My love affair with the Web

. . . and why it's over!

Marco Baroni

Center for Mind/Brain Sciences
University of Trento

BOTWU
Forlì, 24 June 2013

Change photo

Marco Baroni Edit

Tenured Researcher, Center for Mind/Brain Sciences, University of Trento (Italy) Edit

Computational Linguistics - Linguistics - Semantics Edit

Verified email at unitn.it Edit

My profile is private Edit Add homepage

Citation indices

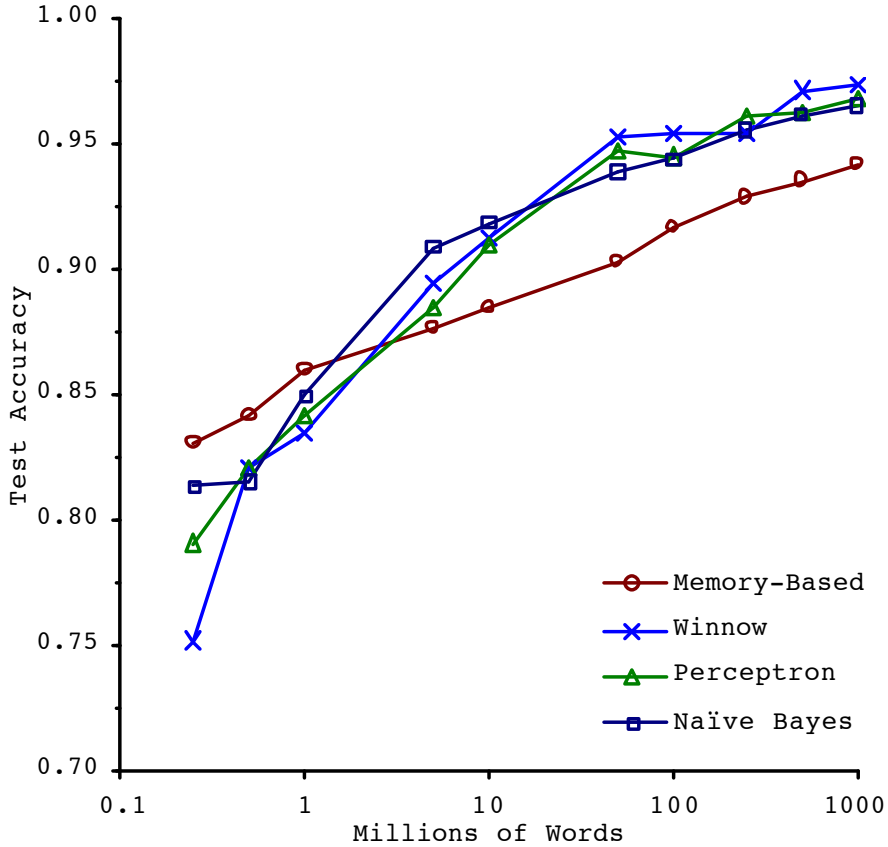
	All	Since 2008
Citations	2185	1735
h-index	24	21
i10-index	45	39

Citations to my articles

Select: All, None Actions Show: 20 **1-20** Next >

Title / Author	Cited by	Year
The WaCky wide web: a collection of very large linguistically processed web-crawled corpora M Baroni, S Bernardini, A Ferraresi, E Zanchetta Language Resources and Evaluation 43 (3), 209-226	227	2009
BootCaT: Bootstrapping corpora and terms from the web M Baroni, S Bernardini Proceedings of LREC 4, 1313-1316	217	2004
Large linguistically-processed web corpora for multiple languages M Baroni, A Kilgarriff Proceedings of the Eleventh Conference of the European Chapter of the ...	113	2006
Introducing and evaluating ukwac, a very large web-derived corpus of english A Ferraresi, E Zanchetta, M Baroni, S Bernardini Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google, 47-54	95	2008

