Lexicon Acquisition

Daniel Zeman
zeman@ufal.mff.cuni.cz
Lexicon Acquisition

• Some hints only (approach must vary greatly depending on language)

• Identify part of speech and inflection pattern

• If affixes restrict possible classes, use it!
  – E.g. in Czech, the following suffixes increase likelihood of an infinitive: -st, -át, -at, -ct, -ci, -ít, -out, -ýt, -ovat, -it, -êt, -et
  – English does not inflect but verb forms and derivational suffixes (-ness, -ity, -able) can help

• Otherwise, syntax might help
  – E.g. if it’s after preposition or an article it’s likely an adjective or a noun
Lexicon Acquisition

• Create word frequency list
• Identify closed-class words
  – Many of them will be very frequent
  – A textbook and/or a bilingual dictionary may help with the rest
  – Parallel corpus + word aligner may supplement the dictionary
• What remains are mostly nouns, adjectives, verbs and adverbs
  – Try to sort it out by iteratively looking at the word list, identifying repeating affixes etc.
  – If there are no repeating bound morphemes
    • then you may not be able to sort out the parts of speech
    • but maybe the morphology of the language is not so interesting after all
English Lexicon (LAB)


- There is also a link from the NPFL094 website
English Lexicon Acquisition

• Example only! Other languages and corpora may require a different approach.

• Input: a plain-text corpus (taken from Penn Treebank)
  – Tokenized (punctuation separated from words)
  – Remove traces (non-word terminal nodes in Penn Treebank): all tokens containing “*”?
  – Lowercase
    • Later we will want to identify proper nouns
    • Complicated by sentence-initial capitalization
When it’s time for their biannual powwow, the nation’s manufacturing titans typically jet off to the sunny confines of resort
English Frequency Wordlist

- Penn Treebank 3 / Wall Street Journal:
  - 49,208 sentences
  - 1,253,013 terminal nodes (tokens and traces)
  - 49,817 word types (opposed to word occurrences) including traces
  - 43,764 lowercased types without traces and some other technical nodes (“error:” etc.)
- The most frequent types often have these (overlapping) properties:
  - stopwords
  - closed-class words
  - short words?
## English Frequency Wordlist

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>60484</td>
</tr>
<tr>
<td>the</td>
<td>59459</td>
</tr>
<tr>
<td>.</td>
<td>48144</td>
</tr>
<tr>
<td>to</td>
<td>29576</td>
</tr>
<tr>
<td>of</td>
<td>28440</td>
</tr>
<tr>
<td>a</td>
<td>24781</td>
</tr>
<tr>
<td>in</td>
<td>21047</td>
</tr>
<tr>
<td>and</td>
<td>20449</td>
</tr>
<tr>
<td>’s</td>
<td>11556</td>
</tr>
<tr>
<td>for</td>
<td>10454</td>
</tr>
<tr>
<td>that</td>
<td>10422</td>
</tr>
<tr>
<td>$</td>
<td>8817</td>
</tr>
<tr>
<td>`` (“)</td>
<td>8735</td>
</tr>
<tr>
<td>is</td>
<td>8539</td>
</tr>
<tr>
<td>&quot; (”)</td>
<td>8506</td>
</tr>
<tr>
<td>it</td>
<td>7195</td>
</tr>
<tr>
<td>said</td>
<td>7141</td>
</tr>
<tr>
<td>on</td>
<td>6646</td>
</tr>
<tr>
<td>%</td>
<td>6121</td>
</tr>
<tr>
<td>at</td>
<td>5770</td>
</tr>
<tr>
<td>by</td>
<td>5705</td>
</tr>
<tr>
<td>as</td>
<td>5701</td>
</tr>
</tbody>
</table>
### Punctuation and Special Characters

