Lecture 8: Corpora

16 Jan 2004
Scribe: Min-Jung Kang

1 Corpora

Use corpora to give lots of data and general syntactic. A corpus is a special collection of text collected according to certain set of criteria. It is important that the corpus is a representative sample of the population of interest.

A good way to easily and quickly check a large amount of words is using a large hash table. It allows one to see efficiently whether the word/token had been seen before to count up.

Balanced corpora are put together so as to give each subtype of text a share of the corpus that is proportional to some predetermined criterion of importance.

Looking at text corpora and analyzing data programs can be used: Text editor: Emacs (for Unix) TextPad(for Windows) BBEdit(for Macintosh)

In corpora, it is important to what it representative and through Statistical NLP work. Regular expression is most general sequence matching and this pattern describe is called regular language. Regular expression can be used in any plain text editors and built-in or libraries in program languages (Perl, C, C++, etc.). To deal with large data and processing it in fast, using C/C++ had been favored.

1.1 Evaluating Usefulness

1.1.1 Size (contains lots of data)

<table>
<thead>
<tr>
<th>Annotated 1M words</th>
<th>Minimum reasonable size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un-annotated 30M words</td>
<td></td>
</tr>
<tr>
<td>Internet bazillions words</td>
<td>Very noisy and disorganized</td>
</tr>
</tbody>
</table>
1.1.2 Quality

Things to look for high quality in Corpora include consistency and balance. Consistency in information, style, and type. Balance (“representation”) is depended on what Corpora are going to be used for. Example: Applying words from Wall Street Journal to write Moby Dick

- Annotated 1M words - Highest quality
- Un-annotated 30M words - Middle quality
- Internet bazillions words - Low quality

2 Tokenization (define a word)

A token ≈ a word/ a number/ a punctuation mark

Tokenizing is a process of breaking down input text into units. Tokenize in natural language.

2.1 Space (separation)

's He’s (pronoun he) (VBS is)
  John’s

n’t wouldn’t
  can’t

. Mr.
  I love the U.S. (word + . )
  U.S. +
  U.S + .
  U.S. +

$12.47 “Twelve dollars and forty-seven cents”
  $ + 12.27
  $ + 12 + . + 27
2.2 Hyphen

- Used in example
- Compound words Disk-drive, New York
- Phone number 123-4567
- Chaining thoughts (dash)
- Line breaks language.
- Link words five-year-old recently-modified

2.3 Are they same word?

<table>
<thead>
<tr>
<th></th>
<th>U.S</th>
<th>U.S.</th>
<th>US</th>
<th>(.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fire-fighter</td>
<td>firefighter</td>
<td>(hyphen)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good-bye</td>
<td>good bye</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e-mail</td>
<td>email</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Was</td>
<td>was</td>
<td>(capital and lower case)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td>brown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n’t</td>
<td>not</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ca</td>
<td>can</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hit</td>
<td>hits</td>
<td>(same words)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lying</td>
<td>lie</td>
<td>(one for two meaning)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.4 Sentences

How to define a sentence? A sentence is not always end with punctuation marks (., ? !).

- ) Sentence inside parenthesis
- ” Sentence inside quotation
- “you,” he said, “are bad.”
- U.S. Can’t make general rule that period means end of the sentence.
- Mr. Roger
- Yahoo! News
3 Mark Up

Markup is a term for putting codes of some sort into a computer file that are not actually part of the text in the file, but explain something of the structure or formatting of that text.

Sentence and paragraph is text markup and other busy markup is the fully syntactic structure in corpora. Most common grammatical markup is coding of words for part of speech.

Common form of grammatical markup is indicating the part of speech is words by adding a part of speech tag to each word.

Tags indicated by devices such as following each word by a slash or underline and then a short code naming the part of speech.

XML

<s> ... </s>
<s id="3.47"> ... </s>

Part of speech tags (see p141–142)

Designing part of speech tag for fuzzy cases -ing (verb or noun?)