Sketch Engine development plans

Miloš Jakubíček



milos.jakubicek@sketchengine.co.uk

7th Sketch Engine Workshop @ LREC Portoroz, May 23, 2016

Sketch Engine Future: Research agenda

https://www.sketchengine.co.uk/research-agenda/

- Parallel and Distributed Processing of Very Large Text Corpora
- Building Very Large Text Corpora from the Web
- Corpus Heterogeneity and Homogeneity
- Corpus Evaluation
- Terminology Extraction
- Corpora and Language Teaching
- Language change over time
- Corpus Data Visualization
- Dictionary from Scratch

Building Very Large Text Corpora from the Web

- well-studied domain
- but many ongoing challenges including:
 - text type identification (genres on the web)
 - spam fighting
 - text normalization and cleaning
 - dealing with low-resourced languages
 - diachronic analysis (timestamping)

Parallel and Distributed Processing of Very Large Corpora

- targeting corpus size of ca 100 billion words
- trivial parallelization often not possible
- compile-time:
 - corpus virtualization
- run-time
 - asynchronous processing all over web pages
 - reimplementation of the database backend (Manatee) in Go language
 - native support for concurrency

Corpus Heterogeneity and Homogeneity

- what is in the corpus?
- how is corpus X similar to corpus Y? (link)
- assumes we know how much X and Y are homogenous
- text type induction, clustering, . . .



Corpus Evaluation

- is corpus X better than corpus Y?
 - assumes: better for a purpose
- 2012: collocation dictionary task (En, Cz)
 - word sketch evaluation
 - sketch grammar vs. parser comparison
 - . . .
- next run of the task to come soon

Terminology Extraction

- given a domain corpus, find all terms in it
- assumes: term as a concept is plausible only within a fixed domain (link)
- so far: Chinese, Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish
- recent work on bilingual terminology extraction (link)

Corpora and Language Teaching

- biggest problem with Sketch Engine: too many buttons
- SkELL http://skell.sketchengine.co.uk
 - English only
 - Russian coming very soon
 - more to come on demand

Language Change over Time

- neologisms finding
- so far: new lexemes (link)
- now: new/changed senses based on word sketches
- data is the problem, not the algorithms



Corpus Data Visualization

- work by Lucia Kocincova
- to be integrated into Sketch Engine and continued
- preview



Dictionary from Scratch

A corpus2dict approach:

- headwords
- senses + definitions
- examples
- labels
- collocations/colligations
- (translations)

Key aspect: how far can we get fully automatically?

Further Sketch Engine development

- moving to Go and to a fully a distributed architecture
- multi-level tokenization
- multi-word thesaurus and sketch-diff
- new look & feel (UI, word list form, ...)
- virtual word sketches
- word sketch clustering
- learner corpus section (+ LCA)
- and many more . . . ©

Conclusions

- many ongoing developments, stay tuned
- as always, we are looking for feedback and suggestions
- going in the spirit of Adam's motto:

Bringing corpora to masses



Conclusions



Conclusions

