Announcements

Be working on HW2
Be working on word segmentation review questions
Midterm on Tuesday, 5/8


Experimental evidence suggests that 8-month-old infants can track statistical information such as the transitional probability between syllables. This can help them solve the task of word segmentation.

Evidence comes from testing children in an artificial language paradigm, with very short exposure time.
Computational model: a program that simulates the mental processes occurring in a child. This requires knowing what the input and output are, and then testing the strategies that can take the given input and transform it into the desired output.

For example, in word segmentation, the input could be a sequence of syllables and the desired output is words (groups of syllables).

Input: “un der stand my po si tion”
Desired Output: “understand my position”

How good is transitional probability on real data?
Gambell & Yang (2006): Computational model goal
Real data, Psychologically plausible learning algorithm

Realistic data is important to use since the experimental study of Saffran, Aslin, & Newport (1996) used artificial language data, and it’s not clear how well the results they found will map to real language.

A psychologically plausible learning algorithm is important since we want to make sure whatever strategy the model uses is something a child could use, too. (Transitional probability would probably work, since Saffran, Aslin, & Newport (1996) showed that infants can track this kind of information in the artificial language.)

How do we measure word segmentation performance?
Perfect word segmentation:
identify all the words in the speech stream (recall)
only identify syllables groups that are actually words (precision)

ɑbɪɡbædɔlf
    ᵁ
ɑ bɪɡ bæd ɔlф
 the big bad wolf
How do we measure word segmentation performance?

Perfect word segmentation:
identify all the words in the speech stream (recall)
only identify syllables groups that are actually words (precision)

Perfect word segmentation:
identify all the words in the speech stream (recall)
only identify syllables groups that are actually words (precision)

Recall calculation:
\[
\frac{\text{# of real words found}}{\text{# of actual words}}
\]
Identified 4 real words: the, big, bad, wolf
Should have identified 4 words: the, big, bad, wolf
Recall Score: 4 words found/4 should have found = 1.0

Recall calculation:
\[
\frac{\text{# of real words found}}{\text{# of actual words}}
\]
Identified 4 real words: the, big, bad, wolf
Should have identified 4 words: the, big, bad, wolf
Recall Score: 4 words found/4 should have found = 1.0

Recall calculation:
\[
\frac{\text{# of real words found}}{\text{# of words guessed}}
\]
Identified 4 real words: the, big, bad, wolf
Identified 4 words total: the, big, bad, wolf
Precision Score: 4 real words found/4 words found = 1.0

Recall calculation:
\[
\frac{\text{# of real words found}}{\text{# of words guessed}}
\]
Identified 4 real words: the, big, bad, wolf
Identified 4 words total: the, big, bad, wolf
Precision Score: 4 real words found/4 words found = 1.0

Recall calculation:
\[
\frac{\text{# of real words found}}{\text{# of words guessed}}
\]
Identified 4 real words: the, big, bad, wolf
Identified 4 words total: the, big, bad, wolf
Precision Score: 4 real words found/4 words found = 1.0

Error

\[
\text{the big bad wolf}
\]

Error

\[
\text{the big bad wolf}
\]
How do we measure word segmentation performance?

Perfect word segmentation:
- identify all the words in the speech stream (recall)
- only identify syllables groups that are actually words (precision)

\[
\text{Precision} = \frac{\text{Identified real words}}{\text{Identified words total}}
\]

Example:

\[
\text{Identified 2 real words: bad, wolf}
\]

\[
\text{Identified 3 words total: thebig, bad, wolf}
\]

\[
\text{Precision Score: } \frac{2}{3} = 0.666...
\]

How do we measure word segmentation performance?

Perfect word segmentation:
- identify all the words in the speech stream (recall)
- only identify syllables groups that are actually words (precision)

Want good scores on both of these measures in order to be sure that word segmentation is really successful.

Where does the realistic data come from?

CHILDES
Child Language Data Exchange System
http://childes.psy.cmu.edu/

Large collection of child-directed speech data (usually parents interacting with their children) transcribed by researchers. Used to see what children’s input is actually like.

Where does the realistic data come from?

Looked at Brown corpus files in CHILDES (226,178 words made up of 263,660 syllables).

Converted the transcriptions to pronunciations using a pronunciation dictionary called the CMU Pronouncing Dictionary.

http://www.speech.cs.cmu.edu/cgi-bin/cmudict

The CMU Pronouncing Dictionary
Where does the realistic data come from?

Converting transcriptions to pronunciations

- Look up words or a sentence (v. 0.7a)
  
  Show Lexical Stress

Gambell and Yang (2006) tried to see if a model learning from transitional probabilities between syllables could correctly segment words from realistic data.

*There is a word boundary AB and CD if TrProb(A --> B) > TrProb(B --> C) < TrProb(C --> D).*

Transitional probability minimum

Segmenting Realistic Data

Gambell and Yang (2006) tried to see if a model learning from transitional probabilities between syllables could correctly segment words from realistic data.

