

Building Large Corpora from the Web

Marco Baroni

SSLMIT, University of Bologna

LCT Colloquium, Bolzano, January 19 2006

Outline

- 1 Introduction
- 2 The procedure
- 3 WaCky corpora
- 4 Conclusion

Corpora: what and why

- Collections of natural text stored on computer
- Useful for:
 - NLP (e.g., speech recognition, text categorization, question answering, machine translation. . .)
 - lexicography, grammar writing, language teaching
 - theoretical linguistics?

Corpora: what and why

- Collections of natural text stored on computer
- Useful for:
 - NLP (e.g., speech recognition, text categorization, question answering, machine translation. . .)
 - lexicography, grammar writing, language teaching
 - theoretical linguistics?
- Typology and size:
 - “Balanced”/representative/reference corpora: BNC, 100M tokens
 - Mono-source: Gigaword, 1B tokens (newswire)

Minimum requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing for fast, sophisticated queries

Minimum requirements for a modern corpus

- POS-tagging and lemmatization
- Indexing for fast, sophisticated queries
- Also desirable:
 - Parsing
 - Categorization of documents
 - ...

An example

```
[lemma="cause" & pos="VV.*"][pos="AT.*"]? [pos="AJ.*"]* [pos="NN.*"];
```

even times less likely to <cause irritation> . For extra protection motor cells , whose axons <cause the muscles> which withdraw the gi nce the spirochaete which <caused syphilis> had been identified and illery close at hand . . . <causing rattling> of windows and shaking " and walking aerobically <causes the lungs> to take in more air wi , any one of which is to <cause a successful jump> , is specified ntensive agriculture that <caused the rhinoceros> , like the elepha tress Sinister . It would <cause a scandal> . Come now , I 've obli n government revenue will <cause a surge> of new bill issues . Las ombed Ljubljana airport , <causing the deaths> of two Austrian jour system the vibrations can <cause a segment> to jump into a hole by s . Cholera toxin did not <cause nausea> or vomiting . Stool volu entifying what it is that <causes the reduction> in utility ie. in ere gas tissue interfaces <cause susceptibility> artefacts which ma rse clip over the mid CBD <causing a total stop> (Fig 1) . An att in the exchange rate will <cause a fall> in imports and a rise in lder . Their absence will <cause a major reshuffle> of the side by rankenstein 's family and <caused the deterioration> of Victor 's h fore the day was out , it <caused a great deal> of laughter . When

An example (cont.)

```
count by lemma on matchend;
```

620	problem
453	damage
259	death
232	harm
174	injury
171	trouble
140	concern
140	difficulty
122	change
110	loss

Why the corpora we have are not enough

- Not many corpora available

Why the corpora we have are not enough

- Not many corpora available
- Zipf/data sparseness

The Web is a corpus!

- The Web is a huge database of documents, mostly text.

The Web is a corpus!

- The Web is a huge database of documents, mostly text.
- Pretty much all written textual typologies and languages are attested on Web, often in huge quantities. . .

The Web is a corpus!

- The Web is a huge database of documents, mostly text.
- Pretty much all written textual typologies and languages are attested on Web, often in huge quantities. . .
- and there are interesting, new forms of computer-mediated communication somewhere between written and oral language.

The WaCky approach

- **Web as Corpus** kool y initiative
- `http://wacky.sslmit.unibo.it` (WaCky wiki, mailing list)
- `http://sslmitdev-online.sslmit.unibo.it/wac/post_processing.php`
- Collaborators outside SSLMIT: Serge Sharoff, Stefan Evert, Adam Kilgarriff, Massimiliano Ciaramita...

The WaCky approach

- **Web as Corpus** **kool y** initiative
- `http://wacky.sslmit.unibo.it` (WaCky wiki, mailing list)
- `http://sslmitdev-online.sslmit.unibo.it/wac/post_processing.php`
- Collaborators outside SSLMIT: Serge Sharoff, Stefan Evert, Adam Kilgarriff, Massimiliano Ciaramita...
- Something simple, but concrete!

The WaCky approach

- **Web as Corpus** **kool y**nitiative
- `http://wacky.sslmit.unibo.it` (WaCky wiki, mailing list)
- `http://sslmitdev-online.sslmit.unibo.it/wac/post_processing.php`
- Collaborators outside SSLMIT: Serge Sharoff, Stefan Evert, Adam Kilgarriff, Massimiliano Ciaramita...
- Something simple, but concrete!
- Emphasis on collaboration, using existing open tools, make developed tools publicly available.

The WaCky approach (cont.)

- Current status:
 - Large corpora built for German, Italian
 - Ongoing work on English, Japanese, Russian, Chinese

Outline

- 1 Introduction
- 2 The procedure**
- 3 WaCky corpora
- 4 Conclusion

Basic steps

- Select “seed” urls.
- Crawl.
- Post-processing.
- Linguistic annotation.
- Indexing.

