

Web Corpus Creation, Cleaning and Evaluation

Web as Corpus Meeting @ eLex 2015

Egon Stemle, Stefan Evert and
The Special Interest Group of the ACL on Web as Corpus



+

Outline

- 1 Introduction
 - The Web as Corpus - Why?
 - Bird's-eye View of the Web – as seen by WaCky-ists
- 2 WaCky Corpus Creation
 - Search Engine Results - let's have more of them!
 - Load'em down - all! - yes, right now!
- 3 Corpus Cleaning
- 4 WaCky Corpus Evaluation

Outline

- 1 Introduction
 - The Web as Corpus - Why?
 - Bird's-eye View of the Web — as seen by WaCky-ists
- 2 WaCky Corpus Creation
 - Search Engine Results - let's have more of them!
 - Load'em down - all! - yes, right now!
- 3 Corpus Cleaning
- 4 WaCky Corpus Evaluation

Why Web Corpora?

Why Web Corpora?

- As replacement for linguistic reference corpora
 - main goal of the early WaC(ky) community
 - cheaper, larger and more up-to-date than traditional corpora
 - Web corpus should be similar to reference corpus

Why Web Corpora?

- As replacement for linguistic reference corpora
 - main goal of the early WaC(ky) community
 - cheaper, larger and more up-to-date than traditional corpora
 - Web corpus should be similar to reference corpus
- Computer-mediated communication (CMC)
 - Twitter, Facebook, chatroom logs, discussion groups, . . .
 - many Web genres share aspects of interactive CMC
 - Web corpus = targeted collection of CMC genres

Why Web Corpora?

because more data are better data

- As replacement for linguistic reference corpora
 - main goal of the early WaC(ky) community
 - cheaper, larger and more up-to-date than traditional corpora
 - Web corpus should be similar to reference corpus
- Computer-mediated communication (CMC)
 - Twitter, Facebook, chatroom logs, discussion groups, . . .
 - many Web genres share aspects of interactive CMC
 - Web corpus = targeted collection of CMC genres
- Scaling up NLP training data
 - 1964: 1 million words (Brown Corpus)
 - 1995: 100 million words (British National Corpus)
 - 2003: 1,000+ million words (English Gigaword, WaCky)
 - 2006: 1,000,000 million words (Google Web 1T 5-Grams)

Why Web Corpora?

because more data are better data

- As replacement for linguistic reference corpora
 - main goal of the early WaC(ky) community
 - cheaper, larger and more up-to-date than traditional corpora
 - Web corpus should be similar to reference corpus
- Computer-mediated communication (CMC)
 - Twitter, Facebook, chatroom logs, discussion groups, . . .
 - many Web genres share aspects of interactive CMC
 - Web corpus = targeted collection of CMC genres
- Scaling up NLP training data
 - 1964: 1 million words (Brown Corpus)
 - 1995: 100 million words (British National Corpus)
 - 2003: 1,000+ million words (English Gigaword, WaCky)
 - 2006: 1,000,000 million words (Google Web 1T 5-Grams)

Is bigger always better?

- From small, clean and well designed . . .
 - British National Corpus (BNC)
 - movie subtitles, newspapers, . . .

Is bigger always better?

- From small, clean and well designed . . .
 - British National Corpus (BNC)
 - movie subtitles, newspapers, . . .
- . . . to large and messy . . .
 - WaCky, WebBase, COW, TenTen, GloWbE, Aranea, . . .
 - sampling frame unclear, lack of metadata
 - boilerplate, duplicates, non-standard language

Is bigger always better?

- From small, clean and well designed . . .
 - British National Corpus (BNC)
 - movie subtitles, newspapers, . . .
- . . . to large and messy . . .
 - WaCky, WebBase, COW, TenTen, GloWbE, Aranea, . . .
 - sampling frame unclear, lack of metadata
 - boilerplate, duplicates, non-standard language
- . . . to huge n-gram databases
 - largest corpora only available as n-gram databases, e.g. Google's 1-trillion-word Web corpus (Web 1T 5-Grams)
 - tend to be even messier, often w/o linguistic annotation
 - lack of context, *and* incomplete because of frequency threshold

