New Format of Word Sketches

Pavel Rychlý

pavel.rychly@sketchengine.co.uk

5th Sketch Engine Workshop
Bolzano, July 14, 2014

Word Sketches

- One page grammatical and collocational behaviour of words.
- Collocations grouped in grammatical relations.
- Based on sketch grammar – shallow parsing via corpus queries.

- New development to optimize word sketch data files.

Word Sketch Data Files

- To provide fast access to word sketches most of the data are precomputed.
- Two groups of files
  - sketch tables with frequencies and scores
  - hit lists for concordances and filtering
- On-the-fly computations
  - filtering on minimum frequency and score, sorting
  - multi-word sketches

Word Sketch Data File Size

The size depends on corpus size and language

- Language – complexity of sketch grammar
  - English – 22 gramrels (43 in BNC, 83 for a lexicography project)
  - Italian – 7 gramrels
  - Japanese – 101 gramrels
- Corpus size – approx. 1 GB per 100 mil. tokens
- English sketches – approx. 500 MB per 100 mil. tokens

New Format Features

It is much more scalable.

- less data files
- smaller data files
- no limits in stored values
- better scores computation

less data files

- old format 18 files + 4 optional
- new format 7 files + 4 optional
- faster access
- compilation time: from 13 to 2 files per run
  – 6 times bigger data processed in one stage

smaller data files

- sketch tables with frequencies and scores
  - 50 % reduction with more information stored
- hit lists for concordances and filtering
  - same size for smaller corpora
  - up to 30 % reduction for 10+ billion corpora

no limits in stored values

- fix size (4B) of all values in previous format
- variable bit length in new format
- bigger values use more space
- really no limits in stored values

better scores computation

- logDice used with
  - $|w_1, gr, w_2|$, $|w_1, gr, *|$, $|*, *, w_2|$ before
  - $|w_1, gr, w_2|$, $|w_1, gr, *|$, $|*, gr, w_2|$ now
- scores are independent from number of gramrels

Conclusions

- We can handle bigger data.
- No limits in corpus size, gramrel count, word list size.