CSE574/EE517 – Lecture 2, Part I

• Data-Issues (cont.): Text PreProcessing

• Statistical language models
  – Preliminaries
  – N-gram models
Text-Preprocessing Concerns

• Data format (getting raw data into the form you need for your tools)
• Tokenization (what counts as a “word”)
• Segmentation (what is a sentence? (or paragraph or ....))
• Text normalization (orthographic form of the word)
Data Format

• Character encoding
• Document markup
  – Document headers, separators
  – Tables, figures and captions
  – URLs, font
• Alignment of multimodal data
Tokenization in English

- Acronyms: I.B.M. vs. IBM vs. I B M
- Hyphenated words: best-seller vs. best seller vs. bestseller
- Punctuation: treat as... a “word,” a feature of a word, ignore
- Multi-word lexical items: San Francisco, you know
- Contractions and possessives: can’t vs. can +not; boy’s vs. boy +possessive
- Fillers (uh, um, hmmm) and word fragments (fl-)
Tokenization in Other Languages

• English: in the country
• Hebrew: בארץ

Easy task in English: spaces separate words. Challenging for Semitic languages, Chinese, ....
Tokenization (cont.)

• Why does tokenization matter?
  – A system trained with one tokenization will have trouble with another (vocabulary mismatch)
  – Two annotations with different tokenizations are hard to align

• Why not just work with characters?
  – Bigger units give better results (in theory & practice)
  – Smaller units are more efficient (more frequent so easier to train, smaller space to search)
Sentence Segmentation

Why is sentence segmentation non-trivial?

• Ambiguity, e.g. for English
  – Period in written text (abbreviation vs. punctuation)
  – Capitalization (sentence start vs. proper noun)
• Different conventions for using punctuation
• Definition of a “sentence” is not straightforward for speech
  – People are not always “grammatical”
  – Automatic detection is non-trivial
• Web text sometimes has characteristics of speech (lack of capitalization, abbreviated forms)
Sentence Segmentation (cont.)

Why does sentence segmentation matter?

• Sentence boundaries (like word boundaries) have information
• Sentences provide more manageable chunks for language processing
• Computer processing of larger units can run into memory problems
• Computer models trained with sentences work better on sentences (e.g. parsing, MT, ….)
• People are used to a sentence-based presentation.
Text Normalization

- **Case:** might want to treat (the, The, THE) as all the same thing, but brown (color) vs. Brown (name) as different.
- **Acronyms:** IBM vs. International Business Machines
- **Numbers:** e.g. 219 → <number>, <ID_number>, two hundred nineteen, two nineteen
  - Aside: collapsing of categories may be useful for non-numeric words as well, e.g. Seattle → <city>
- **Other characters:** money, :) etc.
- **Abbreviations:** St. → street, saint, Dr. → drive, doctor
Case in Other Languages

Named Entity Detection
• English: She saw Jacob ...
• Hebrew: יעקב את ראתה היא

Easy task in English: capitalization is a strong hint. Challenging for Semitic languages, Chinese...