Natural Language Processing

CORPUS-BASED WORK

Updated 05/09
Corpora

- Large databases of text, speech.
- Many types of text corpora exist – plain text, domain specific, tagged, parallel bi-lingual...
- This data allows us to use statistically based techniques to derive the needed probabilities.
- Thus, it needs to be a representative sample of the population of interest.
Formatting Issues

- **Cleaning**: removal of HTML tags, diagrams, tables, foreign words etc.

- **Uppercase/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.
Formatting Issues: Tokenization and Sentences.

- **Form tokens**: divide the input text into units called tokens where each is either a word or something else like a number or a punctuation mark.

- **Mark sentence boundaries**: Can be confused by abbreviations. Most sentences end with ‘.’, ‘?’ or ‘!’.
Formatting Issues: Abbreviations and Morphology

- **Expanding abbreviated words**: J, Jan. or Jan all to January.

- **Morphology**
  - **Stemming**: Strips off affixes and leaves a stem.
  - happy (happy), happier (happy + er), happiest (happy + est).
  - But seed is not see or se + ed.
Example of stemming

raw = ['caresses', 'flies', 'dies', 'mules', 'denied', 'died', 'agreed', 'owned', 'humbled', 'sized', 'meeting', 'stating', 'siezing', 'itemization', 'sensational', 'traditional', 'reference', 'colonizer', 'plotted'];

stemmed = ['caress', 'fli', 'die', 'mule', 'deni', 'die', 'agre', 'own', 'humbl', 'size', 'meet', 'state', 'siez', 'item', 'sensat', 'tradic', 'refer', 'colon', 'plot']
Application Specific Formatting Issues

- **Mark Headings separately/Retain information on size of font:** Search Engines may need this.

- **Aligning parallel corpora.** In machine translation this is essential.
Using a Corpus

- There is a lot of information in the relationships between words. The meaning of a word could be known by the company it keeps.
- Statistical NLP approach seeks to automatically learn lexical and structural preferences from corpora.
Using a Corpus

- **Word Counts:**
  - The most common words in the text.
  - How many words are in the text (word tokens and word types).
  - What the average frequency of each word in the text is.

- **Limitation of word counts:** Most words appear very infrequently and it is hard to predict much about the behavior of words that do not occur often in a corpus.
The Distribution of Words in a Text: Zipf’s Law

- **Zipf’s Law** says that:  \( f \propto 1/r \)
- Zipf’s Law explores the relationship between the frequency of a word, \( f \), and its position in the list, known as its rank, \( r \).
- Significance of Zipf’s Law: For most words, our data about their use will be exceedingly sparse. Only for a few words will we have a lot of examples.
Other things we can Learn from Corpora

- **Collocations**: Certain words co-occur.
- These words together can mean more than their sum of parts (*The Times of India, disk drive*)
- Collocation can be extracted from a text (example, the most common **bigrams** can be extracted).
- Many bigrams are often insignificant (e.g., “at the”, “as a”), they can be **filtered**.
Other Things we can Learn (Cont.)

- **Concordances**: The different contexts in which a given word occurs.