CS497: Learning and NLP
Lec 3: Natural Language and Statistics

Spring 2009

January 28, 2009
Corpora and its analysis
Motivation for statistical approaches
Statistical properties of language (e.g., Zipf’s laws)
Examples of corpora
Corpus linguistics is the study of language as it is expressed in corpora;

Corpora is often annotated with syntactic, semantic or other information;
The study of language is concerned with two basic questions:

- What kinds of things do people say?
- What do these things say/ ask/ request about the world?

The first point covers aspects of the *structure of language*. The second pertains to semantics, pragmatics and discourse – how to connect utterances to the world.
Most of corpus linguistics is about the first point (structure of language)

Hope:
- there is connection between patterns of use and deep understanding;
- by being able to say something about the first point we can make progress on the second

Wittgenstein: The meaning of the word is defined by the circumstances of its use.

We will focus on statistical discoveries in the context of the first question.
Why statistical approaches?

- Chomsky (1957):
  
  1. “Colorless green ideas sleep furiously.
  2. Furiously sleep ideas green colorless.

  It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally ‘remote’ from English. Yet (1), though nonsensical, is grammatical, while (2) is not.”

- What do you thing of this statement?
Reasoning of Chomsky:

- For each sentence it is unlikely that a part of this sentence has ever occurred in the corpus.
- He argues that it follows that any statistical model would assign zero probabilities to both cases.

However:

- This relies on the assumption that any probabilistic model assigns 0 probability to unseen events.
- It would be the case that if statistical models assign 0 probability to unseen events (maximum likelihood estimation).
Why statistical approaches?

- Aggregate bigram model (Saul and Pereira, 99):

  \[
P(w_2 | w_1) = \sum_c P(w_2 | c) P(c | w_1)
  \]  

  \(c\) is a latent class encoding relation between \(w_1\) and \(w_2\) 
  (more on latent variable models later in the course)

- Parameters \(P(w_2 | c), P(c | w_1)\) are estimated from a corpus

- Probability of a sequence

  \[
P(w_1 \ldots w_n) = P(w_1) \prod_{i=2}^{n} P(w_i | w_{i-1})
  \]  

- Compare probability estimates of a model trained on a news corpus with \(C = 16\)

  \[
  \frac{P(\text{Colorless green ideas sleep furiously})}{P(\text{Furiously sleep ideas green colorless})} = 2 \times 10^5
  \]
Chomsky’s example

But, what does it really show?

- Is grammatical/non grammatical the significant distinction?

Do we really need to distinguish them to achieve our goal?

In many cases we cannot expect input to be fully grammatical (e.g., speech processing, analysis of user reviews etc)
(Abney, 1996) provides a very good survey of problems in natural language that non-statistical approaches to linguistics will find hard to handle.

The most important problem is ambiguity
Ambiguity exists in almost any natural language decision

We gave some examples last time, all of them, in principle, can be resolved in several ways, only one of which makes sense to humans.

Traditional approaches attempt first to determine the *structure* of the sentence and then use it to determine other things: Semantic analysis is done, if at all, only after the syntactic analysis.

Can they be decoupled?
Our company is training workers.

[Our company *NP*] [ [is *aux*] [ [training *V*] [workers *NP*] *VP*] *VP*]

Here *training workers* is understood correctly. *is training* is the Verb group.
Our company is training workers.

[Our company $NP$] [ [is $V$] [ [training $V$] [workers $NP$] $VP$] $NP$] $VP$

Here $is$ is the main verb and $training$ workers is used like a 
$gerund$ as in $our$ problem $is$ training workers.
Our company is training workers.

[Our company \textit{NP}] [is \textit{V}] [training \textit{AdjP}] [workers \textit{N} \textit{NP}] \textit{VP}]

Here \textit{is} is also the main verb and \textit{training} modifies \textit{workers} as in \textit{training wheels}. 
This is an example with (at least) three different syntactic analysis (parses).

Examples of these sort exist in almost any non-trivial, or long enough sentence.

Preposition phrases have several possible attachments, only one of which makes sense.
  - I wore the shirt with the short sleeves.
Long sentences may have hundreds different syntactically legitimate parses. The sentence

List the sales of the products produced in 1973 with the products produced in 1972.

is reported to have 455 different parses by one parsing system.
a NLP system is required to be good at making disambiguation decisions (word sense, category, syntactical structure, semantic scope,...).

