The Journey to Word Segmentation

Constantine Lignos

University of Pennsylvania
Department of Computer and Information Science
Institute for Research in Cognitive Science

BUCLD 36
11/5/2011

I. Infant word segmentation

What’s the real task?

- Approximation: unsegmented sounds as input, goal is to insert word boundaries
- But word segmentation is not an end in itself: provides useful units (Peters, 1983) for learning
  - Lexicon, Morphosyntax, Phonology
- We focus on word segmentation as a process that must occur to allow other levels of representation to form

The larger picture

- Experimental insights
  - Which cues (e.g., transitional probabilities, stress) are used and when?
  - How well do infants perform at varying ages?
- Naturalistic behaviors
  - What errors do children make?
  - How does competency develop over time?
- Modeling
  - How can cues be used?
  - What kind of behavior would be predicted during development?

Modeling development?

- Focus: models using TPs, a lexicon, and Bayesian inference (Goldwater et al., 2009; Johnson and Goldwater, 2009)
  - Integrates progress in unsupervised learning (Teh, 2006)
  - High performance. Developmental impact?
  - Transition to less idealized learner model from these is non-trivial (Pearl et al., 2009, 2010)
- Maybe there's a simpler way (Gambell and Yang, 2006; Yang, 2004)
- Modeling of experimental performance (Frank et al., 2010)
  - Matches adult experimental performance with models
  - Can we extend to naturalistic development?

Our modeling goal:

Build the simplest model that:

- Aligns with infants' capabilities
- Replicates infants' behavior in a principled fashion
- Performs reasonably at the task

Model in a nutshell:

1. Use utterance boundaries to help find initial words.
2. Bootstrap from known words.
3. Reward the words that appear to lead to better segmentations, penalize the ones that lead us astray.

II. An algorithm for segmentation

How do infants segment speech?

- Possible strategy: identification of words in isolation (Peters, 1983; Pinker et al., 1984)
  - Unlikely to be sufficient (Aslin et al., 1996), but probably helpful (Brent and Siskind, 2001)
- Attending to multiple cues in the input, most popularly:
  - Bootstrapping from known words (Bortfeld et al., 2005; Dahan and Brent, 1999)
  - Dominant stress patterns (Jusczyk et al, 1999)
  - Transitional probabilities (Saffran et al, 1996 et seq.)
- More easily identify novel words at beginning and ends of utterances at 8 months (Seidl & Johnson, 2006)

Modeling assumptions

- In modeling, assumptions needed to help isolate phenomena at a particular level
  - With goal to relax assumptions as more is known about solution
- Learner is given syllabified input
  - As with artificial language learning (Saffran et al, 1996 et seq.)
  - Learner treats syllables holistically (Jusczyk and Derrah, 1987)
- Able to map acoustic signal to strong/weak stress on syllables (Johnson & Jusczyk, 2001)

Overview of our algorithm

- Segmenter has a lexicon of potential words it builds over time
  - Starts empty, words are added based on segmentation of each utterance
  - Each word has a score
- Operates online
  - Processes one utterance at a time
  - Cannot remember previous utterances or how it segmented them, only lexicon
- Operates left-to-right in each utterance to insert word boundaries between syllables

Subtractive Segmentation

- Use words in the lexicon to break up the utterance
- Increase word's score when it's used
- Add new words to lexicon

Lexicon:
Mommy's
tea
...

