Indexing & Querying Corpora over 20 Billion Tokens

Miloš Jakubíček

Lexical Computing Ltd.
milos.jakubicek@sketchengine.co.uk

SKEW3, 21. 3. 2012

Introduction

Overview of

- currently available and soon coming corpora in the TenTen series
- present challenges in indexing and querying very large corpora
- technology preview of related enhancements

TenTen corpora

a Ten(-to-the-power-of-)Ten corpus = a corpus of target size over 10 G

zhTenTen	2.1 G
enTenTen	3.2 G
deTenTen	2.8 G
itTenTen	3 G
noTenTen	0.7 G
ptTenTen	0.95 G
skTenTen	0.9 G
esTenTen	2.5 G

TenTen corpora

making TenTen's indeed TenTen's

ruTenTen	20.1 G
czTenTen	5.8 G
arTenTen	6.6 G
trTenTen	4.1 G
enClueWeb	70 G

Corpora over 20 G

Challenges in:

- corpus preparation
- indexing and encoding of the corpus
- searching the corpus
- providing appropriate user interface

Indexing

- changes to the core of our database system needed
- speed starts to matter (not days, but months with current technology)
- effective parallelization required in a distributed environment
- main concerns: compilation of the corpus, compilation (evaluation) of sketches

Compilation of Sketches

- parallelized (ready on beta.sketchengine.co.uk)
- effective speed up by factor of 5–10 (enClueWeb: 45 to 6 days)
- parallelization is performed on the level of separate grammatical relations and their queries

Compilation of Main Corpus

- work in progress
- cannot be trivially parallelized due to the need of consistent lexicon (token string to id mapping)
- special handling of unique attributes implemented (doc.url, doc.id, etc.) – speed up by factor of 3–5
- still a present issue (enClueWeb: 120 to 30 days)

Searching Large Corpora

- even with advanced indexation, even optimal algorithms might result into unsatisfying response times
- advanced management of long-running jobs needed
- providing intermediate results to users as soon as they are available $\longrightarrow$ asynchronous query processing (AQP)
- exploiting distributed file systems (Global File System) in server nodes and multi-process and multi-threaded applications

Technology preview: AQP

Principles of AQP:

- concordance user queries are started asynchronously as background jobs in the server infrastructure
- web page response created as soon as the required concordance page is ready
- remaining results are computed in background and ready for later usage

Conclusions

- substantial efforts are being made to make working with large corpora comfortable and accessible
- TenTen's over 20 G are on the way and coming soon
- ... together with advanced user management of resource intensive long-running tasks in Sketch Engine

Thank you!

Thank you for your attention!