Unsupervised phonemic Chinese word segmentation using Adaptor Grammars

Mark Johnson\textsuperscript{1} and Katherine Demuth\textsuperscript{2}

\textsuperscript{1}Department of Computing

\textsuperscript{2}Department of Linguistics

Macquarie University
Sydney
Australia

COLING, August 2010
Talk outline

- Adaptor grammars are a framework for expressing non-parametric hierarchical Bayesian models.
- They can be used to define unsupervised word segmentation models that learn:
  - word-internal structure: how words are composed out of syllables, and
  - inter-word structure: collocational dependencies between words.
- Adaptor Grammars provide state-of-the-art unsupervised segmentation results for English: will they work for Mandarin Chinese?
  - can Adaptor Grammars model lexical tone?
  - does modelling lexical tone improve word segmentation accuracy?
Why study computational models of language acquisition?

- **Hypothesis:** acquisition, comprehension and production are *computational processes*
  - computational models need not be just *descriptions* of language acquisition
  - a computational model should be able to *learn a language*
- **Characterising computational models of acquisition:**
  - the input (information available to learner)
  - the output (generalisations learner can make)
  - the algorithm used to map input to output
- **Bayesian inference algorithms** are optimal learners
  - computational generalisation of “ideal observer” theory
- **Computational models** let us study the effect of
  - changing the information in the input, and
  - altering the kinds of generalisations the learner can acquire in ways that would be impractical or unethical with real children
- May be useful for designing experiments or therapeutic interventions
Unsupervised word segmentation

- Input: phoneme sequences with *sentence boundaries* (Brent)
  - English data produced from orthographic transcripts of child-directed speech by *looking up each word in a pronouncing dictionary*

- Task: identify *word boundaries*, and hence words, in unsegmented utterance (in ARPABET)

\[ y \_\_u\_w \_a\_n\_t\_t\_u\_s\_i\_D\_6\_b\_U\_k \]

- Useful cues for word segmentation:
  - Phonotactics and syllable structure (Fleck)
  - Inter-word dependencies (Goldwater)
CFG models of word segmentation

Words $\rightarrow$ Word
Words $\rightarrow$ Word Words
Word $\rightarrow$ Phons
Phons $\rightarrow$ Phon
Phons $\rightarrow$ Phon Phons
Phon $\rightarrow$ $a \mid b \mid \ldots$

- CFG trees can \textit{describe} segmentation, but
- PCFGs \textit{can't distinguish} good segmentations from bad ones
  - PCFG rules are \textit{too small} a unit of generalisation
  - need to learn e.g., probability that \textit{bUk} is a Word
Towards non-parametric grammars

Words $\rightarrow$ Word
Words $\rightarrow$ Word Words
Word $\rightarrow$ \textit{all possible phoneme sequences}

- Learn probability Word $\rightarrow$ b U k
- But \textit{infinitely many possible Word expansions} \\
  $\Rightarrow$ this grammar is \textit{not a PCFG}

- Given \textit{fixed training data}, only finitely many useful rules \\
  $\Rightarrow$ use data to choose Word rules as well as their probabilities

- Non-parametric models: parameters of model depend on data
From PCFGs to Adaptor grammars

- An adaptor grammar is a PCFG where a subset of the nonterminals are *adapted*

**Adaptor grammar generative process:**
- to expand an *unadapted nonterminal* $B$: (just as in PCFG)
  - select a rule $B \rightarrow \beta \in R$ with prob. $\theta_{B \rightarrow \beta}$, and
    recursively expand nonterminals in $\beta$
- to expand an *adapted nonterminal* $B$:
  - select a *previously generated subtree* $T_B$
    with prob. $\propto$ number of times $T_B$ was generated, or
  - select a rule $B \rightarrow \beta \in R$ with prob. $\propto \alpha_B \theta_{B \rightarrow \beta}$, and
    recursively expand nonterminals in $\beta$
Unigram adaptor grammar (Brent)

- **Word** nonterminal is adapted

⇒ To generate a **Word**:
  - select a previously generated **Word** subtree with prob. \( \propto \) number of times it has been generated
  - expand using **Word** → Phons rule with prob. \( \propto \alpha_{\text{Word}} \) and recursively expand Phons

```
Words → Word
Words → Word Words
Word → Phons
Phons → Phon
Phons → Phon Phons

Words
  | Word
  | Phons
  | Phon
  | Phon
  | Phon
  | U
  | k
```

```
Words
  | Word
  | Phons
  | Phon
  | Phon
  | Phon
  | Phon
  | U
  | Phon
```

```python
Words
  | Word
  | Phons
  | Phon
  | Phon
  | Phon
  | Phon
  | U
  | Phon
```

