Word segmentation
(example paper presentation)

Topics in Cognitive Modelling
Jan. 29, 2013

Sharon Goldwater
School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

Word segmentation
• One of the first problems infants must solve when learning language: where are the word boundaries?

• May be similar to segmenting other kinds of sequences (e.g., actions) and visual scenes.

Cues to word segmentation
• Infants make use of many different cues.
  • Phonotactics (which sound sequences are legal?)
    • sound vs. ndsequence
  • Stress patterns
    • English usually stresses 1st syllable, French always the last.
    • Etc.
• But specifics differ between languages, presenting a chicken-and-egg problem:

  Learn typical stress  Learn some words

Statistical word segmentation
• In any language, words create statistical regularities in the sequences of sounds in the language.
• Experimental work (Saffran et al. 1996) focuses on transitional probabilities between syllables.
  • Idea: $P(syl_i | syl_{i-1})$ is often lower at word boundaries.
  • "pretty baby": $P(by|ba) > P(ba|ty)$

Experimental evidence
• Infants (and adults) can learn word-like units in a nonsense language based on statistics alone.
  
  Lexicon: pabiku dibudo golatu daropi
  Training stimulus: ...
pabiku daropigolatubudodoripitubudogolatu
  pabikudaropigolatupabikutudogolatupabikud
  arotibudo...

  • After training, test: Can subjects distinguish words (pabiku) vs. part-words (kudaro)?

Questions raised
• What statistical information is actually being used?
  • Transitional probabilities or something else?
• Does the mind represent and compute with these statistics directly, or is it doing something else?
• Are listeners finding boundaries or finding words?
• What happens with more realistic linguistic input?
Today's models

- **PARSER** (Perruchet and Vinter, 1998)
  - Humans are not tracking boundary statistics; segmentation results from general properties of attention, perception, and memory.
- **Bayesian model** (Goldwater, Griffiths, and Johnson, 2007)
  - What kind of information would be useful for segmenting from more realistic input? What would result, if humans use the information optimally?
  - Both models focus on words, not boundaries.
  - Both use little or no domain-specific information.

PARSER

- Main thesis: No special mechanism is needed for word segmentation; it results from interaction of perception and internal representation.
  - Initially, input is perceived and chunked randomly into units.
  - Units are encoded in memory.
  - Memory decays rapidly.
  - Uncommon units disappear, common units are reinforced.
  - Units in memory influence perception and encoding of new input (input is segmented into existing units).

PARSER

- Main thesis: No special mechanism is needed for word segmentation; it results from interaction of perception and internal representation.
  - Units are stored in “Percept Shaper” (PS): set of units and their weights (≈ strength in memory).
  - PS starts with set of primitive units (syllables), weight = 1.
  - Units with weight 1 or more can “shape perception”

Representation

- Units are stored in “Percept Shaper” (PS): set of units and their weights (≈ strength in memory).
  - PS starts with set of primitive units (syllables), weight = 1.
  - Units with weight 1 or more can “shape perception”

Over time

- Frequent subsequences reinforce units in PS
- Infrequent subsequences disappear from PS.
- Words are more frequent, so will dominate.
Experiments

- Experiment 1, 2, and 4 show:
  - Using same input stimulus as Saffran et al. experiments, PARSER learns the lexicon.
  - Can also do so while simulating lowered attention (like humans).
  - Predicts that different word lengths should present no problem (since then, this has been verified in humans).

Issues

- Would it work on realistic input data?
  - Discussion suggests not (unless modified).
- Experiment 3: simulating infant study.
  - Uses 4 lexical items instead of 6.
  - Performance actually goes down: pairs of words are found more commonly (pabikutibudo), interfere with single words.
  - Fixes this by changing model parameters – “infants have more limited memory” – but this is done post-hoc.
  - Still predicts that adults would have more trouble with 4 lexical items than 6.

Summary

- PARSER provides a mechanistic account of word segmentation based on general principles of attention, perception, and memory.
- No explicit tracking of statistics is needed.
- Works on experimental stimuli but might need modifications for realistic language.
- Probably would work in other domains.
- Smaller vocabulary is harder than larger one??
- Lots of parameters – how sensitive to these?

