Intrinsic Methods for Comparison of Corpora

Vít Baisa and Vít Suchomel

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

December 6, 2013

A need for comparison of corpora

- There are large textual corpora from the web...
  - but do we know what is inside?
- Which corpus is generally better?
  - Comparison based on inner properties of corpora.
- Which corpus is better for a specific task?
  - Comparison based on external use of corpus data.

Intrinsic vs. extrinsic comparison

- The paper describes 8 intrinsic methods of corpus comparison divided into the following groups:
  - general intrinsic properties,
  - text cleaning and processing,
  - wordlist-based methods and
  - syntactic analysis.
- Extrinsic methods will be explored in a future paper.

Data used in the experiment

- The proposed methods were applied to compare two recent very large web-based Czech text corpora:
  - Hector (Spoustová et al., 2010)
  - czTenTen12 (Suchomel, 2012)
- The majority of presented methods is language independent but both corpora must be in the same language.

Size

- A simple rule: The bigger the better.
- Because *We need very large corpora to provide evidence about rare phenomena.* (Pomikálek et al., 2009)
- The measurement of words, tokens and sentences depends on the means of tokenization and sentence detection algorithms used for processing corpus data.

CORPUS	BYTES	TOKENS	WORDS	SENTENCES
Hector	17 GB	3.285 bn	2.607 bn	219 m
czTenTen12	31 GB	5.437 bn	4.458 bn	303 m

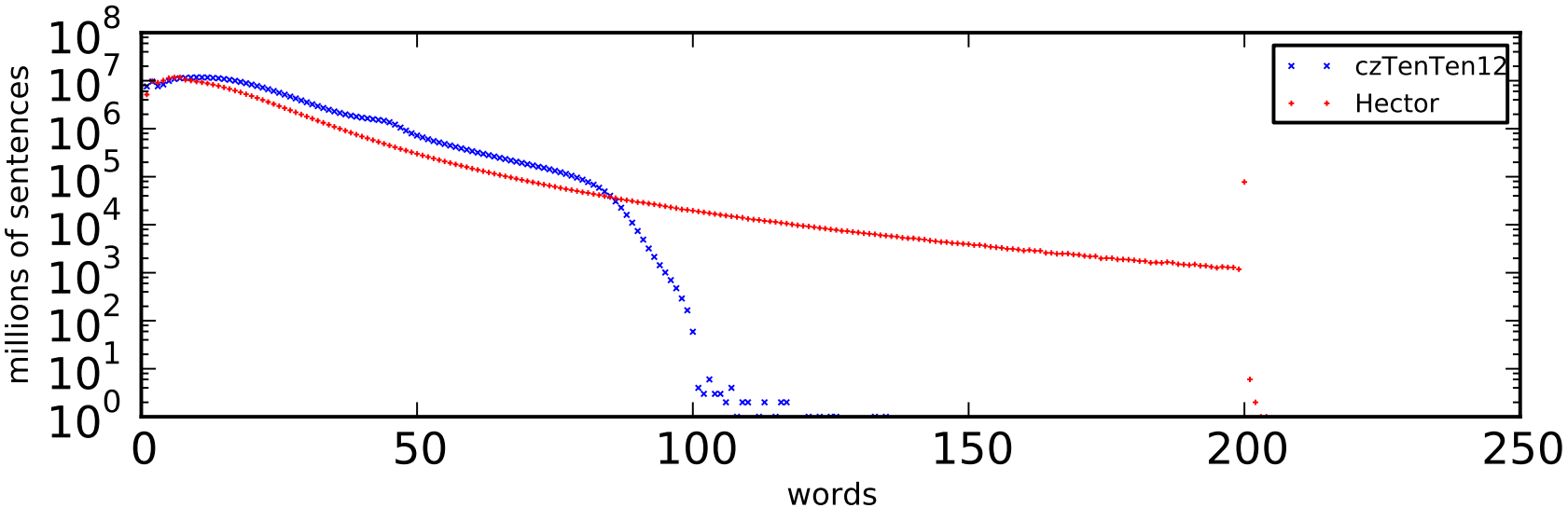
Diversity of sources

- The more diverse source of the data, the better coverage of language by the corpus may be expected.
- Hector: constructed from manually selected web sites with large and good-enough-quality textual content (e.g. news servers, blog sites, discussion fora).
- czTenTen12: a general Czech web crawl.
- Constraining sources of a monolingual corpus to the corresponding national TLD – useful in the case of Czech.

CORPUS	PAGES	DOMAINS	AVG	MED	TLDS
czTenTen12	9,747,315	233,122	42	4	97.6 % cz

Sentence length

- Sentence border detection – different solutions observed.



