COW & texrex: Gigatoken Web Corpora and Tools for Web Corpus Construction

Roland Schäfer & Felix Bildhauer
Freie Universität Berlin
`http://hpsg.fu-berlin.de/cow/`

Sonderforschungsbereich 632
Informationsstruktur:
Die sprachlichen Mittel der Gliederung von Äußerung, Satz und Text
UP HUB FUB

Goals and Data

Target Corpora

- building on work on the **WaCky copora** (Baroni et. al., 2009), introducing incremental improvements
- general purpose and very large, enabling linguistic research of **low-frequency phenomena**
- over 5 billion tokens, better **over 10 billion tokens**; large enough **to derive purpose-specific corpora**
- best possible **random samples from the web** (by top-level domain)
- mostly **free of duplication** on the concordance-level
- containing a considerable amount of **quasi-spontaneous and substandard language** (chats, forums, etc.)
- languages: German, UK and World English, Castilian, Swedish, French; planned: Dutch, Danish, Malay, . . .

Software

- full tool chain for ad-hoc corpus creation **including crawler** (not including linguistic processing)
- **independence of search engine results**; guaranteed **no-cost corpus construction**
- **efficient**, **cross-platform** (written in ObjectPascal with the FreePascal compiler), **open-source**

Data Collection

- for current corpora: long or very long **web crawls** using Heritrix 1.4 (similar to Emerson and O'Neil, 2006)
- **seed URLs** for Heritrix: search engine results (Yahoo, Bing)
- maximum number of documents crawled so far for one TLD (DECOW2012): **130,602,410**

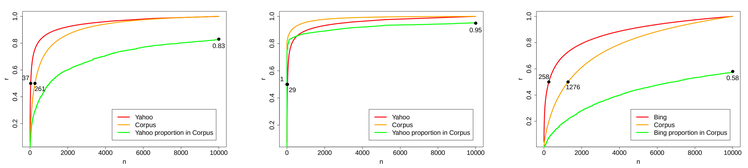
Problems with Established Methods

- **BootCaT method** (Baroni and Bernardini, 2004): relying entirely on search engine results
- **WaCky method** (Baroni et. al., 2009): starting with search engine results and doing **short web crawls**
- random samples from search engine index: a complicated matter, cf. Bar-Yossef and Gurevich (2006)
- random sample from the web: impossible; not the same as a random sample from a search engine
- **problem 1** pseudo-random samples from a search engine not state-of-the-art in web corpus construction
  ⇨ weakness of **mid-frequency tuple query method** (**not** the methods of Bar-Yossef and Gurevich, 2006)
- **problem 2** **shutdown of free search engine API access** for massive URL requesting
- **problem 3** **host bias**: search engines and short (breadth-first) crawls leading to **samples unnecessarily biased towards certain web hosts**

Corpus	Documents	Hosts	Crawl time	**Docs/Hosts**
DEWaC (WaCky)	1,501,076	9,502	10 d	**158**
ESCOW2012	1,295,387	41,900	28 d	**30**
DECOW2012	7,632,384	372,687	28 d	**20**

Host bias plots:

- **how large a proportion** r **of the documents stems from the top** n **hosts**
- . . . both in the **seed set** used and in the **final corpus**
- . . . and how large a proportion of the documents in the final corpus stems from the top n seed hosts



Host bias for ESCOW2011, SECOW2011, DECOW2012

Possible **failure of short crawls**:

- source of **75% of the final SECOW2011 corpus**: `http://www.blogg.se/`
- showing that **content/genre bias** comes easily with **host bias**

Conclusion:

- necessity of long **deep crawls** with immense storage requirements – therefore:
- development of our **own corpus crawler** `heidix`: actively enforcing randomness and avoiding bias, cleansing on-the-fly to keep storage requirements down (no need to keep "bad" corpus documents)
- alternative ways of **seed URL generation**
  ⇨ current experiment: `http://de.wikipedia.org/wiki/Special:Random`

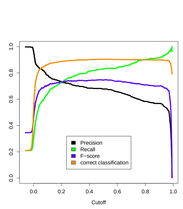
Post-Processing

Basic Cleanup

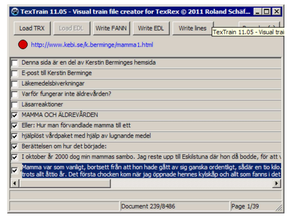
Heritrix ARC file parsing, HTML markup removal, HTML entity conversion, UTF8-to-ISO8859 conversion, some punctuation cleanup, duplicate line removal, primitive paragraph detection, . . . is all performed on-the-fly and in a fault tolerant way by our **`texrex` tool**.

