An NLP Approach to the Evaluation of Web Corpora

Stefan Evert
Corpus Linguistics Group, Department Germanistik & Komparatistik
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
stefan.evert@fau.de

Marburg, 6 March 2014

Collaborators
parts of this presentation are based on two collaborative studies

- Biemann, Chris; Bildhauer, Felix; Evert, Stefan; Goldhahn, Dirk; Quasthoff, Uwe; Schäfer, Roland; Simon, Johannes; Swiezinski, Leonard; Zesch, Torsten (2013). Scalable construction of high-quality Web corpora. Journal for Language Technology and Computational Linguistics (JLCL), 28(2), 23–59.


Introduction
The Web as Corpus

Why Web corpora?
because more data are better data (Church and Mercer 1993)

- Properties of the Web
  - Internet English, distribution of Web genres, hyperlink graph
  - Web corpus = random sample of the (public) WWW

- 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

- 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 corpora

- Scaling up NLP training data (Banko and Brill 2001)
  - 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, . . .

- . . . 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, incomplete because of frequency threshold
The Google Web 1T 5-Gram database
Brants and Franz (2006)

<table>
<thead>
<tr>
<th>word 1</th>
<th>word 2</th>
<th>word 3</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>supplement</td>
<td>depend</td>
<td>on</td>
<td>193</td>
</tr>
<tr>
<td>supplement</td>
<td>depending</td>
<td>on</td>
<td>174</td>
</tr>
<tr>
<td>supplement</td>
<td>depends</td>
<td>entirely</td>
<td>94</td>
</tr>
<tr>
<td>supplement</td>
<td>depends</td>
<td>on</td>
<td>338</td>
</tr>
<tr>
<td>supplement</td>
<td>derived</td>
<td>from</td>
<td>2668</td>
</tr>
<tr>
<td>supplement</td>
<td>des</td>
<td>coups</td>
<td>77</td>
</tr>
<tr>
<td>supplement</td>
<td>described</td>
<td>in</td>
<td>200</td>
</tr>
</tbody>
</table>

excerpt from file 3gm-0088.gz

Web1T5 made Easy
but not for the computer (Evert 2010)

<table>
<thead>
<tr>
<th>word id</th>
<th>depend</th>
<th>6094</th>
</tr>
</thead>
<tbody>
<tr>
<td>depending</td>
<td>3571</td>
<td></td>
</tr>
<tr>
<td>depends</td>
<td>3846</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>supplement</td>
<td>5095</td>
<td></td>
</tr>
</tbody>
</table>

- Use numeric ID coding as in IR / large-corpus query engines
- More efficient to store, index and sort in RDBMS
- Frequency-sorted lexicon is beneficial for variable-length coding of integer IDs (used by SQLite)

This looks very much like a relational database table
So why not just put the data into an off-the-shelf RDBMS?
▶ built-in indexing for quick access
▶ powerful query language SQL

Web1T5-Easy database encoding procedure

- Pre-processing (normalisation, filtering, ...)
- Numeric ID coding & database insertion [1d 23h]
- Collapse duplicate rows (from normalisation) [6d 7h]
- Indexing of each n-gram position [3d 2h]
- Statistical analysis for query optimisation [not useful]
- Build database of co-occurrence frequencies [ca. 3d]

Carried out in spring 2009 on quad-core Opteron 2.6 GHz with 16 GiB RAM
— should be faster on state-of-the-art server with latest version of SQLite.
Querying the database

It's easy to search the database for patterns like

\[ \text{association} \ldots \text{Xal} \ Y \]

with a “simple” SQL query:

```sql
SELECT w3, w4, SUM(f) AS freq FROM ngrams
WHERE w1 IN (SELECT id FROM vocab WHERE w='association')
AND w3 IN (SELECT id FROM vocab WHERE w LIKE '%al')
GROUP BY w3, w4 ORDER BY freq DESC;
```

Web1T5-Easy implements a more user-friendly query language:

\[ \text{association} \ ? \ %al \ * \]