Properties of adaptor grammars

- Probability of regenerating an adapted subtree $T_B$ 
  $\propto$ number of times $T_B$ was previously generated
  - adapted subtrees are *not independent*
    - an adapted subtree can be *more probable* than the rules used to construct it
  - but they are *exchangable* $\Rightarrow$ efficient sampling algorithms
  - “rich get richer” $\Rightarrow$ Zipf power-law distributions

- Each adapted nonterminal is associated with a *Chinese Restaurant Process* or *Pitman-Yor Process*
  - CFG rules define *base distribution* of CRP or PYP

- CRP/PYP parameters (e.g., $\alpha_B$) can themselves be estimated (e.g., slice sampling)
Abbreviatory notation

Words → Word
Words → Word Words
**Word** → Phons
Phons → Phon
Phons → Phon Phons

*is abbreviated as*

Words → Word⁺
**Word** → Phon⁺
Unigram model of word segmentation

- Unigram “bag of words” model (Brent):
  - generate a *dictionary*, i.e., a set of words, where each word is a random sequence of phonemes
  - Bayesian prior prefers smaller dictionaries
  - generate each utterance by choosing each word at random from dictionary

- Brent’s unigram model as an Adaptor Grammar

\[
\text{Words} \rightarrow \text{Word}^+ \\
\text{Word} \rightarrow \text{Phon}^+
\]

- Accuracy of word segmentation learnt: *56% token f-score* (same as Brent model)
- But we can construct many more word segmentation models using AGs
Adaptor grammar learnt from Brent corpus

- **Initial grammar**

  1. $\text{Words} \rightarrow \textbf{Word} \text{ Words}$
  1. $\textbf{Word} \rightarrow \text{Phon}$
  1. $\text{Phons} \rightarrow \text{Phon} \text{ Phons}$
  1. $\text{Phon} \rightarrow D$
  1. $\text{Phon} \rightarrow A$

- **A grammar learnt from Brent corpus**

  16625 $\text{Words} \rightarrow \textbf{Word} \text{ Words}$  9791 $\text{Words} \rightarrow \textbf{Word}$
  1575 $\textbf{Word} \rightarrow \text{Phons}$
  4962 $\text{Phons} \rightarrow \text{Phon} \text{ Phons}$  1575 $\text{Phons} \rightarrow \text{Phon}$
  134 $\text{Phon} \rightarrow D$  41 $\text{Phon} \rightarrow G$
  180 $\text{Phon} \rightarrow A$  152 $\text{Phon} \rightarrow E$
  460 $\textbf{Word} \rightarrow (\text{Phons} (\text{Phon } y) (\text{Phons} (\text{Phon } u)))$
  446 $\textbf{Word} \rightarrow (\text{Phons} (\text{Phon } w) (\text{Phons} (\text{Phon } A) (\text{Phons} (\text{Phon } t))))$
  374 $\textbf{Word} \rightarrow (\text{Phons} (\text{Phon } D) (\text{Phons} (\text{Phon } 6)))$
  372 $\textbf{Word} \rightarrow (\text{Phons} (\text{Phon } \&)) (\text{Phons} (\text{Phon } n) (\text{Phons} (\text{Phon } d)))$
Undersegmentation errors with Unigram model

Words → \textbf{Word}^+ \quad \textbf{Word} → \text{Phon}^+

• Unigram word segmentation model assumes each word is generated independently
• But there are strong inter-word dependencies (collocations)
• Unigram model can only capture such dependencies by analyzing collocations as words (Goldwater 2006)
Collocations ⇒ Words

Sentence → Colloc⁺
Colloc → Word⁺
Word → Phon⁺

A Colloc(ation) consists of one or more words
Both Words and Collocs are adapted (learnt)
Significantly improves word segmentation accuracy over unigram model (74% f-score; ≈ Goldwater’s bigram model)
Collocations ⇒ Words ⇒ Syllables

- Sentence → Colloc
- Word → Syllable
- Word → Syllable Syllable Syllable
- Onset → Consonant
- Nucleus → Vowel

- Colloc → Word
- Word → Syllable Syllable
- Syllable → (Onset) Rhyme
- Rhyme → Nucleus (Coda)
- Coda → Consonant

- Rudimentary syllable model (an improved model might do better)
- With 2 Collocation levels, f-score = 84%
Distinguishing internal onsets/codas helps in English