Bayesian model

- An ideal observer analysis: what words would be learned if statistical information is used optimally, and the learner assumes:
  a) Words are defined as statistically independent units in the input (i.e., randomly ordered, as in experimental stimuli)?
  b) Words are defined as units that help predict other units?
- Is (a) sufficient? I.e., what kind of prior does the learner need?

Two kinds of models

- Unigram model: words are independent.
- Bigram model: words depend on other words.
Bayesian segmentation

• Data: unsegmented corpus (transcriptions).
• Hypotheses: sequences of word tokens.

\[
P(h|d) \propto P(d|h) P(h)
\]

Posterior \hspace{1cm} Likelihood \hspace{1cm} Prior

1 if concatenating words forms corpus, = 0 otherwise. Encodes assumptions of learner.

• Optimal solution is the segmentation with highest prior probability.

Bayesian model

Assumes word \( w_i \) is generated as follows:

1. Is \( w_i \) a novel lexical item?

\[
P(yes) = \frac{\alpha}{n+\alpha}
\]

Fewer word types = Higher probability

\[
P(no) = \frac{n}{n+\alpha}
\]

Bayesian model

Assume word \( w_i \) is generated as follows:

2. If novel, generate phonemic form \( x_1 \ldots x_m \):

\[
P(w_i = x_1 \ldots x_m) = \prod_{i=1}^{m} P(x_i)
\]

Shorter words = Higher probability

If not, choose lexical identity of \( w_i \) from previously occurring words:

\[
P(w_i = w) = \frac{n_w}{n}
\]

Power law = Higher probability

Experiments

• Input: phonemically transcribed infant-directed speech.

Unigram model:

you want to see the book
look there a boy with his hat
and a doggie
you want to look at this
look at this
have a drink
okay now
what is this
what is this
look can you take it out
...

Bigram model:

you want to see the book
look there a boy with his hat
and a doggie
you want to look at this
look at this
have a drink
okay now
what is this
what is this
look can you take it out
...

• Quantitative comparison verifies bigram is better.

Example output

• Optimal segmentation is found using a standard optimization algorithm (Gibbs sampling).
• Compare to bigram model (developed using similar maths).

Data:

lookatthedoggie
seethedoggie
shelookssofriendly
...

Hypotheses:

lookatthedoggie
seethedoggie
shelookssofriendly
...

look at the doggie
see the doggie
she looks so friendly
...

i like pizza
what about you
What's wrong with unigrams?

- Model assumes (false) that words have the same probability regardless of context.
  \[ P(\text{that}) = .024 \quad P(\text{that|whats}) = .46 \quad P(\text{that|to}) = .0019 \]

- Positing amalgams allows the model to capture word-to-word dependencies.
- Paper argues that this is a general property of unigram models, not specific to this one.

Summary

- Good segmentations of naturalistic data can be found using fairly weak/domain-general prior assumptions.
- Utterances are composed of discrete units (words).
- Units tend to be short.
- Some units occur frequently, most do not.
- Units tend to come in predictable patterns.
- More sophisticated use of information works better.
  - But still possible that simpler learner is enough to start learning other language-specific cues.

Issues

- No direct comparison to humans.
  - Is there evidence that human performance is consistent with Bayesian predictions? [Later paper suggests: yes]
  - Are humans able to use bigram information?
- Algorithm iterates multiple times over the entire corpus – are more cognitively plausible algorithms possible?

Conclusion

- Models have different emphasis:
  - PARSER: mechanistic explanation; experimental data.
  - Bayesian model: ideal observer analysis; naturalistic data.
- But some similar ideas/conclusions:
  - Segmentation is about building a lexicon, not finding boundaries.
  - Built on domain-general principles.
- Open questions:
  - Relationship to adult speech processing?
  - Multiple cues?

References


Bayesian learning

- Want to find an explanatory linguistic hypothesis that
  - accounts for the observed data.
  - conforms to prior expectations.

\[ P(h | d) \propto P(d | h)P(h) \]

Two kinds of models

- Unigram model: words are independent.
  - Generate a sentence by generating each word independently.

Two kinds of models

- Bigram model: words predict other words.
  - Generate a sentence by generating each word, conditioned on the previous word.