Outline

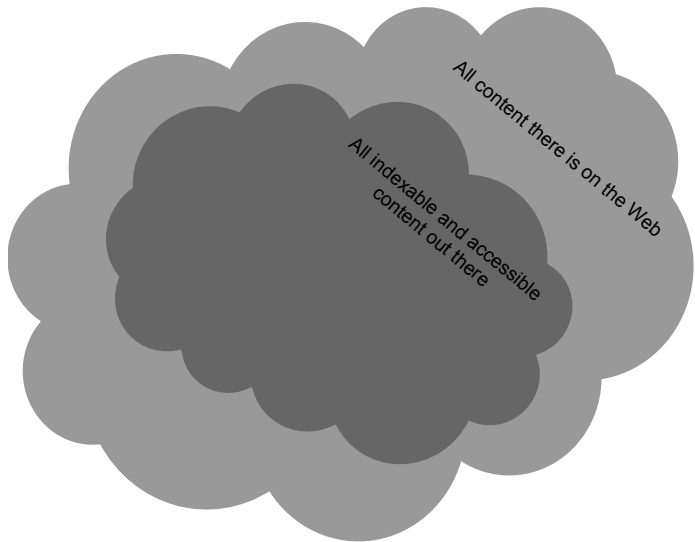
- 1 Introduction
 - The Web as Corpus - Why?
 - **Bird's-eye View of the Web** – as seen by WaCky-ists
- 2 WaCky Corpus Creation
 - Search Engine Results - let's have more of them!
 - Load'em down - all! - yes, right now!
- 3 Corpus Cleaning
- 4 WaCky Corpus Evaluation

