

Statistics used in the Sketch Engine

Lexical Computing Ltd.

January 23, 2012

1 Conventions

N – corpus size,

f_A – number of occurrences of the keyword in the whole corpus (the size of the concordance),

f_B – number of occurrences of the collocate in the whole corpus,

f_{AB} – number of occurrences of the collocate in the concordance (number of co-occurrences)

1.1 With grammatical relations

Terminology follows Dekang Lin, ACL-COLING 1998: “Automatic Retrieval and Clustering of Similar Words.”

We count frequencies for triples of a first word connected by a specific grammatical relation to a second word, written $(word_1, gramrel, word_2)$

$||w_1, R, w_2||$ – number of occurrences of the triple,

$||w_1, R, *||$ – number of occurrences of the first word in the grammatical relation with any second word

$||*, *, w_2||$ – number of occurrences of the second word in any grammatical relation with any first word

$||*, *, *||$ – number of occurrences of any first word in any grammatical relation with any second word: that is, the total number of triples found in the corpus.

2 Word Sketches

Until September 2006 we used a version of MI-Score modified to give greater weight to the frequency of the collocation defined as:

MI-Score

$$\log_2 \frac{f_{ABN}}{f_A f_B}$$

Association score

$$AScore(w_1, R, w_2) = \log \frac{\|w_1, R, w_2\| \cdot \|\ast, \ast, \ast\|}{\|w_1, R, \ast\| \cdot \|\ast, \ast, w_2\|} \cdot \log(\|w_1, R, w_2\| + 1)$$

Since September 2006, noting the scale-dependency of AScore and recent relevant research including Curran 2004 “From Distributional to Semantic Similarity” (PhD Thesis, Edinburgh Univ) we changed the statistic to logDice, based on the Dice coefficient:

Dice

$$\frac{2f_{AB}}{f_A + f_B}$$

logDice

$$14 + \log_2 \frac{2 \cdot \|w_1, R, w_2\|}{\|w_1, R, \ast\| + \|\ast, \ast, w_2\|}$$

3 Thesaurus

To compute a similarity score, we:

- compare w_1 and w_2 ’s word sketches
- ignore contexts that supply no useful information (e. g. Association score < 0)
- find all the overlaps, e. g. where w_1 and w_2 “share a triple” as in *beer* and *wine* “sharing” (*drink, OBJECT, beer/wine*)

$$Dist(w_1, w_2) = \frac{\sum_{(tup_i, tup_j) \in \{tup_{w_1} \cap tup_{w_2}\}} AS_i + AS_j - (AS_i - AS_j)^2 / 50}{\sum_{tup_i \in \{tup_{w_1} \cup tup_{w_2}\}} AS_i}$$

The term $(AS_i - AS_j)^2 / 50$ is subtracted in order to give less weight to shared triples, where the triple is far more salient with w_1 than w_2 or vice versa. We find that this contributes to more readily interpretable results, where words of similar frequency are more often identified as near neighbours of each other.

The constant 50 can be changed using the `-k` option of the `mkthes` command.

4 Other statistics

These are the statistics offered under the “collocations” function accessible from the concordance window; these statistics do not involve grammatical relations.

$$\mathbf{T\text{-Score}} \frac{f_{AB} - \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$$

MI-Score $\log_2 \frac{f_{AB} N}{f_A f_B}$
Church and Hanks, Word Association Norms, Mutual Information, and Lexicography, in Computational Linguistics, 16(1):22-29, 1990

MI³-Score $\log_2 \frac{f_{AB}^3 N}{f_A f_B}$
Oakes, Statistics for Corpus Linguistics, 1998

log-likelihood $2 \cdot (x\log(f_{AB}) + x\log(f_A - f_{AB}) + x\log(f_B - f_{AB}) + x\log(N) + x\log(N + f_{AB} - f_A - f_B) - x\log(f_A) - x\log(f_B) - x\log(N - f_A) - x\log(N - f_B))$
where $x\log(f)$ is $f \ln(f)$
Dunning, Accurate Methods for the Statistics of Surprise and Coincidence, Computational Linguistics 19:1 1993

minimum sensitivity $\min(\frac{f_{AB}}{f_B}, \frac{f_{AB}}{f_A})$
Pedersen, Dependent Bigram Identification, in Proc. Fifteenth National Conference on Artificial Intelligence, 1998

MI-log-prod (formerly called **salience**) $MI - Score \times \log(f_{AB} + 1)$
Kilgariff, Rychly, Smrz, Tugwell, “The Sketch Engine” Proc. Euralex 2004.

$$\mathbf{Dice} \frac{2f_{AB}}{f_A + f_B}$$

$$\mathbf{relative\ freq} \frac{f_{AB}}{f_A} 100$$