Selecting seed urls

- Query Google search engine (via API) for random word combinations, and use urls found in this way as seeds.
- How random are the urls collected in this way? Work with Massimiliano Ciaramita to be presented at EACL 2006 suggests: BNC-like random.

Crawling with Heritrix

- `http://crawler.archive.org/`
- Free/open Java crawler of Internet Archive

Crawling with Heritrix

- `http://crawler.archive.org/`
- Free/open Java crawler of Internet Archive
- Supported by active community...

Crawling with Heritrix

- `http://crawler.archive.org/`
- Free/open Java crawler of Internet Archive
- Supported by active community...
- that includes linguists and machine learning experts

Important in a good crawler

- Honoring robots.txt, politeness

Important in a good crawler

- Honoring robots.txt, politeness
- Robust “Frontier”

Important in a good crawler

- Honoring robots.txt, politeness
- Robust “Frontier”
- Avoid spider traps

Important in a good crawler

- Honoring robots.txt, politeness
- Robust “Frontier”
- Avoid spider traps
- Control over crawl scope, customizable

Important in a good crawler

- Honoring robots.txt, politeness
- Robust “Frontier”
- Avoid spider traps
- Control over crawl scope, customizable
- Intelligent management of downloaded text

Code removal and boilerplate stripping

- Removing HTML and javascript is not enough.

Code removal and boilerplate stripping

- Removing HTML and javascript is not enough.
- “Boilerplate”: links, navigational information, advertisement, etc.

Cleaning with a standard HTML formatter

Blackmore's Night Latest News
Ritchie Blackmore's Bio
Blackmore's Night Band Bios
Blackmore's Night Tour Info
Blackmore's Night Merchandise
Blackmore's Night Photo Gallery
Blackmore's Night Audio Clips

...

Register for
Blackmores Night
Email Updates!
Just enter your
email address in
the box below and
click the 'Sign up' button!

...

RITCHIE BLACKMORE A MUSICAL HISTORY...

1967 - RITCHIE BLACKMORE - who has previously played with such bands as the Outlaws, Screaming Lord Sutch, and Neil Christian & The Crusaders - is invited by ex-Artwoods/The Flowerpot Men keyboardist Jon Lord (who was invited by The Searchers ex-drummer, Chris Curtis) to form a new band. Other musician's would be auditioned from a Melody Maker ad in Deeves Hall in Hertfordshire.

1968- In February, the group would form as Roundabout, consisting of

The HTML density heuristic

- `http://www.smi.ucd.ie/hyppia/`
- We provide more efficient re-implementation

The HTML density heuristic

- `http://www.smi.ucd.ie/hyppia/`
- We provide more efficient re-implementation
- Basic observation: Content-rich section of page tends to occur in low-HTML-density area

The HTML density heuristic

- `http://www.smi.ucd.ie/hyppia/`
- We provide more efficient re-implementation
- Basic observation: Content-rich section of page tends to occur in low-HTML-density area
- Look for stretch that maximizes the quantity:
 $N(TOKEN) - N(TAG)$

Why it (mostly) works

TAG TAG TOKEN TOKEN TAG TAG TAG
TOKEN TAG TAG
TOKEN TAG TAG
TAG TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
TOKEN TAG TOKEN TOKEN TAG TOKEN TOKEN TOKEN
TAG TAG
TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN
TAG TAG TAG TAG TAG
TAG TOKEN TAG TAG TOKEN TAG

Cleaning with HTML density heuristic

- RITCHIE BLACKMORE - who has previously played with such bands as the Outlaws, Screaming Lord Sutch, and Neil Christian & The Crusaders - is invited by ex-Artwoods/The Flowerpot Men keyboardist Jon Lord (who was invited by The Searchers ex-drummer, Chris Curtis) to form a new band. Other musician's would be auditioned from a Melody Maker ad in Deeves Hall in Hertfordshire. 1968 - In February, the group would form as Roundabout, consisting of the three (with Chris Curtis on vocals) along with Dave Curtis on bass and Bobby Woodman on drums. After only a month of uncompromising rehearsals, BLACKMORE and LORD would be the only two remaining, bringing in vocalist Rod Evans (formally of the groups M15 and Maze), bassist Nick Simper (ex-Johnny Kidd & the Pirates) and drummer Ian Paice. In April, the band would change its name to DEEP PURPLE, and forego the name Concrete God. Using Vanilla Fudge as its model, the group records an album and is signed to EMI in the United Kingdom and Tetragrammation (Bill Cosby's label) in the United States. A few months later, the band performs its first major UK performance at the Sunbury Festival. In September, the band would release the first single from the album, "Hush", which reaches #4 on the US charts. The album, itself, Shades of Deep Purple, would reach the Top 25. In December, the band would release their rendition of Neil Diamond's "Kentucky Woman" a single which would also experience considerable chart success. 1969 - Early in the year, the band would release The Book of Taliesyn which would feature a revival of Ike and Tina Turner's "River Deep, Mountain High." In July, the band would release their self-titled album, Deep Purple. Shortly thereafter, their record label would go under and Rod Evans and Nick Simper would both leave the band. They would be replaced by singer Ian Gillan and bassist Roger Glover

