Effective Corpus Virtualization

Miloš Jakubíček, Adam Kilgarriff, Pavel Rychlý

Lexical Computing Ltd.
Brighton, United Kingdom

NLP Centre, Masaryk University,
Brno, Czech Republic

{milos.jakubicek,pavel.rychly,adam.kilgarriff}@sketchengine.co.uk

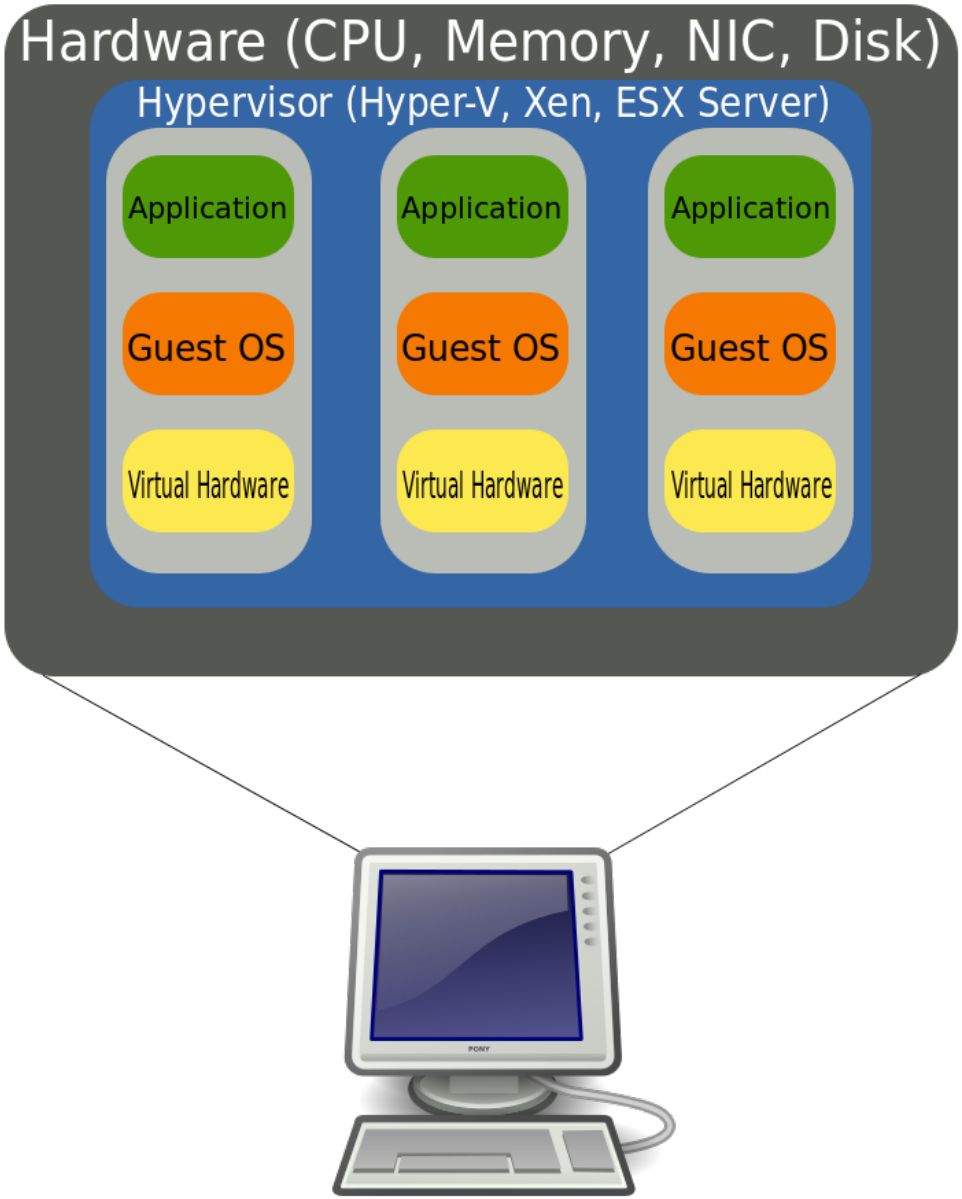
May 31st, 2014

Outline

Historical Overview

Virtualization: established method for postponing commitments to resources

- $<$ 2000: nothing to virtualize
- $\sim$ 2000: lots of resources (computers, RAM, processors, hard drives) available
  - at a moderate price
  - good predictions on resource needs $\Rightarrow$ large savings
  - . . . but good predictions were very hard to obtain on the rapidly changing IT market


