Corpora and Linguistic Annotation

L645
Dept. of Linguistics, Indiana University
Fall 2015

Outline

The goal today is to get us all on the same page about where linguistic data comes from

- ... since we’ll be assuming, and at times using, such data & annotation all semester

Outline

- Corpus design
- Preprocessing
- Annotation & annotation schemes
  - (& linguistic distinctions therein)
- Mark-up encoding

If you enjoy today’s material, consider taking L615, Corpus Linguistics

Corpus design

Representativeness

Representativeness: the extent to which a sample includes the full range of variability in a population

- Distinguishes corpora from archives
- Allows findings to be generalized to a particular variety of language

Corpus: sample of language use (i.e., from a particular population)

- Balance: types of genres
- Sampling: how the text is selected

Balance

What should be covered in a balanced corpus?

- Balanced: covers a range of text categories
  - Definition depends upon the intended uses
  - No real objective measure of balance
  - Usually based on proportional sampling

Sampling

To achieve representativeness & balance, one samples language (since it cannot be exhaustively described)

- The sample should be representative of the larger population

To properly sample, need to define:

- Sampling unit: book, periodical, newspaper, articles, chapters, ...
- Sampling frame: list of all possible units, from which that actual ones are selected
  - Brown sampling frame: list of books & periodicals in Brown University Library & Providence Athenaeum
  - (written English text published in 1961)

Corpus size

Generally speaking, the bigger the corpus the better, but ...

- For annotation, text cannot be as big
  - e.g., not possible to have 100 billion words with high-quality linguistic annotation without many errors
Some issues arise as the corpus is being built:

- **Segmentation:** what counts as a sentence for this corpus?
- **Tokenization:** what counts as a word for this corpus?
- **Lemmatization:** do we include word stems for each word in the corpus?

### Heuristic sentence boundary detection

Manning & Schütze, figure 4.1

- Place putative sentence boundaries after all occurrences of . ? ! (and maybe ; : —)
- Move boundary after following quotation marks, if any
- **Disqualify a period boundary if:**
  - it is preceded by a known abbreviation and not followed by an uppercase word (e.g., etc.)
  - it is preceded by a known abbreviation which is commonly followed by a capitalized word (e.g., Prof.)
- **Disqualify a boundary with a ? or ! if:**
  - it is followed by a lowercase letter (or a known name)
- Regard other putative sentence boundaries as sentence boundaries.

### Preprocessing

- **Lemmatization**
  - Corpora often also contain the lemma for each word, the base form or stem of a word
  - Need to be able to do this accurately: e.g., what is the lemma of leaves? leaf or leave?

**Stemming** is a similar process popular in Information Retrieval (IR)
- Instead of finding the dictionary form, simply strip off affixes
One of the most common forms of annotation is that of part-of-speech classes, or grammatical classes:

- Part-of-speech (POS) tags are a combination of morphological and syntactic information.
- Occasionally require semantic distinctions to be made.
- Tags differ in size and granularity, and there is not always a way to map between tags.

**Brown corpus example**

```
[ The/DT Fulton/NNP County/NNP Grand/NNP Jury/NNP ]
said/VBD [ Friday/NNP ]
[ an/DT investigation/NN ]
of/IN [ Atlanta/NNP 's/POS recent/JJ primary/JJ election/NN ]
produced/VBD ' ' '/'
[ no/DT evidence/NN ]
''/'' that/IN [ any/DT irregularities/NNS ]
took/VBD [ place/NN ]
./.
```

**Comparison of different tagsets**

Manning & Schütze, figure 4.2

<table>
<thead>
<tr>
<th>Sentence</th>
<th>CLAWS c5</th>
<th>Brown</th>
<th>PTB</th>
<th>ICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>she</td>
<td>PNP</td>
<td>PPS</td>
<td>PRP</td>
<td>PRON(pers,sing)</td>
</tr>
<tr>
<td>was</td>
<td>VBD</td>
<td>BEDZ</td>
<td>VBD</td>
<td>AUX(pass.part)</td>
</tr>
<tr>
<td>told</td>
<td>VVN</td>
<td>VBN</td>
<td>VBN</td>
<td>V(ditr,edp)</td>
</tr>
<tr>
<td>that</td>
<td>CJT</td>
<td>CS</td>
<td>IN</td>
<td>CONJUNC(subord)</td>
</tr>
<tr>
<td>the</td>
<td>AT0</td>
<td>AT</td>
<td>DT</td>
<td>ART(def)</td>
</tr>
<tr>
<td>journey</td>
<td>NN1</td>
<td>NN</td>
<td>NN</td>
<td>N(com,sing)</td>
</tr>
<tr>
<td>might</td>
<td>VM0</td>
<td>MD</td>
<td>MD</td>
<td>AUX(modal,past)</td>
</tr>
<tr>
<td>kill</td>
<td>VVI</td>
<td>VB</td>
<td>VB</td>
<td>V(montr,infin)</td>
</tr>
<tr>
<td>her</td>
<td>PNP</td>
<td>PPO</td>
<td>PRP</td>
<td>PRON(poss,sing)</td>
</tr>
<tr>
<td>.</td>
<td>PUN</td>
<td>.</td>
<td>.</td>
<td>PUNC(per)</td>
</tr>
</tbody>
</table>