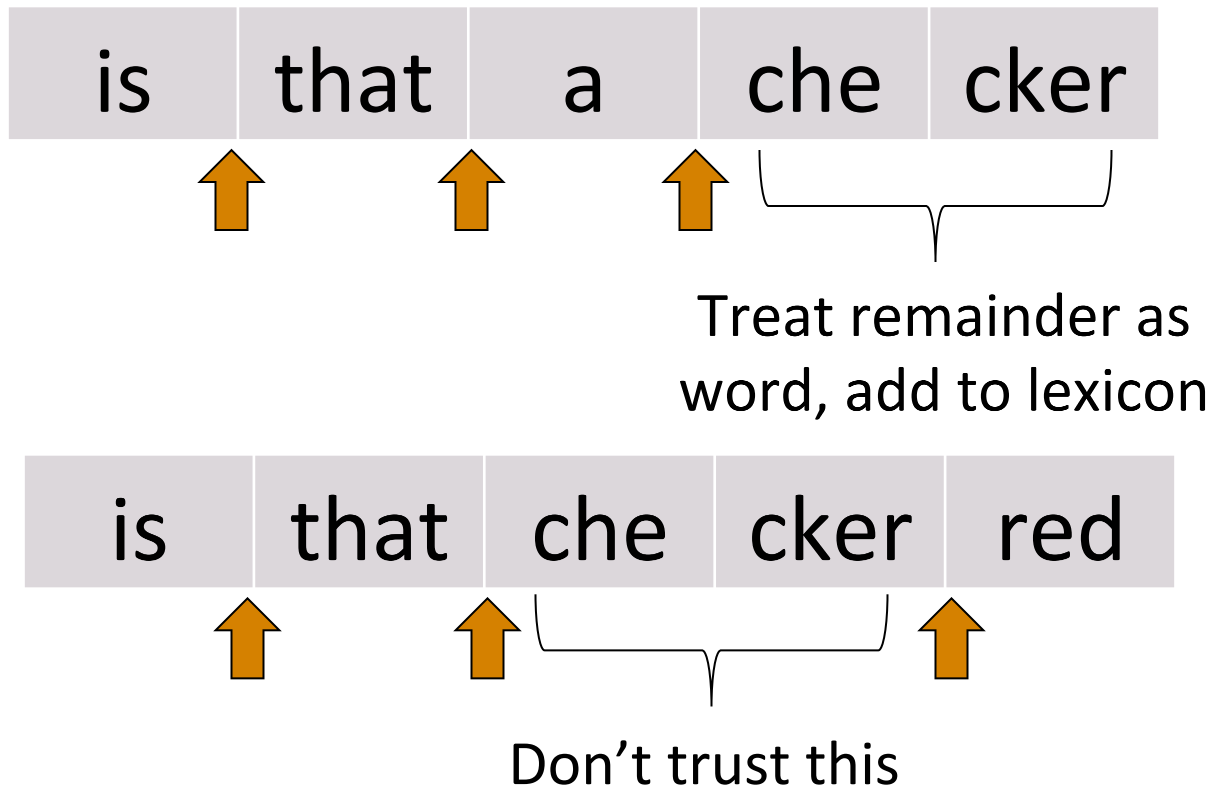
Mo	mmy's	tea

Treat remainder as word, add to lexicon

Trust

- Add new words to lexicon based on whether we *trust* them (touch an utterance boundary)

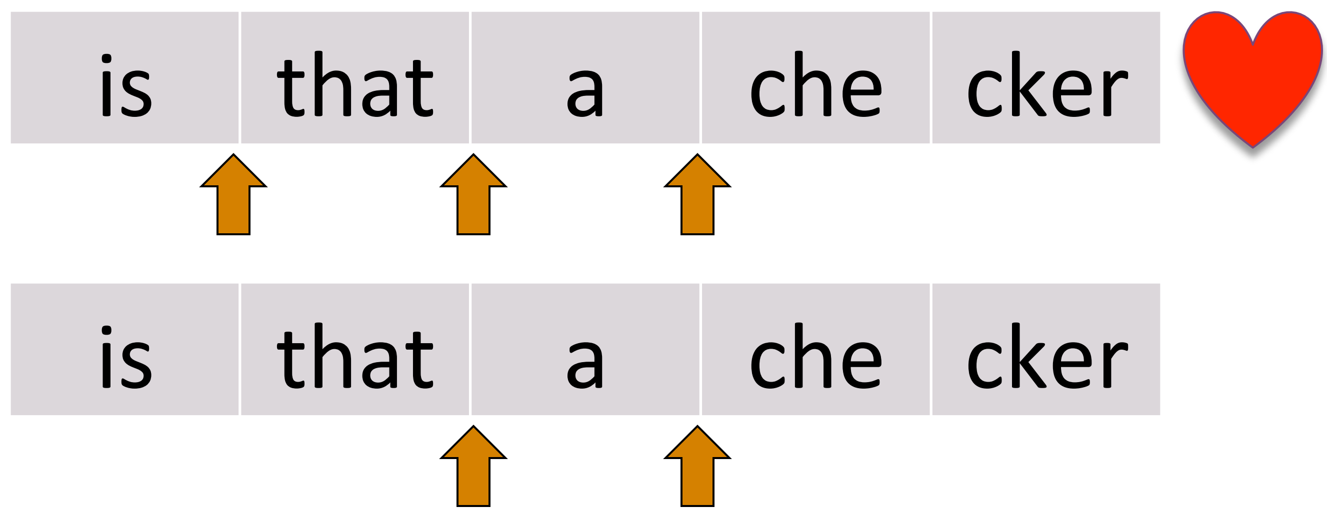
Lexicon:
a
is
that
red
checker
…



Multiple hypotheses

- For multiple possible subtractions two options:
  - Greedy approach (Lignos and Yang, 2010)
  - Pursue two hypotheses (*beam search*)
- Two hypotheses allow for *penalization*: reduce score of word that started losing hypothesis

Lexicon:
a
is
isthat
that
...



