Efficient construction of metadata-enhanced web corpora

WAC-X workshop @ ACL 2016

Adrien Barbaresi

Austrian Academy of Sciences – Berlin-Brandenburg Academy of Sciences

August 12th, 2016

Background

Shift from web *AS* corpus to web *FOR* corpus:

The golden age of web corpora may be behind us (Tanguy 2013)

- The “Web as corpus” paradigm is outdated
- Not enough machine power
- Copyright infringement issues
- No real change of paradigm between traditional and web corpora

The web corpus as a construct

Web corpora are heirs of historical, traditional corpora:

1. A corpus is always a construct implying conceptual and technological decisions
2. Historical importance of representativeness and text categorization
3. Web FOR corpus paradigm: the Web in itself is not suitable for linguistic research

The web corpus as a construct

Web corpora are heirs of historical, traditional corpora:

1. A corpus is always a construct implying conceptual and technological decisions
2. Historical importance of representativeness and text categorization
3. Web FOR corpus paradigm: the Web in itself is not suitable for linguistic research

⇒ The scientific building of web corpora still needs to be established (*Ad hoc and general-purpose corpus construction from web sources*, 2015)

The metadata problem

Potentially erroneous metadata in “one size fits all” web corpora undermine the relevance of web texts for linguistic purposes

e.g. Use of source type and date in lexicography to follow word creation and trends

Relevant issues

1. Relevant web documents (*which kind and where?*)
   - content language, layout, and quality
2. Extraction of text and metadata (*which information?*)
   - boilerplate removal
   - rich metadata

Ex-cursus: Blogs

"a reverse chronological sequences of dated entries" (Kumar et al. 2003)

"The cross-linking that takes place between blogs, through blogrolls, explicit linking, trackbacks, and referrals has helped create a strong sense of community in the weblogging world." (Glance et al. 2004)

Brief chronology

1996	acknowledged beginning of the blog/weblog genre
end of 90s	web diaries in Japan
1999	emergence of several user-friendly publishing tools
2003	WordPress
after 2008	sharp decrease of interest (short message services)

Discovery and construction: state of the art

No comprehensive directory

Corpus size and length of downloads are frequently mentioned as potential obstacles
→ convenient ways through needed (platform, feeds, etc.)

The chosen solution and its advantages

A software (WordPress) & its platform (wordpress.com)

wordpress.com: potentially more than 1,350,000 blogs in German

+ all the self-hosted websites using WordPress (approx. $\frac{1}{4}$ worldwide)

- Host diversity = various user profiles?
- Same software

⇒ Comparable if not same content **structure**
⇒ Potentially identical extraction procedures

Into the wild: discovery of blogs

Detection of WordPress software

- Search for URL patterns in URL lists
  so-called Permalinks on WP: 5 common URL structures
  such as *?p=* or */[year]/[month]/*
- *and/or* HTTP HEAD requests
  no webpage is actually ?seen? during the process, which makes it a lot faster
  - XMLRPC pingback information on the homepage
    + points to the ?real? domain name
  - Look for WP extensions headers
  - HTTP response code and header analysis of /login and /feed

Into the wild: processing pipeline

Repeated use of FLUX (Filtering and Language identification for URL Crawling Seeds)

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

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

Sources

1. URLs from the CommonCrawl
2. the CDX index query frontend of the internet Archive
3. public instances of the metasearch engine Searx

Extraction: Blogs come in all forms and colors

Exotic markup & text genres make it difficult to extract the content.

Filtering problems by "one size fits all" web corpora:

- on website scale:
  lists of cars for sale or addresses of dentists are poor CMC data...
- on webpage scale:
  it is possible to filter out tag clouds, post lists and left/right columns in general, but poor metadata

References of the trilogy on web corpus construction

- Schäfer R., Barbaresi A., & Bildhauer F. (2013). "The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction." in *Proceedings of the 8th Web as Corpus Workshop (WAC8)*
- Barbaresi A. & Würzner K.M. "For a fistful of blogs [...]" in *Proceedings of NLP4CMC 2014*
- Barbaresi A., "For a few points more: Improving decision processes in web corpus construction", DGfS conf. 2014 + PhD thesis 2015

Metadata extraction

Wrappers/scrapers approaches:

- wrapper induction
- sequence labeling
- statistical analysis and heuristics

Common targets:

Title, date, author, content, number of comments, archived link, trackback link, comments

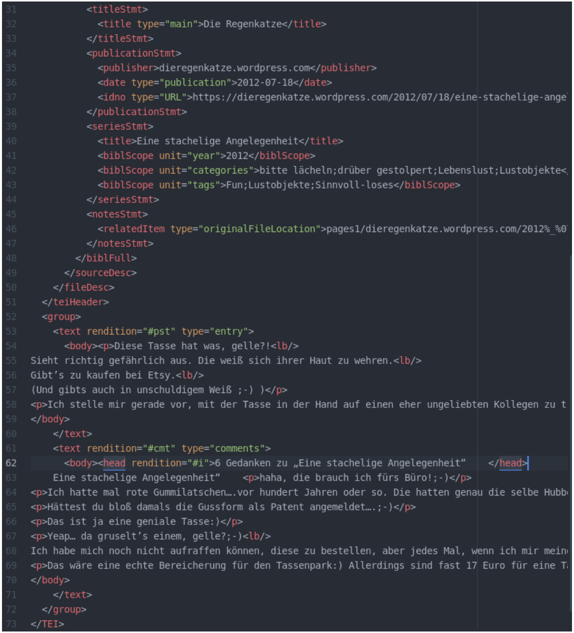
Extraction

1. HTML parse
2. subtree selection with XPATH-expressions
3. tag conversion and pruning
4. output in XML TEI format

Targets

- Title of post, title of blog, date of publication, canonical URL, author, categories, and tags
- Posts // Comments
  (text structuration: titles, paragraphs, bold and italic, no links)

⇒ Extraction as proxy for quality assessment:
Full duplicates, short documents, and documents without date removed



```
          <titleStmt>
            <title type="main">Die Regenkatze</title>
          </titleStmt>
          <publicationStmt>
            <publisher>dieregenkatze.wordpress.com</publisher>
            <date type="publication">2012-07-18</date>
            <idno type="URL">https://dieregenkatze.wordpress.com/2012/07/18/eine-stachelige-ange
          </publicationStmt>
          <seriesStmt>
            <title>Eine stachelige Angelegenheit</title>
            <biblScope unit="year">2012</biblScope>
            <biblScope unit="categories">bitte lächeln;drüber gestolpert;Lebenslust;Lustobjekte<
            <biblScope unit="tags">Fun;Lustobjekte;Sinnvoll-loses</biblScope>
          </seriesStmt>
          <notesStmt>
            <relatedItem type="originalFileLocation">pages1/dieregenkatze.wordpress.com/2012%_%0
          </notesStmt>
        </biblFull>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <group>
    <text rendition="#pst" type="entry">
      <body><p>Diese Tasse hat was, gelle?!<lb/>
Sieht richtig gefährlich aus. Die weiß sich ihrer Haut zu wehren.<lb/>
Gibt’s zu kaufen bei Etsy.<lb/>
(Und gibts auch in unschuldigem Weiß ;-) )</p>
<p>Ich stelle mir gerade vor, mit der Tasse in der Hand auf einen eher ungeliebten Kollegen zu t
</body>
    </text>
    <text rendition="#cmt" type="comments">
      <body><head rendition="#i">6 Gedanken zu „Eine stachelige Angelegenheit“    </head>
    Eine stachelige Angelegenheit“    <p>haha, die brauch ich fürs Büro!;-)</p>
<p>Ich hatte mal rote Gummilatschen….vor hundert Jahren oder so. Die hatten genau die selbe Hubb
<p>Hättest du bloß damals die Gussform als Patent angemeldet….;-)</p>
<p>Das ist ja eine geniale Tasse:)</p>
<p>Yeap… da gruselt’s einem, gelle?;-)<lb/>
Ich habe mich noch nicht aufraffen können, diese zu bestellen, aber jedes Mal, wenn ich mir mein
<p>Das wäre eine echte Bereicherung für den Tassenpark:) Allerdings sind fast 17 Euro für eine T
</body>
    </text>
  </group>
</TEI>
```

Experiment 1: Retrieving German blogs

→ Starting point: 158,719 blogs in German found on wordpress.com (Barbaresi & Würzner 2014)

wordpress.com corpus

- 145,507 websites seen – 141,648 kept
- 6,605,078 documents – 6,024,187 “valid” files
- 390 Gb downloaded – 36 Gb after processing
- 2.11 billion tokens
- Comments extracted for 1,454,752 files (24%)