Hardware (CPU, Memory, NIC, Disk)
Hypervisor (Hyper-V, Xen, ESX Server)
Application
Application
Application
Guest OS
Guest OS
Guest OS
Virtual Hardware
Virtual Hardware
Virtual Hardware

Text corpora

- ten years back: not many available
- now in similar position to IT hardware at 2000
- lots of corpora available, raising many questions
- including: how to effectively organize them into logical units presented to their users

Text corpora

- ten years back: not many available
- now in similar position to IT hardware at 2000
- lots of corpora available, raising many questions
- including: how to effectively organize them into logical units presented to their users
  - clearly not 1 corpus per language
  - granularity driven by user needs
  - copy & create very demanding, both in space and time

Text corpora

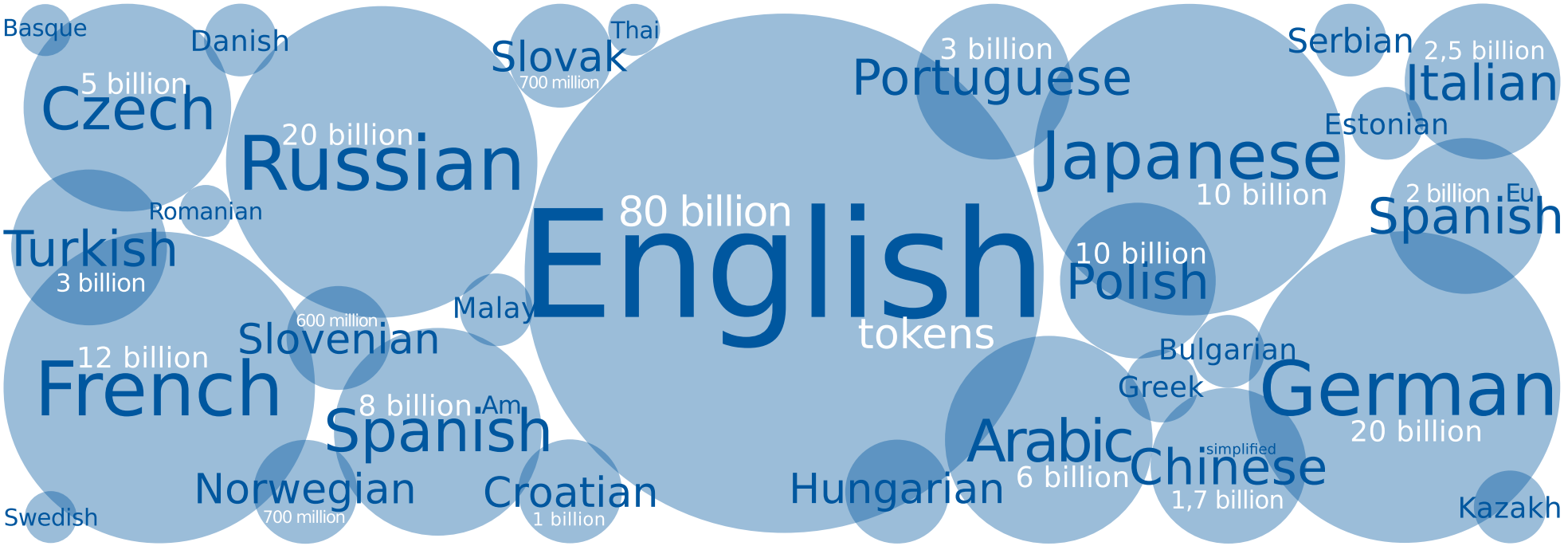
- ten years back: not many available
- now in similar position to IT hardware at 2000
- lots of corpora available, raising many questions
- including: how to effectively organize them into logical units presented to their users
  - clearly not 1 corpus per language
  - granularity driven by user needs
  - copy & create very demanding, both in space and time
- context: Sketch Engine (Kilgarriff, 2004)