Even if one could write down a good set of constrains and preference rules we still need to address
- Scaling up beyond small and domain specific applications.
- Practicality: time consuming to build if we want reasonable coverage
- Brittleness (e.g., in the face of using metaphors)

an instance of the general Knowledge Representation problem in AI, and it is hard to imagine that it is possible to get around this problem without learning. (See Charniak book for more discussion)
Statistics of Language

- What we can learn from just extracting statistics?
- We will use *Tom Sawyer* by Mark Twain as our corpus
- What are simple questions we can ask about language?
What are the most common words?

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</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, verbal infinitive 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>(personal/expletive)</td>
</tr>
<tr>
<td>in</td>
<td>906</td>
<td>preposition</td>
</tr>
<tr>
<td>that</td>
<td>877</td>
<td>complementizer, demonstrative pronoun</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>I</td>
<td>783</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>his</td>
<td>772</td>
<td>(possessive) pronoun</td>
</tr>
<tr>
<td>you</td>
<td>686</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>Tom</td>
<td>679</td>
<td>(proper) noun</td>
</tr>
<tr>
<td>with</td>
<td>642</td>
<td>preposition</td>
</tr>
</tbody>
</table>
What we see...

- Dominated by function words (determiners, prepositions,...)
- Content-dependent words (*Tom*)
- How representative is that?
There are **71,370** word tokens. Is that enough to collect statistics on?

The text takes 0.5 MB (500k characters) is a very small corpus.

There are only **8018** different words.

There is a ratio of 1:9 between word types and tokens.

This is a small ratio – this is a children’s book. For news corpora we would have 1:6.
Two important issues to address here are:

- Does the ratio, #(types)/#(tokens) depend on the size of the corpus?

- Here, on average, words occur 9 times. But, what is the distribution? It turns out that word types have very uneven distribution.
What is the distribution of words?

<table>
<thead>
<tr>
<th>frequency</th>
<th>num of words with this 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>
<tr>
<td>&gt;100</td>
<td>102</td>
</tr>
</tbody>
</table>
If we look more carefully on the data we can see that:

- The most common 12 words (over 700 occurrences) account for 1% of the text.
- The most common 100 words account for 50.9% of the text.
- Almost half (3993/8018=49.8%) of the words occur once!
- Over 90% of the word types occur less than 10 times. (only 540+99+102=741 occur > 10 times)
<table>
<thead>
<tr>
<th>word</th>
<th>freq</th>
<th>rank</th>
<th>freq x rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
<td>3332</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>2</td>
<td>5944</td>
</tr>
<tr>
<td>a</td>
<td>1773</td>
<td>3</td>
<td>5235</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>10</td>
<td>8770</td>
</tr>
<tr>
<td>but</td>
<td>410</td>
<td>20</td>
<td>8400</td>
</tr>
<tr>
<td>be</td>
<td>294</td>
<td>30</td>
<td>8820</td>
</tr>
<tr>
<td>but</td>
<td>410</td>
<td>20</td>
<td>8400</td>
</tr>
<tr>
<td>be</td>
<td>294</td>
<td>30</td>
<td>8820</td>
</tr>
<tr>
<td>there</td>
<td>222</td>
<td>40</td>
<td>8880</td>
</tr>
<tr>
<td>one</td>
<td>172</td>
<td>50</td>
<td>8600</td>
</tr>
<tr>
<td>never</td>
<td>124</td>
<td>80</td>
<td>9920</td>
</tr>
<tr>
<td>turned</td>
<td>51</td>
<td>200</td>
<td>10200</td>
</tr>
<tr>
<td>name</td>
<td>21</td>
<td>400</td>
<td>8400</td>
</tr>
<tr>
<td>friends</td>
<td>10</td>
<td>800</td>
<td>8000</td>
</tr>
<tr>
<td>brushed</td>
<td>4</td>
<td>2000</td>
<td>8000</td>
</tr>
<tr>
<td>applauded</td>
<td>1</td>
<td>8000</td>
<td>8000</td>
</tr>
</tbody>
</table>
First Zipf’s Law (1929)

Let

- $f$ be the frequency of a word type in a large corpus (# of occurrences)
- $r$ be the position of the word in a list of words ranked according to frequency.