<table>
<thead>
<tr>
<th>Character</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>60484</td>
</tr>
<tr>
<td>.</td>
<td>48144</td>
</tr>
<tr>
<td>'s</td>
<td>11556</td>
</tr>
<tr>
<td>$</td>
<td>8817</td>
</tr>
<tr>
<td>`` (&quot;)</td>
<td>8735</td>
</tr>
<tr>
<td>&quot; (&quot;)</td>
<td>8506</td>
</tr>
<tr>
<td>%</td>
<td>6121</td>
</tr>
<tr>
<td>mr. (tokenization?)</td>
<td>4950</td>
</tr>
<tr>
<td>n’t</td>
<td>4006</td>
</tr>
<tr>
<td>--</td>
<td>2585</td>
</tr>
<tr>
<td>u.s.</td>
<td>2056</td>
</tr>
<tr>
<td>third-quarter</td>
<td>333</td>
</tr>
<tr>
<td>buy-out</td>
<td>222</td>
</tr>
<tr>
<td>s&amp;p</td>
<td>164</td>
</tr>
<tr>
<td>3,000</td>
<td>28</td>
</tr>
<tr>
<td>3.7</td>
<td>28</td>
</tr>
<tr>
<td>total types</td>
<td>10607</td>
</tr>
<tr>
<td>the rest</td>
<td>33157</td>
</tr>
</tbody>
</table>

- Caught, OK
- Not caught (but should have been caught)
- Caught (disputable)
- Caught (we want better tokenization)
### Numbers

\[ m/\pN/ \]

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12447</td>
<td>b-2</td>
</tr>
<tr>
<td>10</td>
<td>668</td>
<td>19th</td>
</tr>
<tr>
<td>30</td>
<td>607…</td>
<td>1989-90</td>
</tr>
<tr>
<td>1988</td>
<td>503…</td>
<td>80%-owned</td>
</tr>
<tr>
<td>1,000</td>
<td>111…</td>
<td>xr4ti</td>
</tr>
<tr>
<td>1/2</td>
<td>105…</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>88…</td>
<td>total types</td>
</tr>
<tr>
<td>30-year</td>
<td>79…</td>
<td>the rest</td>
</tr>
<tr>
<td>1980s</td>
<td>53…</td>
<td></td>
</tr>
<tr>
<td>ru-486</td>
<td>15…</td>
<td>no punctuation or numbers</td>
</tr>
<tr>
<td>mid-1980s</td>
<td>12…</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>total types: 6123</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the rest: 37641</td>
</tr>
<tr>
<td></td>
<td></td>
<td>no punctuation or numbers: 32218</td>
</tr>
</tbody>
</table>

22.10.2010
## Real Words

\![m/ \[\pP\pN`\$\]]/

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>59459</td>
<td>on</td>
<td>6646</td>
</tr>
<tr>
<td>of</td>
<td>28440</td>
<td>at</td>
<td>5770</td>
</tr>
<tr>
<td>to</td>
<td>27448</td>
<td>by</td>
<td>5705</td>
</tr>
<tr>
<td>a</td>
<td>24781</td>
<td>as</td>
<td>5701</td>
</tr>
<tr>
<td>in</td>
<td>21047</td>
<td>from</td>
<td>5438</td>
</tr>
<tr>
<td>and</td>
<td>20449</td>
<td>with</td>
<td>5357</td>
</tr>
<tr>
<td>for</td>
<td>10454</td>
<td>million</td>
<td>5335</td>
</tr>
<tr>
<td>that</td>
<td>10422</td>
<td>was</td>
<td>4901</td>
</tr>
<tr>
<td>is</td>
<td>8539</td>
<td>be</td>
<td>4586</td>
</tr>
<tr>
<td>it</td>
<td>7195</td>
<td>its</td>
<td>4571</td>
</tr>
<tr>
<td>said</td>
<td>7141</td>
<td>are</td>
<td>4528</td>
</tr>
</tbody>
</table>

22.10.2010

http://ufal.mff.cuni.cz/course/npfl094
Enumerating Closed-Class Words

- Pronouns / determiners / articles in all cases
  - Personal: I, me, you, he, him, she, her, it, we, us, they, them
  - Impersonal: one (as in “One has to be careful here.”)
  - Reflexive: myself, yourself, himself, herself, itself, ourselves, yourselves, themselves, oneself
  - Possessive: my, mine, your, yours, his, her, hers, its, our, ours, their, theirs
  - Demonstrative: this, these, that, those
  - Article: the, a, an
  - Interrogative / relative: who, whom, whose, what, which
  - Indefinite: some, somebody, someone, something, any, anybody, anyone, anything, every, everybody, everyone, everything, each, all, both; many, much, more, most, too, enough, few, little, fewer, less, least
  - Negative: no, nobody, nothing, none
Enumerating Closed-Class Words