Boilerplate Removal

- decision for each paragraph: **good corpus text** or **boilerplate** (menues, copyright notice, etc.)
- **multi-layer perceptron** using `libfann` (cf. Nissen, 2005)
- input to network: currently 9 values calculated by `texrex`
- output from network: a real between 0 (totally boilerplate) and 1 (totally good text)
- user-settable cutoff to favor **precision** or **recall**
- graphical tool `textrain` to train own networks included



Quality of boilerplate removal depending on cutoff (pre-packaged network on 1,000 unseen paragraphs)



Screenshot of `textrain` program

Connected Text Recognition

- **problem** not simply **foreign language** documents, but **tag clouds**, **lists**, **tables**, etc.
- simple language identification (a textbook matter) insufficient
- **WaCky method**: requiring certain type and token counts of function words
- **problem** depends heavily on document length; is unreliable for long documents with mixed text
- our solution: calculate the **standardized summed negative deviation** $B(d)$ **of frequencies of certain function words per document** d compared to a set T of training documents
- `texprof`: generator for profiles over n tokens $t_{1..n}$ based on a manually selected T
- $\mu(t_i)$: weighted mean for t_i in T; $\sigma^2(t)$: corresponding weighted standard deviation
- for unseen document d in production run, $f(t,d)$: relative frequency of t in d

$$z(t,d) = \frac{\mu(t) - f(t,d)}{\sigma^2(t)} \qquad b(t,d) = \begin{cases} z(t,d) & \text{if } z(t,d) > 0 \\ 0 & \text{else} \end{cases} \qquad B(d) = \sum_{i=1}^{n} b(t_i,d)$$

- for COW corpora: removal of documents with $B(d) > 10$ (very strict, so: Recall < 0.8, Precision > 0.95)

Perfect and Near Duplicate Removal

1. **problem** large amount of **near**-duplication on the web; up to **50% of documents** are (near-)duplicates
2. solution: **w-shingling** (Broder et. al., 1997); estimate **Jaccard coefficient** of two documents' n-gram sets
3. as opposed to BootCaT: **proper implementation of w-shingling** (without clustering)
4. native 64-bit Rabin hash implementation (Rabin, 1981) in `teshi` (configurable parallelization)
5. for COW corpora: shorter of two documents removed if fingerprint overlap is 5% or higher
6. deduping fine-tunable; easy experimenting due to optimal re-use of calculations after settings change

Software Performance

1. not yet fully parallelized; some time lost for decompression/recompression of ARC files
2. benchmark machine: Xeon 5160 at 3.00 GHz, 12 GB RAM
3. for DECOW2012 corpus (9.1 billion tokens in 7.6 million documents):
   - 8.8 CPU texrex days (processing 130,602,410 input documents = **170 documents/s**)
   - 2.3 CPU shingling days (8 threads) (processing input 16,935,226 documents = **85 documents/s**)

Evaluation

Duplication

Reduction in number of documents by w-shingling (5% or higher w-shingling overlap); notice remaining duplication in WaCky corpora due to "simplification" of shingling algorithm:

Corpus	Before	After	% Reduction
DECOW2012	16,935,226	7,632,384	54.9
ESCOW2012	3,498,351	1,295,387	63.0
DEWAC (WaCky)	1,751,903	1,501,076	14.3
FRWAC (WaCky)	2,268,304	1,473,513	35.0

Assessment of remaining duplication:

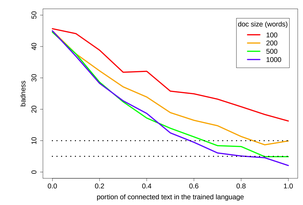
- choose frequent word W
- examine all occurrences of W plus n characters to its left and right; count repetitions
- no distinction between citations and duplicates; multiple counts per document possible; yet conservative estimate of duplication

Results for $n = 60$:

