CS674 Natural Language Processing

- Last class
  - Word sense disambiguation
    - Decision lists approach
    - Weakly supervised
    - Unsupervised learning
    - Dictionary-based approaches

- Today
  - Word sense disambiguation
    - SENSEVAL
  - Noisy channel model
    - Spelling correction
    - Pronunciation variation

SENSEVAL-2

- Three tasks
  - Lexical sample
  - All-words
  - Translation

- 12 languages

- Lexicon
  - SENSEVAL-1: from HECTOR corpus
  - SENSEVAL-2: from WordNet 1.7

- 93 systems from 34 teams

Lexical sample task

- Select a sample of words from the lexicon
- Systems must then tag several instances of the sample words in short extracts of text
- SENSEVAL-1: 35 words, 41 tasks
  - 700001 John Dos Passos wrote a poem that talked of 'the <tag>bitter</tag> beat look, the scorn on the lip.'
  - 700002 The beans almost double in size during roasting. Black beans are over roasted and will have a <tag>bitter</tag> flavour and insufficiently roasted beans are pale and give a colourless, tasteless drink.

Lexical sample task: SENSEVAL-1

<table>
<thead>
<tr>
<th>Nouns</th>
<th>Verbs</th>
<th>Adjectives</th>
<th>Indeterminates</th>
</tr>
</thead>
<tbody>
<tr>
<td>-n</td>
<td>N</td>
<td>-v</td>
<td>-a</td>
</tr>
<tr>
<td>accident</td>
<td>267</td>
<td>amaze</td>
<td>70</td>
</tr>
<tr>
<td>behaviour</td>
<td>279</td>
<td>bet</td>
<td>177</td>
</tr>
<tr>
<td>bet</td>
<td>274</td>
<td>bother</td>
<td>209</td>
</tr>
<tr>
<td>disability</td>
<td>160</td>
<td>bury</td>
<td>201</td>
</tr>
<tr>
<td>excess</td>
<td>186</td>
<td>calculate</td>
<td>217</td>
</tr>
<tr>
<td>float</td>
<td>75</td>
<td>consume</td>
<td>186</td>
</tr>
<tr>
<td>giant</td>
<td>118</td>
<td>derive</td>
<td>216</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>TOTAL</td>
<td>2756</td>
<td>TOTAL</td>
<td>2501</td>
</tr>
</tbody>
</table>
All-words task

- Systems must tag almost all of the content words in a sample of running text
  - sense-tag all predicates, nouns that are heads of noun-phrase arguments to those predicates, and adjectives modifying those nouns
  - ~5,000 running words of text
  - ~2,000 sense-tagged words

Translation task

- SENSEVAL-2 task
  - Only for Japanese
  - word sense is defined according to translation distinction
    - if the head word is translated differently in the given expressional context, then it is treated as constituting a different sense
  - word sense disambiguation involves selecting the appropriate English word/phrase/sentence equivalent for a Japanese word

SENSEVAL-2 results

<table>
<thead>
<tr>
<th>Language</th>
<th>Task</th>
<th>No. of submissions</th>
<th>No. of teams</th>
<th>IAA</th>
<th>Baseline</th>
<th>Best system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>AW</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>94</td>
</tr>
<tr>
<td>Basque</td>
<td>LS</td>
<td>3</td>
<td>2</td>
<td>.75</td>
<td>.65</td>
<td>76</td>
</tr>
<tr>
<td>Estonian</td>
<td>AW</td>
<td>2</td>
<td>2</td>
<td>.72</td>
<td>.85</td>
<td>67</td>
</tr>
<tr>
<td>Italian</td>
<td>LS</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>39</td>
</tr>
<tr>
<td>Korean</td>
<td>LS</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>.71</td>
<td>74</td>
</tr>
<tr>
<td>Spanish</td>
<td>LS</td>
<td>12</td>
<td>5</td>
<td>.64</td>
<td>.48</td>
<td>65</td>
</tr>
<tr>
<td>Swedish</td>
<td>LS</td>
<td>8</td>
<td>5</td>
<td>.95</td>
<td>-</td>
<td>70</td>
</tr>
<tr>
<td>Japanese</td>
<td>LS</td>
<td>7</td>
<td>3</td>
<td>.86</td>
<td>.72</td>
<td>78</td>
</tr>
<tr>
<td>Japanese</td>
<td>TL</td>
<td>9</td>
<td>8</td>
<td>.81</td>
<td>.37</td>
<td>79</td>
</tr>
<tr>
<td>English</td>
<td>AW</td>
<td>21</td>
<td>12</td>
<td>.75</td>
<td>.57</td>
<td>69</td>
</tr>
<tr>
<td>English</td>
<td>LS</td>
<td>26</td>
<td>15</td>
<td>.86</td>
<td>.51/.16</td>
<td>64/40</td>
</tr>
</tbody>
</table>

SENSEVAL plans

- Where next?
  - Supervised ML approaches worked best
    » Looking at the role of feature selection algorithms
  - Need a well-motivated sense inventory
    » Inter-annotator agreement went down when moving to WordNet senses
  - Need to tie WSD to real applications
    » The translation task was a good initial attempt
**CS674 Natural Language Processing**