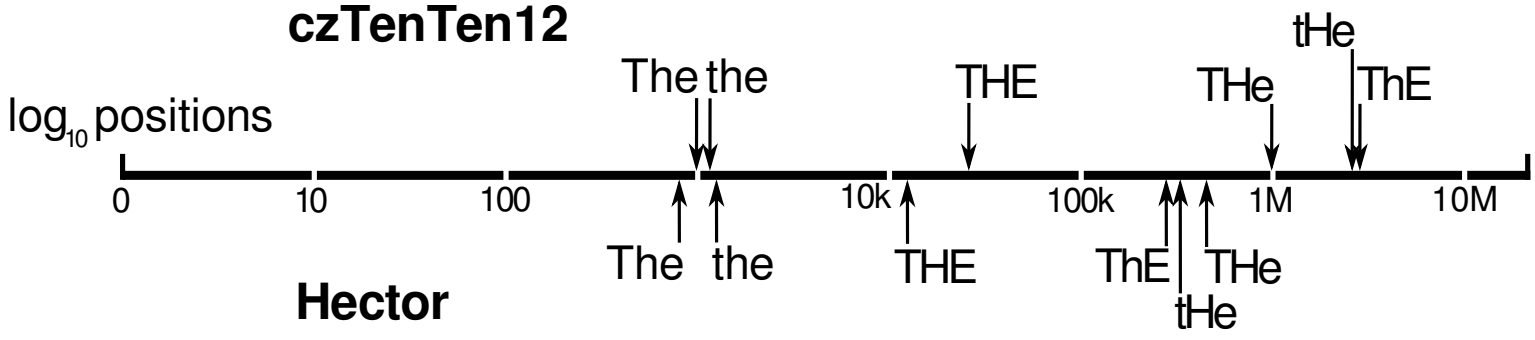
Data duplicity

- The less duplicate texts in a corpus the better. However, a very strict deduplication results in removing usable data needlessly.
- Hector: paragraphs containing more than 30% seen 8-grams were removed
- czTenTen12: paragraphs containing more than 50% seen 7-grams were removed
- onion was used to remove sentences consisting of 50% seen 5-grams of sentences (with smoothing disabled).

CORPUS	BYTES	TOKENS	SENTENCES
Hector	-23.3 %	-25.8 %	-23.6 %
czTenTen12	-17.6 %	-18.7 %	-18.4 %

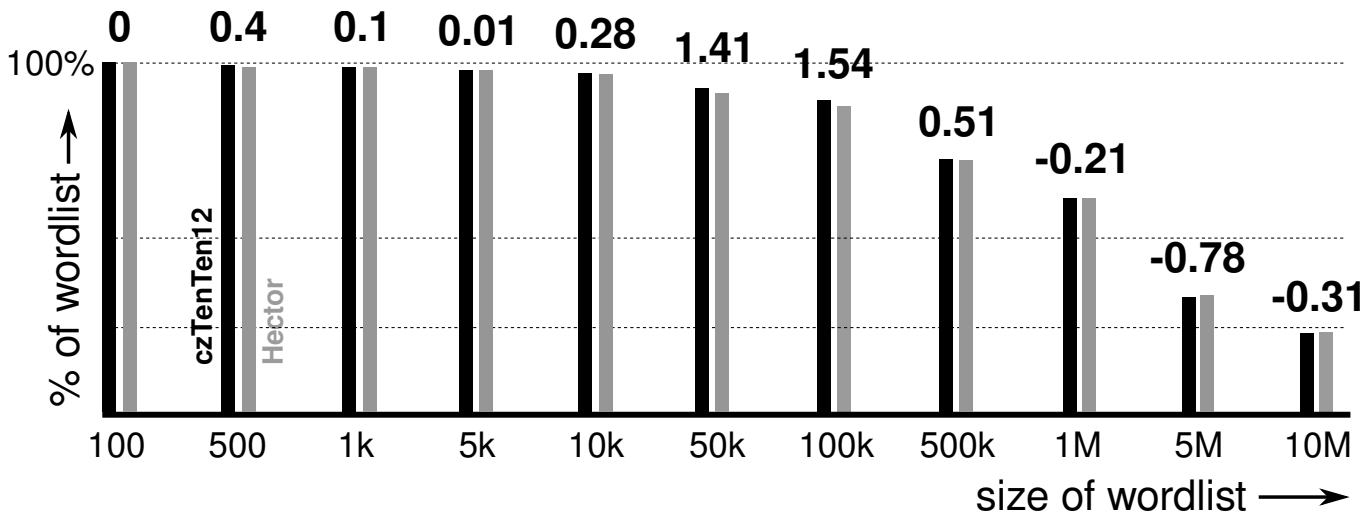
The test

The less paragraphs full of text in unwanted language the better. However, some level of foreign words cannot be avoided, e.g. in developers' blogs, movie or music reviews.



Filtering wordlists

- Unknown words from corpus wordlists were filtered out by a morphologic analyzer. The bigger the size of the rest, the better.
- Czech fast analyser Majka was used.



Keyword comparison

- Following Kilgarrif's work, lowercase keywords were extracted to reveal in which words these corpora differ the most.
- Both recent web corpora contain more data from internet message boards and less news documents than the Czech National Corpus.
- Hector vs. SYN2000: taky, teda, ahoj, holky, mám, fakt, moc, sem, dneska, takže, blog, nevím, máš, super, ráda, ahojky (discussions of women).
- czTenTen12 vs. SYN2000: taky, můžete, moc, děkuji, takže, cca, mám, dobrý, opravdu, dle, ahoj, bych, jestli, díky, hodně, super (discussions).
- SYN2000 vs. Hector and czTenTen12: praha, včera, korun, procent, české, vlády, státní, miliónů, zákona, trhu, ministr, ředitel, výstava, společnost, nato, prezident, čtk (standard language, news, Prague).

Syntactic functions

- Syntactically correct sentences are good. Presence of the main syntactic roles – subject and predicate was checked.
- That rules out web garbage (navigation and labels, tables, program code, SEO keywords, link spam, generated texts,. . .)
- but also syntactically problematic but otherwise quite common and understandable sentences.
- Set was used to carry out the experiment.

CORPUS	NCL	NSEN	PNSEN
Hector	36.6 %	19.0 %	23.7 %
czTenTen12	39.6 %	23.6 %	29.2 %

Future work

- Explore other intrinsic methods:
  - perplexity of language models,
  - finding topics,
  - measuring homogenity and heterogenity.
- Develop extrinsic methods:
  - word sketch evaluation (submitted to LREC 2014),
  - morphological segmentation morfessor.

Conclusion

- Eight methods for a general systematic comparison of text corpora were developed and described.
- The methods were applied to two recent very large Czech web corpora.
- The related tools can be downloaded from the website of the project.

`http://nlp.fi.muni.cz/projekty/corpora_comparison`