Corpora

- A corpus is a body of naturally occurring text, stored in a machine-readable form
- A balanced corpus tries to be representative across a language or other domains

Today

- Corpora and its properties
- Zipf's Law
- Examples of annotated corpora
- Word segmentation algorithm

Word Counts

- What are the most common words in the text?
- How many words are there in the text?
- What are the properties of word distribution in large corpora?

We will consider Mark Twain's *Tom Sawyer*
**Most Common Words**

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>determiner (article)</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>conjunction</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>determiner</td>
</tr>
<tr>
<td>to</td>
<td>1725</td>
<td>preposition, inf. marker</td>
</tr>
<tr>
<td>of</td>
<td>1440</td>
<td>preposition</td>
</tr>
<tr>
<td>was</td>
<td>1161</td>
<td>auxiliary verb</td>
</tr>
<tr>
<td>it</td>
<td>1027</td>
<td>pronoun</td>
</tr>
<tr>
<td>in</td>
<td>906</td>
<td>preposition</td>
</tr>
<tr>
<td>that</td>
<td>877</td>
<td>complementizer</td>
</tr>
<tr>
<td>Tom</td>
<td>678</td>
<td>proper name</td>
</tr>
</tbody>
</table>

Some observations:

- Dominance of function words
- Presence of corpus-dependent items (e.g., “Tom”)

Is it possible to create a truly “representative” sample of English?

**How Many Words Are There?**

They picnicked by the pool, then lay back on the grass and looked at the stars.

- **Type** — number of distinct words in a corpus (vocabulary size)
- **Token** — total number of words in a corpus

Tom Sawyer:

word types — 8,018  
word tokens — 71,370

average frequency — 9

---

**Frequencies of Frequencies**

<table>
<thead>
<tr>
<th>Word Frequency</th>
<th>Frequency of Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3993</td>
</tr>
<tr>
<td>2</td>
<td>1292</td>
</tr>
<tr>
<td>3</td>
<td>664</td>
</tr>
<tr>
<td>4</td>
<td>410</td>
</tr>
<tr>
<td>5</td>
<td>243</td>
</tr>
<tr>
<td>6</td>
<td>199</td>
</tr>
<tr>
<td>7</td>
<td>172</td>
</tr>
<tr>
<td>8</td>
<td>131</td>
</tr>
<tr>
<td>9</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>91</td>
</tr>
<tr>
<td>11-50</td>
<td>540</td>
</tr>
<tr>
<td>51-100</td>
<td>99</td>
</tr>
</tbody>
</table>

Most words in a corpus appear only once!
**Zipf’s Law**

The frequency of use of the $n$th-most-frequently-used word in any natural language is approximately inversely proportional to $n$.

Zipf’s Law captures the relationship between frequency and rank:

$$ f \propto \frac{1}{r} $$

There is a constant $k$ such that:

$$ f \cdot r = k $$

---

**Mandelbrot’s refinement**

$$ f = P(r+p)^{-B} $$

or

$$ \log f = \log P - B \log(r+p) $$

- $P$, $B$, $p$ are parametrized for particular corpora
- Better fit at low and high ranks
Zipf’s Law and Principle of Least Effort

*Human Behavior and the Principle of Least Effort* (Zipf):
“… Zipf argues that he found a unifying principle, the Principle of Least Effort, which underlies essentially the entire human condition (the book even includes some questionable remarks on human sexuality!). The principle argues that people will act so as to minimize their probable average rate of work”. (Manning & Schutze, p.23)

---

Examples of collections approximately obeying Zipf’s law

- Frequency of accesses to web pages
- Sizes of settlements
- Income distribution amongst individuals
- Size of earthquakes
- Notes in musical performances

---

Other laws

- Word sense distribution
- Phonemes distribution
- Word co-occurrence patterns

---

Is Zipf’s Law unique to human language?

(Li 1992): randomly generated text exhibits Zipf’s law

Consider a generator that randomly produces characters from the 26 letters of the alphabet and the blank.

\[ p(w_n) = \left( \frac{26}{27} \right)^n \cdot \frac{1}{27} \]

The words generated by such a generator obey a power law of the Mandelbrot:

- There are 26 times more words of length \( n + 1 \) than words of length \( n \)
There is a constant ratio by which words of length $n$ are more frequent than words of length $n + 1$

**Sparsity**

There is no data like more data

- How often does “kick” occur in 1M words? 58
- How often does “kick a ball” occur in 1M words? 0
- How often does “kick” occur in the web? 6 M
- How often does “kick a ball” occur in the web? 8.000

**Very Very Large Data**

- Brill & Banko 2001: In the task of confusion set disambiguation increase of data size yields significant improvement over the best performing system trained on the standard training corpus size set
  - Task: disambiguate between pairs such as too, to
  - Training Size: varies from one million to one billion
  - Learning methods used for comparison: winnow, perceptron, decision-tree

- Lapata & Keller 2002, 2003: the web can be used as a very very large corpus
  - The counts can be noisy, but for some tasks this is not an issue
The Brown Corpus

Famous early corpus (Made by Nelson Francis and Henry Kucera at Brown University in the 1960s)

- A balanced corpus of written American English
  - Newspaper, novels, non-fiction, academic
- 1 million words, 500 written texts

Do you think this is a large corpus?

