Web 2.0 as corpus:
One decade of textual analysis with Web data

Marco Baroni
Center for Mind/Brain Sciences
University of Trento

JADT 2010

Don’t ask what you can do for the Web...

but what the Web can do for you!

The WaCky Bunch

Eros Zanchetta
Silvia Bernardini
Adriano Ferraresi
Massimo Poesio
Amaç Herdağdelen
Egon Stemle
Marco Brunello

Also starring: Alessandro Lenci, Stefan Evert, Marco Milano,
Giuseppe Attardi, Adam Kilgarriff, Serge Sharoff…
Momentous events in the history of the Web as Corpus

2001: Michele Banko and Eric Brill: Scaling to very very large corpora for natural language disambiguation, ACL

The Web as Corpus kool ynitiative
A WaCky pipeline

- Seed URL selection by queries for random word combinations to Google/Yahoo API
- Crawl with Heritrix
- Post-crawl, pre-linguistic processing
  - Boilerplate stripping
  - Near duplicate detection
  - Language recognition
- Automated linguistic annotation

Web cleaning:
The boilerplate problem

- Seed URL selection by queries for random word combinations to Google/Yahoo API
- Crawl with Heritrix
- Post-crawl, pre-linguistic processing
  - Boilerplate stripping
  - Near duplicate detection
  - Language recognition
- Automated linguistic annotation

Momentous events in the history of the Web as Corpus

1997 Dragomir Radev and Kathleen McKeown: Building a generation knowledge source using Internet-accessible newswire, ANLP

2002 Frank Keller, Mirela Lapata and Olga Ourioupina: Using the Web to overcome data sparseness, best paper award at EMNLP

2003 Special issue of Computational Linguistics on the Web as Corpus

2004 AltaVista NEAR operator stops working

2006 Google trillion-word Web 1T 5-Gram collection

2007 Googleology is bad science! (Adam Kilgarriff, Computational Linguistics)

2008 WaCky site up with Italian, German and English corpora available
Corpus-based distributional models (such as LSA or HAL) have been claimed to capture interesting aspects of word meaning ...

The WaCky unsupervised heuristic method
A. Finn et al., Fact or fiction: Content classification for digital libraries. In Joint DELOS-NSF Workshop, 2001

Web cleaning on steroids: **krdwrld**

Cues from text

Main Index Page

General Ratings Page

One true sign of a truly great band is when said band ardently defies categorisation, that is, when for every "well, they sound like this reggae-influenced heavy metal band playing avantgarde bebop" remark you can have yourself a "funny, I thought they were this raw punk outfit doing acoustic folk" counterproposal. And I don’t simply mean "being diverse" here, I mean "being different". Blazing off every colour of the spectrum. Baring one’s soul in all of its existing aspects. That sort of thing.

READER COMMENTS SECTION

Return to the Index page! NOW!
Web cleaning on steroids: krdwrd
Cues from visual rendering

Web cleaning on steroids
Cross-validation on gold standard

<table>
<thead>
<tr>
<th>features</th>
<th>precision</th>
<th>recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>text heuristic</td>
<td>NA</td>
<td>80%</td>
<td>99%</td>
</tr>
<tr>
<td>DOM cues only</td>
<td>21</td>
<td>92%</td>
<td>93%</td>
</tr>
<tr>
<td>visual cues only</td>
<td>13</td>
<td>89%</td>
<td>91%</td>
</tr>
<tr>
<td>full krdwrd</td>
<td>42</td>
<td>93%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Cleaning effect on English semantic tasks

Rubenstein and Goodenough
Spearman correlation
Accuracies
TOEFL synonyms
Concept clustering

Cleaning effect on Italian Part-of-Speech tagging

Italian tagging
Accuracy * 100
Italian cleaned word count
M words
docs (N*4.4K)

Cleaning effect on Italian Part-of-Speech tagging

Rubenstein and Goodenough
Spearman correlation
Accuracies
TOEFL synonyms
Concept clustering

Cleaning effect on Italian Part-of-Speech tagging

Italian tagging
Accuracy * 100
Italian cleaned word count
M words
docs (N*4.4K)
Web 2.0: A text document repository?

The Twitter case study

The not so momentous impact of the Web on the computational linguistics community

2005 deWaC, 1.3 billion words, 82 downloads
2006 itWaC, 1.6 billion words, 79 downloads
2007 ukWaC, 1.9 billion words, 156 downloads
2009 pukWaC, dependency-parsed, 60 downloads
(also: WaCkyPedia, 820M words, dependency-parsed, 53 downloads)
2010 frWaC, 1.3 billion words, 27 downloads

Web 2.0: A text document repository?

