting the final segmentation.

[5] The model treats word boundary markers like a phoneme for the purposes of storing n-grams (i.e., a word boundary marker may occur anywhere within the n-grams).

for a particular n-gram inside a word is calculated by dividing the total number of occurrences of that n-gram (including in the word we are currently examining) by the total number of n-grams (including those in the current word). The numbers of n-grams are computed with respect to the obtained lexicon, not the corpus, and thus the frequency of lexical items in the corpus does not affect the n-gram counts, just like Brent's unigram phonotactic model and other phonotactic learning models (e.g., Hayes and Wilson, 2008).

We use the joint probability instead of the conditional probability which is often used in computational linguistics (Manning and Schütze, 1999; Jurafsky and Martin, 2000), because of our intuition that the joint probability is truer to the idea that a phonotactically well-formed word is made up of n-grams that occur frequently in the lexicon. On the other hand, the conditional probability is used when one tries to predict the next phoneme that will occur in a word, rather than judging the well-formedness of the word as a whole.[6]

We are able to drop the denominator that was originally in Equation 3, because $P_\Sigma(\#)$ is zero for an n-gram model when $n > 1$. This simple modification allows the model to learn what phonemes are more likely to occur at the beginnings and ends of words, and what combinations of phonemes rarely occur within words.

What is especially interesting about this modification is that the phonotactic learning component estimates the probabilities of the n-grams by using their relative frequencies in the words the segmenter has extracted. The phonotactic learner is guaranteed to see at least two valid patterns in every utterance, as the n-grams that occur at the beginnings and ends of utterances are definitely at the beginnings and ends of words. This allows the learner to provide useful information to the segmenter even early on, and as the segmenter correctly identifies more words, the phonotactic learner has more correct data to learn from. Not only is this mutually beneficial process supported by evidence from language acquisitionists (Mattys et al., 1999; Mattys and Jusczyk, 2001), it also resembles co-training (Blum and Mitchell, 1998). We refer to the extended version of Brent's model described above as MBDP-Phon.

4 Evaluation

4.1 The Corpus

We run all of our experiments on the Bernstein-Ratner (1987) infant-directed speech corpus from the CHILDES database (MacWhinney and Snow, 1985). This is the same corpus that Brent (1999), Goldwater (2007), and Venkataraman (2001) evaluate their models on, and it has become the *de facto* standard for segmentation testing, as unlike other corpora in CHILDES, it was phonetically transcribed.

We examine the transcription system Brent (1999) uses and conclude some unorthodox choices were made when transcribing the corpus. Specifically, some phonemes that are normally considered distinct are combined into one symbol, which we call a bi-phone symbol. These phonemes combinations include diphthongs and vowels followed by /ɹ/. Another seemingly arbitrary decision is the distinction between stressed and unstressed syllabic /ɹ/ sound (i.e., there are different symbols for the /ɹ/ in "butter" and the /ɹ/ in "bird") since stress is not marked elsewhere in the corpus. To see the effect of these decisions, we modified the corpus so that the bi-phone symbols were split into two[7] and the syllabic /ɹ/ symbols were collapsed into one.

4.2 Accuracy

We ran MBDP-1 on the original corpus, and the modified version of the corpus. As illustrated by Figures 1 and 2, MBDP-1 performs worse on the modified corpus with respect to both precision and recall. As MBDP-1 and MBDP-Phon are both iterative learners, we calculate segmentation precision and recall values over 500-utterance blocks. Per Brent (1999) and Goldwater (2007), precision and recall scores reflect correctly segmented words, not correctly identified boundaries.

We also test to see how the addition of an n-gram phonotactic model affects the segmentation accuracy of MBDP-Phon by comparing it to MBDP-1 on our modified corpus.[8] As seen in Figure 3, MBDP-Phon using bigrams (henceforth MBDP-Phon-Bigrams) is consistently more precise in its

[6] This intuition is backed up by preliminary results suggesting MBDP-Phon performs better when using MLEs of the joint probability as opposed to conditional probability. There is an interesting question here, which is beyond the scope of this paper, so we leave it for future investigation.

