Finding viable seed URLs for web corpora

A scouting approach and comparative study of available sources

Adrien Barbaresi

ICAR Lab @ ENS Lyon

9th Web as Corpus Workshop
Gothenburg – April 26, 2014

icar
UMR 5191 - CNRS / Université Lyon 2
Interactions, Corpus, Apprentissages, Représentations

ENS DE LYON

Outline

The "Web as Corpus" paradigm...

The state of the art tools rely heavily on search engines

BootCaT method (Baroni & Bernardini 2004)[a]:
"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[b], Kilgarriff et al. 2010[c])

Until recently, could be used easily and at no cost

---

[a] M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web, *Proceedings of LREC*, p.1313–1316.

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

[c] A. Kilgarriff, S. Reddy, J. Pomikálek, and PVS Avinesh. 2010. A Corpus Factory for Many Languages, *Proceedings of LREC*, p.904–910.

... and its URL seeds problem

- Increasing limitations on the search engines' APIs
- Diverse and partly unknown search biases related to search engine optimization tricks and undocumented PageRank adjustments
- 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)

⇒ This is what I call the post-BootCaT world in web corpus construction[a]

[a] Barbaresi A. 2013. Challenges in web corpus construction for low-resource languages in a post-BootCaT world, *Proceedings of LTC 2013*.

Motivation

- Open question: does the method used so far provides a good overview of a language?
- Research material does not have to rely on a single biased data source
- Black box effect + paid queries really impedes reproducibility of research
- Potential lack of project financing: light-weight approaches
- A preliminary light scouting approach and a full-fledged focused crawler like the COW[1] or Spiderling[2] projects are complementary

[1] R. Schäfer and F. Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Proceedings of LREC, pages 486–493.

[2] V. Suchomel and J. Pomikálek. 2012. Efficient Webcrawling for large text corpora. In Proceedings of the 7th Web as Corpus Workshop, pages 40–44.

Looking for alternatives, what issues do we face? (I)

→ Where may we find web pages which are bound to be interesting for corpus linguists and which in turn contain many links to other interesting web pages?

- Search engines actually yield good results in terms of linguistic quality
- Purely URL-based approaches are a trade-off in favor of speed which sacrifices precision, e.g. by language identification tasks (Baykan et al. 2008)[3]
- Machine-translated content is a major issue, so is text quality in general (Arase & Zhou 2013)[4]

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

[4] Y. Arase and M. Zhou. 2013. Machine Translation Detection from Monolingual Web-Text. In Proceedings of the 51th Annual Meeting of the ACL, pages 1597–1607.

Looking for alternatives, what issues do we face? (II)

- Mixed-language documents slow down text gathering processes (King & Abney 2013) [5]
- Link diversity is also an (underestimated) problem
- The resource is constantly moving (redirections, content injected from other sources, etc.)

[5] B. King and S. Abney. 2013. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In Proceedings of NAACL-HLT, pages 1110–1119.

Aim of the study

Web text gathering

Find viable alternative data sources

→ What is a "good" URL base and where do we find that?

Light scout

Exploration step that could eventually lead to full-fledged crawls and linguistic processing
Discover resources and build a language-classified URL directory

→ What does it reveal about the linguistic nature of the afferent resources and about the challenges to address?

Languages studied

Several language-dependent web spaces which ought to have different if not incompatible characteristics

"Speaker to website quantity" ratio is probably extremely different

This affects greatly link discovery and corpus construction processes[6]

1. Dutch (spoken in several countries)
2. French (spoken on several continents)
3. Indonesian (surprisingly rare during random crawls)
4. Swedish (smaller, arguably more homogeneous speaker community)

[6] Barbaresi 2013, *op.cit.*

Data sources (I)

Search engine queries

URLs retrieved using BootCaT seed method on meta-engine E-Tools[a] at the end of 2012

[a] http://www.etools.ch/
URL extraction by COW project

Social networks

FriendFeed, identi.ca and Reddit
Data from a previous study[a]
URLs re-examined and relevant metadata added to allow for a comparison

[a] A. Barbaresi. 2013b. Crawling microblogging services to gather language-classified URLs. Workflow and case study. *Proceedings of the ACL SRW*, p.9–15.

Data sources (II)

Open Directory Project

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

[a] http://www.dmoz.org/
DMOZ URL extraction courtesy of Roland Schäfer

Wikipedia

Free encyclopedia as another spam-resilient data source
Content quality is expected to be high
What is the profile of the links pointing to the outside world?

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

Processing pipeline

1. URL harvesting: queries or archive/dump traversal, filtering of obvious spam and non-text documents
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
- URLs sampled by domain name prior to the download
- Web pages discarded if they are too short or too long

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
- Number of in-domain and out-domain links

Language identification with langid.py [7]

- Pre-trained statistical model, covers 97 languages
- Used as a web service (distributed work)
- Underlying classification (texts without surprises)
- Can encounter difficulties with rare encodings (not part of this study)

Corollaries

Enables finding "regular" texts in terms of statistical properties and excluding certain types of irregularities such as encoding problems.
Web text collections are smoothed out in relation to the statistical model applied for each language target: a partly destructive but interesting feature.