More data trump better algorithms



Michele Banko and Eric Brill, ACL 2001

Momentous events in the history of the Web as Corpus

- 1997 Radev and McKeown: Building a generation knowledge source using Internet-accessible newswire, ANLP
- 2002 Keller, Lapata and 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 First version of BootCaT available! :-)
- 2006 WAC2, ACL Special Interest Group on Web-as-Corpus is born
- 2006 Google trillion-word Web 1T 5-Gram collection
- 2007 Googleology is bad science! (Adam Kilgarriff, Computational Linguistics)
- 2008 WaCky site up with billion-word Italian, German and English corpora freely available

A shocking revelation

BootCaT was invented in 2001 at CMU:

Rayid Ghani, Rosie Jones and Dunja Mladenić: Mining the web to create minority language corpora. CIKM 2001

Futurology

Web-as-Corpus 2014 as seen from 2004

- We’ve got the linguist search engine
- Data retrieval and pre-processing are huge topics in computational linguistics
- Cleaning Web data is a solved problem
- The Web is exactly the same as in 2004, just bigger

Son of BootCaT at ACL 2013

Flavio De Benedictis, Stefano Faralli and Roberto Navigli:
GlossBoot: Bootstrapping Multilingual Domain Glossaries from the Web

We present GlossBoot, an effective minimally-supervised approach to acquiring wide-coverage domain glossaries for many languages. For each language of interest, given a small number of hypernymy relation seeds concerning a target domain, we bootstrap a glossary from the Web for that domain by means of iteratively acquired term/gloss extraction patterns. Our experiments show high performance in the acquisition of domain terminologies and glossaries for three different languages.

Cleaning the Web

The boilerplate challenge

Courtesy of Egon Stemle

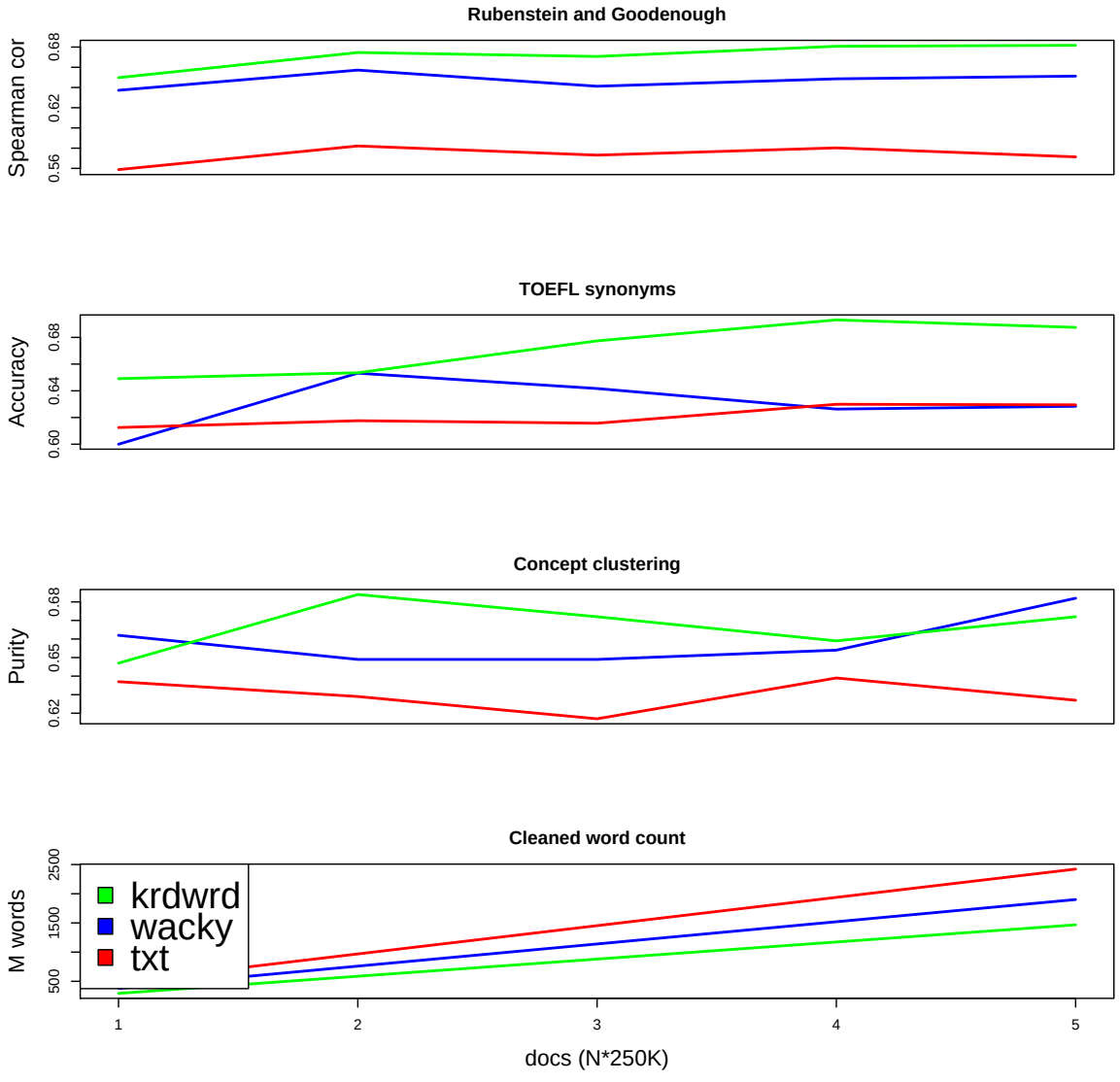
Cleaning, a solved problem?