Sentence → Colloc\(^+\)
Word → SyllableIF
Word → SyllableI Syllable SyllableF
OnsetI → Consonant\(^+\)
Nucleus → Vowel\(^+\)

\[\text{Colloc} \rightarrow \text{Word}^+\]
\[\text{Word} \rightarrow \text{SyllableI SyllableF}\]
\[\text{SyllableIF} \rightarrow (\text{OnsetI}) \text{ RhymeF}\]
\[\text{RhymeF} \rightarrow \text{Nucleus (CodaF)}\]
\[\text{CodaF} \rightarrow \text{Consonant}^+\]

- With 2 Collocation levels, not distinguishing initial/final clusters, f-score = 84%
- With 3 Collocation levels, distinguishing initial/final clusters, f-score = 87%
Collocations² ⇒ Words ⇒ Syllables
Summary so far

- Word segmentation accuracy depends on the kinds of generalisations learnt.

<table>
<thead>
<tr>
<th>Generalization</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>words as units (unigram)</td>
<td>56%</td>
</tr>
<tr>
<td>+ associations between words (collocations)</td>
<td>79%</td>
</tr>
<tr>
<td>+ syllable structure</td>
<td>87%</td>
</tr>
</tbody>
</table>

- **Word segmentation accuracy improves when you learn other things as well**
  - *explain away* potentially misleading generalizations
Tone in Mandarin Chinese word segmentation

- Tone in Mandarin Chinese provides an additional dimension of information to the language learner
- It is necessary in order to distinguish lexical items, but how important is it for word segmentation?

Approach:
- construct a pair of otherwise identical corpora, one that contains tone and one that does not
- run identical learning algorithms on both corpora
- compare the accuracy with which each learns word segmentation
Mandarin Chinese corpus

- Used Tardif (1993) “Beijing” corpus (in Pinyin format)
  - deleted all “Child” utterances, and utterances with codes $INTERJ, $UNINT, $VOC and $PRMPT
  - corpus contains 50,118 utterances, consisting of 187,533 word tokens

zen3me gei3 ta1 bei1 shang4 lai2 (1.) ?
ta1: (.) a1yi2 gei3 de (. ) ta1 gei3 de .
hen3 jian3dan1 .

- Used a Pinyin to IPA translation program to produce IPA format

  tsən 214 mɤ kei 214 tʰ a 55 pei 55 ʂəŋ 51 lai 35
  tʰ a 55 a 55 i 35 kei 214 tɤ tʰ a 55 kei 214 tɤ
  xən 214 tɕiɛn 214 tan 55

- Moved tones from end of syllable to preceding vowel

  ts ə 214 n mɤ kei 214 tʰ a 55 pei 55 ʂə 51 ŋ l a i 35
  tʰ a 55 a 55 i 35 kei 214 tɤ tʰ a 55 kei 214 tɤ
  x ə 214 n tɕ iɛ 214 n t a 55 n

- (Optionally delete tones)
Unigram word segmentation adaptor grammar

```
Words → Words Word
Words → Word
Word → Phons
Phons → Phon
Phons → Phons Phon
Phons → Phons Tone
Phon → ai | o | ... | ʂ | tʂʰ | ... |
Tone → 35 | 55 | 214 | ...  Word
      | Phonemes
      | Phonemes
      | Phonemes Tone
      | Phonemes 35
      | Phoneme u
      | Phonemes
      | Phoneme 51
      | Phoneme a
      | Phoneme
      | p
      | Phoneme
      | kʰ
```
Collocation adaptor grammars

- Adaptor grammars with one level of collocation:
  \[
  \text{Collocs} \rightarrow \text{Colloc}^+ \quad \text{Colloc} \rightarrow \text{Words} \quad \text{Words} \rightarrow \text{Word}^+
  \]

- Adaptor grammars with two levels of collocation:
  \[
  \text{Colloc2s} \rightarrow \text{Colloc2}^+ \quad \text{Colloc2} \rightarrow \text{Collocs}^+
  \]
  \[
  \text{Collocs} \rightarrow \text{Colloc}^+ \quad \text{Colloc} \rightarrow \text{Words} \quad \text{Words} \rightarrow \text{Word}^+
  \]

- We experiment with up to three levels of collocation here
Syllable structure adaptor grammars