The Twitter case study

What's happening?

Real-time results for humid

leahcwr @chadhpry good! ran a 10K with my dad this weekend, it was super humid so didn't go so great. but we'll get there, this is week 1 of training.

Alli_Renee Glad idnt have to deal with Ga's hot nasty humid summer days anymore! (:

tom39526 ok just went outside and its FUCKING HUMID out,waiting for the storms

sim_samra Cold, windy and now its humid and sunny. Not cool mother nature, make up your mind.
Collocates of dog in different Web corpora

ukwac wikipedia twitter
sled scootering reservoir
cat sniffer stray
marshmallow sled bounty
sniffer samoyed bark
bark purebred obedience
sledding mongrel breed
campfire shaggy bladder
agility whelk whisperer
prairie leash yorkie
stray breed parasite

Collocates of mouse in different Web corpora

ukwac wikipedia twitter
trackballs orthologue mickey
ftr soledons imperator
parthenogenetic transgenic modest
transgenic mrl minnie
rat pinkie cordless
hover vole clubhouse
keyboard desmin logitech
pointer rat optical
catnip trackball gyration
vole musculus magic

Modeling similarity judgments: word and user space

Rubenstein & Goodenough (65 pairs)

<table>
<thead>
<tr>
<th>corpus</th>
<th>space</th>
<th>ρ%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ukwac</td>
<td>word</td>
<td>72</td>
</tr>
<tr>
<td>wikipedia</td>
<td>word</td>
<td>71</td>
</tr>
<tr>
<td>twitter</td>
<td>word</td>
<td>58</td>
</tr>
<tr>
<td>twitter</td>
<td>user</td>
<td>41</td>
</tr>
</tbody>
</table>

WordSim Similarity (203 pairs)

<table>
<thead>
<tr>
<th>corpus</th>
<th>space</th>
<th>ρ%</th>
</tr>
</thead>
<tbody>
<tr>
<td>wikipedia</td>
<td>word</td>
<td>60</td>
</tr>
<tr>
<td>ukwac</td>
<td>word</td>
<td>57</td>
</tr>
<tr>
<td>twitter</td>
<td>word</td>
<td>42</td>
</tr>
<tr>
<td>twitter</td>
<td>user</td>
<td>25</td>
</tr>
</tbody>
</table>

WordSim Relatedness (252 pairs)

<table>
<thead>
<tr>
<th>corpus</th>
<th>space</th>
<th>ρ%</th>
</tr>
</thead>
<tbody>
<tr>
<td>wikipedia</td>
<td>word</td>
<td>35</td>
</tr>
<tr>
<td>twitter</td>
<td>word</td>
<td>33</td>
</tr>
<tr>
<td>ukwac</td>
<td>word</td>
<td>32</td>
</tr>
<tr>
<td>twitter</td>
<td>user</td>
<td>21</td>
</tr>
</tbody>
</table>

Deeper into Web2.0: Let the People build just the corpus you need!
Phrase Detectives
http://anawiki.essex.ac.uk/phrasedetectives/

Coming full circle:
Web surfers validating Web harvests

Training Data
Web Mining
Web-based Games
Corpora
Candidate assertions
Players

The Concept Game
http://apps.facebook.com/conceptgame

Score points each time you find an assertion that makes sense.

Points Collected
4

Previous Round
Your answer:
Correct answer:
Time left: 00:27

Tight-tipped Fellow of the Librarian
Next Level
Collect 2550 points (2242 more to go).

Eating your mind?
If you think you can beat me in Concept Game accept my challenge.

Welcome | Scoreboard | Your Status | Hall of Fame | Top 10
Your Level

Do you have common sense? Are you fast, calm, and lucky?
Concept Game is a simple yet intense game. Each month, we award top three players with Amazon gift cards (up to $25). Check the Scoreboard page for details.

pressure

Publish this story to your Facebook Wall and your friends’ home pages?

Challenge your friends. Bookmark Concept Game and access it quickly next time.

A logo design by Dana Cam Ombregt; Sound samples by Ear August.
Boring talks?

Play the Concept Game instead!

apps.facebook.com/conceptgame

That's all, folks!

Thank you!

marco.baroni@unitn.it