Word Segmentation as General Chunking

Daniel Hewlett and Paul Cohen

Introduction

The ability to extract words from fluent speech appears early in human development, as early as seven months (Jusczyk, 1999). Many models of this word segmentation ability have emerged, coming from such diverse fields as linguistics, psychology, and computer science. Here, we examine a representative set of computational models in light of what is known about the segmentation ability that children possess. Specifically, we explore the possibility that children could use general-purpose chunking mechanisms to perform word segmentation. We argue that such a model is consistent with key experimental results and offers a more parsimonious alternative to models that posit special-purpose linguistic mechanisms to explain word segmentation.

Word Segmentation Strategies

Considering only the general outcome that children successfully learn to segment words, a wide range of segmentation strategies are plausible. One common approach is to take advantage of the semi-supervised nature of the problem: Each utterance is implicitly surrounded by two boundaries. When a sequence of phonemes encountered as a short utterance is discovered within a larger utterance, these boundaries can be placed into the larger utterance, thus splitting the large utterance. When these utterance segments are used to split future utterances, the resulting self-sustaining process is called *bootstrapping*.

Brent (1999) proposed a model of word segmentation in children that operates on similar principles, in the form of the MBDP-1 algorithm. MBDP-1 achieves its robust form of bootstrapping through the use of Bayesian maximum likelihood estimation of a language model. Since MBDP-1, several other language-modeling have been proposed (Venkataraman, 2001; Fleck, 2008; Goldwater, Griffiths, & Johnson, 2008).

However, because such accounts depend intrinsically on the input being structured as a series of bounded utterances, they cannot explain key experimental results in the child language learning literature, such as the seminal series of studies by Saffran et al. (1996). In these studies, Saffran et al. show that both adults and 8-month-old infants can extract words from a continuous speech stream in an artifical language. Speech segmentation experiments with cotton-top tamarins (Hauser, Newport, & Aslin, 2001) have yielded similar results to Saffran's experiments with human infants, suggesting that this ability might be innate. As bootstrapping algorithms cannot produce this result, Saffran et al. concluded that statistical analysis of the properties of the speech stream forms the basis for the infant's ability to infer word boundaries. In particular, they proposed that infants attend to transitional probabilities (TP) between syllables, and posit boundaries at places of low transitional probability.

While this simple TP model is sufficient to explain the results of Saffran et al.'s 1996 study, it performs very poorly on actual child-directed speech, regardless of whether the probabilities are calculated between phonemes (Brent, 1999) or syllables (Gambell & Yang, 2006). In the case of syllables, preferred by Saffran et al., there is a further problem of correct syllabification, which requires knowledge of language-specific phonotactic constraints. In response to this failure of TP to generalize to the full complexity of natural language, Gambell and Yang (2004, 2006) suggest that perhaps children's ability to segment sequences based on statistical analysis is not the driving force behind word segmentation in a natural setting. Instead, they propose a model that combines innate constraints on linguistic stress with simple bootstrapping.

Here, we explore an alternative explanation: Slightly more sophisticated statistical methods, particularly those based on entropy, may provide a natural explanation of the effects Saffran et al. observed, and still perform well on natural language input. Within the context of language, the form of entropy most often explored is *boundary entropy* (also called *branching entropy*), which is the entropy of the set of continuations of a sequence, typically a sequence of phonemes or syllables. For example, the boundary entropy of `victo` is very low, because the next letter is almost certainly `r`, but the boundary entropy of `th` is high, as many letters can follow `th` in English.

Harris was one of the first to propose that boundary entropy at the phoneme level could serve as an indicator for morpheme boundaries (and therefore word boundaries as well). He observed that morpheme boundaries tended to occur at points where boundary entropy increased relative to the previous position (Harris, 1955). For example, the boundary entropy of `victor` is greater than that of `victo`, and thus a morpheme boundary is likely to occur after `victor`. This proposal was implemented computationally by Tanaka-Ishii and Jin (2006), who found that the method was able to segment phonemic English and Chinese corpora with performance comparable to other unsupervised algorithms.

Chunking

Here, we consider a slightly more complex entropic method of segmentation, in the form of the VOTING EX-

PERTS algorithm (VE) for finding chunks. A *chunk* is a sequence with the property that the elements within it predict one another, but do not predict elements outside the sequence. In information-theoretic terms, chunks have low *internal entropy* (also called *surprisal*), and high boundary entropy. VE is a local, greedy algorithm that works by moving a small sliding window along the input, and examining only sequences within the window. For further details of the VE algorithm, see Cohen (2001).

Importantly for the present discussion, the entropic properties of chunks that enable VE to succeed are present in a variety of domains, including word segmentation. Cohen and Adams (2001) explored word segmentation in a variety of languages, as well as segmenting sequences of robot actions. Miller and Stoytchev (2008) also used VE twice to perform a vision task similar to OCR: First to chunk columns of pixels into images letters, and then to chunk sequences of these discovered letters into words.

In Figure 1, we compare VE to a bootstrapping algorithm proposed by Gambell and Yang (2006) that is endowed with innate knowledge of stress patterns (the "Unique Stress Constraint," or USC), allowing it to eliminate many potential boundary locations. Operating on syllabified input, both algorithms outperform simple transitional probability (TP), but the performance of VE demonstrates that successful segmentation does not require the innate constraints suggested by Gambell and Yang.