Recent Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Domain</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA News Corpus</td>
<td>600 million</td>
<td>newswire</td>
<td>American English</td>
</tr>
<tr>
<td>British National Corpus</td>
<td>100 million</td>
<td>balanced</td>
<td>British English</td>
</tr>
<tr>
<td>EU proceedings</td>
<td>20 million</td>
<td>legal</td>
<td>10 language pairs</td>
</tr>
<tr>
<td>Penn Treebank</td>
<td>2 million</td>
<td>newswire</td>
<td>American English</td>
</tr>
<tr>
<td>Broadcast News</td>
<td></td>
<td>spoken</td>
<td>7 languages</td>
</tr>
<tr>
<td>SwitchBoard</td>
<td>2.4 million</td>
<td>spoken</td>
<td>American English</td>
</tr>
</tbody>
</table>

For more corpora, check the Linguistic Data Consortium:
http://www.ldc.upenn.edu/

Corpus Content

- Genre:
  - newswires, novels, broadcast, spontaneous conversations
- Media: text, audio, video
- Annotations: tokenization, syntactic trees, semantic senses, translations

Example of Annotations: POS Tagging

POS tags encode simple grammatical functions

- Several tag sets:
  - Penn tag set (45 tags)
  - Brown tag set (87 tags)
  - CLAWS2 tag set (132 tags)

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
<th>Claws c5</th>
<th>Brown</th>
<th>Penn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adverb</td>
<td>often, badly</td>
<td>AJ0</td>
<td>JJ</td>
<td>JJ</td>
</tr>
<tr>
<td>Noun singular</td>
<td>table, rose</td>
<td>NN1</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td>Noun plural</td>
<td>tables, roses</td>
<td>NN2</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td>Noun proper singular</td>
<td>Boston, Leslie</td>
<td>NP0</td>
<td>NP</td>
<td>NNP</td>
</tr>
</tbody>
</table>
Issues in Annotations

- Different annotation schemes for the same task are common
- In some cases, there is a direct mapping between schemes; in other cases, they do not exhibit any regular relation
- Choice of annotation is motivated by the linguistic, the computational and/or the task requirements

Tokenization

Goal: divide text into a sequence of words
- Word is a string of contiguous alphanumerical characters with space on either side; may include hyphens and apostrophes but no other punctuation marks (Kucera and Francis)

Is tokenization easy?

What’s a word?

- English:
  - “Wash. vs wash”
  - “won’t”, “John’s”
  - “pro-Arab”, “the idea of a child-as-required-yuppie-possession must be motivating them”, “85-year-old grandmother”
- East Asian languages:
  - words are not separated by white spaces

Word Segmentation

- Rule-based approach: morphological analysis based on lexical and grammatical knowledge
- Corpus-based approach: learn from corpora (Ando & Lee, 2000)

Issues to consider: coverage, ambiguity, accuracy
**Motivation for Statistical Segmentation**

- Unknown words problem:
  - presence of domain terms and proper names
- Grammatical constraints may not be sufficient
  - Example: alternative segmentation of noun phrases

<table>
<thead>
<tr>
<th>Segmentation</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>sha-choh/ken/gyoh/mu/bu-choh</td>
<td>&quot;president/and/business/general/manager&quot;</td>
</tr>
<tr>
<td>sha-choh/ken-gyoh/mu/bu-choh</td>
<td>&quot;president/subsidiary business/Tsutomi[a name]/general manager&quot;</td>
</tr>
</tbody>
</table>

**Algorithm for Word Segmentation**

- $s^n_i$: non-straddling n-grams to the left of location $k$
- $s^n_2$: non-straddling n-grams to the right of location $k$
- $t^n_j$: straddling n-gram with $j$ characters to the right of location $k$
- $I_{\geq}(y, z)$: indicator function that is 1 when $y \geq z$, and 0 otherwise.

1. Calculate the fraction of affirmative answers for each $n$ in $N$:

$$v_n(k) = \frac{1}{2 + (n - 1)} \sum_{i=1}^{n-1} \sum_{j=1}^{2} I_{\geq}((s^n_i), (t^n_j))$$

2. Average the contributions of each $n$-gram order

$$v_N(k) = \frac{1}{N} \sum_{n \in N} v_n(k)$$

**Algorithm for Word Segmentation (Cont.)**

Place boundary at all locations $l$ such that either:

- $l$ is a local maximum: $v_N(l) > v_N(l - 1)$ and $v_N(l) > v_N(l + 1)$
- $v_N(l) \geq t$, a threshold parameter

For $N = 4$, consider the 6 questions of the form:

"Is $\#(s_i) \geq \#(t_j)$?", where $\#(x)$ is the number of occurrences of $x$

Example: Is “TING” more frequent in the corpus than "INGE"?
Experimental Framework

- Corpus: 150 megabytes of 1993 Nikkei newswire
- Manual annotations: 50 sequences for development set (parameter tuning) and 50 sequences for test set
- Baseline algorithms: Chasen and Juman morphological analyzers (115,000 and 231,000 words)

Evaluation Measures

\[ P = \frac{tp}{tp + fp} \]
\[ R = \frac{tp}{tp + fn} \]
\[ F = \frac{2 \cdot PR}{(R + P)} \]

Word precision (P) is the percentage of proposed brackets that match word-level brackets in the annotation;
Word recall (R) is the percentage of word-level brackets that are proposed by the algorithm.

Conclusions

- Corpora widely used in text processing
- Corpora used either annotated or raw
- Zipf’s law and its connection to natural language
- Sparsity is a major problem for corpus processing methods

Next time: Language modeling