Corpus Linguistics
(L615)
Application #2: Collocations

Markus Dickinson
Department of Linguistics, Indiana University
Spring 2013

Working with collocations

Where we finished:

- What significant collocations are there that start with the word sweet?
- Specifically, what nouns tend to co-occur after sweet?

What do your intuitions say?

Calculating collocations: web interface

Can work with an online concordancer of the BNC, http://corpus.byu.edu/bnc/

1. Enter sweet in the Search String box.
   - Enter * in the COLLOCATE box
   - Change the Sort option to be done by RELEVANCE
     - This calculates & sorts collocates by MI scores
2. On the left side, check COMPARE WORDS
   - Enter sweet and some other word (e.g., sour)
   - This compares the collocates between the two words

Calculating collocations: Perl script

We could write a Perl script which does the following:

1. Reads in a corpus file (could be changed to read over a directory of files, if need be)
2. Stores unigram and bigram counts as it reads the file in
3. Loops over all bigrams
4. For each bigram, calculates the pointwise mutual information score

N-grams Statistics Package (NSP)

“The Ngram Statistics Package (NSP) is a suite of programs that aids in analyzing Ngrams in text files.” (from README)

http://www.d.umn.edu/~tpederse/nsp.html

Two main files:

- count.pl: takes regular text files and generates a list of ngrams & their frequencies
- statistic.pl: takes ngram lists (output from count.pl) & runs a measure of association

Some example cases uses found here:
http://search.cpan.org/~tpederse/Text-NSP/doc/USAGE.pod

- see also the README which comes with the software
**N-grams Statistics Package (NSP)**

**Other programs**:
- `rank.pl`: takes 2 output files from `statistic.pl` & compares the orderings
- `kocos.pl`: kth order co-occurrences of words (e.g., due to New York & New Jack, York & Jack are 2nd order co-occurrences)
- `combigr.pl`: ignores order (e.g., fine wine & wine fine treated the same)
- `huge-count.pl`: for larger corpora

You can type the name of a program with `--help` to see the options & usage notes

---

**NSP**

**Counting: count.pl**

Count bigrams occurring 5 or more times & storing a histogram

```
> count.pl -frequency 5 -hist leaves-5.hist leaves-of-grass.txt
> more leaves-5.hist
```

Total ngrams = 148857
Number of n-grams ... 1 time(s) = 59199 (39.77 percent)
Number of n-grams ... 2 time(s) = 8261 (11.10 percent)
Number of n-grams ... 3 time(s) = 3005 (6.06 percent)
Number of n-grams ... 4 time(s) = 1409 (3.79 percent)

A list of stop words to exclude can be given and indicated with `-stop`

---

**NSP**

**Collocations: statistic.pl**

Score bigram lists: log-likelihood ratios

```
> statistic.pl -score 6.00 -frequency 5 leaves.ll
> more leaves.ll
```

148857,
> The<>1 2815.3163 978 17936 1298
of<>the<>2 2368.0242 1221 4126 8814
I<>see<>3 1868.6223 291 2932 393
,...
these<>States<>42 326.7150 34 241 86
Give<>me<>43 313.4514 32 33 982

NB: there happen to be a lot of ties here
Corpus Linguistics
Application #2:
Collocations

Web interface
Perl
NSP
count.pl
statistic.pl
rank.pl
kocos.pl
UCS

Available bigram measures:
- Dice Coefficient (dice)
- Fishers exact test - left sided (leftFisher)
- Fishers exact test - right sided (rightFisher)
- Fishers twotailed test - right sided (twotailed)
- Jaccard Coefficient (jaccard)
- Log-likelihood ratio (ll)
- Mutual Information (tmi)
- Odds Ratio (odds)
- Pointwise Mutual Information (pmi)
- Phi Coefficient (phi)
- Pearson’s Chi Squared Test (x2)
- Poisson Stirling Measure (ps)
- T-score (tscore)

Comparing methods: rank.pl

rank.pl “computes the Spearman’s rank correlation coefficient on the Ngrams that are common to both files” (from README)

Comparing two collocation files:

> rank.pl -precision 3 leaves.c5.pmi leaves.c5.ll

Rank correlation coefficient = 0.781

Another nice package is the UCS toolkit
- http://www.collocations.de/software.html

There are interfaces for both Perl and R
- The R interface provides more graphical output
- Tutorials & thorough documentation are available for both

UCS

kth order co-occurrences: kocos.pl

> kocos.pl --literal soul --order 2 \
  --trace soul.trace leaves.c5.ll > soul.k2
> more soul.k2

own<> poems<> body<> life<> love<> eyes<> heart<> 

...
Exercises
Testing our intuitions

Some things you can do to test your intuitions about specific collocations:

▶ Try different corpora on miller
▶ Hypothesize better/worse collocations
▶ Try to implement a POS filter in the Perl code
▶ ...

Web interface
Perl
NSP
count.pl
statistic.pl
rank.pl
kocos.pl
UCS