Finding words and morphemes with Minimum Description Length

(1) Review: we have now looked at a couple ways of trying to segment text into word-like units

- **Predictability strategies:** insert a boundary when the following segment would not be predicted based on what you’ve seen so far
  - \(n\)-grams: at each point in the text, consider how surprising (unlikely) the next segment is. Posit boundaries at very surprising/very low probability transitions.
  - Successor counts: at each point in the text, consider how many different continuations there could be for the next segment. The more possible continuations there are, the more likely it is that we are at a boundary.
  - Why are successor counts for possible prefixes better than simple \(n\)-gram transition probabilities at finding word-like units?
  - What does a model based on successor counts “know”?
- **Utterance-boundary strategies:** insert a boundary after sequences that typically occur at the ends of utterances
  - Pre-boundary transitional probabilities: makes use of some known boundaries to guess about other, unknown boundaries (bootstrapping)
  - Cross-word vs. within-word transitional probabilities: makes use of known boundaries in a training text, to make guesses about boundaries in a new, unseen text.

(2) Today’s class: an approach that tries to finds words more directly, by looking for chunks of material that occur over and over.

- **Word recognition strategy:** rather than looking through a long list of possible utterances every time you want to know if two segments are in the same word, just learn words and try to recognize them
- **Intuitive appeal:**
  - This is precisely what we, as adult speakers, seem to be doing when we read running text with no boundaries: *whatareyoudoingsilly?*
  - Children eventually need to learn words anyway, so why not try to start right away?

(3) An actual example of child-directed speech
ftp://lsrganon:lsrganon@cs.wustl.edu/pub/web/chat/f1-h31may97.cha

- *whatareyoudoingsilly?*
- *noyoudontplaywiththestovedillon.*
- *no.*
- *youknowbetterthanthat.*
- *up.*
- *what?*
- *itsadishwasher.*
- *dishwasher.*
- *yeah.*
- *what?*
- *ohreally?*
- *isthatso?*
- *yougottyslipper?*
- *what?*
- *yoursmokingmeover.*
- *iseeyou.*
- *hi.*
- *holdon!*
- *imalmostdone.*
- *ijustgottarinseoffthebottlecleanoffthecountertopandthetableandwellbealldone.*
- *yeah.*
- *ihopeeforgetthatpanoverthere.*
- *alrightgottadosapantoo.*
- *yeah.*
- *idontlikedothemetheredillon.*

- How might we go about trying to find word boundaries using just a text like this one?
One useful source of information: recurring substrings

noyoudontplaywiththestovedillon.
nodillon.
no.

• Step 1: Start by considering each utterance to be a single, unanalyzed word (even if it’s a very, very long word)

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 noyoudontplaywiththestovedillon</td>
<td>noyoudontplaywiththestovedillon (1)</td>
</tr>
<tr>
<td>2 nodillon</td>
<td>nodillon (2)</td>
</tr>
<tr>
<td>3 no</td>
<td>no (3)</td>
</tr>
</tbody>
</table>

• Step 2: look for short utterances that are contained within longer utterances
  - The last utterance, no, is also contained in the previous utterance (nodillon)
  - Thus, we could posit that the utterance is actually composed of two words: no dillon
  - Furthermore: the first utterance is also composed of two words:
    no youdontplaywiththestovedillon

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 youdontplaywiththestovedillon</td>
<td>no youdontplaywiththestovedillon (3 1)</td>
</tr>
<tr>
<td>2 dillon</td>
<td>no dillon (3 2)</td>
</tr>
<tr>
<td>3 no</td>
<td>no (3)</td>
</tr>
</tbody>
</table>

• Step 3: now dillon is a separate word, and we can find it in the first utterance, too:

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 youdontplaywiththestove</td>
<td>no youdontplaywiththestove dillon (3 1 2)</td>
</tr>
<tr>
<td>2 dillon</td>
<td>no dillon (3 2)</td>
</tr>
<tr>
<td>3 no</td>
<td>no (3)</td>
</tr>
</tbody>
</table>

And so on...

This approach is simple, but inadequate:

• itsapiano
  - What problem does this utterance pose?
  - A similar example from last time: selling we bsitesoftware...