Desired word segmentation

```
the big bad wolf
DH AH0 B IH1 G B AE1 D W UH1 L F
```

Modeling Results for Transitional Probability

Precision: 41.6%
Recall: 23.3%

A learner relying only on transitional probability does not reliably segment words such as those in child-directed English.

About 60% of the words posited by the transitional probability learner are not actually words (41.6% precision) and almost 80% of the actual words are not extracted (23.3% recall).
Why such poor performance?

“We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason... a sequence of monosyllabic words requires a word boundary after each syllable; a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those surrounding the sequences]...” - Gambell & Yang (2006)
Why such poor performance?

“We were surprised by the low level of performance. Upon close examination of the learning data, however, it is not difficult to understand the reason…a sequence of monosyllabic words requires a word boundary after each syllable; a [transitional probability] learner, on the other hand, will only place a word boundary between two sequences of syllables for which the [transitional probabilities] within [those sequences] are higher than [those surrounding the sequences]...” - Gambell & Yang (2006)

learner posits one word boundary at minimum TriProb

\[ \begin{align*}
\alpha & \quad \beta & \quad \gamma & \quad \omega & \\
0.6 & 0.3 & 0.7 & \\
\end{align*} \]

\[ 0.6 > 0.3, \quad 0.3 < 0.7 \]

Why such poor performance?

…but nowhere else

\[ \begin{align*}
\alpha & \quad \beta \quad \gamma & \quad \omega & \\
0.6 & 0.3 & 0.7 & \\
\end{align*} \]

\[ 0.6 > 0.3, \quad 0.3 < 0.7 \]

Precision for this sequence: 0 words correct out of 2 found
Recall: 0 words correct out of 4 that should have been found
Why such poor performance?

“More specifically, a monosyllabic word is followed by another monosyllabic word 85% of the time. As long as this is the case, [a transitional probability learner] cannot work.” - Gambell & Yang (2006)

Additional Learning Bias

Gambell & Yang (2006) idea
Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Hypothesis: Unique Stress Constraint (USC)
Children think a word can bear at most one primary stress.

Learner gains knowledge:
These must be separate words

Get these boundaries because stressed (strong) syllables are next to each other.
Additional Learning Bias

Gambell & Yang (2006) idea
Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Hypothesis: Unique Stress Constraint (USC)
Children think a word can bear at most one primary stress.

Additional Learning Bias

Gambell & Yang (2006) idea
Children are sensitive to the properties of their native language like stress patterns very early on. Maybe they can use those sensitivities to help them solve the word segmentation problem.

Hypothesis: Unique Stress Constraint (USC)
Children think a word can bear at most one primary stress.

USC + Transitional Probabilities

Precision: 73.5%
Recall: 71.2%

A learner relying on transitional probability but who also has knowledge of the Unique Stress Constraint does a much better job at segmenting words such as those in child-directed English.

Only about 25% of the words posited by the transitional probability learner are not actually words (73.5% precision) and about 30% of the actual words are not extracted (71.2% recall).

Another Strategy

Using words you recognize to help you figure out words you don’t recognize
Another Strategy: Algebraic Learning

Algebraic Learning (Gambell & Yang (2003))

Subtraction process of figuring out unknown words.

“Look, honey - it’s a big goblin!”

big

big = big (familiar word)

bíggabílm

big

biggáblím

= (new word)

Evidence of Algebraic Learning in Children

“Behave yourself!”

“I was have!”

(beh-have = be + have)

“Was there an adult there?”

“No, there were two dults.”

(a-dult = a + dult)

“Did she have the hicups?”

“Yeah, she was hiccing-up.”

(hicc-up = hicc + up)

Experimental Evidence of Algebraic Learning

Experimental studies show young infants can use familiar words to segment novel words from their language

- Bortfeld, Morgan, Golinkoff, & Rathbun 2005: 6-month-old English infants use their own name or Mommy/Mama
- Halle, Durand, Bartsies, & de Boysson 2008: 11-month-old French infants use French articles like le, les, and la
- Shi, Werker, & Cutler 2006: 11-month-old English infants use English articles like her, its, and the
- Shi, Cutler, Werker, & Cruickshank 2006: 11-month-old English infants (but not 8-month-old English infants) use the English article the

Using Algebraic Learning + USC

WeakSyl | StrongSyl | StrongSyl | StrongSyl
---------|-----------|-----------|-----------
the       | big        | bad       | wolf
ðə       | bɪg        | bæd       | wɔlф
```
"the big bad wolf"
```
Using Algebraic Learning + USC

Familiar word: “the” (algebraic learning)

USC says these must be separate words

A learner relying on algebraic learning and who also has knowledge of the Unique Stress Constraint does a really great job at segmenting words such as those in child-directed English - even better than one relying on the transitional probability between syllables.

Only about 5% of the words posited by the transitional probability learner are not actually words (95.9% precision) and about 7% of the actual words are not extracted (93.4% recall).
Gambell & Yang 2006 Summary
Learning from transitional probabilities alone doesn’t work so well on realistic data, even though experimental research suggests that infants are capable of tracking and learning from this information.

