Helping BootCaT to catch the Babel fish: Getting encoding, content and language right

Nikola Ljubešić

Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb

BOTWU 2013, Forlì, 24 Juli 2013

Motivation

1. support for western-European languages / encodings only
2. BTE content extractor rather old, many more sophisticated extractors available
    - DOM parser
    - multiple heuristics
3. Bing bad in language identification

- basic idea – take mature tools/API-s for all three problems and rewrite the script for retrieving documents and extracting text
- what programming language? – Perl, Java? – Java!

Encoding guessing and content extraction

- BoilerPipe, `http://code.google.com/p/boilerpipe/`, Apache License 2.0
- decision trees on number of words and link density of blocks
- alternative in Python – chared and justext
- evaluation from Ljubešić and Erjavec (2011)

	precision	recall	F1
ContentExtractor	0.979	0.707	0.821
BTE	0.570	0.955	0.713
BoilerPipe	0.847	0.921	0.882
justext	0.778	0.914	0.841

Language identification

- language-detection, `http://code.google.com/p/language-detection/`, Apache License 2.0
- language profiles as distributions of n-graphs, Naïve Bayes classifier
- 55 language profiles out-of-the-box
- simple to add new language profiles or remove existing ones
- 100% accuracy on most predefined languages
- confusion on Danish (da=179, no=14, en=7), Norwegian (no=199, da=1)

Demo

Experiments

- domain
    - health corpus in English – BTE vs. Boilerpipe
    - corpus of ICT domain in Croatian – BootCaT vs. hrWaC

    1. define one general language corpus and one domain corpus
    2. extract single-word domain terms with CollTerm – seed terms
    3. collect corpora with various BootCaT settings
    4. calculate "domainness" – corpora as tf-idf models, dice similarity to initial domain corpus
- language identification
    - Bing API evaluation
    - Slovene vs. Croatian
    - Croatian vs. Serbian

Health corpus

- 100Mw of ukWaC as reference corpus, 4Mw from health.com as domain corpus
- extract terms from the health.com corpus with CollTerm via tf-idf, reference corpus for idf statistic
- take 100 strongest terms, create 500 trigram queries – entry point for BootCaT
- collect URL-s – 4793 URL-s after cleanup
- retrieve corpora with old and new tool
- measure distance to the initial domain corpus - tf-idf + dice

	#ofDocs	#ofTokens	avg#Tokens	distDom
BTE	4,269	6,225,680	1458	0.738
BP.Default	4,567	6,145,941	1346	0.743
BP.Article	4,577	4,508,944	985	0.758

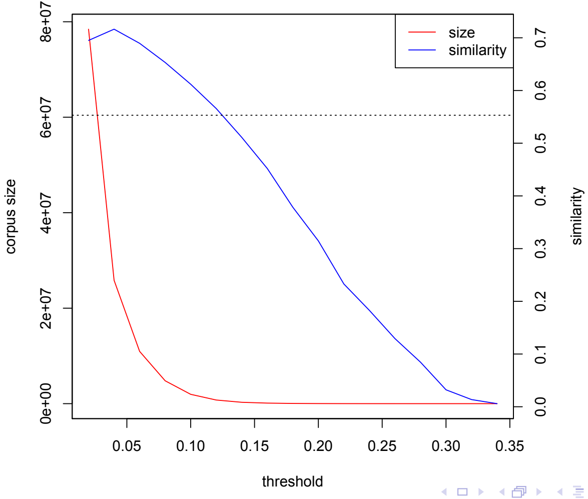
Croatian corpus of ICT

- 35Mw from vecernji.hr as reference corpus
- 10Mw from bug.hr as domain corpus

	#ofDocs	#ofTokens	avg#Tokens	distDom
BTE	3,593	3,098,422	862	0.459
BP.Default	4,261	9,513,564	2232	0.517
BP.Default.hr	3,114	7,664,719	2461	0.553
BP.Article	4,264	6,318,224	1482	0.483
BP.Article.hr	3,171	5,108,200	1610	0.515

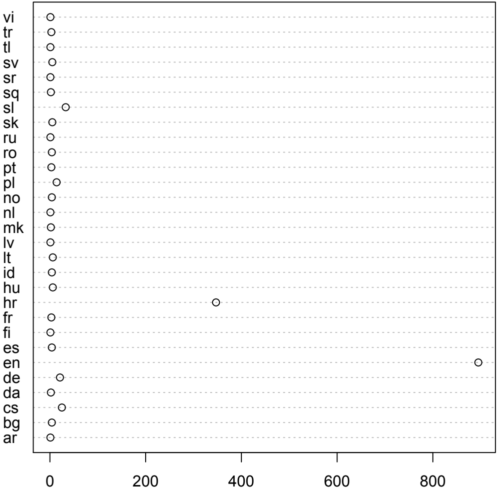
BootCaT vs. WaC

- extract documents containing seed terms from hrWaC



Language identification of Bing API

- "nasty" task for Bing API – germanisms in Croatian



Discriminating between similar languages with langdetect

1. Croatian and Slovene
    - vecernji.hr urls-a from hrWaC and delo.si url-s from slWaC, 3000 documents each
    - use built-in language profiles
    - only 2 Slovene documents annotated as Croatian – near 100% accuracy
2. Croatian and Serbian
    - vecernji.hr urls-a from hrWaC and b92.rs url-s from srWaC
    - two new language profiles – parallel data from SETimes
    - accuracy – Croatian 98.8%, Serbian 99%

Conclusion

- improvements on BootCaT
- encoding guessing – non-Western European languages
- improved content extraction – better domain coverage, cleaner corpora as well
- language identification
  - no need for frequent words lists
  - discrimination between similar languages
  - simple control of profiles
  - adding new profiles
- BootCaT vs. WaC – building whole web corpora becoming quite simple
- SpiderLing + chared + justext + onion
- domain corpora as subcorpora of web corpora

Helping BootCaT to catch the Babel fish: Getting encoding, content and language right

Nikola Ljubešić

Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb

BOTWU 2013, Forlì, 24 Juli 2013