	DEWAC (WaCky)	DECOW2011	DECOW2012
N tokens	1,627,169,557	1,200,246,297	9,108,097,177
W	% **duplicated**		
hat	12.01	4.05	**6.52**
haben	11.28	5.33	**6.24**
ist	11.36	4.56	**5.69**
sind	11.44	4.46	**6.28**

Quality of Connected Text Detection

- as a simple language identifier on 105 German and 15 non-German/dialect test documents:
  Precision $= 1$, Recall $= 0.97$, $F = 0.99$
- promising results as connected text identifier, but **problem of defining what a good document is**
- **goal**: provide a set of training and test documents to define a **gold standard**
- right: $B(d)$ for synthetic German test documents of different sizes containing different amounts of tag cloud material



Genres and Text Types

- classification scheme based on Sharoff (2006), with modifications
- manual coding of 200 documents per corpus
- substantial to almost perfect inter-coder agreement (measured for German corpus only)
- *CI* given for 90% confidence level, $n = 200$

Variable	% Agreement	Cohen's κ
Authorship	89.0	.85
Mode	98.0	.94
Audience	88.0	.64
Aim	73.0	.61
Domain	86.0	.82

	DECOW2012		ESCOW2012	
Type	Percentage	CI ±%	Percentage	CI ±%
Authorship				
Single, female	**6.0**	2.8	**5.0**	2.5
Single, male	**11.5**	3.7	**16.5**	4.3
Multiple	**36.0**	5.6	**16.5**	4.3
Corporate	**21.0**	4.7	**20.5**	4.7
Unknown	**25.5**	5.0	**41.5**	5.7
Mode				
Written	**71.0**	5.0	**86.0**	4.0
Spoken	**1.0**	3.0	**2.5**	1.8
Quasi-Spontaneous	**22.5**	4.9	**3.5**	2.1
Blogmix	**4.5**	2.4	**8.0**	3.2
Audience				
General	**75.5**	5.0	**94.0**	2.8
Informed	**17.0**	4.4	**2.5**	1.8
Professional	**7.5**	3.0	**3.5**	2.1

	DECOW2012		ESCOW2012	
Type	Percentage	CI ±%	Percentage	CI ±%
Aim				
Recommendation	**12.5**	3.8	**7.0**	3.0
Instruction	**4.5**	2.4	**6.0**	2.8
Information	**36.0**	5.5	**41.5**	5.7
Discussion	**47.0**	5.8	**44.5**	5.8
Fiction	**0.0**	0.0	**1.0**	1.2
Domain				
Science	**2.5**	1.8	**5.0**	2.5
Technology	**14.0**	4.0	**6.5**	2.9
Medical	**4.5**	2.4	**4.0**	2.3
Pol., Soc., Hist.	**21.5**	4.8	**21.0**	4.7
Business, Law	**10.0**	3.5	**12.5**	3.8
Arts	**8.5**	3.2	**8.5**	3.2
Beliefs	**5.0**	2.5	**3.0**	2.0
Life, Leisure	**34.0**	5.5	**39.5**	5.7

References

Bar-Yossef, Ziv and Gurevich, Maxim. 2006. Random Sampling from a Search Engine's Index. In *Proceedings of WWW 2006*, Edinburgh.

Baroni, Marco and Bernardini, Silvia. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316.

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano and Zanchetta, Eros. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3), 209–226.

Broder, Andrei Z., Glassman, Steven C., Manasse, Mark S. and Zweig, Geoffrey. 1997. Syntactic Clustering of the Web. Technical Note 1997-115, SRC, Palo Alto.

Emerson, Thomas and O'Neil, John. 2006. Experience Building a Large Corpus for Chinese Lexicon Construction. In Marco Baroni and Silvia Bernardini (eds.), *Wacky! Working papers on the Web as Corpus*, Bologna: GEDIT.

Nissen, Steffen. 2005. Neural Networks made simple. *Software 2.0* 2, 14–19.

Rabin, Michael O. 1981. Fingerprinting by Random Polynomials. Technical Report TR-CSE-03-01, Center for Research in Computing Technology, Harvard University, Harvard.

Sharoff, Serge. 2006. Creating General-Purpose Corpora Using Automated Search Engine Queries. In Marco Baroni and Silvia Bernardini (eds.), *Wacky! Working papers on the Web as Corpus*, Bologna: GEDIT.