Term Finding

Vojtěch Kovář, Vít Suchomel

Lexical Computing Ltd.
{vit.suchomel, vojtech.kovar}@sketchengine.co.uk

SKEW 3
March 22, 2012

Keyword extraction

Finding the words most specific for a corpus

- take a word (lemma, lemma + POS) in the given corpus
- calculate relative corpus frequency (occurences per million) of the word
- calculate relative corpus frequency of the word in the reference corpus identical tokenization required
- calculate $score = \frac{this\ corpus\ rel.\ frequency\ +\ simple\ math\ parameter}{reference\ corpus\ rel.\ frequency\ +\ simple\ math\ parameter}$
- the higher the simple maths parameter, the less important are low frequency words
- sort all words in the given corpus by score
- words with the highest score are most likely to be keywords in the given corpus

Term extraction

Key collocates extraction using word sketches

- take a wordsketch triple (lemma + POS, grammatical relation, lemma + POS) in the given corpus
- calculate relative corpus frequency (occurences per million) of the triple
- calculate relative corpus frequency of the triple in the reference corpus identical tokenization and sketch grammar required
- calculate $score = \frac{this\ \ corpus\ \ rel.\ \ frequency\ +\ simple\ \ math\ \ parameter}{reference\ \ corpus\ \ rel.\ \ frequency\ +\ simple\ \ math\ \ parameter}$
- the higher the simple maths parameter, the less important are low frequency words
- sort all triples in the given corpus by score
- triples with the highest score are most likely to be terms in the given corpus

Term finding in SketchEngine

Currently developed functions

- keyword extraction and term finding in preloaded corpora
- keyword extraction and term finding in user corpora
- keyword extraction and term finding in subcorpora[1]

Plans for improvements

- extra sketch grammar for term finding
- commonest match for term finding

[1] Keyword extraction and term finding in subcorpora is available for preloaded corpora in both directions: the source can be a subcorpus as well as the reference can be a subcorpus. This functionality (in both directions) is also available for subcorpora of user built corpora by selecting the user corpus in Corpus Architect and clicking on "Open in SkE" to open it in Bonito.