The big Picture

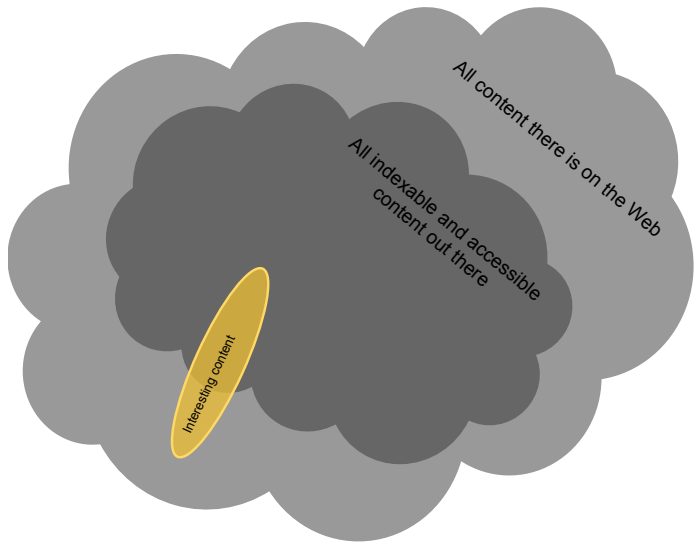


All content there is on the Web

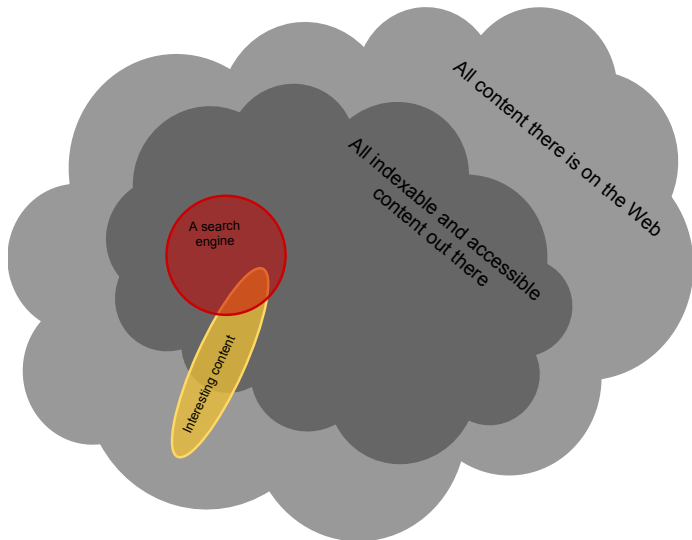
The big Picture



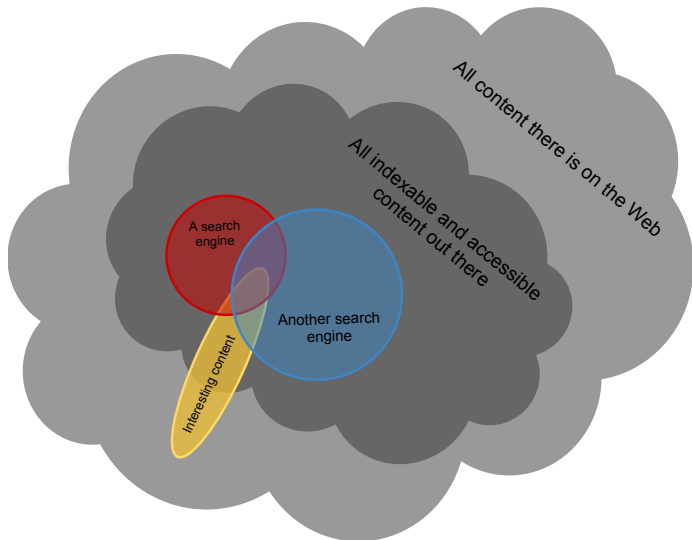
The big Picture



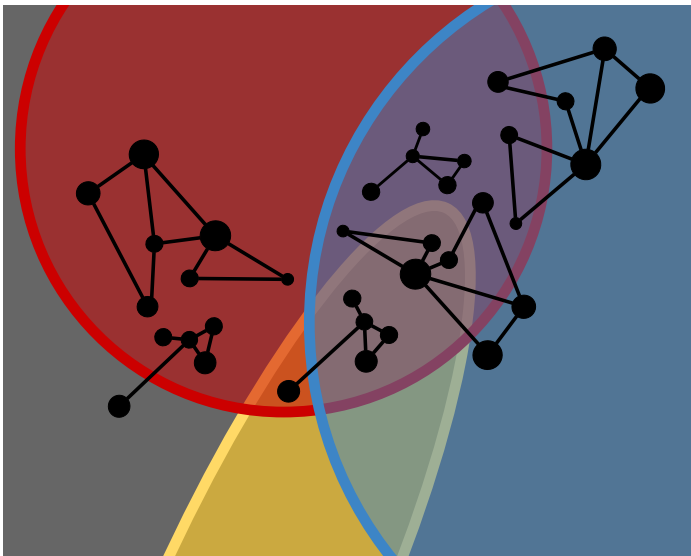
The big Picture



The big Picture

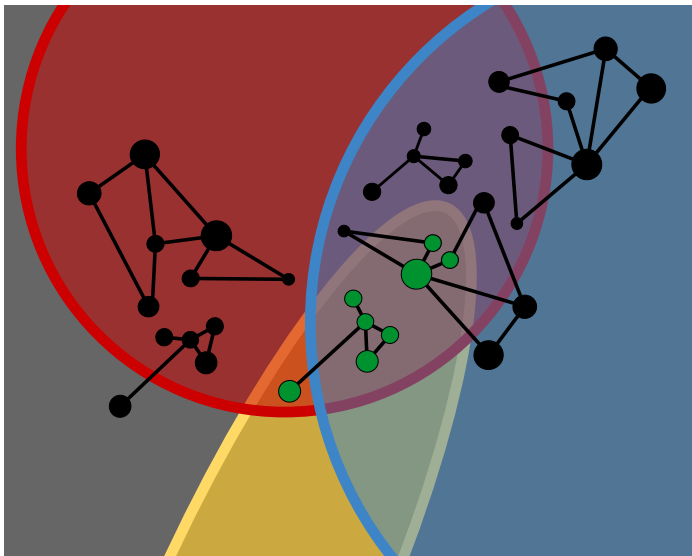


The close-up of the big Picture



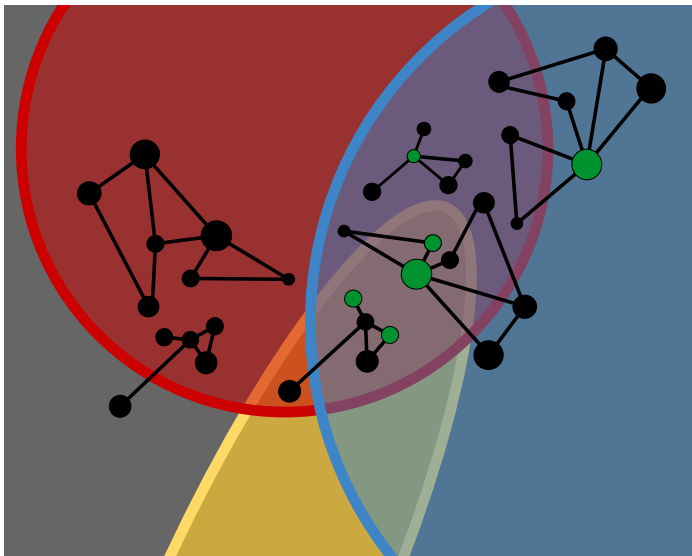
cross-linked structure of web pages

The close-up of the big Picture



the interesting web pages

The close-up of the big Picture




web pages returned by some means of query

(Some) Things can go wrong – and if they can they will. . .

(Some) Things can go wrong – and if they can they will. . .

Documents might turn out to

- 1 be in the wrong language (or of the wrong genre)



Open your mind.
INVEST YOUR TALENT IN A CREATIVE CONTEXT

INTERNATIONAL DEGREE SEEKING STUDENTS

- THE AREA
- THE UNIVERSITY
- STUDENTS' LIFE
- FEES AND FUNDING
- ENROLMENTS
- LEISURE TIME AND SPORTS

[Home](#) > The Rector Welcome

The Rector Welcome



The University of Trento is a dynamic and medium-sized research university, with high quality teaching and research facilities, tailored to respond to the needs of high quality students and researchers.

