Processing Web-derived Text

Or: Working with very messy data

Sebastian Hoffmann
University of Zurich
sebhoff@es.unizh.ch

GENERAL APPROACH

At the current point in time, treating the web as corpus raises a whole range of methodological and technical questions. Until solutions are found and implemented, web-based data can be downloaded and post-processed to complement existing standard corpora.

THE STATESMAN ARCHIVE

- available at http://www.thestatesman.org
- news articles - many with focus on local Indian topics
- automatic download with Perl scripts (module: LWP)

CONVERSION:

Isolation of relevant text from Hypertext Mark-up Language (HTML) code with the help of Perl-scripts. Removal of dupicalte entries.

STATISTICS:

Period covered:	January 2002 - March 2005
Total number of news articles:	81,150
Total number of words:	approx. 31.5 million

VERB-COMPLEMENTATION PATTERNS IN INDIAN ENGLISH:

(1) The first obligation of these media institutions will be to inform the country the facts and events without prejudice or conscious partisanship. (2003-11-07)

USENET

- Dates back to 1979
- Offers access to thousands of newsgroups with a large range of topics
- Discussions are "interactive", i.e. users can respond to previous posts

DOWNLOAD:

- Creation of a primitive newsreader written in Perl (with Net::NNTP module)
- Access via commercial newsserver Giganews
- MySQL database is used to keep track of downloaded material

Table 1: Newsgroups downloaded

Name	N messages	N words
alt.alien.research	58,189	13,978,358
alt.coffee	68,056	12,076,503
alt.fan.noam-chomsky	62,703	25,683,845
alt.games.warcraft	23,849	3,896,371
alt.music.oasis	21,883	1,934,548
alt.support.marriage	76,183	20,757,171
news.software.nntp	5,033	1,177,785
news.software.readers	31,874	4,790,531
rec.audio.tech	24,645	5,247,679
rec.gambling.sports	34,848	4,955,744
rec.music.classical.recordings	147,866	28,188,853
rec.photo.digital	279,846	46,358,859
rec.sport.swimming	15,980	3,360,974
Total:	849,955	172,407,221

CONVERSION:

- Convert files into format that can be used with standard concordancing software
- Take care of quoted text and assign correct author

PROBLEM 1:

Quoted text is standardly indicated with angle brackets - but this is user-definable.

Solution: none; only consider text quoted with angle brackets

PROBLEM 2:

Line breaks are handled inconsistently by different newsreaders. As a result, quoted text may receive an incorrect number of angle brackets:

```
> > They packed them off into some shit-hole of a settlement in the occupied
> > territories. Same for some South American Indians they converted and
> shipped
> > over as well.


>>>>>>>>>>>>>>>Are you giving 'props' to Gary?
>>>>>>>>>>>>>>>Yo!!! Give Gary Burnore a chance to defend himself! I'm
sure
>>>>>he'll be
>>>>>>>>>>>>>>>back soon. He's probably out sniffing bicycle seats!!!!
>>>>>>>>>>>>>>>:-)
```

SOLUTION:

- Isolate “root” messages
- Then match quotes of “level 1” messages to text in root messages
- Work through all levels successively, matching quotes to converted text of previous level

INDIVIDUAL STEPS:

- First try matching whole lines
  - If line matches: annotate with Msg-ID of “parent”
  - If line doesn't match, proceed with steps below
- Allow optional line break after each word
- Allow last character of line to be cut off
- Allow elisions of text (<SNIP>, [...], etc.)
- ... ...

For various reasons, quote assignment can never be 100 per cent accurate ("spurious" quotes, correction of spelling mistakes, deliberate changes in quoted text). Optimization is possible but time-consuming - and may also introduce new levels of imprecision.

MESSINESS OF DATA:

- Clean-up process does not result in a well-balanced corpus
- Next to no metatextual data is available
- Authorship is very difficult to determine
- Influence of spam and troll posts

Extract 1: A converted message from alt.fan.noam-chomsky

```
<header>
  Message-ID: <MPG.196bfcab94cf0a8989718@news.supernews.net>
  From:       Dan Swartzendruber <dswartz@druber.com>
  Subject:    Re: What is the most dangerous false belief in the
              world today ?
  Date:       Tue, 1 Jul 2003 21:17:20 -0400
  Root MsgID: <a1333567.0307010632.744e81cd@posting.google.com>
  Level:      4
</header>

<body>
  <4_MPG.196bfcab94cf0a8989718@news.supernews.net> In article
      <bdtbh7$jhi$1@news6.svr.pol.co.uk>,
      agamemnon@hello.to.NO_SPAM says...
  <3_bdtbh7$jhi$1@news6.svr.pol.co.uk> "Dan Swartzendruber"
      <dswartz@druber.com> wrote in message
      news:MPG.196bf1e0fe42036a989717@news.supernews.net...
  <2_MPG.196bf1e0fe42036a989717@news.supernews.net> In article
      <bdt8ab$ojs$1@newsg4.svr.pol.co.uk>,
      agamemnon@hello.to.NO_SPAM says...
  <1_bdt8ab$ojs$1@newsg4.svr.pol.co.uk>    "Zardoz"
      <zardoz07@myfastmail.com> wrote in message
      news:a1333567.0307010632.744e81cd@posting.google.com...
  <0_a1333567.0307010632.744e81cd@posting.google.com> What is,
      in your opinion, the most influential and dangerous
      false belief in today's world? By false belief, I mean
      something that had been refuted by the experts beyoun a
      reasonable doubt, but is still held by the general
      public (or a part thereof) as true.
  <1_bdt8ab$ojs$1@newsg4.svr.pol.co.uk> Neo-Conservatism,
      Zionism, and Islam.
  <2_MPG.196bf1e0fe42036a989717@news.supernews.net> I guess
      Marxism is not a valid choice, since the "still held by
      the general public" is no longer true :)
  <3_bdtbh7$jhi$1@news6.svr.pol.co.uk>Marx was a Zionist.
  <4_MPG.196bfcab94cf0a8989718@news.supernews.net> Which is
      irrelevant, as far as I can tell. Certainly 99.999% of
      the people who purported to follow Marxism weren't.
</body>
```

Table 2: Success rates for quote assignment

Name	messages	with quotes	% un-assigned	% unassigned per level
rec.gambling.sports	34,848	17,628	26.9%	17.2%
alt.fan.noam-chomsky	62,703	52,801	10.9%	6.1%
alt.alien.research	58,189	48,189	18.3%	6%
rec.music.classical.recordings	147,866	119,269	8.4%	5.1%
rec.sport.swimming	15,980	12,799	6.9%	4.4%
alt.support.marriage	76,183	69,955	7.1%	3.8%
rec.photo.digital	279,846	234,497	6.1%	3.6%
news.software.readers	31,874	26,480	5.5%	3.5%
rec.audio.tech	24,645	18,868	5%	3.2%
alt.music.oasis	21,883	14,989	4.7%	3.1%
alt.coffee	68,056	50,243	4.4%	3%
news.software.nntp	5,033	3,709	4.6%	2.9%
alt.games.warcraft	23,849	20,135	2.6%	1.4%
Total:	849,955	689,562		

BIBLIOGRAPHY

Hoffmann, Sebastian. In press. "From Web-Page to Mega-Corpus: The CNN Transcripts." In: Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Mukherjee, Joybrato and Sebastian Hoffmann. In preparation. "Describing Verb-Complementational Profiles of New Englishes: A Pilot Study of Indian English."

NOTES