Cognition 112 (2009) 21–54

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

A Bayesian framework for word segmentation: Exploring the effects of context

Sharon Goldwater [a,*], Thomas L. Griffiths [b], Mark Johnson [c]

[a] School of Informatics, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, UK
[b] Department of Psychology, University of California, Berkeley, CA, United States
[c] Department of Cognitive and Linguistic Sciences, Brown University, United States

ARTICLE INFO

Article history:
Received 30 May 2008
Revised 11 March 2009
Accepted 13 March 2009

Keywords:
Computational modeling
Bayesian
Language acquisition
Word segmentation

ABSTRACT

Since the experiments of Saffran et al. [Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning in 8-month-old infants. *Science, 274*, 1926–1928], there has been a great deal of interest in the question of how statistical regularities in the speech stream might be used by infants to begin to identify individual words. In this work, we use computational modeling to explore the effects of different assumptions the learner might make regarding the nature of words – in particular, how these assumptions affect the kinds of words that are segmented from a corpus of transcribed child-directed speech. We develop several models within a Bayesian ideal observer framework, and use them to examine the consequences of assuming either that words are independent units, or units that help to predict other units. We show through empirical and theoretical results that the assumption of independence causes the learner to undersegment the corpus, with many two- and three-word sequences (e.g. *what's that*, *do you*, *in the house*) misidentified as individual words. In contrast, when the learner assumes that words are predictive, the resulting segmentation is far more accurate. These results indicate that taking context into account is important for a statistical word segmentation strategy to be successful, and raise the possibility that even young infants may be able to exploit more subtle statistical patterns than have usually been considered.

1. Introduction

One of the first problems infants must solve as they are acquiring language is word segmentation: identifying word boundaries in continuous speech. About 9% of utterances directed at English-learning infants consist of isolated words (Brent & Siskind, 2001), but there is no obvious way for children to know from the outset which utterances these are. Since multi-word utterances generally have no apparent pauses between words, children must be using other cues to identify word boundaries. In fact, there is evidence that infants use a wide range of weak cues for word segmentation. These cues include phonotactics (Mattys, Jusczyk, Luce, & Morgan, 1999), allophonic variation (Jusczyk, Hohne, & Bauman, 1999), metrical (stress) patterns (Jusczyk, Houston, & Newsome, 1999; Morgan, Bonamo, & Travis, 1995), effects of coarticulation (Johnson & Jusczyk, 2001), and statistical regularities in the sequences of syllables found in speech (Saffran, Aslin, & Newport, 1996). This last source of information can be used in a language-independent way, and seems to be used by infants earlier than most other cues, by the age of 7 months (Thiessen & Saffran, 2003). These facts have caused some researchers to propose that strategies based on statistical sequencing information are a crucial first step in bootstrapping word segmentation (Thiessen & Saffran, 2003), and have provoked a great deal of interest in these strategies (Aslin, Saffran, & Newport, 1998; Saffran, Newport, & Aslin, 1996; Saffran et al., 1996; Toro, Sinnett, & Soto-Faraco, 2005). In this paper, we use computational

* Corresponding author. Tel.: +44 131 651 5609.
E-mail addresses: sgoldwat@inf.ed.ac.uk, sgwater@gmail.com (S. Goldwater).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.03.008

modeling techniques to examine some of the assumptions underlying much of the research on statistical word segmentation.

Most previous work on statistical word segmentation is based on the observation that transitions from one syllable or phoneme to the next tend to be less predictable at word boundaries than within words (Harris, 1955; Saffran et al., 1996). Behavioral research has shown that infants are indeed sensitive to this kind of predictability, as measured by statistics such as transitional probabilities (Aslin et al., 1998; Saffran et al., 1996). This research, however, is agnostic as to the mechanisms by which infants use statistical patterns to perform word segmentation. A number of researchers in both cognitive science and computer science have developed algorithms based on transitional probabilities, mutual information, and similar statistics of predictability in order to clarify how these statistics can be used procedurally to identify words or word boundaries (Ando & Lee, 2000; Cohen & Adams, 2001; Feng, Chen, Deng, & Zheng, 2004; Swingley, 2005). Here, we take a different approach: we seek to identify the *assumptions* the learner must make about the nature of language in order to correctly segment natural language input.