Sketch Engine

- corpus query system
- web service (including API)
- widely used for
  - lexicography purposes
    - Oxford University Press, Cambridge University Press, Harper Collins, Macmillan, ...
  - linguistic and language technology teaching and research at universities
    - about 100 academic institutions worldwide
    - thousands of individuals

Sketch Engine features

- **concordancing**, sorting, sampling, wordlists, collocation lists
- full **regular-expression** searching
- support for **parallel corpora**, virtual sub- and supercorpora
- handles **billion-word (80 G+)** corpora smoothly
- **word sketches**: one-page summaries of a word's grammatical and collocational behaviour
- distributional **thesaurus**
- **keywords** extraction, **terms** extraction
- **Corpus Architect**: user corpora
  - uploaded by users
  - created by WebBootCaT


Basque
Danish
5 billion
Czech
20 billion
Russian
Romanian
Turkish
3 billion
Slovak
700 million
Thai
80 billion
English
tokens
Malay
600 million
Slovenian
12 billion
French
8 billion
Am.
Spanish
Norwegian
700 million
Croatian
1 billion
Swedish
3 billion
Portuguese
Hungarian
Arabic
6 billion
Serbian
2,5 billion
Italian
Estonian
Japanese
10 billion
2 billion
Eu.
Spanish
10 billion
Polish
Bulgarian
Greek
German
20 billion
simplified
Chinese
1,7 billion
Kazakh

Sketch Engine languages

By May 2014 more than **500 corpora** for **70 languages**:

- 38 languages with corpora having than 100 million tokens
- 18 languages with corpora having more than 1 billion tokens
  - In 2010 a series of TenTen (10^{10}) corpora started
- 56 languages with a PoS-tagged corpus
- 36 languages with word sketches
- 21 languages with integrated tagger for tagging user corpora

Manatee corpus scheme

Back-end corpus database management system (Rychlý, 1999, 2007)

- Corpus
- Subcorpus
  - set of corpus segments (usually defined by text types)
  - ad-hoc filtering solutions required

Manatee corpus scheme

Back-end corpus database management system (Rychlý, 1999, 2007)

- Corpus
- Subcorpus
  - set of corpus segments (usually defined by text types)
  - ad-hoc filtering solutions required
- Supercorpus = Virtual Corpus
  - set of (sub)corpora
  - first-class corpus entity

Virtual corpus

- first-class corpus entity within Manatee
- balanced approach between compile-time and run-time processing
- $\Rightarrow$ fast compilation (just lexicon harmonization), negligible impact on query evaluation (< 10 %)
- large space savings

Virtual corpus – esTenTen11

corpus	number of tokens (billions)	database size (gigabytes)
esAmTenTen11	8.7	217
esEuTenTen11	2.4	35
esTenTen11	11.1	252

Virtual corpus – esTenTen11

	virtual	**regular**
space occupied	13 GB	252 GB
compilation time	3.4 hrs	30.6 hrs

Virtual corpus – esTenTen11

	virtual	**regular**
space occupied	13 GB	252 GB
compilation time	3.4 hrs	30.6 hrs

- 20 × less space
- 10 × faster

Virtual corpus preparation

1. definition file

```
=corpus1
0,1000000
2000000,3000000
=corpus2
0,$
```

2. placing path to it into the `VIRTUAL` directive instead of providing `VERTICAL` source texts
3. running `mkvirt CORPUS`

Ongoing work

1. Faster word sketch compilation for virtual corpora
2. Exploitation for parallel corpus compilation
   - compilation on n parts $\rightarrow$ virtualization $\rightarrow$ devirtualization
   - 80 % speedup so far

Conclusions

- growing amount of text corpora raises issues concerning their management, composition and structuring into logical units
- Manatee now offers effective and flexible methods for this task
- all the developments are part of the open-source version released within NoSketch Engine at `http://nlp.fi.muni.cz/trac/noske`