[7] We only split diphthongs whose first phoneme can occur in isolation in English, so the vowels in "bay" and "boat" were not split.

[8] We also compare MBDP-Phon to MBDP-1 on the original corpus. The results are given in Tables 1 and 2.

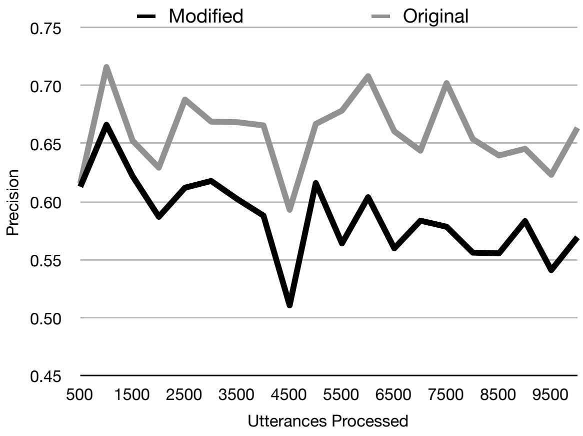

Figure 1: Precision of MBDP-1 on both corpora.

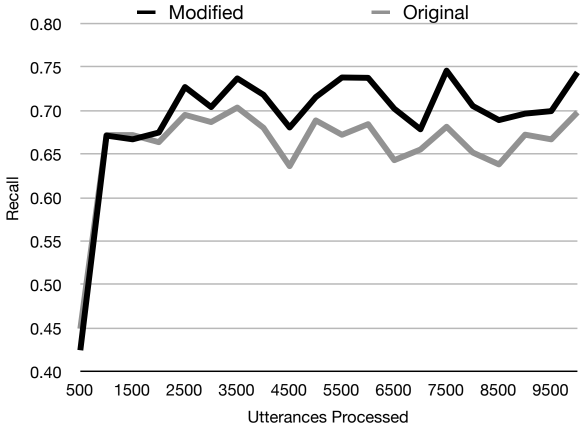

Figure 2: Recall of MBDP-1 on both corpora.

segmentation than MBDP-1, and bests it by $\sim 18\%$ in the last block. Furthermore, MBDP-Phon-Bigrams significantly outpaces MBDP-1 with respect to recall only after seeing 1000 utterances, and finishes the corpus $\sim 10\%$ ahead of MBDP-1 (see Figure 4). MBDP-Phon-Trigrams does not fair as well in our tests, falling behind MBDP-1 and MBDP-Phon-Bigrams in recall, and MBDP-Phon-Bigrams in precision. We attribute this poor performance to the fact that we are not currently smoothing the n-gram models in any way, which leads to data sparsity issues when using trigrams. We discuss a potential solution to this problem in §5.

Having established that MBDP-Phon-Bigrams significantly outperforms MBDP-1, we compare its segmentation accuracy to those of Goldwater (2007) and Venkataraman (2001).[9] As before, we

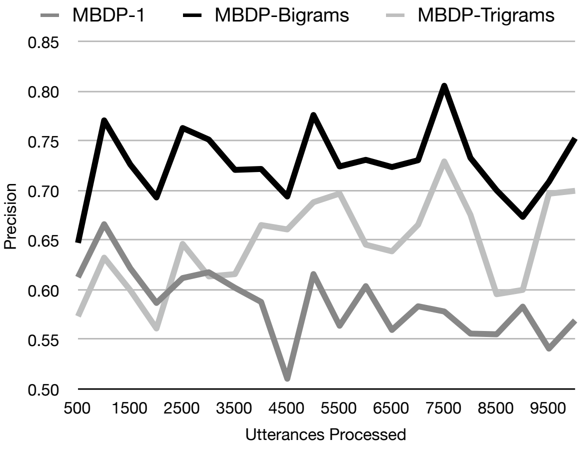

Figure 3: Precision of MBDP-1 and MBDP-Phon on modified corpus.