HTML density heuristic: pros and cons

- Pros:
 - Does not require global statistics
 - Works across different page formats

HTML density heuristic: pros and cons

- Pros:
 - Does not require global statistics
 - Works across different page formats
- Cons:
 - Not able to identify discontinuous content-rich chunks
 - Requires tokenization: problem with East Asian languages

Zipfian filtering

- If at least 25% of the words in a document are not from short list of function words in target language, then document
 - is not in target language or
 - does not contain a high proportion of connected text

Near-duplicate detection

- Perfect duplicates trivial (compare fingerprints)

Near-duplicate detection

- Perfect duplicates trivial (compare fingerprints)
- Near-duplicates very common online (dynamically generated pages with slightly different contents, same document on different sites, etc.)

Near-duplicate detection

- Perfect duplicates trivial (compare fingerprints)
- Near-duplicates very common online (dynamically generated pages with slightly different contents, same document on different sites, etc.)
- One could compare all possible sequences of length n (n -grams), and measure overlap between two documents:

this is a short toy document
and this is a short toy document too

Near-duplicate detection

- Perfect duplicates trivial (compare fingerprints)
- Near-duplicates very common online (dynamically generated pages with slightly different contents, same document on different sites, etc.)
- One could compare all possible sequences of length n (n -grams), and measure overlap between two documents:

this is a short toy document
and this is a short toy document too

- Safe, but not efficient

The Shingling Algorithm

- Simplified version of method of: Broder, Glassman, Manasse and Zweig (1997). Syntactic Clustering of the Web. Sixth International World-Wide Web Conference.

The Shingling Algorithm

- Simplified version of method of: Broder, Glassman, Manasse and Zweig (1997). Syntactic Clustering of the Web. Sixth International World-Wide Web Conference.
- Steps:
 - For each document, remove function words.
 - Randomly select fixed number of n-grams from document.
 - Look for pairs of documents that share at least X of the randomly sampled n-grams.

The Shingling Algorithm

- Simplified version of method of: Broder, Glassman, Manasse and Zweig (1997). Syntactic Clustering of the Web. Sixth International World-Wide Web Conference.
- Steps:
 - For each document, remove function words.
 - Randomly select fixed number of n-grams from document.
 - Look for pairs of documents that share at least X of the randomly sampled n-grams.
- Unbiased estimate of overlap between pages.

The Shingling Algorithm

- Simplified version of method of: Broder, Glassman, Manasse and Zweig (1997). Syntactic Clustering of the Web. Sixth International World-Wide Web Conference.
- Steps:
 - For each document, remove function words.
 - Randomly select fixed number of n-grams from document.
 - Look for pairs of documents that share at least X of the randomly sampled n-grams.
- Unbiased estimate of overlap between pages.
- Our parameters: 25 5-grams; maximum acceptable overlap: $1/25$

POS tagging (and lemmatization)

- In principle, nothing special about Web data, but targeted training data needed (and serious tokenization problems)

POS tagging (and lemmatization)

- In principle, nothing special about Web data, but targeted training data needed (and serious tokenization problems)
- We used standard TreeTagger for German, re-trained TreeTagger with our own morphological resources for Italian

POS tagging (and lemmatization)

- In principle, nothing special about Web data, but targeted training data needed (and serious tokenization problems)
- We used standard TreeTagger for German, re-trained TreeTagger with our own morphological resources for Italian
- A few problems:
 - finanziert/VVPP ganz/ADJD erheblich/ADJD mit./NE
 - das/ART buch/VVIMP

Indexing and querying

- Once we have a corpus, we would like to use it,

Indexing and querying

- Once we have a corpus, we would like to use it,
- and share it via Web interface allowing linguists to do serious research

Indexing and querying

- Once we have a corpus, we would like to use it,
- and share it via Web interface allowing linguists to do serious research
- Need for flexible annotation-aware queries rules out:

Indexing and querying

- Once we have a corpus, we would like to use it,
- and share it via Web interface allowing linguists to do serious research
- Need for flexible annotation-aware queries rules out:
 - Standard relational db based solution (not flexible)