Internationalisation is a structural feature of the University itself.

These profile results have been achieved thanks to a serious and coherent strategy focusing on the areas in which the University has attained the highest standards. Along these lines, the University has invested in solid international relations, in order to enrich its educational offer, and to give students good opportunities to study and work abroad, and grow in an international environment (*Internationalisation at home*)

- WELCOME SERVICES
- MEET US IN YOUR COUNTRY
- DOWNLOAD YOUR BROCHURE
- PHOTO GALLERY
- CONTACTS

(Some) Things can go wrong – and if they can they will. . .

Documents might turn out to

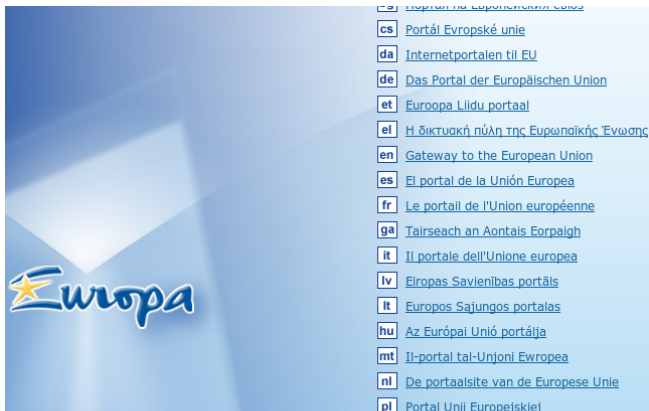
- 1 be in the wrong language (or of the wrong genre)
- 2 contain gibberish (characters, words, . . .)



(Some) Things can go wrong – and if they can they will. . .

Documents might turn out to

- 1 be in the wrong language (or of the wrong genre)
- 2 contain gibberish (characters, words, . . .)



(Some) Things can go wrong – and if they can they will. . .

Documents might turn out to

- 1 be in the wrong language (or of the wrong genre)
- 2 contain gibberish (characters, words, . . .)

Wikipedia

Wikipedia [[WEEV^akiE^{pe}Edia](#)] (auch: *die Wikipedia*) ist ein am 15. Januar 2001 gegründetes [freies Online](#) *Wikipedia* ist ein [Kofferwort](#), das sich aus [WIKI](#) ([hawaiisch](#) für [schnell](#)) und [Encyclopædie](#) zusammensetzt. Die englischsprachige Wikipedia ist mit weit über drei Millionen Artikeln die größte [deutschsprachigen Wikipedia](#) mit über einer Million Artikeln.^[1]

Die Einträge ([Artikel](#) u. a.) der Wikipedia werden von individuellen Autoren [seltener](#) von [kollektiv](#) [konzipiert](#), geschrieben und nach der Veröffentlichung [gemeinschaftlich korrigiert](#), [erweitert](#) und [aktualisiert](#).

Das Ziel von Wikipedia ist es, eine frei lizenzierte und qualitativ hochstehende Enzyklopädie zu schaffen. Wikipedia nicht nur lesen, sondern auch als [Autor](#) mitwirken. Um Inhalte zu verändern, ist eine [Anmeldung](#) oder Pseudonym [erwünscht](#). In einem offenen Bearbeitungsprozess hat Bestand, was von der [Gemeinschaft](#) [angenommen](#) wird. Bisher haben international etwa 1.016.000 angemeldete und eine unbekannte Zahl nicht angemeldeter Nutzer.

(Some) Things can go wrong – and if they can they will. . .

Documents might turn out to

- 1 be in the wrong language (or of the wrong genre)
- 2 contain gibberish (characters, words, . . .)
- 3 contain *undesireable* content

(Some) Things can go wrong – and if they can they will. . .

