Distributional Cues to Word Boundaries: Context is Important

By Sharon Goldwater, Thomas Griffiths and Mark Johnson

Word Segmentation Again

- Most work in this area has taken transitional probabilities between syllables to be the driving force of word segmentation- that is, there is an assumption that a word is a unit statistically independent of other words
- This work takes a different approach, they want to look at a word as statistically dependent on other words surrounding it.
- So, they're going to make two Bayesian models that characterize each of these two assumptions, and test them on child directed speech to see which method is better at segmenting speech.

Two Sorts of Assumptions

- Unigram model: A word is statistically independent of other units
  - Looks for the independent units in a speech stream, and counts that as a word
  - Transitional probabilities over syllables is an example of this sort of approach
- Bigram model: A word helps predict other words
  - Pay attention to context, that is, words surrounding the information you want to segment can help give you information about the best way to segment words.

The Bayesian approach

The Bayesian approach will help us to combine expectations about the structure of language with information provided by linguistic data.

The units of representation are phonemes, the data given to the model are unsegmented utterances of child directed speech directed at 13-23-month-olds, and the hypothesis space is just the set of all possible segmentations of the data.

The learner needs to identify the posterior distribution of the segments given the observed data- choose the best hypothesis.

The Bayesian Approach

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)} \propto P(d \mid h)$$

- In our models, *d* is an input corpus of unsegmented words, each hypothesis *h* is a possible segmentation of *d*. P(*d*|*h*)=1, so the posterior probability of a segmentation is directly proportional to its prior probability.
- So, the learner prefers segmentations that correspond to their assumptions- that is their idea of what is “linguistically natural.” So we can use this to figure out which assumptions better segment child directed speech.

Brent’s model

- The goal of the learner is to identify the segmentation of the input corpus with the highest posterior probability (equivalently-find the segmentation with the highest prior probability)
  - Prior probability of a segmentation is defined in terms of four properties of the segmentation
    - The number of distinct lexical types in the segment
    - The phonetic form of each type
    - The frequency of each type
    - The probability of the particular ordering of word tokens in the segment
  - This model assumes a uniform distribution over token orderings, so the probability of any orderings of a particular set of tokens is the same as the probability of any other ordering, that is, word ordering is irrelevant, so we’ve got a unigram model

Brent’s Model

- Problems:
  - It is not clear how to replace the unigram assumption with a bigram assumption.
  - There is no known algorithm that can efficiently identify the best segmentation of the input. For any corpora of decent size, this model can only identify segmentations with high probability, and cannot guarantee finding the optimal segmentation.

So, let’s see if we can do better.

Unigram Model

- Here we will look at the model that characterizes the assumption that words are statistically independent units. And we’ll see how it does compared to Brent’s model.

Unigram Model

The model assumes that the corpus was generated by generating a sequence of words $w_1 \ldots w_N$ in order, and removing the spaces, and the ith word in the sequence, w_i is generated as follows:

1. Decide if w_i is a novel lexical item.
2. a. If so, generate a phonetic form (phonemes $x_1 \ldots x_M$) for w_i,

   b. If not, choose an existing lexical form *l* for w_i.

The probabilities associated with this are:

1. $P(w_i \text{ is novel}) = \frac{\alpha}{n+\alpha} P(w_i \text{ is not novel}) = \frac{n}{n+\alpha}$

2. a. $P(w_i = x_1 \ldots x_M \mid w_i \text{ is not novel}) = \prod_{j=1}^{M} P(x_j)$

   b. $P(w_i = l \mid w_i \text{ is not novel}) = \frac{n_l}{n}$

Where α is a parameter of the model, *n* is the number of previously generated words, and n_l is the number of times *l* has occurred.

Unigram Word Segmentation

The probability that a given sequence of phonemes is a novel word decreases as the number of words in the lexicon increases- this works to keep the number of words at a respectable level

The probability that a novel lexical item is a word is just the product of the probability of each of its phonemes (so, long strings of phonemes are dispreferred)

The probability that a given non-novel word is a particular word, is proportional to the number of times that word has appeared- that is, this model assumes we're going to see tokens of the same word type appearing quite frequently.

Results

	P	R	BP	BR	LP	LR
Brent	**67.0**	**69.4**	80.3	**84.3**	53.6	51.3
GGJ	61.9	47.6	**92.4**	62.2	**57.0**	**57.5**

Where P and R are precision and recall on word tokens, BP, and BR on boundaries, and LP and LR on lexical entries

So, the unigram learner did very well when it posited word boundaries, but it didn't posit enough of them.

Unigram Model

- Why are the results so poor?
  - Well, the assumption was that words are independent of context- but in our corpus this is clearly violated! The probability of the word "that" on this model is .024, but following "what's" it rises to .46, and following "to" it's only .0019!
  - This model tends to group together words that frequently appear together.

Why were Brent's results better?

- Brent's got some issues of his own. Brent's algorithm finds a segmentation that is far from optimal under his own model. They compared the negative log probabilities of 3 segmentations, the true segmentation, and the 2 the models found, and Brent's did worse overall than either the true segmentation or the segmentation the GGC model found.

Seg:	True	Brent	GGJ
Brent	208.2	217.0	**189.8**
GGJ	222.4	231.2	**200.6**

What's a modeler to do?