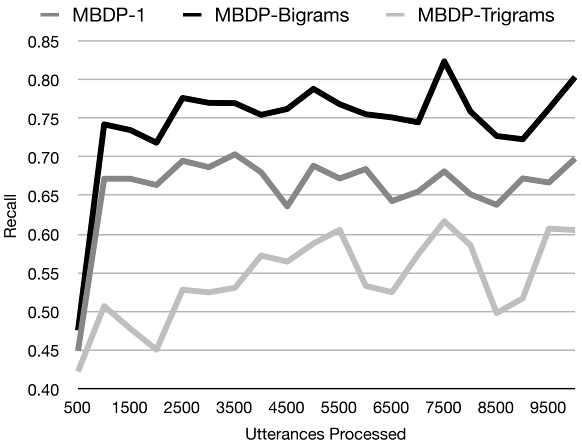

Figure 4: Recall of MBDP-1 and MBDP-Phon on modified corpus.

run the models on the entire corpus, and then measure their performance over 500-utterance blocks.

MBDP-Phon-Bigrams edges out Goldwater's model in precision on our modified corpus, with an average precision of 72.79% vs. Goldwater's 70.73% (Table 1). If we drop the first 500-utterance block for MBDP-Phon-Bigrams because the model is still in the early learning stages, whereas Goldwater's has seen the entire corpus, its average precision increases to 73.21% (Table 1). When considering the recall scores in Table 2, it becomes clear that MBDP-Phon-Bigrams has a clear advantage over the other models. Its average recall is higher than or nearly equal to both of the other models' maximum scores. Since Venkataraman's (2001) model performs similarly to MBDP-1, it is no surprise that MBDP-Phon-Bigrams achieves higher precision and recall.

[9] We only examine Venkataraman's unigram model, as his bigram and trigram models perform better on precision, but worse on recall.

	MBDP-Phon-Bigrams	Venkataraman	Goldwater
	Original: Utterances 0 to 9790		
Avg.	72.84%	67.46%	67.87%
Max.	79.91%	71.79%	71.98%
Min.	63.97%	61.77%	61.87%
	Modified: Utterances 0 to 9790		
Avg.	72.79%	59.64%	70.73%
Max.	80.60%	66.84%	74.61%
Min.	64.78%	52.54%	65.29%
	Modified: Utterances 500 to 9790		
Avg.	73.21%	59.54%	70.59%
Max.	80.60%	66.84%	74.61%
Min.	67.40%	52.54%	65.29%

Table 1: Precision statistics for MBDP-Phon-Bigrams, Goldwater, and Venkataraman on both corpora over 500-utterance blocks.

	MBDP-Phon-Bigrams	Venkataraman	Goldwater
	Original: Utterances 0 to 9790		
Avg.	72.03%	70.02%	71.02%
Max.	79.31%	75.59%	76.79%
Min.	44.71%	42.57%	64.32%
	Modified: Utterances 0 to 9790		
Avg.	74.63%	66.24%	70.48%
Max.	82.45%	70.47%	74.79%
Min.	47.63%	44.71%	63.74%
	Modified: Utterances 500 to 9790		
Avg.	76.05%	67.37%	70.28%
Max.	82.45%	70.47%	74.79%
Min.	71.92%	63.86%	63.74%

Table 2: Recall statistics for MBDP-Phon-Bigrams, Goldwater, and Venkataraman on both corpora over 500-utterance blocks.

The only metric by which MBDP-Phon-Bigrams does not outperform the other algorithms is lexical precision, as shown in Table 3. Lexical precision is the ratio of the number of correctly identified words in the lexicon to the total number of words in the lexicon (Brent, 1999; Venkataraman, 2001).[10] The relatively poor performance of MBDP-Phon-Bigrams is due to the incremental nature of the MBDP algorithm. Initially, it makes numerous incorrect guesses that are added to the lexicon, and there is no point at which the lexicon is purged of earlier erroneous guesses (c.f. the improved lexical precision when omitting the first block in Table 3). On the other hand, Goldwater's algorithm runs over the corpus multiple times, and only produces output when it settles on a final segmentation.