Documents might turn out to

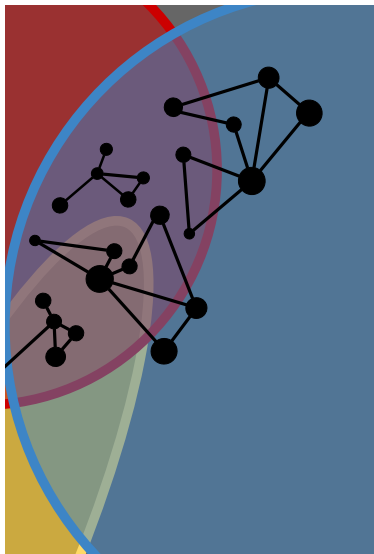
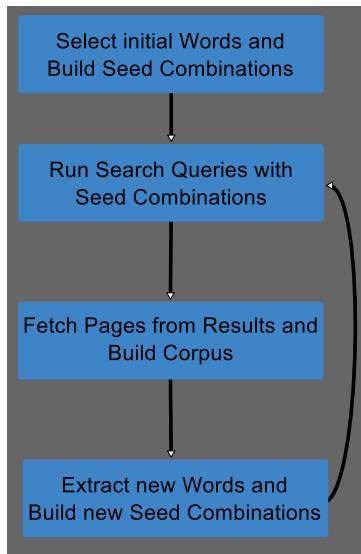
- 1 be in the wrong language (or of the wrong genre)
 - 2 contain gibberish (characters, words, . . .)
 - 3 contain *undesireable* content
 - parts of a page (e.g. boilerplate)
 - whole pages (e.g. duplicates or near-duplicates)
 - whole sites (e.g. bot-traps)
- . . .

The screenshot shows the WEB.DE website interface. At the top left is the WEB.DE logo. Below it is a navigation menu with categories like Auto, Digitale Welt, WEB.DE DSL, EM 2008, Exklusiv, Finanzen, Games, Gesundheit, and Horoskop. The main content area features a search bar with the text 'Suche' and a search button. Below the search bar, there are search results for 'Ein schweres Erdbeben der Stärke 7,8 hat 80 Prozent aller Häuser der chinesischen Provinz Sichuan zerstört.' and 'Sex And The City'. There are also advertisements for 'Last-Minute-Auktionen zu EUR 1,-' and a 'Tierwelt' section with a digital clock showing '00:00'.

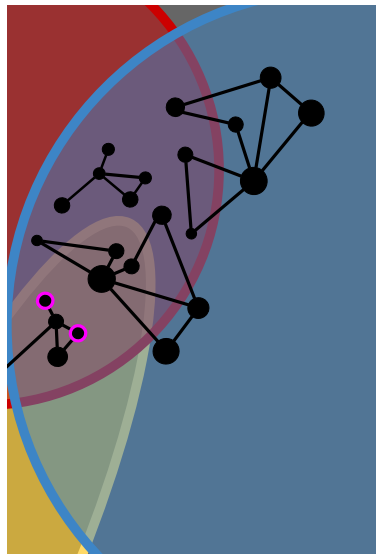
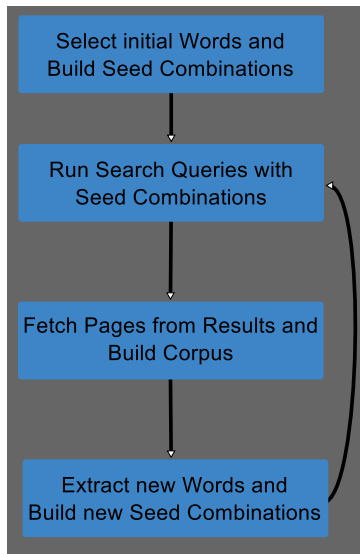
Outline

- 1 Introduction
 - The Web as Corpus - Why?
 - Bird's-eye View of the Web — as seen by WaCky-ists
- 2 **WaCky Corpus Creation**
 - **Search Engine Results - let's have more of them!**
 - Load'em down - all! - yes, right now!
- 3 Corpus Cleaning
- 4 WaCky Corpus Evaluation

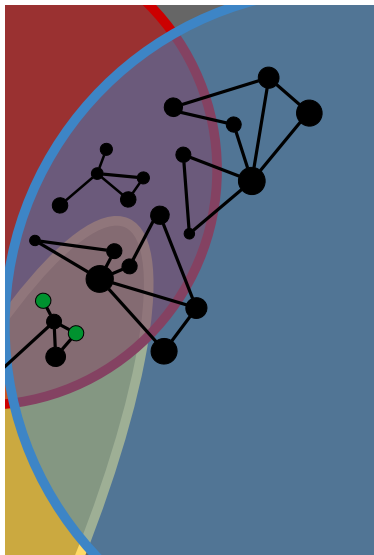
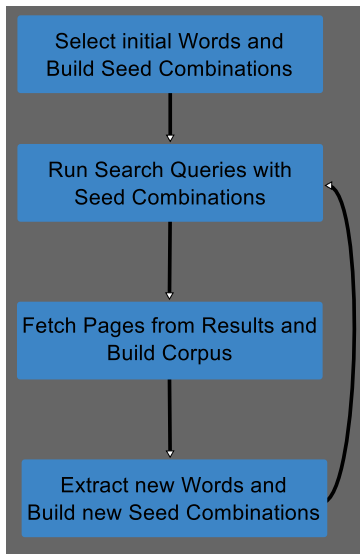
The BootCaT Idea



