Introduction to Natural Language Processing I [Statistické metody zpracování přirozených jazyků I] (NPFL067)

http://ufal.mff.cuni.cz/courses/npfl067

prof. RNDr. Jan Hajič, Dr. / doc. RNDr. Pavel Pecina, Ph.D.
ÚFAL MFF UK
{hajic,pecina}@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/jan-hajic
http://ufal.mff.cuni.cz/~pecina/index.html

Words and the Company They Keep

Motivation

- Environment:
  - mostly “not a full analysis (sentence/text parsing)”
- Tasks where “words & company” are important:
  - word sense disambiguation (MT, IR, TD, IE)
  - lexical entries: subdivision & definitions (lexicography)
  - language modeling (generalization, [kind of] smoothing)
  - word/phrase/term translation (MT, Multilingual IR)
  - NL generation (“natural” phrases) (Generation, MT)
  - parsing (lexically-based selectional preferences)

Collocations

- Collocation
  - Firth: “word is characterized by the company it keeps”; collocations of a given word are statements of the habitual or customary places of that word.
  - non-compositionality of meaning
    - **cannot be derived directly from its parts (heavy rain)**
  - non-substitutability in context
    - **for parts (red light)**
  - non-modifiability (& non-transformability)
    - **kick the ~~yellow~~ bucket; take exceptions ~~s~~ to**

Association and Co-occurence; Terms

- Does not fall under “collocation”, but:
- Interesting just because it does often [rarely] appear together or in the same (or similar) context:
  - **(doctors, nurses)**
  - **(hardware,software)**
  - **(gas, fuel)**
  - **(hammer, nail)**
  - **(communism, free speech)**
- Terms:
  - need not be > 1 word (notebook, washer)

Collocations of Special Interest

- Idioms: really fixed phrases
  - **kick the bucket, birds-of-a-feather, run for office**
- Proper names: difficult to recognize even with lists
  - **Tuesday (person's name), May, Winston Churchill, IBM, Inc.**
- Numerical expressions
  - containing "ordinary" words
    - **Monday Oct 04 1999, two thousand seven hundred fifty**
- Phrasal verbs
  - Separable parts:
    - **look up, take off**

Further Notions

- Synonymy: different form/word, same meaning:
  - **notebook / laptop**
- Antonymy: opposite meaning:
  - **new/old, black/white, start/stop**
- Homonymy: same form/word, different meaning:
  - **“true” (random, unrelated): can (aux. verb / can of Coke)**
  - **related: polysemy; notebook, shift, grade, ...**
- Other:
  - **Hyperonymy/Hyponymy: general vs. special: vehicle/car**
  - **Meronymy/Holonymy: whole vs. part: body/leg**

How to Find Collocations?

- Frequency
  - plain
  - filtered
- Hypothesis testing
  - *t* test
  - χ^2 test
- Pointwise ("poor man's") Mutual Information
- (Average) Mutual Information

Frequency

- Simple
  - Count n-grams; high frequency n-grams are candidates:
    - **mostly function words**
    - **frequent names**
- Filtered
  - Stop list: words/forms which (we think) cannot be a part of a collocation
    - **a, the, and, or, but, not, ...**
  - Part of Speech (possible collocation patterns)
    - **A+N, N+N, N+of+N, ...**

Hypothesis Testing

- Hypothesis
  - something we test (against)
- Most often:
  - compare possibly interesting thing vs. “random” chance
  - “Null hypothesis”:
    - **something occurs by chance (that’s what we suppose).**
    - **Assuming this, prove that the probabilty of the “real world” is then too low (typically < 0.05, also 0.005, 0.001)... therefore reject the null hypothesis (thus confirming “interesting” things are happening!)**
    - **Otherwise, it’s possibile there is nothing interesting.**

t test (Student's *t* test)

- Significance of difference
  - compute "magic" number against normal distribution (mean μ)
  - using real-world data: (x' real data mean, s^2 variance, N size):
    - **$t = (x' - \mu) / \sqrt{s^2 / N}$**
  - find in tables (see MS, p. 609):
    - **d.f. = degrees of freedom (parameters which are not determined by other parameters)**
    - **percentile level p = 0.05 (or better)**
  - the bigger t:
    - **the better chances that there is the interesting feature we hope for (i.e. we can reject the null hypothesis)**
    - **t: at least the value from the table(s)**

t test on words

- null hypothesis: independence
  - **mean μ: $p(w_1)\ p(w_2)$**
- data estimates:
  - **x' = MLE of joint probability from data**
  - **s^2 is p(1-p), i.e. almost p for small p; N is the data size**
- Example: (d.f. ~ sample size)
  - **'general term' (homework corpus): c(general) = 108, c(term) = 40**
  - **c(general,term) = 2; expected p(general)p(term) = 8.8E-8**
  - **$t = (9.0E\text{-}6 - 8.8E\text{-}8) / (9.0E\text{-}6 / 221097)^{1/2} = 1.40$ (not > 2.576) thus 'general term' is <u>not</u> a collocation with confidence 0.005**
  - **'true species': (84/1779/9): t = 2.774 > 2.576 !!**

Pearson's Chi-square test

- χ^2 test (general formula): $\Sigma_{i,j} (O_{ij}-E_{ij})^2 / E_{ij}$
  - where O_{ij}/E_{ij} is the observed/expected count of events i, j
- for two-outcomes-only events:

w_{right} \ w_{left}	= true	≠ true
= species	9	1,770
≠ species	75	219,243

$\chi^2 =$ $221097(219243x9-75x1770)^2/1779x84x221013x219318 = 103.39 > 7.88$

(at .005 thus we can reject the independence assumption)

Pointwise Mutual Information

- This is ***NOT*** the MI as defined in Information Theory
  - (IT: average of the following; not of ***values***)
- ...but might be useful:

  $I'(a,b) = \log_2 (p(a,b) / p(a)p(b)) = \log_2 (p(a|b) / p(a))$
- Example (same):

  $I'(true,species) = \log_2 (4.1e\text{-}5 / 3.8e\text{-}4 \times 8.0e\text{-}3) = 3.74$

  $I'(general,term) = \log_2 (9.0e\text{-}6 / 1.8e\text{-}4 \times 4.9e\text{-}4) = 6.68$
  - **measured in bits but it is difficult to give it an interpretation**
  - **used for ranking (~~$\sim$~~ the null hypothesis tests)**