Year	Docs.
2003	1,746
2004	4,993
2005	13,916
2006	62,901
2007	191,898
2008	377,271
2009	575,923
2010	733,397
2011	871,108
2012	1,066,996
2013	1,108,495
2014	717,861
2015	362,633
rest	6,068

Table : Distribution of documents among plausible years in the first experiment

	Name	**Freq.**	**Translation**
1	Fotografie	35,910	*photography*
2	Berlin	34,553	
3	Deutschland	30,351	*Germany*
4	Leben	29,597	*life*
5	Politik	26,315	*politics*
6	Musik	26,221	*music*
7	Foto	26,202	
8	Liebe	24,865	*love*
9	Kunst	24,382	*art*
10	USA	21,059	
11	Fotos	20,829	*pictures*
12	Natur	17,490	*nature*
13	Gedanken	16,542	*thoughts*
14	Weihnachten	16,344	*christmas*
15	Video	16,329	

Table : Most frequent tags in the first experiment

Experiment 2: Targeting the .at-domain

→ .at-domain 32th TLD, about 3,7 million hosts reported

- 5,664 different domain names – 7,275 after
- 2,589,674 files – about 2 million "valid" files
- 159 Gb – 14 Gb
- 550 million tokens
- Comments extracted for 181,246 files (7%)

	Name	Freq.	Translation
1	Allgemein	28,005	*general*
2	Blu-ray	10,445	*(laser disc standard)*
3	MedienFamilie	9,662	*media-family*
4	Blog	9,652	
5	Familienleben	9,278	*family life*
6	News	8,857	*(also German)*
7	Film	8,222	*movies*
8	Absolut-Reisen	6,964	*absolute travels*
9	Buch	6,146	*book*
10	Schule	6,108	*school*
11	Spiele	5,939	*games*
12	Familienpolitik	5,781	*family policies*
13	Gewinnspiel	5,607	*competition*
14	In eigener Sache	5,463	*in our own cause*
15	Uncategorized	5,150	*(defaut category)*

Table : Most frequent categories in the second experiment

	Name	Freq.	Translation
1	Wien	18,973	*Vienna*
2	Deutschland	18,895	*Germany*
3	Usermeldungen	14,409	*user reports*
4	Österreich	10,886	*Austria*
5	Angebot aus DE	10,155	*offer from Germany*
6	sex	10,112	
7	Frauen	9,541	*women*
8	Kinder	8,968	*children*
9	USA	8,013	
10	Urlaub	7,767	*holiday*
11	homemade	7,666	
12	amateur	7,660	
13	mydirtyhobby	7,635	
14	Recht	7,611	*law*
15	Arbeitsrecht	7,294	*labor legislation*

Table : Most frequent tags in the second experiment

Conclusions on the experiments

All blogs in a formal sense, but strong differences

Typological gap between original and current studies as well as between users of a platform and users of a content management system

Relatively few interlinking and interaction

Sketches the typical profile of a **passive internet consumer, a "prosumer" at best**, which should be taken in consideration

Attempt at a typology

- Blogs **mimic existing text types**, audiences, and motivations, with a focus on information (general, specialized, or community-based) as well as on promotional goals
- Websites whose **finality is to sell** information, entertainment, or concrete products and services

Conclusions on the corpora

The resulting corpus **complies with formal requirements** on metadata-enhanced corpora and on weblogs considered as a series of dated entries.

The trade-off to gain metadata using focused downloads following strict rules seems to get enough traction to build larger web corpora:
A total of **550 Gb** of actually downloaded material leads to about **2.7 billion tokens** with rich metadata.

This comparatively high yield is a step towards more efficiency with respect to machine power and “Hi-Fi” web corpora, which could help promoting **web sources and updates** of research methodology.

tl;dr

1. a method to find and download large amounts of WordPress pages
2. a targeted extraction of content featuring much needed metadata
3. an analysis of the documents in the corpus with insights of actual uses of the blog genre

Thank you for your attention!

✉ adrien.barbaresi@oeaw.ac.at

@adbarbaresi

http://www.oeaw.ac.at/ac
http://adrien.barbaresi.eu/blog/