	features	*precision*	*recall*	*F*
wacky heuristic	NA	80%	99%	88%
text cues only	21	92%	93%	92%
DOM cues only	13	89%	91%	90%
visual cues only	8	90%	93%	91%
full krdwrd	42	93%	92%	92%

Courtesy of Egon Stemle

Does cleaning really help?

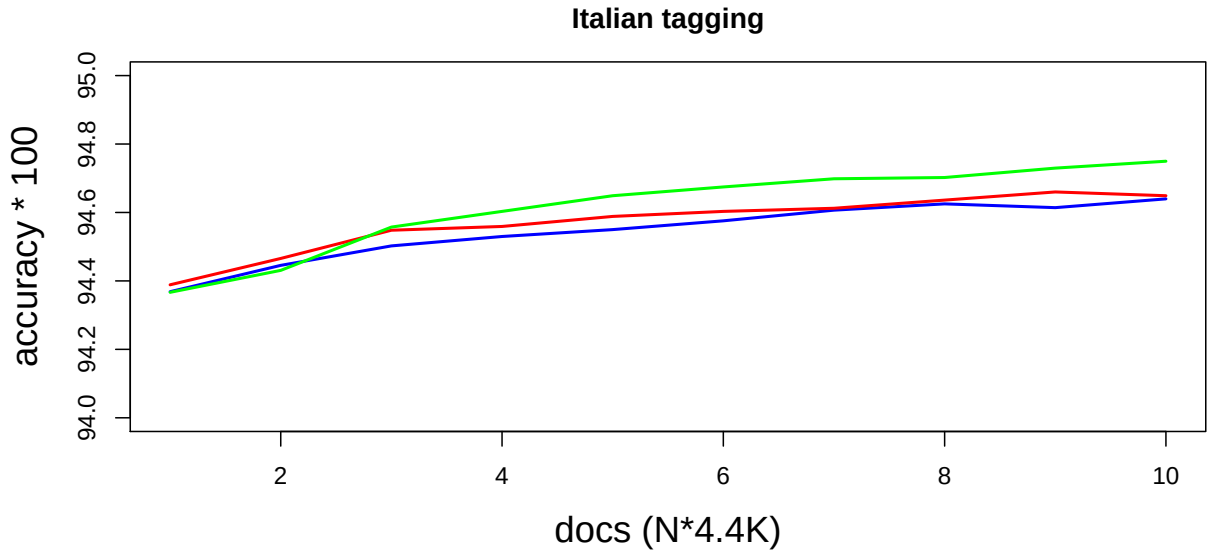
Various English semantic tasks

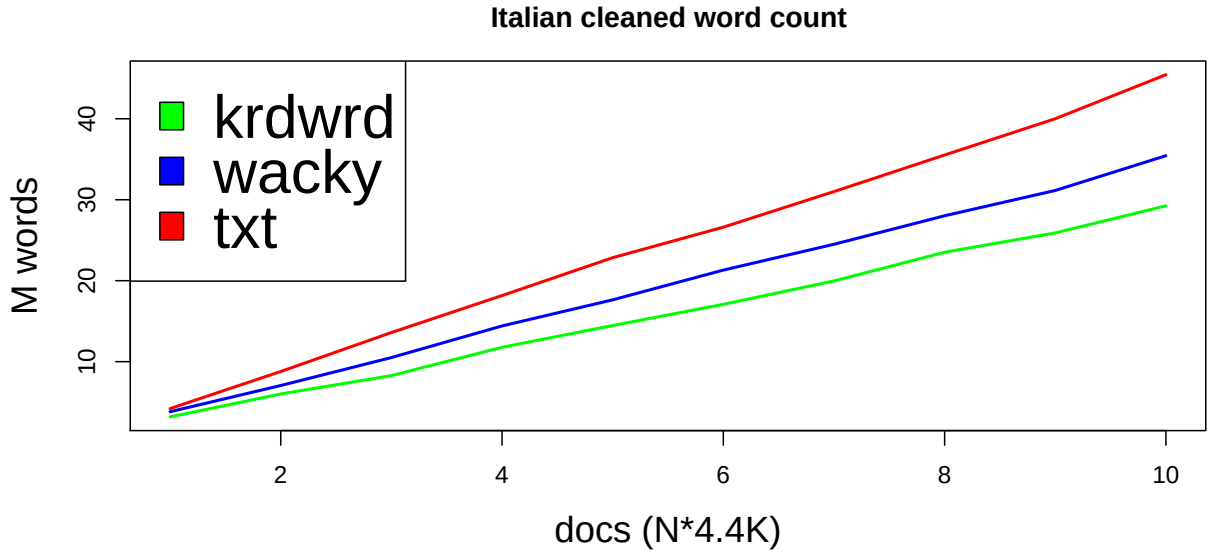


Collaboration with Egon Stemle

Does cleaning really help?

Italian Part-of-Speech tagging





Collaboration with Egon Stemle

The Web is no longer what it used to be!

Marco Baroni

http://clic.cimec.unitn.it/marco/

Latest Headlines

Marco Baroni | :) on Flickr – Photo Sharing!

Marco Baroni's Homepage

Tyrolean pride! (photo by Motis)

I am a tenured researcher in the CLIC group of CIMeC, the Center for Mind/Brain Sciences of the University of Trento, and a member of the DISCoF Department.

Research

My current research focuses on two related themes.

First, I am exploring the crazy idea that human conceptual knowledge is mostly the result of the extraction of simple distributional information from large amounts of verbal input, i.e., that we think of *barking*, *having a tail* and *being a pet* as salient properties of dogs because we heard/read lots of sentences/phrases like "her dog barks all the time", "the dog wagged its tail" or "dogs and other pets".

While this sounds counter-intuitive, I am encouraged by the large body of work in corpus-based Natural Language Processing that has shown how simple pattern extraction techniques can harvest very rich and multi-faceted knowledge from raw text -- way beyond what is learned by models based on visual cues, non-verbal interaction and other sources.

My most representative work in this direction is summarized here and here.