Indexing and querying

- Once we have a corpus, we would like to use it,
- and share it via Web interface allowing linguists to do serious research
- Need for flexible annotation-aware queries rules out:
 - Standard relational db based solution (not flexible)
 - Lucene/Nutch full text search engine (not annotation-aware)

Indexing and querying

- Once we have a corpus, we would like to use it,
- and share it via Web interface allowing linguists to do serious research
- Need for flexible annotation-aware queries rules out:
 - Standard relational db based solution (not flexible)
 - Lucene/Nutch full text search engine (not annotation-aware)
- Current experiments with specialized software: IMS Corpus WorkBench, Word SketchEngine

Indexing and querying

- Once we have a corpus, we would like to use it,
- and share it via Web interface allowing linguists to do serious research
- Need for flexible annotation-aware queries rules out:
 - Standard relational db based solution (not flexible)
 - Lucene/Nutch full text search engine (not annotation-aware)
- Current experiments with specialized software: IMS Corpus WorkBench, Word SketchEngine
- At the moment, neither solution satisfactory

Outline

- 1 Introduction
- 2 The procedure
- 3 WaCky corpora**
- 4 Conclusion

deWaC corpus

- Seeded from random Google queries for SDZ and basic vocabulary pairs.
- Crawl limited to `.at/ .de` domains, with URL-based regular expression to focus on HTML.
- On dedicated server running RH Fedora Core 3 with 4 GB RAM, Dual Xeon 4.3 GHz CPUs, about 2.5 TB hard disk space
- crawl took about 10 days, post-processing 1 week, post-processing 5 days, near-duplicate detection 4 days, annotation about 3 days (Italian processing slightly faster)
- Crawl output: 85GB *compressed* data
- Cleaned corpus: 1.65B tokens, 1.76M documents, 30GB of data *uncompressed and with annotation*

Comparison with APA corpus

Using Log-Likelihood Ratio score

deWaC		APA	
ich	hier	APA	NATO
dass	wir	Schluß	EU
und	man	Prozent	Forts
sie	nicht	Mill	AFP
ist	das	MRD	Dollar
oder	sind	Wien	Reuters
kann	so	Kosovo	Dienstag
du	mir	DPA	Mittwoch
wenn	ein	US	Donnerstag
was	da	am	sei

An application: potentially separable prefixes

Work done with Matiasek, Neubarth, Trost

- Identifying behaviour of potentially separable prefixes based on unambiguous contexts
- Pilot study of set of 250 *durch*- verbs:

disambiguated in DUDEN	198
disambiguated in APA	95 (only 1 not in DUDEN)
disambiguated in deWaC	241

- Crawl performed by Eros Zanchetta
- Seeds from Google queries for terms extracted from *la Repubblica* corpus and basic vocabulary list
- Procedure similar to deWaC
- 81GB gzipped archives from crawl
- Cleaned corpus: 1.9B tokens, 1.87M documents, 31GB of data including annotation

Comparison with *la Repubblica*

Log-Likelihood Ratio, function words only

itWaC		Repubblica	
ed	hai	ha	una
perchè	tali	ieri	due
delle	tuo	ma	il
tale	vi	un	suo
ti	nn	aveva	dopo
cui	nonché	hanno	non
presso	di	era	fa
ciao	tua	che	lui
tu	possono	più	si
te	ovvero	perché	adesso

An application: *ri-*

*ri-*repelling verbs in *la Repubblica*:

arrivare
sembrare
restare
continuare
prevedere
rimanere
capire
rispondere
bisognare
raggiungere

An application: *ri-*

itWaC *ri-*frequencies of same verbs

rispondere	18
capire	10
raggiungere	4
arrivare	4
prevedere	3
continuare	2
rimanere	1
sembrare	0
restare	0
bisognare	0

Outline

- 1 Introduction
- 2 The procedure
- 3 WaCky corpora
- 4 Conclusion**

Conclusion

- Building a large corpus by crawling is quite straightforward. . .

Conclusion

- Building a large corpus by crawling is quite straightforward. . .
- but devil is in the (terabytes of) details.

Some open issues

- Indexing/query system/Web interface for very large corpora

Some open issues

- Indexing/query system/Web interface for very large corpora
- Scaling up

Some open issues

- Indexing/query system/Web interface for very large corpora
- Scaling up
- Categorization

Some open issues

- Indexing/query system/Web interface for very large corpora
- Scaling up
- Categorization
- Universal post-processing (encoding hell! East Asian languages!)

Some open issues

- Indexing/query system/Web interface for very large corpora
- Scaling up
- Categorization
- Universal post-processing (encoding hell! East Asian languages!)
- Tuning linguistic tools to Web data (tokenization!)

Please join us!!!

- Hot topic, active community, plenty of unsolved problems, computational expertise needed!
- WaCky corpora available to whoever is interested in using them.
- EACL06 WaC Workshop in Trento!