Challenges in web corpus construction for low-resource languages in a post-BootCaT world

LTC'13, Less Resourced Languages special track

Adrien Barbaresi

ICAR Lab / ENS Lyon

Poznań – December 8, 2013

Outline

The "Web as Corpus" paradigm...

The state of the art tools rely heavily on search engines

BootCaT method (Baroni & Bernardini 2004): repeated search engines queries using several word seeds that are randomly combined.

First coming from an initial list and later from unigram extraction in the corpus itself.

⇒ "seed URLs" used as a starting point for web crawlers.

Approach not limited to English, has been used for major world languages (Baroni et al. 2009, Kilgarriff et al. 2010)

... and its URL seeds problem

- Diverse and partly unknown search biases related to search engine optimization tricks and undocumented PageRank adjustments
- Diverse sources of URL seeds could at least ensure that there is not a single bias, but several ones
- Evolving web document structure
- Shift from "web AS corpus" to "web FOR corpus" (increasing number of web pages and the necessity to use sampling methods)

$\Rightarrow$ This is what we call the post-BootCaT world in web corpus construction.

Peculiarities of lesser-known languages

Lack of interest and project financing when dealing with certain low-resource languages → necessary to use light-weight approaches where costs are lowered as much as possible (Scannell 2007)

URL classification problems make a proper language identification of the content necessary: especially for lesser-known languages, it is not so easy to find working patterns (Baykan et al. 2008)

Crawling without expert knowledge is "doomed to failure" (Scannell 2007)

Aim of the study

Post-BootCaT web text gathering

→ What are viable alternative data sources for low-resource languages?

Follows previous work on social network exploration (see references).

Discovering approach

First exploration step that could eventually lead to full-fledged crawls and linguistic processing and annotation:
Light scout: discover resources and build a language-classified URL directory

→ How far is it possible to go using different types of sources?
→ What does it reveal about the linguistic nature of the afferent resources and about the challenges to address?

Languages studied

> *"Large standard languages – those with numbers of native speakers in the hundreds and tens of millions and having a long tradition of writing – are not necessarily high- or even medium-density languages" (Borin 2009)*

Indonesian (*Bahasa Indonesia*) is a good example

Population of 237,424,363 of which 25.90 % are internet users[a]

Multiethnicity in Southeast Asia, but still more than 60 million Internet users in Indonesia alone.

Hypothesis: the Indonesian web is not well connected to the western world (technicalities and cultural interlinking).

[a] 2011, official Indonesian statistics institute (http://www.bps.go.id)

Languages studied (II)

All studies performed on Indonesian, some on Malaysian
Interpretation aware of the language pair.

Comparison with a Scandinavian language pair:
Danish and Swedish
Medium-resourced languages, impact on production processes (epilinguistic knowledge) and on language identification.

Data sources

Open Directory Project

DMOZ[a]: selection of links curated according to their language and/or topic.

→ What are these URLs worth for language studies and web corpus construction ?

[a] http://www.dmoz.org/

DMOZ URL extraction courtesy of Roland Schäfer (FU Berlin).

Wikipedia

free encyclopedia as another spam-resilient data source

→ Do the links from a particular edition point to relevant web sites (with respect to the language of the documents they contain) ?

Processing pipeline

1. URL harvesting: archive/dump traversal, obvious spam and non-text documents filtering
2. Operations on the URL queue: redirection checks, sampling by domain name
3. Download of the web documents and analysis: collection of host- and markup-based data, HTML code stripping, document validity check, language identification

Important characteristics

- Links pointing to media documents were excluded
- Spam filtering on URL (regex) and at domain name level (blacklist)
- Inefficient to crawl the web very broadly: parallel threads + results are merged in the end of each step (Scannell 2007)
- Domain name sampling

Web page and corpus size metrics

- Web page length in characters was used as a discriminating factor Before and after HTML sampling
- Total number of tokens of a web corpus estimated
- IPs recorded: host diversity
- Hashing on text level: basic duplicate detection

Language identification with langid.py

Lui, M. and Baldwin, T., 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the ACL*.
https://github.com/saffsd/langid.py

