Practical Web Crawling for Text Corpora

Vít Suchomel, Jan Pomikálek
NLP Centre
Masaryk University
Brno, Czech Republic

RASLAN 2011

Introduction

- a need for large text corpora (lexicographers, linguists)
- general web crawlers (Heritrix)
- SpiderLing – a new web crawler for text corpora

Selected corpora sizes

Current state (LCL, FI)

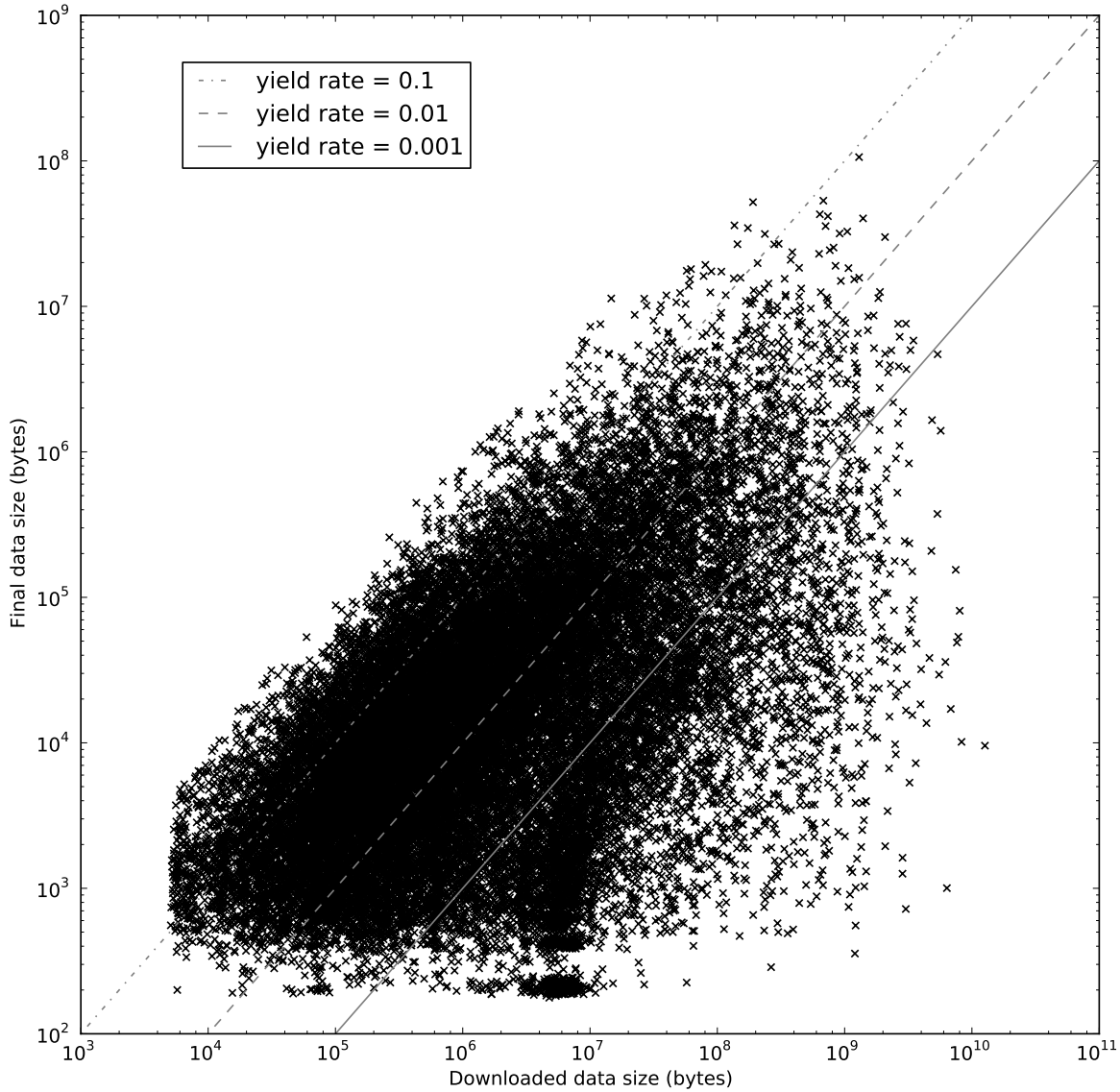
language	millions of tokens
Arabic (Arabic web)	174
Chinese (zhTenTen)	2,107
Czech (czes)	465
English (enTenTen)	3,269
English (enClueWeb)	18,095 (70,000)
German (deTenTen)	2,845
Japanese (JpWaC)	409
Russian (Russian web)	188
Slovak (skTenTen)	876
Spanish (esTenTen)	2,459
Tajic	–