Then, $f$ is proportional to $1/r$. Equivalently, there exists some constant $K$, s.t

$$f \cdot r = K$$
Zipf-Mandelbrot’s Law

\[ f = \frac{h}{(r + q)^s} \]

- \( h, q, s \) depends on type of text (corpus), language, etc
- Fits better low ranks
Effort minimization

(Zipf) Both speaker and listener are trying to minimize their effort:

- The *speaker*: by using a small vocabulary of common words
- The *hearer*, by having a large vocabulary of rare words, to reduce ambiguity.

Argued as a very general principle
Regardless of arguments’ validity, important:
- For most words, our data about their use will be exceedingly sparse.
Universality of Zipf’s law

Other examples of applicability of Zipf’s law:
- Chords in music
- Income
- Views at youtube.com

Lesson: not all the empirical distribution are so light-tail as the Normal distribution

But why Zipf’s law?
Is Zipf’s law so surprising?

Generate words according to the following model:

Uniformly choose one of 27 characters (26 letters + blank).

\[
\text{Prob}[\text{word of } n \text{ characters is generated}] \approx= \frac{1}{27} \left(\frac{26}{27}\right)^n
\]

We get Zipf-Mandelbrot’s law:

- 26 times more words of length \( n + 1 \) then words of length \( n \), and
- words of length \( n \) are more frequent than words of length \( n+1 \).
Other Zipf’s laws

- 2nd Law about *word meaning*: If \( m \) is the number of meanings a word has then \( m \) is proportional to the square root of \( f \).
  - words of rank 10,000 average 2.1 meanings, words of rank 5,000 average 3 meanings, words of rank 2,000 average 4.6 meanings and so on. \((m \text{ behaves like } \sqrt{f} \approx 1/\sqrt{r})\)

- 3rd Law about distances between words in text:
  - Measure, for each content word, the number of words between consecutive occurrences of them in the text.
  - If \( F \) is the frequency of intervals lengths, and the interval length is \( I \), then \( F \) is proportional to the inverse of \( I \).
  - That is, content words tend to occur near each other.

- There exist other Zipf’s laws...
Should we move beyond words?

- Let us try to learn about the language by acquiring statistics from longer sequences of words.
- A *Collocation* is a phrase which is formed by a combination of “parts” and that has an existence beyond the sum of its parts. Examples may include:
  - Compounds. e.g., disk drive or
  - Phrasal verbs e.g., make up or
  - Phrases e.g., bacon and eggs.
- Any phrase that people repeat because they have heard others using it, is a candidate for a collection.
More about Collocations

- Important in Machine Translation and Information Extraction since their meaning, in most cases, is not composed from the meaning of the words they are formed of.
- Collocations can be long and discontinuous (e.g., put [something] on).
- The problem is that the definition we gave is not constructive
  - What if we look just in frequency of all the bigrams?
<table>
<thead>
<tr>
<th>Frequency</th>
<th>Word 1</th>
<th>Word 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>80871</td>
<td>of</td>
<td>the</td>
</tr>
<tr>
<td>58841</td>
<td>in</td>
<td>the</td>
</tr>
<tr>
<td>26430</td>
<td>to</td>
<td>the</td>
</tr>
<tr>
<td>21842</td>
<td>on</td>
<td>the</td>
</tr>
<tr>
<td>21839</td>
<td>for</td>
<td>the</td>
</tr>
<tr>
<td>18568</td>
<td>and</td>
<td>the</td>
</tr>
<tr>
<td>16121</td>
<td>that</td>
<td>the</td>
</tr>
<tr>
<td>15630</td>
<td>at</td>
<td>the</td>
</tr>
<tr>
<td>15494</td>
<td>to</td>
<td>be</td>
</tr>
<tr>
<td>13899</td>
<td>in</td>
<td>a</td>
</tr>
<tr>
<td>13689</td>
<td>of</td>
<td>a</td>
</tr>
<tr>
<td>13361</td>
<td>by</td>
<td>the</td>
</tr>
<tr>
<td>13183</td>
<td>with</td>
<td>the</td>
</tr>
<tr>
<td>12622</td>
<td>from</td>
<td>the</td>
</tr>
<tr>
<td>11428</td>
<td>New</td>
<td>York</td>
</tr>
<tr>
<td>10007</td>
<td>he</td>
<td>said</td>
</tr>
</tbody>
</table>
Observations

- Frequent bigrams extremely common words.
  - Follow the form [preposition, determiner].
- Hard to say that we have gained something by looking at these pairs.
What we can do ...