Chunking and Bootstrapping

The framework provided by the VOTING EXPERTS algorithm does not preclude the possibility of bootstrapping. BOOTSTRAP VOTING EXPERTS (BVE) is an extension to VOTING EXPERTS that incorporates knowledge gained from prior segmentation attempts when segmenting new input (Hewlett & Cohen, 2009). However, unlike bootstrapping algorithms such as MBDP-1, BVE stores statistics describing the beginnings and endings of chunks. In the word segmentation domain, these statistics effectively correspond to phonotactic constraints that are inferred from hypothesized segmentations. Inferred boundaries are stored in a data structure called a *knowledge trie* (shown in Figure 3). BVE achieved a higher level of performance on phonemically-encoded corpora of child-directed speech taken from the CHILDES database (MacWhinney & Snow, 1985). These results from Hewlett and Cohen are reproduced in Figure 2.

Over time, BVE refines the quality of the boundary information stored in the knowledge trie. Thus, simply by storing the beginnings and endings of segments, the knowledge trie comes to store sequences like `#cat#`, where `#` represents a word boundary. The set of such bounded sequences constitutes a simple, but accurate, emergent lexicon. After segmenting a corpus of child-directed speech, the ten most frequent words of this lexicon are *you, the, that, what, is, it, this, what's, to*, and *look*. Of the 100 most frequent words, 93 are correct. The 7 errors include splitting off morphemes such as *ing*, and merging frequently co-occurring word pairs such as *do you*.

Artificial Language Results

To simulate the input children heard during Saffran et al.'s 1996 experiment, we generated a corpus of 400 words, each chosen from the four artificial words from that experiment (`dapiku`, `tilado`, `burobi`, and `pagotu`). Like the original study, the only condition imposed on the random sequence was that no word would appear twice in succession. Voting Experts achieves an F-score of 1.0 whether the input is syllabified or considered simply as a stream of phonemes, suggesting that a child equipped with a chunking ability similar to VE could succeed even without syllabification.

Conclusion

We have argued that the existence of an innate, domain-independent chunking ability provides an explanation for the statistical segmentation ability of infants reported by Saffran et al. When segmenting natural language, chunking can be coupled with simple bootstrapping to produce a segmentation ability that is the most accurate of the computational models studied here, when evaluated against corpora of child-directed speech from CHILDES.

Algorithm	Precision	Recall	BF
Transitional Probability (TP)	0.416	0.233	0.298
TP with USC	0.735	0.712	0.723
Bootstrapping with USC	0.959	0.934	0.946
Voting Experts	0.918	0.992	0.953
All Locations	0.839	1.000	0.913

Figure 1: Performance of various algorithms on syllabified input from the Brown corpus (CHILDES), as measured by boundary F-score. Other than VOTING EXPERTS and ALL-LOCATIONS, values are taken from (Gambell & Yang, 2006).

Algorithm	BP	BR	BF	WP	WR	WF
VE	0.867	0.854	0.860	0.652	0.642	0.647
BVE	0.928	**0.905**	**0.916**	**0.791**	**0.794**	**0.793**
MBDP-1[i]	0.803	0.843	0.823	0.670	0.694	0.682
HDP[i]	0.903	0.808	0.852	0.752	0.696	0.723
WordEnds[i]	**0.946**	0.737	0.829	-	-	0.707
All Locations	0.258	1.000	0.411	0.013	0.051	0.021

Figure 2: Results obtained for the Bernstein Ratner corpus (Ratner, 1987), as well as published results from selected other algorithms. Performance is measured in boundary/word precision, recall, and F-score. Reproduced from Hewlett and Cohen (2009).

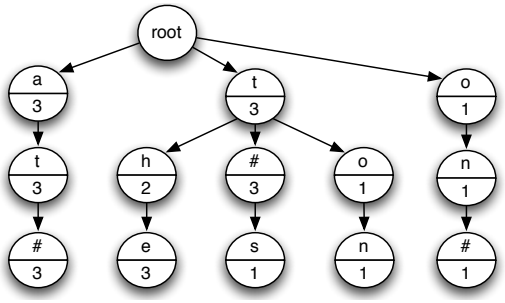

Figure 3: A portion of the knowledge trie built from `#the#cat#sat#on#the#mat#`. Numbers within each node are frequency counts.

References

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

Cohen, P., & Adams, N. (2001). An algorithm for segmenting categorical time series into meaningful episodes. In *Proceedings of the fourth symposium on intelligent data analysis.*

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. In *ACL 2008 proceedings.*

Gambell, T., & Yang, C. (2004). Statistics learning and universal grammar: Modeling word segmentation. In *COLING 2004, workshop on psychocomputational models of human language acquisition.*

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty.* (Manuscript, Yale University)

Goldwater, S., Griffiths, T. L., & Johnson, M. (2008). *A bayesian framework for word segmentation.* (Submitted)

Harris, Z. (1955). From phoneme to morpheme. *Language*, *1*, 190–192.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a nonhuman primate: statistical learning in cotton-top tamarins. *Cognition*, *78*, B53–B64.

Hewlett, D., & Cohen, P. (2009). Bootstrap voting experts. In *Proceedings of the twenty-first international joint conference on artificial intelligence (IJCAI-09).*

Jusczyk, P. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, *3*, 323–328.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*, 271–192.

Miller, M., & Stoytchev, A. (2008). Hierarchical voting experts: An unsupervised algorithm for hierarchical sequence segmentation. In *Proceedings of the 7th ieee international conference on development and learning (icdl).*

Ratner, N. B. (1987). The phonology of parent child speech. In K. Nelson & A. van Kleek (Eds.), *Children's language* (Vol. 6). Hillsdale, NJ: Erlbaum.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Tanaka-Ishii, K., & Jin, Z. (2006). From phoneme to morpheme: Another verification using a corpus. In *ICCPOL 2006* (pp. 234–244).

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, *27*, 351–372.