Some desirable goals in word/morpheme segmentation:

• Don't leave long strings of material unanalyzed
  - youdontplaywiththestove is not a plausible word of English

• On the other hand, don't break things up too much
  - y o u d o n t p l a y w i t h t h e s t o v e is not plausible for English either

• Break utterances into words that occur elsewhere
  - nodillon is better analyzed as no dillon than as nodillon
• When there is more than one possible segmentation, prefer the one that uses more likely words
  - Consider the example of iwanttoastonishhim example from “The Yellow Wallpaper”
  - Suppose we already have found the words i, want, to, toast, on, and him
  - Possible segmentations:
    * i want to astonish him
    * i want toast on ishhim
  - Both segmentations require positing a new, previously unseen word (astonish or ishhim). However, if we choose the first segmentation, the rest of the sentence is composed of very frequent/likely words (i, want, to, him), whereas the second option involves a rather infrequent/unlikely word (toast)

• Other criteria that would be helpful
  - Check words to make sure they are phonotactically probably (astonish is just a better-looking word than ishhim; hh is extremely rare in English orthographic words—threshold)
  - Of course once you know something about word meanings and syntax, these would also play a role—we’re assuming for now that you don’t.

(7) Word segmentation as optimization (Brent and Cartwright 1996)
Want to find a lexicon of words/morphemes that
  • Minimizes the length of words (shorter words are more plausible, to a limit)
  • Minimizes the number of different words (try to make use of words you already know)
  • Maximizes probability of each words (words recur as often as possible)

(8) Making this concrete with an example:
doyouseethekitty
seethekitty
doyoulikethekitty

Two possible analyses:
• Analysis 1: long words, simple parses

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 doyouseethekitty</td>
<td>1</td>
</tr>
<tr>
<td>2 seethekitty</td>
<td>2</td>
</tr>
<tr>
<td>3 doyoulikethekitty</td>
<td>3</td>
</tr>
</tbody>
</table>

• Analysis 2: simple words, complex parses

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 d 5 s 9 k</td>
<td>1 2 3 2 4 5 6 6 7 8 9 10 7 3</td>
</tr>
<tr>
<td>2 o 6 e 10 i</td>
<td>5 6 6 7 8 9 10 7 7 3</td>
</tr>
<tr>
<td>3 y 7 t 11 l</td>
<td>1 2 3 2 4 11 10 9 6 7 8 6 9 10 7 7 3</td>
</tr>
<tr>
<td>4 u 8 h</td>
<td></td>
</tr>
</tbody>
</table>
• Analysis 1:
  – Lexicon: 47 symbols (count them to verify this!)
  – Parses: 3 symbols
  – Total description length: 50
• Analysis 2:
  – Lexicon: 24 symbols
  – Parses: 49 symbols
  – Total description length: 73 symbols

(11) Now compare two other analyses:
  • Analysis 3
    
    | Lexicon  | Segmentation |
    |----------|-------------|
    | 1 do     | 1 2 3 4 6   |
    | 2 you    | 3 4 6       |
    | 3 see    | 1 2 5 4 6   |

  Description Length: ___
  
  • Analysis 4
    
    | Lexicon  | Segmentation |
    |----------|-------------|
    | 1 do     | 1 2 3 4     |
    | 2 you    | 3 4         |
    | 3 see    | 1 2 5 4     |

  Description Length: ___

• Which analysis has a shorter description length?

(12) The Minimum Description Length (MDL) approach:
  • Knowledge of the world can be divided into 2 parts:
    – Things you know and can understand/analyze
    – Things you know, but can't analyze
  ➢ Your knowledge about the world is the sum of both of these things
  • Applied to the segmentation problem:
    – Things you know and can analyze = utterances are parsed into smaller units
      * E.g., you understand that nodillon is no + Dillon
    – Things you know but can't analyze = unparsed units = individual words
      * E.g., Dillon can't be understood by breaking it down any more; you just have to memorize it as such
      * Venerable linguistic concept: the lexicon = unpredictable information
  • Goal of learning: understand the text in the shortest/most compact way possible
    – Shortest combination of lexicon + analyses ⇒ Minimum Description Length

(13) How does an MDL approach capture the goals in (7)?

References