Models of children that have additional knowledge about the stress patterns of words seem to have a much better chance of succeeding at word segmentation if they learn via transitional probabilities.

However, models of children that use algebraic learning and have additional knowledge about the stress patterns of words perform even better at word segmentation than any of the models learning from the transitional probability between syllables.

Gambell & Yang 2006 Critiques
Do children have access to the Unique Stress Constraint (USC)?
- Children definitely use TPs & Algebraic Learning

Does dictionary stress really match actual stress patterns?

Gambell & Yang: the big bad wolf
Typical speech: the big bad wolf

It’s unclear how well this algorithm works with real stress patterns...

Pearl, Goldwater, & Steyvers 2011
What if children are capable of tracking more sophisticated distributional information (that is, they’re not just restricted to transitional probability minima)? In that case, how well do they do on realistic data, if all they’re using is statistical learning (no stress information)?

Pearl, Goldwater, & Steyvers 2011
What if children can use Bayesian inference? Human cognitive behavior is consistent with this kind of reasoning. (Tenenbaum & Griffiths 2001, Griffiths & Tenenbaum 2005, Xu & Tenenbaum 2007)

Bayesian inference is a sophisticated kind of probabilistic reasoning that tries to find hypotheses that
(1) are consistent with the observed data
(2) conform to a child’s prior expectations
Bayesian inference for word segmentation
What kind of hypotheses might a child have for word segmentation?

Observed data: "to the castle beyond the goblin city"

Hypothesis = sequence of vocabulary items producing this observable data

Hypothesis 1: "to the castle beyond the goblin city"
Items: to, the, castle, beyond, goblin, city

Hypothesis 2: "to the castle beyond the goblin city"
Items: to, the, castle, beyond, goblin, city

Note: the is used twice

Bayesian model: Pearl et al. 2010
Learner expectations about word segmentation:
(1) Words tend to be shorter rather than longer
(2) Vocabulary tends to be small rather than large

Bayesian model: Pearl et al. 2011
Learner expectations about word segmentation:
(1) Words tend to be shorter rather than longer
(2) Vocabulary tends to be small rather than large

Comparing hypotheses - which is most likely?
Hypothesis 1: longer words, but fewer words
How long are words? Between 4 and 9 letters
How large is the vocabulary? 5 words

Hypothesis 2: shorter words, but more words
How long are words? Between 3 and 6 letters
How large is the vocabulary? 6 words

A Bayesian learner makes a decision based on how important each of its expectations is (in this case, it’s a balance of the two constraints: fewer words vs. shorter words).
Bayesian model: Pearl et al. 2010

Comparing hypotheses - which is most likely?

Hypothesis 1: longer words, but fewer words
How long are words? Between 4 and 9 letters
How large is the vocabulary? 5 words

Hypothesis 2: shorter words, but more words
How long are words? Between 3 and 6 letters
How large is the vocabulary? 6 words

There will be some probability the Bayesian learner assigns to each hypothesis. The most probable hypothesis will be the one the learner chooses.

Bayesian model: Pearl et al. 2010

Comparing hypotheses - which is most likely?

Hypothesis 1: longer words, but fewer words
How long are words? Between 4 and 9 letters
How large is the vocabulary? 5 words

Hypothesis 2: shorter words, but more words
How long are words? Between 3 and 6 letters
How large is the vocabulary? 6 words

There will be some probability the Bayesian learner assigns to each hypothesis. The most probable hypothesis will be the one the learner chooses.

Realistic Bayesian Learners: Pearl et al. 2011

Pearl et al. 2011 tested their Bayesian learners on realistic data: 9790 utterances of child-directed speech from the Bernstein-Ratner corpus in CHILDES. (Average utterance length: 3.4 words)

Best performance by a Bayesian learner:

Precision: 72%
Recall: 74%

This is much better than what we found for a learner that hypothesizes a word boundary at a transitional probability minimum (41.6% precision, 23.3% recall). Statistical learning by itself isn’t always so bad after all!
Model Comparison

So which model performs better?
- Rate only based on Recall and Precision scores?

Any model makes assumptions which should be included in analysis!

Gambell & Yang:
- Syllables, TPs, USC, Lexicon, Algebraic Learning, Dictionary Stress
- Less processing power

Pearl et al:
- Phonemes, TPs, Lexicon, Bayesian Inference, Bias for shorter/fewer words
- More processing power

Statistical Learning for Word Segmentation

Saffran et al. (1996) found that human infants are capable of tracking transitional probability between syllables and using that information to accomplish word segmentation in an artificial language.

Gambell & Yang (2006) found that this same statistical learning strategy (positing word boundaries at transitional probability minima) failed on realistic child-directed speech data.

Pearl et al. (2011) found that more sophisticated statistical learning (Bayesian inference) did much better on realistic child-directed speech data, suggesting that children may be able to use statistical learning to help them with word segmentation - even if they don't use other strategies like lexical stress.

Questions?

Use the remaining time to work on HW2 and the review questions for word segmentation.