# Building Large Linguistic Corpora by Web Crawling

Marco Baroni

CIMeC Language, Interaction and Computation group
Unversity of Trento

17/5/2007, Torino

## Collaborative work

- Eros Zanchetta, Silvia Bernardini, Adriano Ferraresi (SSLMIT, University of Bologna)
- Serge Sharoff (University of Leeds)
- Adam Kilgarriff (Lexical Computing)
- Stefan Evert (University of Osnabrück)

## Outline

Introduction

Googleology

The WaCky corpus construction pipeline

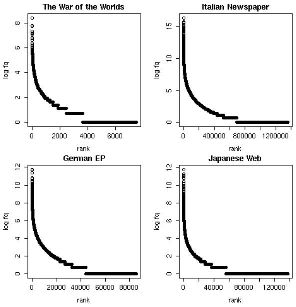
The WaCky corpora

CLEANEVAL

## Corpora in language research

- Linguistic corpora: large collections of “natural” textual data, typically enriched with linguistic annotation
- Useful for:
  - Training statistical algorithms (Machine Translation, Speech Recognition. . . )
  - Finding examples of real linguistic usage in linguistics and lexicography
  - Language teaching
  - Simulating linguistic input when modeling human language acquisition

## The need for (more) data



## The need for (more) data

- Current “national” corpora often around 100 million words, however. . .
- More data is better data
- Banko and Brill 2001 ACL (and many others afterwards):
  - It is better to train a naive algorithm on a lot of data than a smart algorithm on less data
  - After 1 billion words of training, statistical learners have not yet reached asymptote
- Not only most breadth, but also more depth needed
  - For most “minor” languages, not even small corpora available
  - For special topics/genres, not even small corpora available

## The Web as a source of linguistic data

- 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
- Oxford: from the BNC to the mostly WWW-based Oxford English Corpus

## Outline

Introduction

Googleology

The WaCky corpus construction pipeline
    Seeding the crawl
    Crawling
    Post-processing
    Linguistic annotation
    Indexing

The WaCky corpora

CLEANEVAL

## Googleology
Kilgarriff 2007

- Increasingly clever ways to gather frequency information about words and word sequences from Google or other commercial search engine
- E.g., Turney 2001, Keller and Lapata 2003, Nakov and Hearst 2005
- (and even Baroni and Bisi 2004, Baroni and Vegnaduzzo 2004)

## Googleology is bad science!

- You don't know what you are querying
- Queries bound to word-forms
- Linguist satisfaction in not high on the list of SE company priorities:
  - AltaVista removes NEAR operator in 2004
  - Google * wildcard mystery in 2006
  - Google stops supporting SOAP API in 2007

## Googleology is bad science!

- NLP researchers become googleologists
- Tricks to extract needed information in face of brittleness
  - E.g., add negation nonsense string to Google query to force engine to return full db count (Malvina Nissim)
- Effort spent by NLP community in developing Google-skills would be better spent building our own Google-sized corpora

## The WaCky initiative

- Web as Corpus kool ynitiative
- `http://wacky.sslmit.unibo.it`
- Now sponsored by SIGWAC (Special Interest Group on the Web as Corpus of the Association for Computational Linguistics)
- Something simple, but concrete
- Emphasis on collaboration, using existing open tools, make developed tools publicly available

## Outline

Introduction

Googleology

The WaCky corpus construction pipeline
- Seeding the crawl
- Crawling
- Post-processing
- Linguistic annotation
- Indexing

The WaCky corpora

CLEANEVAL

## Selecting “seed” URLs

- Query Google search engine (via API) for random word combinations, and use URLs found in this way as seeds
- Which random words?
  - Mid frequency words from general/newspaper corpus (“public”)
  - Basic vocabulary list (“private”)
- Ciaramita and Baroni (EACL 2006):
  - Method to select seeds that lead to most “random-like” corpus
  - Based on comparing frequency distribution of pages retrieved with seed set to frequency distributions of deliberately biased corpora

## Crawling with Heritrix

- `http://crawler.archive.org/`
- Free/open Java crawler of Internet Archive
- Adaptive crawling in the future?

## Code removal and boilerplate stripping

- Removing HTML and javascript is not enough
- “Boilerplate”: links, navigational information, advertisement, etc.
- HTML density heuristic to spot and remove boilerplate