In sum, MBDP-Phon-Bigrams significantly improves the accuracy of MBDP-1, and achieves better performance than the models described in Venkataraman (2001) and Goldwater (2007).

5 Future Work

There are many ways to implement phonotactic learning. One idea is to to use n-grams over phonological features, as per Hayes and Wilson (2008). Preliminary results have shown that we need to add smoothing to our n-gram model, and we plan to use Modified Kneser-Ney smoothing (Chen and Goodman, 1998).

Another approach would be to develop a syllable-based phonotactic model (Coleman and Pierrehumbert, 1997). Johnson (2008b) achieves impressive segmentation results by adding a syllable level with Adaptor grammars.

Some languages (e.g., Finnish, and Navajo) contain long-distance phonotactic constraints that cannot be learned by n-gram learners (Heinz, 2007). Heinz (2007) shows that precedence-based learners—which work like a bigram model, but without the restriction that the elements in the bigram be adjacent—can handle many long-distance agreement patterns (e.g., vowel and consonantal harmony) in the world's languages. We posit that adding such a learner to MBDP-Phon would allow it to handle a greater variety of languages.

Since none of these approaches to phonotactic learning depend on MBDP-1, it is also of interest to integrate phonotactic learners with other word segmentation strategies.

In addition to evaluating segmentation models integrated with phonotactic learning on their segmentation performance, it would be interesting to evaluate the quality of the phonotactic grammars obtained. A good point of comparison for English are the constraints obtained by Hayes and Wilson (2008), since the data with which they tested their phonotactic learner is publicly available.

Finally, we are looking forward to investigat-

[10] See Brent (1999) for a discussion of the meaning of this statistic.

	MBDP-Phon-Bigrams	Venkataraman	Goldwater
	Original: Utterances 0 to 9790		
Avg.	47.69%	49.78%	56.50%
Max.	49.71%	52.95%	63.09%
Min.	46.30%	41.83%	55.33%
	Modified: Utterances 0 to 9790		
Avg.	48.31%	45.98%	58.03%
Max.	50.42%	48.90%	65.58%
Min.	41.74%	36.57%	56.43%
	Modified: Utterances 500 to 9790		
Avg.	54.34%	53.06%	57.95%
Max.	63.76%	54.35%	62.30%
Min.	51.31%	51.95%	56.52%

Table 3: Lexical precision statistics for MBDP-Phon-Bigrams, Goldwater, and Venkataraman on both corpora over 500-utterance blocks.

ing the abilities of these segmenters on corpora of different languages. Fleck (2008) tests her segmenter on a number of corpora, including Arabic and Spanish, and Johnson (2008a) applies his segmenter to a corpus of Sesotho.

6 Conclusion

From the results established in §4, we can conclude that MBDP-Phon using a bigram phonotactic model is more accurate than the models described in Brent (1999), Venkataraman (2001), and Goldwater (2007). The n-gram phonotactic model improves overall performance, and is especially useful for corpora that do not encode diphthongs with bi-phone symbols. The main reason there is such a marked improvement with MBDP-Phon vs. MBDP-1 when the bi-phone symbols were removed from the original corpus is that these bi-phone symbols effectively allow MBDP-1 to have a select few bigrams in the cases where it would otherwise over-segment.

The success of MBDP-Phon is not clear evidence that the INCDROP framework (Brent, 1997) is superior to Venkataraman or Goldwater's models. We imagine that adding a phonotactic learning component to either of their models would also improve their performance.

We also tentatively conclude that phonotactic patterns can be learned from unsegmented text. However, the phonotactic patterns learned by our model ought to be studied in detail to see how well they match the phonotactic patterns of English.