**Syntactic annotation**

- A common layer above POS is that of syntactic (phrase structure) annotation.
- Corpus with syntactic annotation is often called a treebank.
- Again, you may have to consult the guidelines for the annotation scheme.
- Treebanks generally shoot for “theory-neutrality”.
- Dependency & function labeling are also quite common.

**Phrase structure tree**

```
S
  /----\   
 NP    VP
    /     |
   D     N
     /   |
    V    PP
     /   |
    P    NP
     /   |
    to   N
      |
      heaven
```

**Dependency tree**

```
ROOT
   /----\   
 DET  SUBJ  JCT  PDIR
    /     |
   All  dogs  to  heaven
      |
      D    N   V   P   N
```
One popular form of semantic annotation is to label
sentences with verbal semantic relations and the
arguments of that relation.

There are different kinds of semantic/pragmatic
annotation one could in principle have:

- Word sense annotation, anaphoric relations, etc.
- One popular form of semantic annotation is to label
sentences with verbal semantic relations and the
arguments of that relation.

PropBank example (standoff)

Semantic annotation

File it refers to

Brown part of PTB
To add linguistic properties to a corpus requires mark-up

- XML (eXtensible Markup Language)
- In contrast to HTML, XML does not have built-in “meaning” for labels: you must define your own tags

Benefits of XML:
- Pretty standard internationally and across fields
- Good tools available & convenient for a lot of programs
- Widely used in real text processing
- Handles multilinguality very well

XML tags
Several things to note in the previous example:
- Needs to be one root tag, e.g., `<corpus>` in this case
- Every (opening) tag needs a closing tag:
  - `<t>` is legitimate;
  - `<t>` is not valid

So, structure ends up like:
```xml
<root>
  <child>
    <subchild>.....</subchild>
  </child>
</root>
```

Attributes
Each tag (or element) can have a variety of attributes (provided such attributes have been declared; see below)
- Attributes are noted within an element tag; there can be multiple attributes & can be put in different orders
- Attribute values are put in quotes after the attribute
  ```xml
  <t XYZ />
  ```
  is the same as
  ```xml
  <t XYZ></t>
  ```

Child & Attribute possibilities
```
<terminal tag="NN">dog</terminal>
<terminal word="dog" tag="NN"></terminal>
<!-- or: <terminal word="dog" tag="NN"/> -->
<terminal>
  <word>dog</word>
  <tag>NN</tag>
</terminal>
```

XML for “raw” text
```
<DOC id="NYT19940701.0001" type="story" >
  <HEADLINE>
    WITNESS SAYS O.J. SIMPSON BOUGHT KNIFE WEEKS BEFORE SLAYINGS
  </HEADLINE>
  <DATELINE>
    LOS ANGELES (BC-SIMPSON-KILLINGS-1stLd-3Takes-Writethru-LADN)
  </DATELINE>
  <TEXT>
    With the nation’s attention riveted again on a Los Angeles courtroom, a knife dealer testified that O.J. Simpson bought a 15-inch knife five weeks before the slaying deaths of his ex-wife and her friend.
    ...''She frequented the restaurant quite often,'’ DeBello said.
    (STORY CAN END HERE. OPTIONAL 2ND TAKE FOLLOWS.)
  </TEXT>
</DOC>
```
Follow-up material we'll zip through:
http://cl.indiana.edu/~md7/14/715/slides/03-nlp/03-nlp.pdf

And some material to specifically help us with the Stanford CoreNLP tools: http://cl.indiana.edu/~md7/14/715/slides/03-nlp/03b-practical.pdf