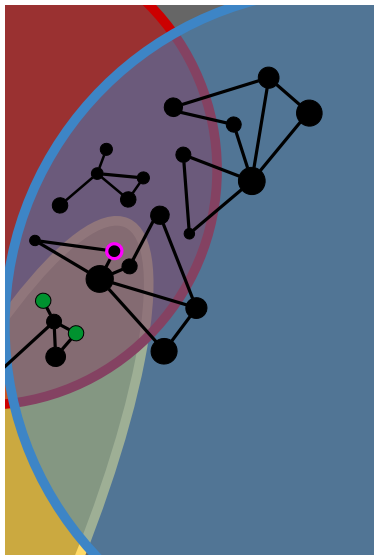
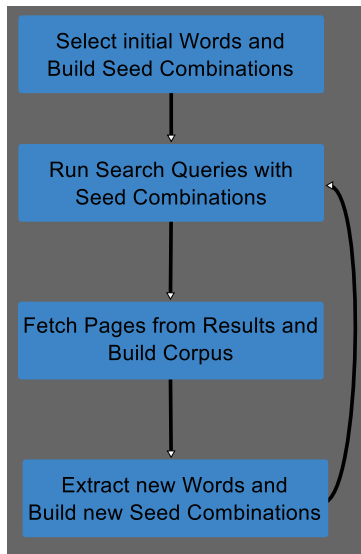
The BootCaT Idea



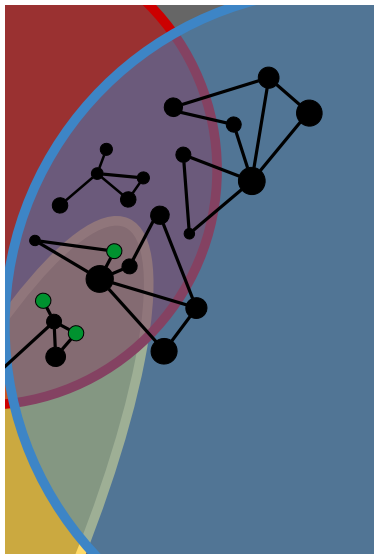
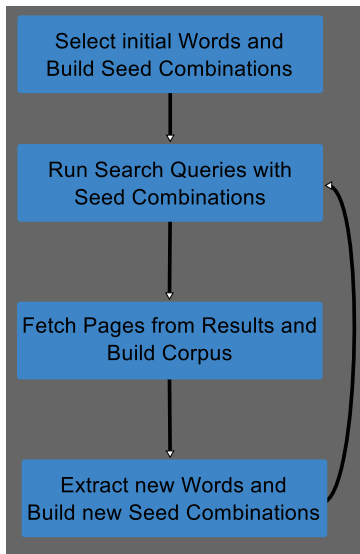
The BootCaT Idea



The BootCaT Idea



The BootCaT Idea



How to select Words for the initial Seed Combinations

... or: it's good to have a corpus to build one

Use a small list (in the 5-to-15 range) of middle-frequency words from a general corpus.

How to select Words for the initial Seed Combinations

... or: it's good to have a corpus to build one

Use a small list (in the 5-to-15 range) of middle-frequency words from a general corpus.

Digression: For a *specialized corpus* words that are expected to be representative of this very domain can be used, e.g. names of rock bands.

Application Programming Interface (API)

An API is a means for software to interact with other software.

Application Programming Interface (API)

An API is a means for software to interact with other software.

Major search engines (e.g. Google, Yahoo!, Bing, Ask.com) provide APIs that let you specify (some of) the following features:

- the language of the result pages
- the country (or region) to which to restrict your search results, i.e. only results on web sites within this country (or region) are returned
- the Creative Commons license of the contents

Outline

- 1 Introduction
 - The Web as Corpus - Why?
 - Bird's-eye View of the Web – as seen by WaCky-ists
- 2 **WaCky Corpus Creation**
 - Search Engine Results - let's have more of them!
 - **Load'em down - all! - yes, right now!**
- 3 Corpus Cleaning
- 4 WaCky Corpus Evaluation

Web Crawler

A (web) crawler is a software agent (also: robot, bot, spider) that browses the World Wide Web in a methodical, automated manner. It visits an initial list of seed URLs, identifies all the hyperlinks in the pages and adds them to a list of URLs still to visit.

Web Crawler

A (web) crawler is a software agent (also: robot, bot, spider) that browses the World Wide Web in a methodical, automated manner. It visits an initial list of seed URLs, identifies all the hyperlinks in the pages and adds them to a list of URLs still to visit.