Target: $\geq 10^{10}$ words for main languages

SpiderLing – a new web crawler for text corpora

- asynchronous communication design
- partial processing of downloaded data
  - lanuage filters (trigram model, wordlist)
  - encoding detection by chared
  - boilerplate removal by justext
  - simple deduplication
- three kinds of subprocesses
  - evaluator
  - downloader
  - data processors
- written in Python
- aimed to improve the crawling efficiency
- $\textit{yield rate} = \frac{\textit{final data}}{\textit{downloaded data}}$
- focus on the text–rich web domains

Yield rate by web domains (Heritrix, Portuguese web)



Websites' yield rate threshold function in SpiderLing

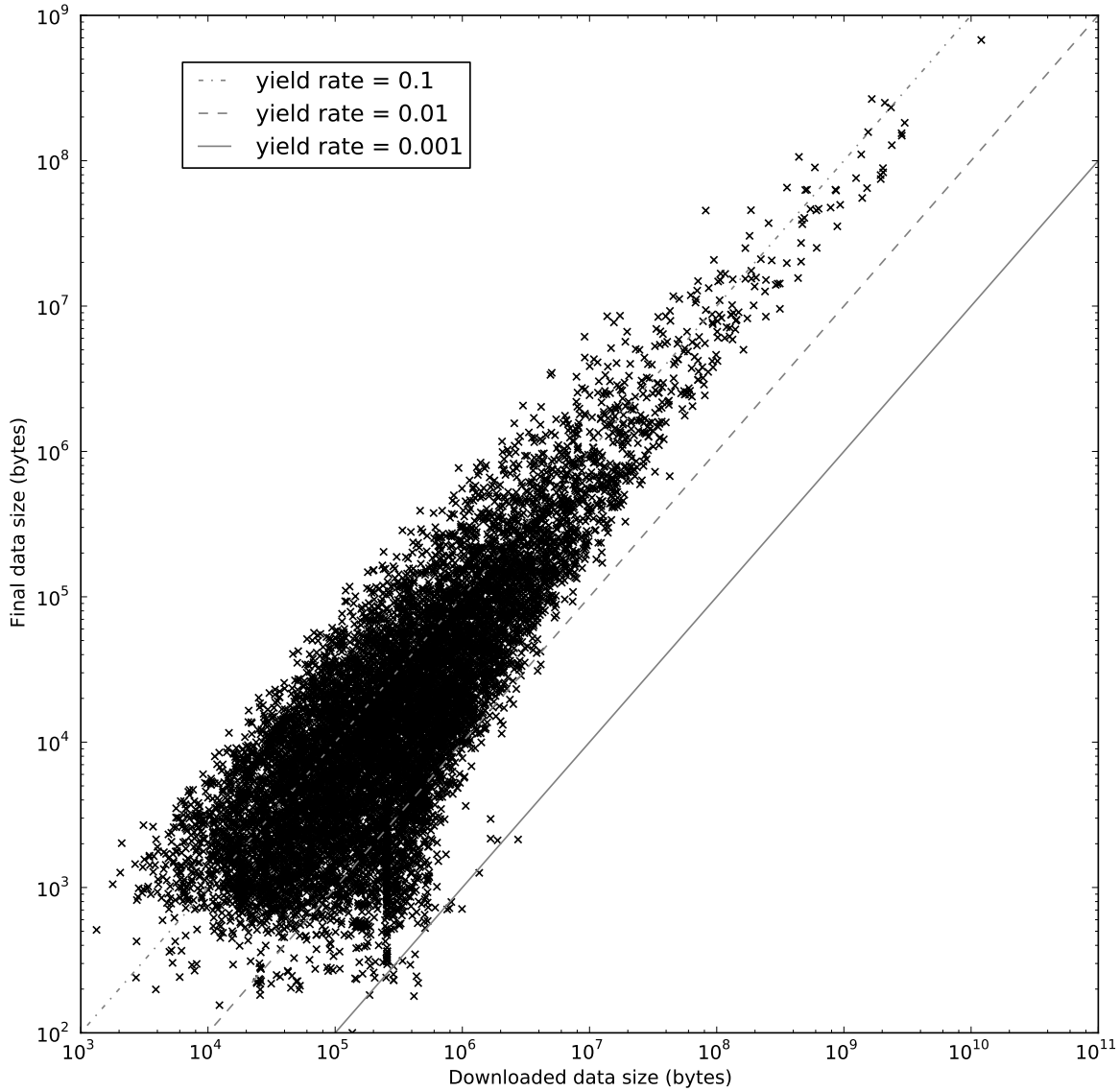
The yield rate threshold for a domain is computed using the following function:

$$t(n) = 0.01 \cdot (\log_{10}(n) - 1)$$

where n is the number of documents downloaded from the domain.