Observations about predictability at word boundaries are consistent with two different kinds of assumptions about what constitutes a *word*: either a word is a unit that is statistically independent of other units, or it is a unit that helps to predict other units (but to a lesser degree than the beginning of a word predicts its end). In most artificial language experiments on word segmentation, the first assumption is adopted implicitly by creating stimuli through random (or near-random) concatenation of nonce words. This kind of random concatenation is often necessary for controlled experiments with human subjects, and has been useful in demonstrating that humans are sensitive to the statistical regularities in such randomly generated sequences. However, it obviously abstracts away from many of the complexities of natural language, where regularities exist not only in the relationships between sub-word units, but also in the relationships between words themselves. We know that humans are able to use sub-word regularities to begin to extract words; it is natural to ask whether attending to these kinds of regularities is sufficient for a statistical learner to succeed with word segmentation in a more naturalistic setting. In this paper, we use computer simulations to examine learning from natural, rather than artificial, language input. We ask what kinds of words are identified by a learner who assumes that words are statistically independent, or (alternatively) by a learner who assumes as well that words are predictive of later words. We investigate this question by developing two different Bayesian models of word segmentation incorporating each of these two different assumptions. These models can be seen as *ideal learners*: they are designed to behave optimally given the available input data, in this case a corpus of phonemically transcribed child-directed speech.

Using our ideal learning approach, we find in our first set of simulations that the learner who assumes that words are statistically independent units tends to undersegment the corpus, identifying commonly co-occurring sequences of words as single words. These results seem to conflict with those of several earlier models (Batchelder, 2002; Brent, 1999; Venkataraman, 2001), where systematic undersegmentation was not found even when words were assumed to be independent. However, we argue here that these previous results are misleading. Although each of these learners is based on a probabilistic model that defines an optimal solution to the segmentation problem, we provide both empirical and analytical evidence that the segmentations found by these learners are not the optimal ones. Rather, they are the result of limitations imposed by the particular learning algorithms employed. Further mathematical analysis shows that undersegmentation is the optimal solution to the learning problem for *any* reasonably defined model that assumes statistical independence between words.

Moving on to our second set of simulations, we find that permitting the learner to gather information about word-to-word dependencies greatly reduces the problem of undersegmentation. The corpus is segmented in a much more accurate, adult-like way. These results indicate that, for an ideal learner to identify words based on statistical patterns of phonemes or syllables, it is important to take into account that frequent predictable patterns may occur *either* within words *or* across words. This kind of dual patterning is a result of the hierarchical structure of language, where predictable patterns occur at many different levels. A learner who considers predictability at only one level (sub-word units within words) will be less successful than a learner who considers also the predictability of larger units (words) within their sentential context. The second, more nuanced interpretation of the statistical patterns in the input leads to better learning.

Our work has important implications for the understanding of human word segmentation. We show that successful segmentation depends crucially on the assumptions that the learner makes about the nature of words. These assumptions constrain the kinds of inferences that are made when the learner is presented with naturalistic input. Our ideal learning analysis allows us to examine the kinds of constraints that are needed to successfully identify words, and suggests that infants or young children may need to account for more subtle statistical effects than have typically been discussed in the literature. To date, there is little direct evidence that very young language learners approximate ideal learners. Nevertheless, this suggestion is not completely unfounded, given the accumulating evidence in favor of humans as ideal learners in other domains or at other ages (Frank, Goldwater, Mansinghka, Griffiths, & Tenenbaum, 2007; Schulz, Bonawitz, & Griffiths, 2007; Xu & Tenenbaum, 2007). In order to further examine whether infants behave as ideal learners, or the ways in which they depart from the ideal, it is important to first understand what behavior to expect from an ideal learner. The theoretical results presented here provide a characterization of this behavior, and we hope that they will provide inspiration for future experimental work investigating the relationship between human learners and ideal learners.