• Numerals
  – Cardinal
    • zero, one, two, three, four, five, six, seven, eight, nine, ten
    • eleven, twelve, thirteen, …, nineteen
    • twenty, thirty, forty, sixty, seventy, eighty, ninety
    • hundred, thousand, million, billion
  – Ordinal
    • first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth ⇒ morphology “-th”
  – In some languages written as one word, i.e. nice morph. exercise:
    • 361,972
    • en: three hundred sixty-one thousand nine hundred and seventy-two
    • de: dreihunderteinundsechzigtausendneunhundertzweiundsiebzig
Enumerating Closed-Class Words

- Auxiliary and modal verbs
  - *be, am, are, is, was, were, been, being, ’m, ’s, ’re*
  - *have, has, had, having, ’ve, ’s, ’d*
  - *will, would, (willing), ’ll, ’d*
  - *can, cannot, could*
  - *shall, should*
  - *may, might*
  - *must*
  - *do, does, did, done, doing*
Enumerating Closed-Class Words

- Pronominal adverbs
  - Demonstrative: here, there, now, then
  - Interrogative / relative: where, when, how, why
  - Indefinite: somewhere, sometime, somehow, anywhere, anytime, anyhow, anyway, everywhere, always
  - Negative: nowhere, never
Enumerating Closed-Class Words

• Prepositions (>60; tagged corpus?)
  – aboard, about, above, across, after, against, ago, along, alongside, amid, among, amongst, around, as, astride, at, atop, before, behind, below, beneath, beside, besides, between, beyond, by, despite, de, down, during, en, except, for, from, in, inside, into, lest, like, minus, near, next, notwithstanding, of, off, on, onto, opposite, out, outside, over, par, past, per, plus, post, since, through, throughout, ’til, till, to, toward, towards, under, underneath, unlike, until, unto, up, upon, versus, via, vs., with, within, without, worth
  – grep 'IN' wsj.mrg | perl -pe 's/^.*?\b(IN (.*)\b).*$/\1/; $_=lc($_)' | sort -u | more
Enumerating Closed-Class Words

• Conjunctions
  – Coordinating: *and, both, but, either, et, less, minus, ’n, ’n’, neither, nor, or, plus, so, times, v., versus, vs., yet*
  – Subordinating: *albeit, although, because, ’cause, if, neither, since, so, than, that, though, ’til, till, unless, until, whereas, whether, which, while*

• Particles
  – *yes, no, not, n’t, to* (infinitival)

• Found in corpus:
  – 256 closed-class types (out of 307 anticipated)
  – 413,914 occurrences (33% of total tokens)
Open-Class Words

• Now there is a nice list of some 32,000 open-class words. What remains is to read them all and sort them out manually 😊
  – Nouns (including proper nouns)
  – Adjectives (including those derived from proper nouns)
  – Verbs (except for auxiliaries and modals)
  – Adverbs
  – (Interjections)

• What else can help us?
Most Frequent OC Words

- said 7141
- new 3257
- company 3078
- year 2753
- market 2648
- says 2467
- stock 2002
- also 1867
- other 1808
- share 1798
- last 1482
- shares 1444
- president 1431
- years 1426
- trading 1415
- sales 1331
- only 1188
- business 1171
- such 1164
- york 1129
- group 1102
- time 1032
Most Frequent OC Words

- said 7141  
- new 3257  
- company 3078  
- year 2753  
- market 2648  
- says 2467  
- stock 2002  
- also 1867  
- other 1808  
- share 1798  
- last 1482  
- shares 1444  
- president 1431  
- years 1426  
- trading 1415  
- sales 1331  
- only 1188  
- business 1171  
- such 1164  
- york 1129  
- group 1102  
- time 1032
## Plurals / 3rd Person Verbs