# of documents	yr threshold
10	0.00
100	0.01
1000	0.02
10000	0.03

Yield rate by web domains (SpiderLing, Czech web)

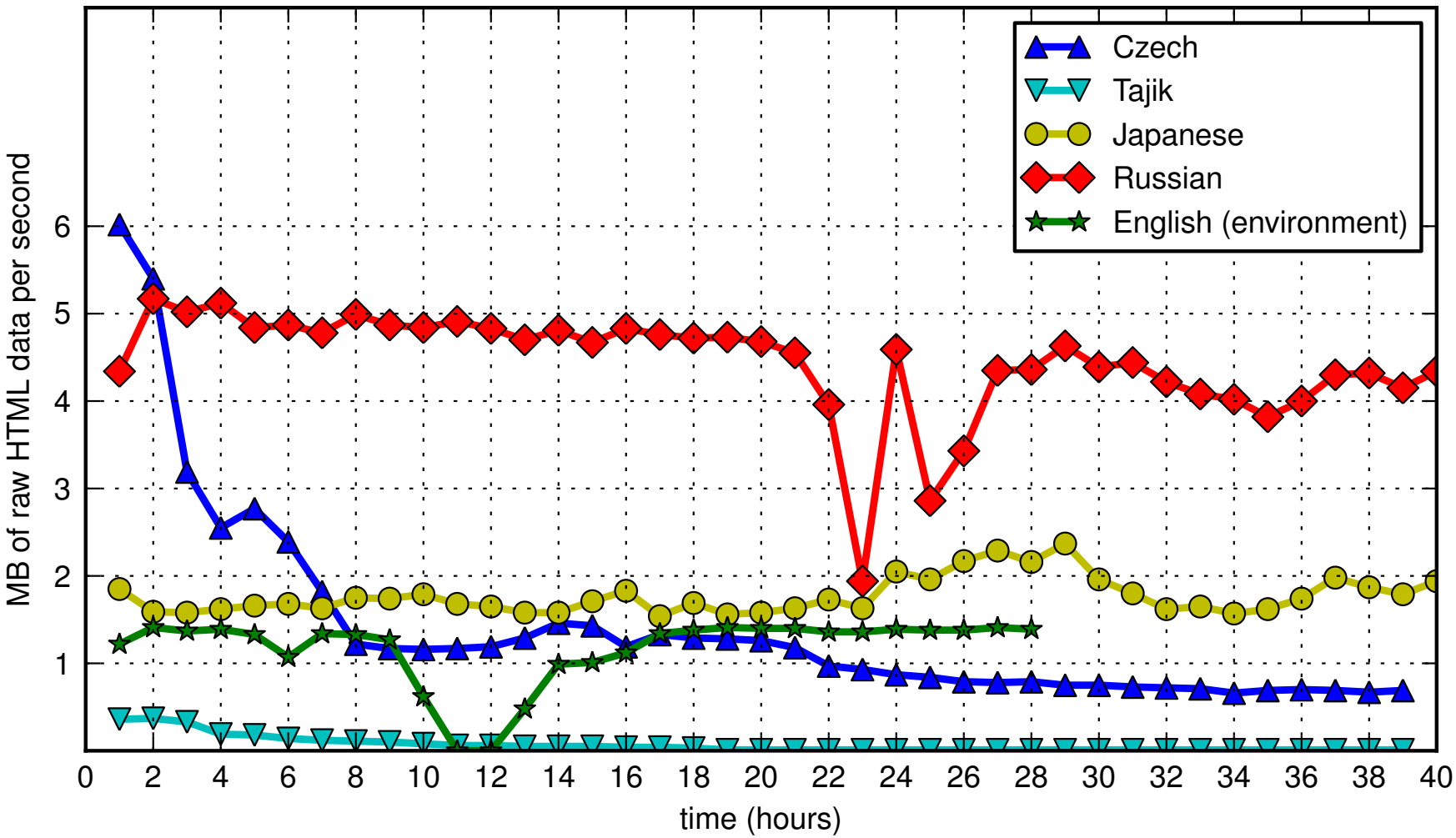


Textual data downloaded by SpiderLing so far

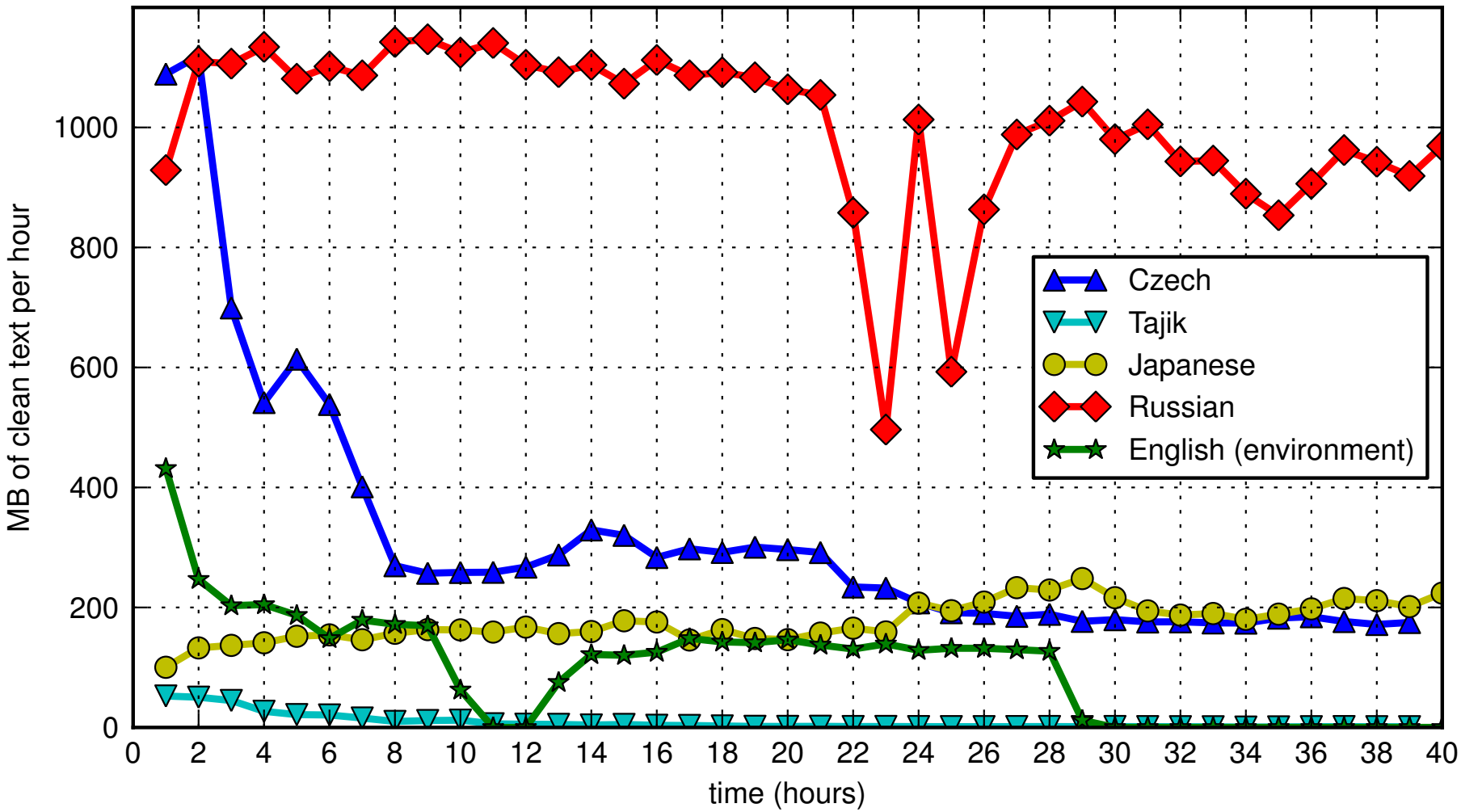
language	raw data	pre-proc. data	yield rate	10^6 tokens
Czech (several test runs)	ca. 4,000	ca. 105	0.026	ca. 4,000
Tajic	9.1	0.32	0.035	ca. 30
Japanese (in progress)	1,180	41	0.035	?
Russian (in progress)	2,230	136	0.061	?
English – environment	114	4.0	0.035	652

data sizes in GB

Crawling speed of SpiderLing (raw HTML data)



Crawling speed of SpiderLing (preprocessed text)



Topic oriented downloading with SpiderLing

- a topic defined by a list of words
- crawling the pages containing the topic words
- the yield rate threshold mechanism ensures the crawler prefers the topic related websites
- currently building an environment related corpus for a customer of LCL

Conclusion & future work

Conclusion

- a new web crawler for linguistic needs
- effective avoiding of web data not suitable for text corpora
- significantly improved yield rate of the downloaded content
- built large Czech and Tajik corpora; Japanese, Russian, Turkish. . . on the way

Future work

- test the crawler on other big languages (Turkish, Arabic)
- adaptation for sparse sources languages (like Tajic)
- optimizing the crawling constraints to achieve a higher crawling speed
- analyzing the topics and genres of the downloaded texts (eventually ballancing the downloaded content in this respect)
- more tuning of the topic sensitive crawling