Statistical Natural Language Processing:

Corpus Tagging and Collocations

Steve Engels, January 31, 2000

Topics

- Marked-up Data (4.3)
  - Tagging Parts of Speech (POS)
  - Tag Sets
- Collocations (5.1-5.3)
  - Definition
  - Finding Collocations
    - Frequency
    - Mean & Variance
    - Hypothesis Testing
      - T-test
      - Chi-square test
      - Likelihood Ratios
      - Relative Frequency Ratios

Marked-up Data

- Markup = Adding structure to the text of a corpus
- Different Levels of structure:
    - sentence & paragraph boundaries
    - full syntactic structure (e.g. Penn Treebank, Susanne corpus)
    - parts of speech (most common)

Tagging Parts of Speech (POS)

- Tags are added after each word to denote parts of speech
- Most common form: SGML (Standard Generalized Markup Language)
  - metalanguage which defines grammar for markup
  - HTML is a language encoded with SGML
  - Document Type Definition defines grammar for legal structures in document
    - e.g. paragraphs must contain one or more sentences

Tag Sets

- Major Corpora (in order of size):
  - London-Lund 197
  - CLAWS2 166
  - CLAWS1 132
  - Brown * 87
  - CLAWS c5 62
  - Penn Treebank * 45
- Words assigned tags corresponding to a category
  - noun, proper, singular: "London", "Michael",...
  - verb, auxiliary be, present, 3SG: "is"

Tag Sets (cont'd)

- Categories can be broken down into more specific subcategories (I.e. plural, subject, possessive)
- Negatives add *
  - e.g. "**is**" : BEZ --> "**isn't**" : BEZ*
- Combined tags joined by +
  - e.g. "**she'll**" : PPS + MD
- Possessives add $
  - e.g. "**children's**" : NNS$
- Foreign words add "FW-" prefix

Tag Sets (cont'd)

- Tags sets for different corpora are different due to the size of the sets, and the interpretations of important categories
- Tags should encode info about grammatical features of the category, and how it affects other words.
- Tagging conflicts occue between *semantic* categorization, *syntactic* categorization and *morphological* categorization

Chap. 5: Collocations

- Collocation = "Two or more words combined together to express a conventional way of saying things"
- Nouns: "strong tea", "time of day"
- Verbs: "to put up", "to kick the bucket"
- Others: "the rich and famous"

Collocations (cont'd)

- Collocations have limited compositionality
  - meaning of expression cannot be completely predicted from meaning of parts
- *Technical terms* and *terminological phrases*
  - combine words to define technical phrases (e.g. fuzzy logic)
  - overlap with domain of collocations
- Collocations affected by context
    - social setting
    - spoken discourse
    - surrounding structure)
- Some collocations are more grammatical than others (e.g. strong tea vs. powerful drugs)

Finding Collocations: Frequency

- Collocations found by counting bigrams and trigrams that occur most often
- Must be filtered to avoid uninteresting results
  - e.g. “of the”, “in the”, “to the”, etc.
- Filter using parts of speech tags
  - AN, AAN, NAN, etc.
- Filters illustrate different collocation contexts for “strong” (intrinsic qualities) and “powerful” (extrinsic qualities)

Finding Collocations: Mean & Variance

- Frequency approach works for fixed phrases
- Phrases like, "I knocked … door" relate *knock* to *door*, despite their separation
- Mean and variance are used to characterize distances between words and deviations from the mean
- Can still find local pairings ($\mu = 1.0$, $\sigma = 0$)
- Low deviation indicates an interesting phrase
    - e.g. "hundreds"/"dollars": "hundreds of Canadian dollars"
- Means for unrelated words = 0

Finding Collocations: Hypothesis Testing

- High frequency and low variance can be accidental
- Hypothesis testing
  - calculate probability of event based on hypothesis that components of event are unrelated
  - reject hypothesis if calculated probability is too low
    - I.e. $P(w^1w^2) > P(w^1)P(w^2)$

Hypothesis Testing (cont'd)

- T-test: $$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$
- Used to determine if difference between sample mean x and normal mean μ is significant enough to not have occurred by chance, given sample size N and sample variance s^2
- Different t values correspond to different confidence levels. T > 2.576 corresponds to 99.5% confidence

Hypothesis Testing (cont'd)

- Pearson's chi-square test
  - Chi-square test (X^2) test:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_i j)^2}{E_{ij}}$$

  - Represents difference between observed and expected means for collocated words together and separately
    - e.g. "new companies", "<non-new> companies", "new <non-companies", "<non-new> <non-companies>"

Hypothesis Testing (cont'd)

- Application: Translation pairs

	cow	!cow
vache	59	6
!vache	8	570934

- $\chi 2$ test yields result of 456400, which is much higher than result obtained for independent words
- Problems occur when numbers in table are small (I.e. when any number in the table is 5 or less)

Hypothesis Testing: Likelihood Ratios

- Propose two hypotheses for bigram w^1w^2:
  - Hypothesis #1 : $P(w^2|w^1) = p = P(w^2|\neg w^1)$
  - Hypothesis #2 : $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$
- If first hypothesis is more accurate, w^2 and w^1 are independent; if second is true, collocations is likely
- Advantages:
  - Better for sparse data
  - More interpretable than χ^2 test

Relative Frequencies

- Ratios used to determine collocations that are characteristic of a single corpus
- Relative frequencies identify the collocations that are subject-specific
  - E.g. New York Times corpus for 1989 vs 1990
  - Ratios not equal to one identify subjects specific to that year (East Berliners, Karim Obeid, Miss Manners)