Scoring hypotheses

- Prefer the hypothesis that uses the higher-scoring words
  - Winner is rewarded, word scores will go up: "rich get richer"
- Geometric mean of scores of words used:

$$\arg\max_H \Big(\prod_{w_i \in H} score(w_i) \Big)^{\frac{1}{n}}$$

  - Useful for compound splitting (Koehn and Knight, 2003; Lignos, 2010)
  - Doesn't penalize for having more words
  - Assume new words occur just once (hapax assumption)
  - Wide range of scoring/smoothing functions is possible

Predictions

- Default assumption of utterance = word → infants will start with oversized units and words in isolation
- Rich-get-richer scoring → As the learner is exposed to more data, learner will tend to use high-frequency elements
- Penalization → Use of collocations will decrease with time

III. Results and analysis

Our evaluation corpus

- Constructed from Brown (1973) subset of CHILDES English (Adam, Eve, Sarah), ~60k utterances
- Pronunciations and stress for each word from CMUDICT, algorithmically syllabified
- Stress modified to better reflect natural speech
  - No adjacent primary stresses (Liberman & Prince, 1977; Selkirk 1984)
- Sample input:

B.IH0.G|D.R.AH1.M
HH.AO1.R.S
HH.UW0|IH0.Z|DH.AE1.T

Evaluation

- F-score and A’ calculated over syllable boundaries
  - F-score: balance of precision (how often a boundary is correct) and recall (how many correct boundaries were found)
  - A’: Balance of hit rate and false alarm rate
- Evaluated segmenter in three forms:
  - Subtractive segmentation
  - Subtractive segmentation with *trust*, only adding words to the lexicon if they touch an utterance boundary
  - Subtractive segmentation with trust and *multiple hypotheses*, considering two hypothetical segmentations and penalizing the loser

Performance

Method	F-score	Error Reduction
Baseline: syllable = word	.8991	
Subtractive Segmentation	.9166	-17.34%
+Trust	.9377	-25.30%
+Multiple Hypotheses	.9392	-2.41%

- Syllable baseline: 82% of syllable boundaries are word boundaries!
- Using utterance boundaries trust improves performance
- Multiple hypotheses help only a small amount but lead to the right behavior…

Performance

Method	Hit Rate	FA Rate	A’
Baseline: syllable = word	100%	100%	0.0
Subtractive Segmentation	98.7%	74.1%	0.150
+Trust	96.0%	38.8%	0.191
+Multiple Hypotheses	95.4%	34.4%	0.196

- Trust and multiple hypotheses work to reduce FA rate
- Evaluating A’ and F-score with imperfect memory/syllable identification yield similar results

Errors over time and predictions

- Early:
  - "Big drum" as "Bigdrum" [First utterance in corpus]
    - Learner's lexicon is empty, thus no segmentation occurs. Predicts early-stage one-word/one-collocation utterances
  - "How many trucks?" as "Howmany trucks?"
    - Frequent function words collocations are treated as single words, resulting in a lack of productivity (Brown, 1973)
- Late:
  - "Want me to take it away from you" as
"Want me to take it a way from you"
    - Function word *a* mistakenly segmented off away, predicts attested behave/be have (Peters, 1983) and tulips/two lips (Yang, 2006) errors

Most frequent error tokens

- Divided into early (first 10k utterances) and late (last 10k utterances) stages of learning
- Coded most frequent incorrect words in output as:
  - Function: Overuse of function word (a way→ a way)
  - Function collocation: Two function words(that's a → that'sa)
  - Content collocation: Content and content/function word (a ball → aball)
  - Other
- Distribution changes across time (Chi-squared $p < .0001$)

Time	Function	Func. Colloc.	Cont. Colloc.	Other
Early	340	350	468	107
Late	675	1050	17	21

Most frequent error tokens

(Converted to orthography for easier reading)

Early		Late	
Item	**Frequency**	**Item**	**Frequency**
oh	209	a	441
a	184	oh	194
thats-a	101	some	101
thank-you	45	any	77
some	39	all	67
all	31	every	60
any	31	in	57
it's-a	30	on	53
why-don't	28	tee	41
don't-know	26	be	40
at-the	24	more	39
put-the	24	huh	37
take-the	24	ta	28
where's-the	24	an	27

How do we correct these errors?

- What's the force that prevents the learner from oversegmenting?
  - Always statistical answers (length priors, hierarchical processes, etc.)
  - Alternative: feedback from other levels (morphosyntax, meanings of words)

What can the learner do with its lexicon?

- Identify stress pattern in the language
  - Multi-syllable words in the learner's acquired lexicon have stress-initial rate of 86.3%
  - Taking advantage of this bias in learning, the learner can further reduce errors by 27.80%
- Use the lexicon to differentiate in-word and between-word transitional probabilities
  - Turns using TPs into semi-supervised problem
- Other information about words can be learned from the lexicon
  - Morphology, phonotactics

Conclusions

- Simple reward-based model can lead to the changes in unit size seen in children
- Language-universal approach can efficiently build the lexicon and allow-language specific segmentation strategies to form
- Hypothesis selection can allow for multiple cues to be integrated in the future
- Further investigations:
  - Need gold standard of segmentation errors to compare against
  - Testing in multiple languages, but segmentation standard is harder (clitics, etc.)
  - Broad evaluation of other systems' performance across time

Thanks to:
Charles Yang
Mitch Marcus
NSF IGERT #50504487

Constantine Lignos
lignos@cis.upenn.edu
http://www.seas.upenn.edu/~lignos
Code/data:
https://github.com/ConstantineLignos/LanguageLearning