- **Last class**
  - Word sense disambiguation
    - Finish decision lists approach
    - Weakly supervised
    - Unsupervised learning
    - Dictionary-based approaches
- **Today**
  - Word sense disambiguation
    - SENSEVAL
    - Noisy channel model
    - Spelling correction
    - Pronunciation variation

---

**Correction of spelling errors**

- Frequency of spelling errors in human typed text varies from
  - 0.05% of the words in carefully edited newswire, to
  - 38% in difficult applications like telephone directory lookup
- Optical character recognition
  - Higher error rates than human typists
  - Make different kinds of errors, “D” → “O”; “ri” → “n”
- On-line handwriting recognition

---

**Types of spelling correction**

- **Non-word error detection**
  - Detecting spelling errors that result in non-words
    - "graffe" → "giraffe"
- **Isolated-word error correction:**
  - Correcting spelling errors that result in non-words
    - Correcting "graffe" to "giraffe", but looking only at the word in isolation

---

**Types of spelling correction**

- **Context-dependent error detection and correction**
  - Using the context to help detect and correct spelling errors
  - Some of these may accidentally result in an actual word (real-word errors)
    - Typographical errors
      - e.g. "there" for "three"
    - Homonym or near-homonym
      - e.g. "dessert" for "desert", or "piece" for "peace"

---

Kukich, 1992
Fixing non-word errors

- Detecting non-words
  - Use a dictionary
  - Usually include models of morphology
  - For other types of spelling correction, we'll need a model of spelling variation.

Proposing candidate corrections

- Simplifying assumption: the correct word will differ from the misspelling by a single insertion, deletion, substitution, or transposition
  - Handles most spelling errors in human typed text
- Generate the candidates by applying any single transformation that results in a word in an on-line dictionary

Candidate corrections for *acress*

<table>
<thead>
<tr>
<th>Error</th>
<th>Correction</th>
<th>Correct Letter</th>
<th>Error Letter</th>
<th>Position (Letter #)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>across</td>
<td>actress</td>
<td>t</td>
<td>–</td>
<td>2</td>
<td>deletion</td>
</tr>
<tr>
<td>across</td>
<td>cress</td>
<td>–</td>
<td>a</td>
<td>0</td>
<td>insertion</td>
</tr>
<tr>
<td>across</td>
<td>caress</td>
<td>ca</td>
<td>ac</td>
<td>0</td>
<td>transposition</td>
</tr>
<tr>
<td>across</td>
<td>access</td>
<td>c</td>
<td>r</td>
<td>2</td>
<td>substitution</td>
</tr>
<tr>
<td>across</td>
<td>across</td>
<td>o</td>
<td>e</td>
<td>3</td>
<td>substitution</td>
</tr>
<tr>
<td>across</td>
<td>acres</td>
<td>–</td>
<td>2</td>
<td>5</td>
<td>insertion</td>
</tr>
<tr>
<td>across</td>
<td>acres</td>
<td>–</td>
<td>2</td>
<td>4</td>
<td>insertion</td>
</tr>
</tbody>
</table>

The pronunciation subproblem

- *[spooky music][music stops]*
- **Head Knight of Ni**: Ni!
- **Knights of Ni**: Ni! Ni! Ni! Ni! Ni! Ni!
- **Arthur**: Who are you?
- **Head Knight**: We are the Knights Who Say…’Ni’! …
- We are the keepers of the sacred words: ‘Ni’, ‘Peng’, and ‘Neee-wom’!
The pronunciation subproblem

- Given a series of phones, compute the most probable word that generated them.
- **Simplifications**
  - Given the correct string of phones
    » Speech recognizer relies on probabilistic estimators for each phone, so it's never entirely sure about the identification of any particular phone
  - Given word boundaries
- “I [ni]...”
  - [ni] → the, neat, need, new, knee, to, and you
  - Based on the (transcribed) Switchboard corpus
- Contextually-induced pronunciation variation

No candidate generation

- Use corpus to expand each pronunciation in advance with all possible variants
- [ni] is stored with the list of words that can generate it

Probabilistic transduction

- surface representation → lexical representation
- sequence of letters in a mis-spelled word → sequence of letters in correctly spelled words
  - acres → actress, cress, acres
- string of symbols representing the pronunciation of a word in context → string of symbols representing the dictionary pronunciation
  - [er] → her, were, are, their, your
  - exacerbated by pronunciation variation
    » the pronounced as THEE or THUH
    » some aspects of this variation are systematic, like spelling error patterns

Noisy channel model

- Channel introduces noise which makes it hard to recognize the true word.
- **Goal**: build a model of the channel so that we can figure out how it modified the true word...so that we can recover it.
Decoding algorithm

- Special case of **Bayesian inference**
  - Bayesian classification
    » Given observation, determine which of a set of classes it belongs to.
    » Observation
      ◆ string of phones or string of letters
    » Classify into
      ◆ words