hrWaC and slWac: Web Corpora for Croatian and Slovene

Nikola Ljubešić[1], Tomaž Erjavec[2]

[1] Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
[2] Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

SlaviCorp 2011, Dubrovnik, 12 September 2011

Motivation

- web corpora are an attractive source of linguistic content
- WaCky initiative (Baroni, 2009) has popularized the concept of "Web as Corpus"
- primarily built for "large" languages such as English, German, French, Italian
- in recent years smaller languages follow using mostly the WaCky pipeline (Norwegian, Czech, Slovak...)
- Webs of smaller languages - drastically smaller amount of information available (number of documents via Google and 5 most frequent tokens - .uk 4.7B, .de 1.2B, .no 261M, .hr 70M, .si 82M)
- modified WaCky pipeline for such languages

Overview

Standard WaCky pipeline with modifications

1. crawling
2. physical deduplication
3. content extraction - new algorithm
4. language identification
5. near-duplicate removal - more gentle approach
6. content filtering
7. linguistic processing

Collecting seeds, crawling and physical deduplication

- collecting data only from .hr and .si domains
- seeds collected via Yahoo! Search BOSS API
- random bigrams (mid-frequency tokens, rank 1,000-10,000) from 100-million-token newspaper corpora
- around 50,000 URLs collected for each language
- small webs - sampling would not produce enough data - collecting the "population"
- seed URL-s - home pages of the 50,000 URL-s
- crawling - breadth-first multi-threaded crawler, collecting "text/html" documents, size 50-500 kB
- physical de-duplication with SHA224 algorithm

Numerical overview

	hrWaC	slWaC
# of seed domains	12,033	11,493
# of domains crawled	16,398	18,418
# of crawled documents	15,747,585	9,247,341
# of documents after deduplication	14,654,394	9,022,716
# of extracted documents		
# of language identified documents		
# of non-filtered documents		
# of tokens		

Content extraction

- crucial step in collecting corpora from the Web - extracting only the linguistically interesting part of the web page
- evaluating algorithms on 200 online newspaper documents downloaded from 20 different news portals
- comparing our algorithm (ContentExtractor) with BTE (PotaModule.pm), BoilerPipe 1.1.0 and justext 1.2

	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

Numerical overview

	hrWaC	slWaC
# of seed domains	12,033	11,493
# of domains crawled	16,398	18,418
# of crawled documents	15,747,585	9,247,341
# of documents after deduplication	14,654,394	9,022,716
# of extracted documents	3,924,194	1,598,011
# of language identified documents		
# of non-filtered documents		
# of tokens		

Language identification, filtering and PoS tagging

- language identification with a combination of a second-order Markov chain model and a function word filter for 22 languages, paragraph level
- additional filtering eliminating too short documents, encoding errors and high amount of punctuation
- no near-duplicate detection, but removing duplicate paragraphs during content extraction
- PoS tagging of Croatian with CroTag (Agić and Tadić, 2006), lemmatization with CST lemmatizer
- PoS tagging and lemmatization of Slovene with ToTaLe (Erjavec et al. 2005)

Numerical overview

	hrWaC	slWaC
# of seed domains	12,033	11,493
# of domains crawled	16,398	18,418
# of crawled documents	15,747,585	9,247,341
# of documents after deduplication	14,654,394	9,022,716
# of extracted documents	3,924,194	1,598,011
# of language identified documents	3,607,054	1,337,286
# of non-filtered documents	3,409,226	1,287,895
# of tokens	1,186,795,086	380,299,844

Comparing the content of the corpora

- explore the content of the web corpora through the topic modeling (LDA) (Blei et al., 2003; Sharoff, 2010) with the MALLET tool
- building topic models only on sample of 10-50% of original data - experiments show that topics remain the same, 20 topics per corpus
- build topics for hrWaC (1.2G), slWaC (380M), Gigafida (1G) and rtvslo.si (32M)
- compare topic models with ukWaC and BNC topics (Sharoff, 2010)

Example of hrWaC and slWaC topic models

Lg	Topic name	Size	Words with highest probability
sl	intl. politics	4.7%	leto država vojna človek predsednik oblast zda napad vojska
hr	**reg. politics**	5.9%	zemlja srbija predsjednik godina država rat vlada hrvatska
sl	**reg. politics**	3.6%	država slovenija eu leto članica minister predsednik hrvaška
hr	**dom. politics**	6.2%	predsjednik vlada stranka izbor sanader ministar pitanje
sl	**dom. politics**	4.9%	vlada predsednik zakon stranka slovenija minister sodišče
hr	**law**	3.0%	zakon tema odluka pravo postupak sud članak osoba ugovor
sl	**law**	4.4%	zakon podatek primer pravica člen oseba plačilo storitev dan
hr	crime	5.3%	policija sud godina osoba slučaj zatvor kazna sat policajac
hr	**finance**	7.1%	godina kuna milijun tvrtka cijena banka euro tržište dionica
sl	**finance**	5.2%	leto evro odstotek podjetje milijon družba banka cena trg
hr	**sports**	2.0%	utrka mjesto godina natjecanje prvenstvo vrijeme sezona
sl	**sports**	4.0%	tekma minuta igra leto točka prvenstvo igralec ekipa mesto
hr	soccer	4.8%	utakmica igrač klub momčad minuta pobjeda liga sezona

The results

- hrWaC and slWaC very similar, ukWaC somewhat less
- web corpora in general very similar, these of closely related languages / cultures (webs of similar age and size?) even more similar
- ukWaC and BNC (Sharoff, 2010) - smaller degree of similarity than that of ukWaC and {hr,sl}WaC
- rtvslo.si has more culture, sports, politics; slWaC has more technology, private affairs; Gigafida has more education, lifestyle
- intersection of most probable keywords in topics that align best

	gigaFida	rtvslo.si
slWaC	.508	.435
gigaFida		.443

Conclusion and further steps

- presented a modified WaCky pipeline for smaller languages / webs
- analyzing content by topic modeling shows high similarity in web corpora, especially of closely related languages
- large, balanced corpora more similar to web corpora than newspaper corpora
- hrWaC and slWaC 2.0
  - collect more Slovene data
  - new crawler, further experiments with near-duplicate removal
  - use search engines for finding new domains / islands of documents written in language of interest
  - methodology for assessing the quality of corpora
  - make available corpora for on-line searches (noSketchEngine)

hrWaC and slWac: Web Corpora for Croatian and Slovene

Nikola Ljubešić[1], Tomaž Erjavec[2]

[1] Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
[2] Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

SlaviCorp 2011, Dubrovnik, 12 September 2011