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>years</td>
<td>1426</td>
</tr>
<tr>
<td>company</td>
<td>companies</td>
<td>1020</td>
</tr>
<tr>
<td>new</td>
<td>news</td>
<td>423</td>
</tr>
<tr>
<td>say</td>
<td>says</td>
<td>2467</td>
</tr>
<tr>
<td>market</td>
<td>markets</td>
<td>621</td>
</tr>
<tr>
<td>stock</td>
<td>stocks</td>
<td>800</td>
</tr>
<tr>
<td>other</td>
<td>others</td>
<td>263</td>
</tr>
<tr>
<td>price</td>
<td>prices</td>
<td>1016</td>
</tr>
<tr>
<td>sale</td>
<td>sales</td>
<td>1331</td>
</tr>
<tr>
<td>last</td>
<td>lasts</td>
<td>8</td>
</tr>
<tr>
<td>month</td>
<td>months</td>
<td>844</td>
</tr>
<tr>
<td>president</td>
<td>presidents</td>
<td>22</td>
</tr>
<tr>
<td>business</td>
<td>businesses</td>
<td>267</td>
</tr>
</tbody>
</table>

Total 4448 pairs
Gerunds / Present Participles

- market 2648 marketing 211 2859
- stock 2002 stocking 2 2004
- trade 525 trading 1415 1940
- share 1798 sharing 9 1807
- last 1482 lasting 9 1491
- bank 955 banking 220 1175
- time 1032 timing 33 1065
- say 878 saying 172 1050
- make 739 making 286 1025
- price 929 pricing 59 988
- sell 603 selling 353 956 ...
- even 905 evening 35 940 ...
- get 572 getting 201 773

Total 1927 pairs
Tagged Corpus Available?

• Having a tagged corpus does not necessarily mean we have a morphological analyzer, so it still could make sense to construct one
• Now it’s trivial to distinguish nouns from verbs, adjectives etc., even if they overlap
• Still, we may need some information not encoded in the tags
• Example: declension class (“pattern”) of Czech nouns:
  – **NNF** = feminine noun ⇒ 4 declension classes:
    • „žena“ -a, -y, -ě, -u, -o, -ě, -ou, -y, -0, -ám, -y, -y, -ách, -ami
    • „růže“ -e, -e, -i, -i, -e, -i, -í, -e, -í, -ím, -e, -e, -ích, -emi
    • „píseň“ -0, -ě, -i, -0, -i, -i, -i, -ě, -í, -ím, -ě, -ě, -ích, -ěmi
    • „kost“ -0, -i, -i, -0, -i, -i, -i, -i, -em, -i, -i, -ech, -mi
And So On…

• Using similar heuristics, gradually classify more and more word forms.
  – Obviously, not everything can be captured this way
    • Some sets of pairs have multiple interpretations
    • For some words no heuristics exist
    • Or the other member of the pair has not occurred in the corpus

• Semi-supervised:
  – You don’t know what word form belongs where
  – However, you know how the suffixes look like

• Unsupervised:
  – You don’t even know the set of affixes
  – However, you know (or assume) that the morphology is concatenative
    (prefix* stem+ suffix*)
  – Look at the corpus, try to find regularities
Unsupervised Morphemic Segmentation

- Morpho Challenge (shared task) since 2005
- Linguistica (John A. Goldsmith)
  (http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/)
- Morfessor (Mathias Creutz & Krista Lagus)
  (http://www.cis.hut.fi/projects/morpho/)
- ParaMor (Christian Monson)
  (http://www.cslu.ogi.edu/~monsonc/ParaMor.html)
- Affisix (Michal Hrušecký, MFF)
- Morseus (Dan Zeman, MFF)
  (http://ufal.mff.cuni.cz/~zeman/projekty/morseus/)
- And many others…
Homework

• Pick a language, get data
• Extract lexicon
• Details and data: http://ufal.mff.cuni.cz/~zeman/vyuka/morfosynt/lab-lexicon/index.html
• Deadline: Friday November 15, 23:59