- No distinction between word-internal and word-peripheral syllables

\[
\begin{align*}
\text{Word} & \rightarrow \text{Syll} \\
\text{Word} & \rightarrow \text{Syll Syll Syll} \\
\text{Syll} & \rightarrow (\text{Onset})? \text{ Rhy} \\
\text{Rhy} & \rightarrow \text{Nucleus} (\text{Coda})? \\
\text{Coda} & \rightarrow C^+ \\
V & \rightarrow ai | o | \ldots
\end{align*}
\]

- Distinguishing word-internal and word-peripheral syllables

\[
\begin{align*}
\text{Word} & \rightarrow \text{SyllIF} \\
\text{Word} & \rightarrow \text{SyllI Syll SyllF} \\
\text{SyllIF} & \rightarrow (\text{OnsetI})? \text{ RhyF} \\
\text{SyllIF} & \rightarrow (\text{OnsetI})? \text{ RhyF} \\
\text{OnsetI} & \rightarrow C^+ \\
\text{CodaF} & \rightarrow C^+
\end{align*}
\]
Mandarin Chinese word segmentation results

- Word segmentation accuracy when input *contains tones*

<table>
<thead>
<tr>
<th>Syllables</th>
<th>None</th>
<th>General</th>
<th>Specialised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>0.57</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Colloc</td>
<td>0.69</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>Colloc(^2)</td>
<td>0.72</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Colloc(^3)</td>
<td>0.64</td>
<td><strong>0.77</strong></td>
<td><strong>0.77</strong></td>
</tr>
</tbody>
</table>

- Word segmentation accuracy when *tones are removed* from input

<table>
<thead>
<tr>
<th>Syllables</th>
<th>None</th>
<th>General</th>
<th>Specialised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>0.56</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>Colloc</td>
<td>0.70</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>Colloc(^2)</td>
<td>0.74</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>Colloc(^3)</td>
<td>0.75</td>
<td>0.76</td>
<td><strong>0.77</strong></td>
</tr>
</tbody>
</table>
Comparable English results

- English word segmentation results

<table>
<thead>
<tr>
<th></th>
<th>Syllables</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>0.56</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>General</td>
<td>0.74</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Specialised</td>
<td>0.79</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Unigram</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colloc</td>
<td></td>
<td>0.74</td>
<td>0.82</td>
<td>0.87</td>
</tr>
<tr>
<td>Colloc²</td>
<td></td>
<td>0.79</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Colloc³</td>
<td></td>
<td>0.74</td>
<td>0.82</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Discussion of Mandarin Chinese word segmentation results

- Mandarin Chinese word segmentation results broadly consistent with English results
  - unigram segmentation accuracies are similar
  - results for other models are lower than corresponding English results
- General improvement in accuracy as number of collocation levels increases
- Caveats: the English and Mandarin Chinese corpora are not directly comparable
  - Discourse context for Mandarin Chinese corpus was far more diverse than for English corpus
  - Mandarin Chinese children were older than English children
Syllable structure and word segmentation

- Syllable structure and phonotactic constraints are very useful for English word segmentation, but are much less useful in Mandarin Chinese
  - perhaps surprising, because Mandarin Chinese has a very regular syllable structure
  - but perhaps this very predictability makes it less useful for identifying words?
  - not surprising that distinguishing word-peripheral syllables does not help, as Mandarin Chinese does not distinguish these
Tone and word segmentation

• Tones only have a small impact on segmentation accuracy
  ▶ surprising, as they are required for lexical disambiguation
  ▶ tones make a small improvement to simpler models (Unigram, Colloc) but no improvement with the more complex ones
    – perhaps tone is redundant given the inter-word context modelled by the Colloc$^{2-3}$ grammars?

• *Perhaps there’s a better way to represent tones in the input, or use tones in the model?*
  ▶ Neutral tones more common on function words — perhaps this can improve segmentation accuracy?
  ▶ Tone sandhi may give information about phonological word boundaries
Conclusion and future work

- The adaptor grammar approach to word segmentation generalises to Mandarin Chinese
- Modelling inter-word dependencies (collocations) greatly improves word segmentation accuracy in Mandarin Chinese (as in English)
- Modelling syllable structure improves segmentation accuracy by a smaller amount in Mandarin Chinese (compared to English)
- Modelling tones improves segmentation accuracy of simpler models, but not of more complex models
- Future work:
  - Comparable multi-lingual corpora of infant-directed speech
  - More realistic, richer corpora (including multi-stratal input representations)
  - Model context-sensitive dependencies (e.g., phonological rules)
Interested in computational linguistics or its applications?

We’re recruiting PhD students!

Contact Mark.Johnson@mq.edu.au or Katherine.Demuth@mq.edu.au for more information.