- Pre-trained statistical model and covers 97 languages
- Used as a web service (distributed work)
- Underlying classification (texts without surprises)

Results: DMOZ

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
ID	2,336	1,088	71.0	5,573	3,922	540,371	81.5
MS	298	111	59.5	4,571	3,430	36,447	80.3
DA	36,000	16,789	89.6	2,805	1,652	5,465,464	32.6
SV	27,293	11,316	91.1	3,008	1,838	3,877,588	44.8

dropped URLs ratio + IP diversity indicator highlight resource quality

Indonesian–Malay pair: about 15 % each time in the concurrent language

far more text to be found for Danish and Swedish

Results: Wikipedia

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
ID	204,784	45,934	9.5	6,055	4,070	3,335,740	46.3
MS	90,839	21,064	3.5	6,064	3,812	548,222	59.1
DA	161,514	33,573	28.3	4,286	2,193	5,329,206	38.1
SV	320,887	62,773	29.7	4,058	2,257	8,388,239	32.7

Retained URLs/analyzed URLs ratio lower, but still constant

Clear case for scarcity of resources in Indonesian and Malay (English: 70 %)

The potential corpora based on Wikipedia contain more text

Results: Crawling experiments

Source	Depth	URLs		% in target	Length		Tokens (total)	% IP diversity
		analyzed	retained		mean	median		
DMOZ	3	32,036	14,893	34.7	6,637	4,330	4,320,137	34.0
Wiki	4	95,512	35,897	24.3	6,754	3,772	7,296,482	28.8

DMOZ and Wikipedia are good starting points to begin a web crawl, a reasonable amount of URLs is to be achieved in three or four steps.

Even when focused: web texts written in Indonesian seem relatively hard to find.
Web crawling is definitely an option, the number of tokens increases notably.
Target-specific strategies are necessary.

Discussion

Issues

Major issues: page access delays, server-related biases, and unexpected web space topography.

Cloaking obstacle: different content according to location and user-agent
→ Spoof the server location accordingly could improve both retrieval speed and content language

Other fine-grained instruments are needed, more linguistically relevant text quality indicator

Language documentation and web corpora

Corpus building in a way similar to language documentation (Austin 2010):
Scientific approach to the environmental factors required during information capture, data processing, archiving, and mobilization
Ensure proper conditions of information capture, data archiving and mobilization for web corpora, through the exploratory steps presented here.

Conclusion

- A possible way to go in order to gather a corpus using different sources.
- A plea for a technicalities-aware web corpus creation: a minimum of web science knowledge among the corpus linguistics community is necessary.
- A step towards reproducible research: the toolchain used to perform these experiments is open-source and can be found online:
  FLUX: Filtering and Language-identification for URL Crawling Seeds
  https://github.com/adbar/flux-toolchain

References

Austin, P. K., 2010. Current issues in language documentation. *Language documentation and description*, 7:12–33.

Barbaresi, A., 2013. Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the Annual Meeting of the ACL, Student Research Workshop*.

Baroni, M. and Bernardini, S., 2004. BootCaT: Bootstrapping corpora and terms from the web. In Proceedings of LREC.

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E., 2009. The WaCky Wide Web: A collection of very large linguistically processed web- crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baykan, E., Henzinger, M., and Weber, I., 2008. Web Page Language Identification Based on URLs. *Proceedings of the VLDB Endowment*, 1(1):176–187.

Borin, L., 2009. Linguistic diversity in the information society. In *Proceedings of the SALTMIL 2009 Workshop on Information Retrieval and Information Extraction for Less Resourced Languages*.

Kilgarriff, A., Reddy, S., Pomikalek, J., and Avinesh, PVS, 2010. A Corpus Factory for Many Languages. In *Proceedings of LREC*.

Scannell, K. P., 2007. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, vol. 4.

Contact: adrien.barbaresi@ens-lyon.fr

http://perso.ens-lyon.fr/adrien.barbaresi/

https://github.com/adbar/flux-toolchain

This work has been partially funded by an internal grant of the FU Berlin, COW (COrpora from the Web) project at the German Grammar Dept.