Corpus-Based Work

(M&S Ch 4)

Corpus-Based Work

- Text Corpora are usually big. They also need to be representative samples of the population of interest.
- Corpus-Based work involves collecting a large number of counts from corpora that need to be access quickly.
- There exists some software for processing corpora (see useful links on course homepage).

Looking at Text (I)
Low-Level Formatting Issues

- ***Junk formatting/Content***. Examples: document headers and separators, typesetter codes, table and diagrams, garbled data in the computer file. Also other problems if data was retrieved through OCR (unrecognized words). Often one needs a filter to remove junk content before any processing begins.
- ***Uppercase and Lowercase:*** should we keep the case or not? *The*, *the* and THE should all be treated the same but "brown" in "George Brown" and "brown dog" should be treated separately.

Looking at Text (II): Tokenization
What is a Word?

- An early step of processing is to divide the input text into units called ***tokens*** where each is either a word or something else like a number or a punctuation mark.
- Periods: ***haplologies*** or end of sentence?
- Single apostrophes
- Hyphenation
- Homographs --> two lexemes

Looking at Text (III): Tokenization
What is a Word (Cont'd)?

- Word Segmentation in other languages: no whitespace ==> words segmentation is hard
- whitespace not indicating a word break.
- variant coding of information of a certain semantic type.
- Speech corpora.

Morphology

- Stemming: Strips off affixes.
- Lemmatization: transforms into base form
- Not always helpful in English (from an IR point of view) which has very little morphology.
- Perhaps more useful in other contexts.

Sentences: What is a sentence?”

- Something ending with a ‘.’, ‘?’ or ‘!’. True in 90% of the cases.
- Sometimes, however, sentences are split up by other punctuation marks or quotes.
- Often, solutions involve heuristic methods. However, these solutions are hand-coded. Some effort to automate the sentence-boundary process have also been done.

End-of-Sentence Detection (I)

- Place EOS after all . ? ! (maybe ;:-)
- Move EOS after quotation marks, if any
- Disqualify a period boundary if:
  - Preceeded by known abbreviation followed by upper case letter, not normally sentence-final: e.g., Prof. vs. Mr.

End-of-Sentence Detection (II)

  - Precedeed by a known abbreviation not followed by upper case: e.g., Jr. etc. (abbreviation that is sentence-final or medial)

- Disqualify a sentence boundary with ? or ! If follwed by a lower case (or a known name)
- Keep all the rest as EOS

Marked-Up Data I: Mark-up Schemes

- Schemes developed to mark up the structure of text
- Different Mark-up schemes:
  - COCOA format (older, and rather ad-hoc)
  - SGML [other related encodings: HTML, TEI, XML]

Marked-Up Data II: Grammatical Coding

- Tagging indicates the various conventional parts of speech. Tagging can be done automatically (we will talk about that in Week 9).
- Different Tag Sets have been used: e.g., Brown Tag Set, Penn Treebank Tag Set.
- The Design of a Tag Set: Target Features versus Predictive Features.