I am also interested in the applied offshoots of our research on concept induction -- in particular, I am currently the recipient of a Google Research Award to study convergences between text- and vision-based semantic spaces and of an Italian PRIN grant to model semantic cognition in the blind with corpus-based distributional models. I am also collaborating to projects on semantic fields in pedagogical lexicography, on the creation of a concept data-base for clinical work with patients with verbal and conceptual deficits and to the mighty LiveMemories project.

The second active strand of my research pertains to the creation of the electronic resources (such as corpora and lexica) and computational tools that enable the sort of quantitative computer-based simulations I describe above. My tools and resources page provides links to some stuff I (co-)developed.

Please take a look at my publications page for more topics I have been working on.

By the way: Are you on Facebook? Then please help us collecting better semantic data by playing with us!

Teaching

As a linguist, I received a "humanist" training, i.e., no math, very little programming, very basic statistics. I am still trying to fully recover from this.

Consequently, my main interest, as a teacher, is to provide my students -- (computational) linguists, philosophers, cognitive scientists -- with a decent quantitative and computational background.

The most concrete step in this direction has been setting up, together with my CLIC colleagues and the Philosophy section of the School of Humanities, a "Philosophy and Informatics" major within the BA- and MA-level Philosophy degrees. Basic information (in Italian) about these programs is available available here and here.

Current teaching activities at the University of Trento include:

- A doctoral level course on statistics in R.
- Courses in text processing and introductory Perl for text processing (class materials) offered to students of the School of Humanities and Philosophy, of the International Master in Cognitive Science and of the Master in Human Language Technologies and Interfaces.
- I belong to the faculty of the Doctoral School in Cognitive and Brain Sciences and I participate in various activities of the School, most notably co-organizing Topic and Research Seminars with my CLIC colleagues.

I also had several chances to teach "statistics for linguists" mini-courses, mostly with Stefan Evert; see the SIGHL page.

Stuff on this site:

- Publications and presentations, with papers available for download.
- Freely available linguistic tools and resources developed by me and various collaborators.

Done

:) on Flickr – Photo Sharing!

http://www.flickr.com/photos/motofot

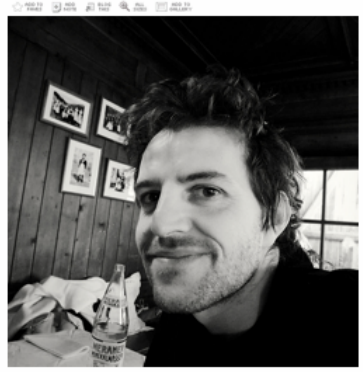
Latest Headlines

Marco Baroni | :) on Flickr – Photo Sharing!

flickr

Home | You | Organize & Create | Contacts | Groups | Explore | Search

:)



Comments

kumo36 says:
かっこいい！
写真から愛情が溢れてますね。やっぱり写真はこうあるべきですね！！

shotam says:
marcoさん、かっこいい
いつかお会いしたいものです。

magic fly paula says:
Happy birdhay to him*

fujiko says:
この笑顔でコチラも:))

*reesie says:
Happy happy birthday!!

wertheim says:
Great portrait. Tanjobi omedetto!

yusuf_alioglu says:
Happy Birthday.

yusuf_alioglu says:
Hi, I'm an admin for a group called FREE EARTH - FREE PEOPLES ©, and we'd love to have this added to the group!

hamapenguin says:
nice portrait!

motocchio's photostream

1,856 uploads

This photo also belongs to:

portraits (Set) 27 items

marco (Set) 54 items

FREE EARTH - FREE PEOPLE © (Pool)

1 person calls this photo a favorite

People in this photo

Add a person

Tags

marco, portrait, 2009, happy birthday, buon compleanno, クマ太, BW, monochrome, square, gx200

Add a tag

Additional Information

All rights reserved

Anyone can see this photo

Taken with a Ricoh GX200. More properties

Taken on November 1, 2009

Viewed 624 times

Replaced on November 14, 2009

Flag this photo

Done

And this is the end. . .

Thank you!