- We're going to make the data fit the assumption the unigram learner makes, and mix up the word order. When we randomize the word order, we get far better results. So, it seems the problem may just be in our assumption.

Results from the permuted corpus:

	P	R	BP	BR	LP	LR
Brent	77.0	86.1	83.7	97.7	60.8	53.0
GGJ	**94.2**	**97.1**	**95.7**	**99.8**	**86.5**	**62.2**

Bigram Word Segmentation

- So, we've seen that, under the assumption that context doesn't matter, we don't really get great results, and it may be because of the assumption we based our unigram model on. So we'll adjust our assumption.
- Context matters- previous words in a sequence help predict subsequent words.

Bigram Model

The bigram model is similar to the unigram model, but takes into account the previous word generated in the sequence, here's how it works:

1. Decide whether the pair (w_{i-1},w_i) will be a novel bigram type
2. a. If so,
   - i. Decide whether w_i will be a novel unigram type.
   - ii. a. if so, generate a phonetic form (phonemes $x_1 \ldots x_M$) for w_i

     b. if not, choose an existing lexical form l for w_i

   b. If not, choose a lexical form l for w_i from among those that have been already observed following w_{i-1}

Bigram Model

Probabilities associated with the model, where β,γ are parameters of the model, *l'* is the lexical form of w_{i-1}, $n_{l'}$ *and* $n_{(l',l)}$ are the number of occurrences of the first *i*-1 words of the unigram *l'* and the bigram (*l,l'*), *b* is the number of bigram types in the first *i*-1 words, and b_l is the number of those types whose second word is *l*:

1. $P\left((w_{i-1},w_i) \text{ is a novel bigram} \mid w_{i-1} = l'\right) = \frac{\beta}{n_{l'} + \beta}$

   $P\left((w_{i-1},w_i) \text{ is not a novel bigram} \mid w_{i-1} = l'\right) = \frac{n_{l'}}{n_{l'} + \beta}$

2. a. i. $P\left(w_i \text{ is a novel word} \mid (w_{i-1},w_i) \text{ is a novel bigram}\right) = \frac{\gamma}{b + \gamma}$

   $P\left(w_i \text{ is not a novel word} \mid (w_{i-1},w_i) \text{ is a novel bigram}\right) = \frac{b}{b + \gamma}$

   ii. a. $P\left(w_i = x_1 \ldots x_M \mid w_i \text{ is not novel}\right) = \prod_{j=1}^{M} P\left(x_j\right)$

   b. $P\left(w_i = l \mid w_i \text{ is not novel}\right) = \frac{b_l}{b}$

   b. $P\left(w_i = l \mid (w_{i-1},w_i) \text{ is not a novel bigram and } w_{i-1} = l'\right) = \frac{n_{(l',l)}}{n_{l'}}$

Bigram Model

This model is similar to the unigram model in its setup- chooses word boundaries from a sequence of words using probabilistic methods, but it differs in that it takes the previous word found into account.

the probability that a bigram is novel given that the first word in the bigram is a particular word decreases as the number of times that word has appeared increases. This limits the number of total bigrams.

Given a novel bigram, the probability that the second word in the bigram is novel decreases as the number of total bigrams increase. The idea is that some words combine more promiscuously than others into bigrams.

The probabilities for novel and non-novel words work similarly to the unigram model.

Given a non-novel bigram, the the probability the second word is proportional to how many times that word has appeared in the second place, with the first word- this favors frequently appearing bigrams.

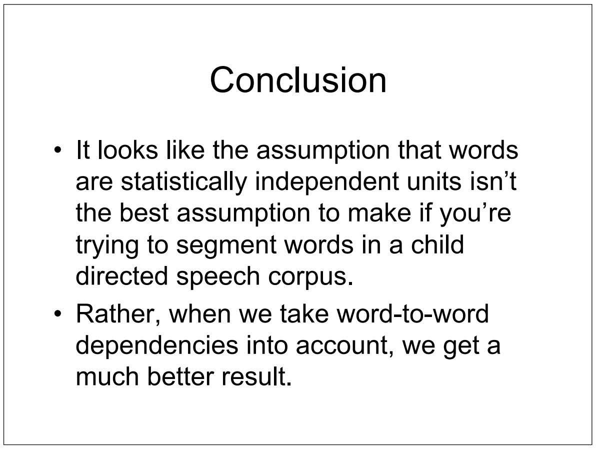
Results

- The bigram learner does quite well compared to Brent's model and the unigram model

	P	R	BP	BR	LP	LR
Brent	67.0	69.4	80.3	**84.3**	53.6	51.3
uni	61.9	47.6	**92.4**	62.2	57.0	57.5
bi	**79.4**	**74.0**	**92.4**	83.5	**67.9**	**58.9**

Simulations and Discussion

- The bigram model proposes more word boundaries than the unigram model, and is just as accurate
- Errors fall into two categories
  - Some multi-word sequences are treated as single words
  - Oversegmentation occurs at morpheme boundaries
    - 100 most frequent lexical items found by the model include plural, possessive, past tense endings, which isn't surprising given the statistical similarities between word boundaries and morpheme boundaries- some additional information might need to be used for the learner to figure out this distinction.



Conclusion

- It looks like the assumption that words are statistically independent units isn't the best assumption to make if you're trying to segment words in a child directed speech corpus.
- Rather, when we take word-to-word dependencies into account, we get a much better result.