Statistics used in the

Lexical Computing Ltd.

July 8, 2015

1 General reference

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014): The Sketch Engine: ten years on. In *Lexicography* 1(1): 7–36. DOI: 10.1007/s40607-014-0009-9. ISSN 2197-4292

2 Conventions

This documents describes statistics used in the Sketch Engine system. Following conventions apply unless specified otherwise:

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)

2.1 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.

3 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

$$\log_2 \frac{f_{AB}N}{f_A f_B}$$

Association score

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

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

$$\text{Dice}(f_A, f_B) = \frac{2\frac{f_A}{N}\frac{f_B}{N}}{\frac{f_A}{N} + \frac{f_B}{N}} \simeq \frac{2\frac{f_{AB}}{N}}{\frac{f_A}{N} + \frac{f_B}{N}} = \frac{2f_{AB}}{f_A + f_B}$$

logDice

$$14 + \log_2 \text{Dice}\left(\frac{||w_1, R, w_2||}{||w_1, R, *||}, \frac{||w_1, R, w_2||}{||*, *, w_2||}\right) = 14 + \log_2 \frac{2 \cdot ||w_1, R, w_2||}{||w_1, R, *|| + ||*, *, w_2||}$$

For more information on logDice, see: Rychlý, P. (2008). A lexicographer-friendly association score. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, 6–9.
Since June 2015 (word sketch format 4, Manatee version 2.125) the indices were modified so that the score is (more correctly) computed as follows:

logDice general word sketch score (apples in all cases except those listed below)

$$14 + \log_2 \text{Dice}\left(\frac{||w_1, R, w_2||}{||w_1, R, *||}, \frac{||w_1, R, w_2||}{||*, R, w_2||}\right) = 14 + \log_2 \frac{2 \cdot ||w_1, R, w_2||}{||w_1, R, *|| + ||*, R, w_2||}$$

score for word sketch triples of UNARY grammatical relations

$$\frac{||w_1, R, w_1||}{f_{w_1}}$$

score for a given grammatical relation R as such

$$\frac{||w_1, R, *||}{f_{w_1}}$$

score for word sketch display with unified grammatical relations

$$14 + \log_2 \text{Dice}\left(\frac{||w_1, R, w_2||}{f_{w_1}}, \frac{||w_1, R, w_2||}{f_{w_2}}\right) = 14 + \log_2 \frac{2 \cdot ||w_1, R, w_2||}{f_{w_1} + f_{w_2}}$$

4 Thesaurus

To compute a similarity score between word w_1 and word w_2, we compare w_1 and w_2's word sketches in this way:

- find all the overlaps, i. e. where w_1 and w_2 share a collocation in the same grammatical relation, e. g.: $(beer/wine, OBJECT\_OF, drink)$, where the association score > 0,
- let ws_{w_1} and ws_{w_2} be the set of all word sketch triples $(headword, relation, collocation)$ for w_1 and w_2, respectively, where the association score > 0,
- let $ctx(w_1) = \{(r, c)|(w_1, r, c) \in ws_{w_1}\}$,
- let AS_i be the association score of a word sketch triple (since September 2006, logDice is used),
- then the distance between w_1 and w_2 is computed as:

$$Dist(w_1, w_2) = \frac{\sum_{(r,c)\in ctx(w_1)\cap ctx(w_2)} AS_{(w_1,r,c)} + AS_{(w_2,r,c)} - (AS_{(w_1,r,c)} - AS_{(w_2,r,c)})^2/50}{\sum_{i\in ws_1} AS_i + \sum_{i\in ws_2} AS_i}$$

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.

5 Key words, key terms, comparing corpora

Key words are words typical of a focus corpus (a corpus we are interested in) in contrast to a reference corpus (usually a general corpus in the same language as the focus corpus).
The keyness score of a word is calculated according to the following formula:

$$\frac{fpm_{focus} + n}{fpm_{ref} + n}$$

where fpm_{focus} is the normalized (per million) frequency of the word in the focus corpus, fpm_{ref} is the normalized (per million) frequency of the word in the reference corpus, n is the simple Maths (smoothing) parameter ($n = 1$ is the default value).

The top key words reflect the domain of the focus corpus very well and can be used to explore differences between corpora in Sketch Engine as shown in Kilgarriff: "Getting to know your corpus". Proceedings of Text, Speech and Dialogue 2012, Lecture Notes in Computer Science. Springer, 2012.

Key terms are multi word noun phrases typical of a corpus. They are defined using term definition rules (similarly to word sketch relations). The keyness score for terms is the same as for words, corpus frequencies of whole term phrases are taken into account in this case.

6 Other statistics

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

T-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

MI^3-Score

$$\log_2 \frac{f_{AB}^3 N}{f_A f_B}$$

Oakes, Statistics for Corpus Linguistics, 1998

log-likelihood

$$2 \cdot (xlx(f_{AB}) + xlx(f_A - f_{AB}) + xlx(f_B - f_{AB}) + xlx(N)$$
$$+xlx(N + f_{AB} - f_A - f_B) - xlx(f_A) - xlx(f_B) - xlx(N - f_A) - xlx(N - f_B))$$

where $xlx(f)$ 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-f (formerly called **salience**)

$$\text{MI-Score} \cdot \ln(f_{AB} + 1)$$

Kilgarriff, Rychlý, Smrž, Tugwell, "The Sketch Engine" Proc. Euralex 2004.

Dice

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

relative freq

$$\frac{f_{AB}}{f_A} \cdot 100$$