MBDP-Phon's performance reinforces the theory put forward by language acquisition researchers that phonotactic knowledge is a cue for word segmentation (Mattys et al., 1999; Mattys and Jusczyk, 2001). Furthermore, our results indicate that learning phonotactic patterns can occur simultaneously with word segmentation. Finally, further investigation of the simultaneous acquisition of phonotactics and word segmentation appears fruitful for theoretical and computational linguists, as well as acquisitionists.

Acknoledgements

We are grateful to Roberta Golinkoff who inspired this project. We also thank Vijay Shanker for valuable discussion, Michael Brent for the corpus, and Sharon Goldwater for the latest version of her code.

References

Batchelder, Eleanor Olds. 2002. Bootstrapping the lexicon: a computational model of infant speech segmentation. *Cognition*, 83(2):167–206.

Bernstein-Ratner, Nan. 1987. *The phonology of parent child speech*, volume 6. Erlbaum, Hillsdale, NJ.

Blum, Avrim and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Workshop on Computational Learning Theory*, pages 92–100.

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

Brent, Michael R. 1997. Towards a unified model of lexical acquisition and lexical access. *Journal of Psycholinguistic Research*, 26(3):363–375.

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

Brent, Michael R and Xiaopeng Tao. 2001. Chinese text segmentation with mbdp-1: Making the most of training corpora. In *39th Annual Meeting of the ACL*, pages 82–89.

Chen, Stanley F and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Center for Research in Computing Technology, Harvard University.

Cole, Ronald and Jola Jakimik. 1980. *A model of speech perception*, pages 136–163. Lawrence Erlbaum Associates, Hillsdale, NJ.

Coleman, John and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Third Meeting of the ACL SIGPHON*, pages 49–56. ACL, Somerset, NJ.

Cutler, Anne and David Carter. 1987. The predominance of strong initial syllables in the english vocabulary. *Computer Speech and Language*, 2(3-4):133–142.

de Marcken, Carl. 1995. Acquiring a lexicon from unsegmented speech. In *33rd Annual Meeting of the ACL*, pages 311–313.

Fleck, Margaret M. 2008. Lexicalized phonotactic word segmentation. In *46th Annual Meeting of the ACL*, pages 130–138. ACL, Morristown, NJ.

Goldwater, Sharon. 2007. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University, Department of Cognitive and Linguistic Sciences.

Harris, Zellig. 1954. Distributional structure. *Word*, 10(2/3):146–62.

Hayes, Bruce and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*.

Heinz, Jeffrey. 2007. *Inductive Learning of Phonotactic Patterns*. Ph.D. thesis, University of California, Los Angeles, Department of Linguistics.

Johnson, Mark. 2008a. Unsupervised word segmentation for sesotho using adaptor grammars. In *Tenth Meeting of ACL SIGMORPHON*, pages 20–27. ACL, Morristown, NJ.

Johnson, Mark. 2008b. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *46th Annual Meeting of the ACL*, pages 398–406. ACL, Morristown, NJ.

Jurafsky, Daniel and James Martin. 2000. *Speech and Language Processing*. Prentice-Hall.

MacWhinney, Brian and Catherine Snow. 1985. The child language data exchange system. *Journal of child language*, 12(2):271–95.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Mattys, Sven and Peter Jusczyk. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78:91–121.

Mattys, Sven, Peter Jusczyk, Paul Luce, and James Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38:465–494.

Olivier, Donald. 1968. *Stochastic Grammars and Language Acquisition Mechanisms*. Ph.D. thesis, Harvard Univerity.

Scholes, Robert. 1966. *Phonotactic Grammaticality*. Mouton, The Hague.

Teahan, W. J., Rodger McNab, Yingying Wen, and Ian H. Witten. 2000. A compression-based algorithm for chinese word segmentation. *Computational Linguistics*, 26(3):375–393.

Venkataraman, Anand. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):352–372.