## 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
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 TOKEN TOKEN
TAG TAG TAG TAG TAG
TAG TOKEN TAG TAG TOKEN TAG
```

## Language 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
- Also (mildly) useful fro discarding Web spam

## 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 |

- Not feasible for very large data-sets
- Need for sampling-based approach

## Near-duplicate detection

- Simplified version of “shingling” method of Broder et al, WWW-1997
- Select random sample of n-grams from each document, and measure overlap
- Our parameters:
  - 25 5-grams
  - maximum acceptable overlap: 1/25

## (Philosophical aside)

- Boilerplate and duplicates are part of our everyday linguistic experience as humans
- Advertisement, street signs, books on our shelf, train announcements at the station, friends who tell the same joke over and over…
- Why is it wrong to have them in a corpus?

## Tokenization and Web data

- Tokenization is the most overlooked phase of linguistic annotation
- Wrongly so, because tokenization problems will affect all other components
- Web text:
  - `tempo bello!!!;-)))`
  - `nmap -sS -v -oG /<path>/ file host bersaglio -oA`
  - `finanziert ganz erheblich mit.  das schlimmste ist wohl`

## POS-tagging and lemmatization

- In principle, nothing special about tagging/lemmatizing Web data, but:
  - targeted training data needed (current POS taggers are typically trained on newspaper data)
  - out-of-lexicon lemma guessing of fundamental importance

## Categorization by topic

Ongoing work by Serge Sharoff

- Dynamic approach to topic (and genre)
- Extract keywords from each corpus document
- Cluster documents by their shared keywords
- Currently seeking funding for WebDoc project

## Indexing and querying

- Once we have a corpus, we would like to use it,
- and share it via Web interface that allows 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)
- Currently, WaCky corpora split into set of IMS Corpus WorkBench sub-corpora
- Scripts to query multiple sub-corpora
- Long-term solution still missing

## Outline

Introduction

Googleology

The WaCky corpus construction pipeline
- Seeding the crawl
- Crawling
- Post-processing
- Linguistic annotation
- Indexing

The WaCky corpora

CLEANEVAL

## Current status

- Large corpora (between 1.5 and 2.25 billion tokens) built for:
  - English
  - Italian
  - German
- All available for download in simple XML format
- (Very approximate) estimates:
  - Italian corpus is about 6.8% of Italian indexed by Google (as of early 2006)
  - German corpus is about 3.1% of German indexed by Google

## itWaC details

- Seeds from Google queries for terms extracted from *la Repubblica* corpus and basic vocabulary list
- Crawl limited to `.it` 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, near-duplicate detection 1 day, annotation about 3 days
- 81GB gzipped archives from crawl
- Cleaned corpus: 1.9B tokens, 1.87M documents, 10GB of data before annotation

## Comparison of Italian Web corpus with *la Repubblica*

Log-Likelihood Ratio, function words only

| Web corpus | | | |
|---|---|---|---|
| ed | “and” | hai | “you have (sg.)” |
| perchè | “because/why” | tali | “such (pl.)” |
| delle | “of the (f. pl.)” | tuo | “your (m. sg.)” |
| tale | “such (sg.)” | vi | “you (acc./dat. pl.)” |
| ti | “you (acc./dat. sg.)” | nn | “not” |
| cui | “which” | nonché | “as well as” |
| presso | “at” | di | “of” |
| ciao | “hi” | tua | “your (f. sg.)” |
| tu | “you” | possono | “they can” |
| te | “you (acc./dat. sg.)” | ovvero | “or rather” |

| la Repubblica | | | |
|---|---|---|---|
| ha | “has” | una | “a/one (f.)” |
| ieri | “yesterday” | due | “two” |
| ma | “but” | il | “the (m. sg.)” |
| un | “a/one (m.)” | suo | “his/her (m. sg.)” |
| aveva | “had” | dopo | “after” |
| hanno | “they have” | non | “not” |
| era | “was” | fa | “makes” |
| che | “that” | lui | “he” |
| più | “more” | si | “it/her/himself” |
| perché | “why/because” | adesso | “now” |

## An application: compound interpretation

Baroni, Guevara and Pirrelli SLI 2006

- Induce semantic/grammatical properties of compounds by looking at direct paraphrases of compounds:
  - fair dedicated to horses → *subordinate relation*
  - singer and songwriter → *coordinate relation*
- Extraction of at least 1 potential paraphrase from *la Repubblica*: succeeds in 55% of cases
- Extraction of at least 1 potential paraphrase from *itWaC*: succeeds in 92% of cases

## Ongoing ukWaC work

- Comparison of BNC and ukWaC in lexicographic task (Adriano Ferraresi, Silvia Bernardini)
- Ad interim results of comparison of concept properties induced from corpora and generated by subjects (Massimo Poesio, Brian Murphy, Eduard Barbu, Luigi Lombardi)
- ukWaC outperforms (albeit mildly):
  - Googleology
  - 4-times larger newsgroup corpus
  - BNC was not even an option

## More data is better data?

SVD-based semantic spaces; ongoing work with Alessandro Lenci

- Nearest neighbours of *pensare*
- From *la Repubblica*:
  - credere, immaginare, sapere, illudere, domandare, sperare, accorgere, ricordare, sognare, progettare
- From *itWaC*:
  - credere, sapere, xkè, cazzo, piacere, cosa, siccomo, anka, sbagliare, cmq

## The copyright question

- Corpus creation
  - Is the CreativeCommons-licensed portion of the Web large and varied enough, yet?
- Web interface
  - Not worse than Google caching
  - Provide opt-out option
- Distributing the corpus
  - The family-business model
  - The Google-5-Gram approach
- Dynamic corpora
  - Distribute URL list and tools to harvest, clean, annotate them
  - Corpus should be random sample: distribute tools to sample, harvest, clean, annotate comparable corpora

## Outline

Introduction

Googleology

The WaCky corpus construction pipeline
- Seeding the crawl
- Crawling
- Post-processing
- Linguistic annotation
- Indexing

The WaCky corpora

CLEANEVAL

## CLEANEVAL

- Data play increasingly central role in NLP and related fields
- Move data acquisition and cleaning from "trivial" pre-processing step to center-stage
- CLEANEVAL: shared competitive task on automatically identifying contentful paragraphs in Web pages
- First CLEANEVAL workshop: Sept 15-16 2007, part of WAC3, Louvain-la-Neuve, Belgium
- A SIGWAC initiative
- `http://cleaneval.sigwac.org.uk`

## The End

Thank you!