Some characteristics of the web make crawling very difficult - crawlers take care of

- obeying politeness policies (visits, re-visits, parallelization, . . .)
- URL normalization
- naïve de-duplication (sometimes)

Outline

- 1 Introduction
 - The Web as Corpus - Why?
 - Bird's-eye View of the Web – as seen by WaCky-ists
- 2 WaCky Corpus Creation
 - Search Engine Results - let's have more of them!
 - Load'em down - all! - yes, right now!
- 3 **Corpus Cleaning**
- 4 WaCky Corpus Evaluation

Observation

However, after crawling content from the web the subsequent steps, namely, language identification, tokenising, lemmatising, part-of-speech tagging, indexing, etc. suffer from

'large and messy' training corpora [. . .] and interesting [. . .] regularities may easily be lost among the countless duplicates, index and directory pages, web spam, open or disguised advertising, and boilerplate.

The big Picture

Observation

However, after crawling content from the web the subsequent steps, namely, language identification, tokenising, lemmatising, part-of-speech tagging, indexing, etc. suffer from

'large and messy' training corpora [. . .] and interesting [. . .] regularities may easily be lost among the countless duplicates, index and directory pages, web spam, open or disguised advertising, and boilerplate.

The Problem

Thorough pre-processing and cleaning of web corpora is crucial in order to obtain reliable frequency data.

What is a 'clean' Page?

The screenshot shows the WEB.DE website in a Mozilla Firefox browser window. The browser's address bar displays the URL `https://brdwd.org/pages/dat/test/input/000.html`. The website's header is red and contains the WEB.DE logo and navigation links. A search bar is prominently displayed in the center, with the text "Suche" and "Suchen". Below the search bar, there are several news snippets. The first snippet has a yellow caption "Suche nach:" and green text "Ein schweres Erdbeben der Stärke 7,8 hat 30 Prozent aller Häuser der chinesischen Provinz Sichuan zerstört, verschütteten". The second snippet has a yellow caption "Saisisches Meer:" and green text "Bald kommt 'Sex And The City' in die deutschen Kinos. Schon jetzt überschlagen sich die Spekulationen". The third snippet has a yellow caption "Aktuell" and green text "Mitarbeiterin öffnet Barma langsam für ausländische Hilfe". The fourth snippet has a yellow caption "Aktuell" and green text "Präsidententochter Jenna Bush heiratet Harry Hager". The fifth snippet has a yellow caption "Aktuell" and green text "Low kann aufatmen: 1.840 Naot:te posieren". To the right of the news snippets, there is a yellow banner for "Last-Minute-Auktionen zu EUR 1,-" with a digital timer showing "00:00" and a "Hier!" button. Below the banner, there are advertisements for "PARTSHIP.de" and "Immobilien". At the bottom of the page, there is a "WEB.DE Surf & Phone" advertisement for "DSL DOPPEL FLAT".

Red: unwanted boilerplate; Yellow: Captions (titles, sub-titles, headings, etc.); Green: wanted running text.

Cleaning a Page with an 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 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,

...

Cleaning a Page with Finn's BTE Heuristic I

- Basic observation: Content-rich section of page tends to occur in low-HTML-density area
- Look for stretch that maximizes the quantity:
 $N(\text{TOKEN}) - N(\text{TAG})$

Cleaning a Page with Finn's BTE Heuristic II

```
<h2><a name="...">Background and motivation</a></h2>
<div class="level2">
<p>
<a href="link"></a>
</p>
<p>
Corpus-based distributional models (such as LSA or HAL)
have been claimed to capture interesting aspects of word meaning
...
</p>
```


Cleaning a Page with Finn's BTE Heuristic II

TAG TAG TOKEN TOKEN TOKEN TAG TAG

TAG

TAG

TAG TAG TAG

TAG

TAG

TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN

TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN TOKEN

...

TAG

Duplicates

- Exact duplicates are exact copies – and easy to identify

Duplicates

- Exact duplicates are exact copies – and easy to identify
- Near-duplicates are identical in terms of *content* but differ in a small portion of the document such as e.g., advertisement, counters, or date – and are more difficult to identify

A Page and a Page and a Page

Duplicates

- Exact duplicates are exact copies – and easy to identify
- Near-duplicates are identical in terms of *content* but differ in a small portion of the document such as e.g., advertisement, counters, or date – and are more difficult to identify

De-duplication

- 1 Use a dimensionality reduction technique to map web page content to small sized fingerprints

A Page and a Page and a Page

Duplicates

- Exact duplicates are exact copies – and easy to identify
- Near-duplicates are identical in terms of *content* but differ in a small portion of the document such as e.g., advertisement, counters, or date – and are more difficult to identify

De-duplication

- 1 Use a dimensionality reduction technique to map web page content to small sized fingerprints
- 2 Use *fingerprinting* that computes similar values for similar documents