- Take into account the frequency of each of the words.
- Remove POS sequences that are not interesting (e.g., [preposition, determiner])
- Keep only POS sequences of interest (e.g., [adj-noun], [noun-noun])

But, doing it this way, we need to be able to tag the text for part-of-speech. Only in order to gather reasonable statistics.
<table>
<thead>
<tr>
<th>Frequency</th>
<th>Word 1</th>
<th>Word 2</th>
<th>POS pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>11487</td>
<td>New</td>
<td>York</td>
<td>AN</td>
</tr>
<tr>
<td>7261</td>
<td>United</td>
<td>States</td>
<td>AN</td>
</tr>
<tr>
<td>5412</td>
<td>Los</td>
<td>Angeles</td>
<td>NN</td>
</tr>
<tr>
<td>3301</td>
<td>last</td>
<td>year</td>
<td>AN</td>
</tr>
<tr>
<td>3191</td>
<td>Saudi</td>
<td>Arabia</td>
<td>NN</td>
</tr>
<tr>
<td>2699</td>
<td>last</td>
<td>week</td>
<td>AN</td>
</tr>
<tr>
<td>2514</td>
<td>vice</td>
<td>president</td>
<td>AN</td>
</tr>
<tr>
<td>2378</td>
<td>Persian</td>
<td>Gulf</td>
<td>AN</td>
</tr>
<tr>
<td>2161</td>
<td>San</td>
<td>Francisco</td>
<td>NN</td>
</tr>
<tr>
<td>2106</td>
<td>President</td>
<td>Bush</td>
<td>NN</td>
</tr>
<tr>
<td>2001</td>
<td>Middle</td>
<td>East</td>
<td>AN</td>
</tr>
<tr>
<td>1942</td>
<td>Sadam</td>
<td>Hussein</td>
<td>NN</td>
</tr>
<tr>
<td>1867</td>
<td>Soviet</td>
<td>Union</td>
<td>AN</td>
</tr>
<tr>
<td>1850</td>
<td>White</td>
<td>House</td>
<td>AN</td>
</tr>
<tr>
<td>1633</td>
<td>United</td>
<td>Nations</td>
<td>AN</td>
</tr>
<tr>
<td>1337</td>
<td>York</td>
<td>City</td>
<td>NN</td>
</tr>
</tbody>
</table>
Most words are content words (Multi-Word expression)
Note: this appear hard even to collect statistics
Anyway, can we estimate probability of ngram?
**Sparsity Issues**

- How frequently words occur in 1M of text?
  - *kick*: 58
  - *ball*: 87
- How frequently *kick a ball* occurs?
  - 0!
- There is no data like more data
  - We need data from the domain of interest!
  - I hope to cover domain-variable towards the end of the course.
What if we use the web as corpus?

- *kick*: 130,000,000 occurrences
- *ball*: 350,000,000 occurrences
- em kick a ball: 171,000 occurrences
Brown Corpus

- Early corpus (a paper about the corpus in the recommended list)
- Balanced (in some way)
  - Press, Reviews, Religion, Humor, Love stories, etc
- 1 mln words
- Small by modern standards
Examples of More Recent Corpora

- British National Corpus (BNC): 600 mln words (balanced)
- Wall Street Journal (WSJ): 2 mln (newswire)
- Switchboard: 2.4 mln (spoken)
- Many others exist, some free, some not (Reuters, EU proceedings, etc)
Recent Trends

- Use of Web as Corpora (e.g., Lapata and Keller, 2002)
  - Google N-grams: 1,024,908,267,229 words were processed
- Use of Wikipedia (and available structured labels and relations)
Types of Corpora

- Different languages
- Different genres

Annotation:
- Tokenization
- Syntactic: part-of-speech (POS) tags, full syntactic trees
- Semantic annotation: predicate-argument structure
- Discourse: relations between sentences
- Translations: parallel corpora
Annotations are not Compatible

- Different granularity:

- Different Formalisms:
  - E.g., for syntax: constituent formalism for Penn Treebank vs. dependency formalism of Prague Treebank

- Possible to convert but hard
  - Competitions and shared tasks bring treebanks in similar format (e.g., CoNLL 2006-07: multilingual dependency parsing, CoNLL 2009: multilingual semantic role-labeling
Critical Reviews

- First critical reviews are due on Feb, 13
- Requirements will be on the web this week
- Register for at least one talk by next Wednesday
- I may not be able to give exact dates for remote dates (need some flexibility in changing the set of lectures)