[7] M. Lui and T. Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the 50th Annual Meeting of the ACL*, p.25–30. https://github.com/saffsd/langid.py

Results: search engine queries

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
FR	16,763	4,215	70.2	47,634	8,518	19,865,833	50.5
ID	110,333	11,386	66.9	49,731	8,634	50,339,311	18.6
NL	12,839	1,577	84.6	27,153	3,600	5,325,275	73.1
SV	179,658	24,456	88.9	24,221	9,994	75,328,265	20.0

Domain name reduction has a substantial impact on the set of URLs

Not all links in the target language (bias of the querying methodology?)

Remarkable length of web documents and size of potential corpus sizes
$\rightarrow$ explains why this method has been considered to be successful

Diversity in terms of hosts seems to get gradually lower as the URL list becomes larger (strong tendency)

Results: social networks

	% in target	URLs retained	Length		Tokens (total)	Different IPs (%)
			mean	median		
FR	5.9	4,320	11,170	5,126	7,512,962	49.7
ID	0.5	336	6,682	4,818	292,967	50.9
NL	0.6	465	7,560	4,162	470,841	68.8
SV	1.1	817	13,807	7,059	1,881,970	58.5

Potential impurity of the source + no targeted language

Mean and median lengths are clearly lower, host diversity comparable

Probably a good candidate for web corpora, but they require a focused approach of microtext

Results: DMOZ

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
FR	225,569	80,150	90.7	3,635	1,915	35,243,024	33.4
ID	2,336	1,088	71.0	5,573	3,922	540,371	81.5
NL	86,333	39,627	94.0	2,845	1,846	13,895,320	43.2
SV	27,293	11,316	91.1	3,008	1,838	3,877,588	44.8

Best ratio in this study between unfiltered and remaining URLs

Adequacy of the web pages with respect to their language is excellent

Document length is the problem
eventual "real" text content somewhere else behind the homepage
the sheer quantity of listed URLs compensates for this fact

Relative diversity of IPs: wide range of websites

Much easier to crawl

Results: Wikipedia

	URLs		% in target	Length		Tokens (total)	Different IPs (%)
	analyzed	retained		mean	median		
FR	1,472,202	201,471	39.4	5,939	2,710	64,329,516	29.5
ID	204,784	45,934	9.5	6,055	4,070	3,335,740	46.3
NL	489,506	91,007	31.3	4,055	2,305	15,398,721	43.1
SV	320,887	62,773	29.7	4,058	2,257	8,388,239	32.7

Also compares favorably to search engines or social networks when it comes to the sampling operation and page availability

Major source of URLs

Often point to web pages written in a foreign language
large number of concise homepages in the total

Insight into the concentration of web hosting providers on the market

A word on in- and outlinks

No clear winner regarding the ratio in-/outlinks

Tendency towards a hierarchy:

1. Search engines
2. Social networks
3. Wikipedia
4. DMOZ

Discussion: Results

The search engines results are probably closer to the maximum amount of text that is to be found on a given website than the other sources

The relatively low performance of Wikipedia and DMOZ is compensated by the ease of URL extraction
In the course of time, DMOZ and Wikipedia are likely to improve over time concerning the number of URLs they reference

A combined approach could take the best of all worlds and provide a web crawler with distinct and distant starting points

Other fine-grained instruments as well as further decisions processes (Schäfer et al. 2013)[8] are needed
Future work could include a few more linguistically relevant text quality indicators

[8] Roland R. Schäfer, A. Barbaresi, and F. Bildhauer. 2013. The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. *Proceedings of the 8th Web as Corpus Workshop*, p.7–15.

Discussion: Language documentation and web corpora

Corpus building in a way similar to language documentation (Austin 2010)[9]

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

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

Conclusion

- I presented metrics to help qualify URL seeds
  + a possible way to go in order to gather a corpus using different sources
- No clear winner, complementary approaches are called for
  → It seems possible to replace or at least to complement BootCaT
- Search engine queries are a useful data source, but they lack diversity at some point
  → This may provide sufficient impetus to look for alternatives
- The first step of web corpus construction in itself can change a lot of parameters
  → Plea for a technicalities-aware web corpus creation and a minimum of web science knowledge among the corpus linguistics community

The FLUX-toolchain

- A step towards reproducible research, the toolchain is open-source and can be found online:
  FLUX: Filtering and Language-identification for URL Crawling Seeds
  https://github.com/adbar/flux-toolchain
- Allows for fast regular updates of the URL directory
  Could also take advantage of the strengths of other tools in order to suit other needs

Thank you for your attention

Contact: adrien.barbaresi@ens-lyon.fr

http://purl.org/adrien-barbaresi

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

This work has been partially supported by machine power provided by the COW (COrpora from the Web) project at the German Grammar Department.