A Page and a Page and a Page

Duplicates

- Exact duplicates are exact copies – and easy to identify
- Near-duplicates are identical in terms of *content* but differ in a small portion of the document such as e.g., advertisement, counters, or date – and are more difficult to identify

De-duplication

- 1 Use a dimensionality reduction technique to map web page content to small sized fingerprints
- 2 Use *fingerprinting* that computes similar values for similar documents
- 3 Consider 'similar enough' fingerprints to represent similar documents

Challenges 2.0

World Wide Web 2.0

The term "Web 2.0" was coined in January 1999 by Darcy DiNucci, a consultant on electronic information design (information architecture). In her article, "Fragmented Future", DiNucci writes:

“The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop. The Web will be understood not as screenfuls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will [...] appear on your computer screen, [...] on your TV set [...] your car dashboard [...] your cell phone [...] hand-held game machines [...] maybe even your microwave oven.”

Outline

- 1 Introduction
 - The Web as Corpus - Why?
 - Bird's-eye View of the Web – as seen by WaCky-ists
- 2 WaCky Corpus Creation
 - Search Engine Results - let's have more of them!
 - Load'em down - all! - yes, right now!
- 3 Corpus Cleaning
- 4 **WaCky Corpus Evaluation**

Evaluating the “quality” of Web corpora

- Statistical properties
 - type-token distributions, n-gram frequencies, other markers
 - representativeness (as sample of the Web)
 - genre distribution (traditional vs. Web genres)

Evaluating the “quality” of Web corpora

- Statistical properties
 - type-token distributions, n-gram frequencies, other markers
 - representativeness (as sample of the Web)
 - genre distribution (traditional vs. Web genres)
- Corpus comparison
 - between Web corpora (→ reliability)
 - between Web corpus and reference corpus
 - compared to within-corpus variation

Evaluating the “quality” of Web corpora

- Statistical properties
 - type-token distributions, n-gram frequencies, other markers
 - representativeness (as sample of the Web)
 - genre distribution (traditional vs. Web genres)
- Corpus comparison
 - between Web corpora (→ reliability)
 - between Web corpus and reference corpus
 - compared to within-corpus variation
- Training data for NLP application
 - larger amount of training data is often beneficial
 - confounding factors (NLP algorithm, training regime, . . .)

Evaluating the “quality” of Web corpora

- Statistical properties
 - type-token distributions, n-gram frequencies, other markers
 - representativeness (as sample of the Web)
 - genre distribution (traditional vs. Web genres)
- Corpus comparison
 - between Web corpora (→ reliability)
 - between Web corpus and reference corpus
 - compared to within-corpus variation
- Training data for NLP application
 - larger amount of training data is often beneficial
 - confounding factors (NLP algorithm, training regime, . . .)
- Linguistic evaluation of Web corpora
 - as substitute for / extension of reference corpus
 - need linguistic tasks that can be judged quantitatively and that make immediate use of corpus frequency data

Evaluating the “quality” of Web corpora

- Statistical properties
 - type-token distributions, n-gram frequencies, other markers
 - representativeness (as sample of the Web)
 - genre distribution (traditional vs. Web genres)
- Corpus comparison
 - between Web corpora (→ reliability)
 - between Web corpus and reference corpus
 - compared to within-corpus variation
- Training data for NLP application
 - larger amount of training data is often beneficial
 - confounding factors (NLP algorithm, training regime, . . .)
- Linguistic evaluation of Web corpora
 - as substitute for / extension of reference corpus
 - need linguistic tasks that can be judged quantitatively and that make immediate use of corpus frequency data

① Frequency comparison

- “good” Web corpora should agree with reference corpus on core phenomena → correlation between frequency counts
- e.g. Basic English vocabulary, compound nouns, . . .

1 Frequency comparison

- “good” Web corpora should agree with reference corpus on core phenomena → correlation between frequency counts
- e.g. Basic English vocabulary, compound nouns, . . .

2 Identification of multiword expressions (MWE)

- well-know NLP task based on co-occurrence statistics
- some gold standard data sets available
- e.g. “phrasal verbs”, lexical collocations, . . .

1 Frequency comparison

- “good” Web corpora should agree with reference corpus on core phenomena → correlation between frequency counts
- e.g. Basic English vocabulary, compound nouns, . . .

2 Identification of multiword expressions (MWE)

- well-know NLP task based on co-occurrence statistics
- some gold standard data sets available
- e.g. “phrasal verbs”, lexical collocations, . . .

3 Distributional semantic models (DSM)

- hypothesis: semantic similarity \sim distributional similarity
- distribution quantified by co-occurrences with other words
- DSMs can be evaluated in various shared tasks