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

Web1T5-Easy query performance

<table>
<thead>
<tr>
<th>Web1T5-Easy query</th>
<th>cold cache</th>
<th>warm cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpus linguistics</td>
<td>0.11s</td>
<td>0.01s</td>
</tr>
<tr>
<td>web as corpus</td>
<td>1.29s</td>
<td>0.44s</td>
</tr>
<tr>
<td>time of *</td>
<td>2.71s</td>
<td>1.09s</td>
</tr>
<tr>
<td>%ly good fun</td>
<td>181.03s</td>
<td>24.37s</td>
</tr>
<tr>
<td>[sit,sits,sat,sitting] * ? chair</td>
<td>1.16s</td>
<td>0.31s</td>
</tr>
<tr>
<td>* linguistics (association ranking)</td>
<td>11.42s</td>
<td>0.05s</td>
</tr>
<tr>
<td>university of * (association ranking)</td>
<td>1.48s</td>
<td>0.48s</td>
</tr>
</tbody>
</table>

(64-bit Linux server with 2.6 GHz AMD Opteron CPUs, 16 GiB RAM and fast local hard disk; based on timing information from the public Web interface.)
Linguistic evaluation of Web corpora

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-known 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

Research questions

- Are English Web corpora a substitute for the BNC?
- What are the differences between Web corpora?
- Does size matter more than content?
- How useful are n-gram databases?
  - esp. negative effects of frequency thresholds
- How important is (automatic) linguistic annotation?
- Do Web corpora offer better coverage?

Corpora in the evaluation

- **Reference**: British National Corpus
  - 0.1 G
- English Movie Subtitles (DESC v2)
  - 0.1 G
- Gigaword (2nd edition)
  - 2.0 G
- Wackypedia subset (WP500)
  - 0.2 G
- English Wackypedia
  - 1.0 G
- ukWaC
  - 2.0 G
- WebBase
  - 3.0 G
- UKCOW 2012
  - 4.0 G
- Joint Web corpus
  - 10.0 G
- Web 1T 5-Grams
  - 1000.0 G
- LCC n-gram database
  - 1.0 G
Comparison of frequency counts

- Scatterplots of (log) frequencies in BNC vs. other corpora
  - Pearson correlation $r$ from regression $f_{WiWIC} \sim \beta \cdot f_{BNC}$ etc.
  - only consider items that occur in both corpora
    (+ low coverage is not penalized directly)

- Test data sets
  - Basic English words (lemmatized)
  - inflected forms of Basic English words
  - binary compound nouns extracted from WordNet 3.0

- Morphological query expansion for unannotated n-grams

<table>
<thead>
<tr>
<th>query</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>hear sound</td>
<td>36,304</td>
</tr>
<tr>
<td>[hear, hears, heard, hearing]</td>
<td>95,453</td>
</tr>
</tbody>
</table>

(dashed lines indicate acceptable frequency difference within one order of magnitude)
Collocations

- Collocation: frequent co-occurrence within short span of up to 5 words (Firth 1957; Sinclair 1966, 1991)
  - plays important role in lexicography, corpus linguistics, language description, word sense disambiguation, ...
  - key feature for MWE identification
  - collocation database is also a sparse representation of a distributional semantic model (term-term matrix)

- Web1T5 only provides exact co-occurrence frequencies for immediately adjacent bigrams (e.g. * day and day *)
- Approximate counts for distance $n$ from $n + 1$-gram table

  $\text{day ? ? * and * ? ? day}$

  $\Rightarrow$ quasi-collocations

Quasi-collocations database

- Web1T5-Easy: pre-compiled database of quasi-collocations
  - brute-force, multi-pass algorithm
  - runtime approx. 3 days on server with 16 GiB RAM

- Flexible collocational span L4, ..., L1 / R1, ..., R4
  - separate count for each collocate and position
  - co-occurrence frequency in user-defined span and association scores are calculated on the fly
  - benefits from tight integration of Perl & SQLite

- Standard association measures: $X^2$, $G^2$, $t$, MI, Dice

Quasi-collocations demo
Evaluation on English VPC extraction task
(Baldwin 2008)

- English verb-particle constructions (VPC) consisting of head verb + one obligatory prepositional particle
  - hand in, back off, wake up, set aside, carry on, ...
- Data set of 3,078 candidate VPC types
  - extracted from written part of BNC with combination of tagger-, chunker-, and parser-based methods
- Manually annotated as compositional / non-compositional
  - baseline: 14.3% non-compositional VPC (440 / 3078)
  - compositional: carry around, fly away, refer back, ...
  - further distinction of transitive/intransitive VPC not used
- Evaluation: candidate ranking based on each corpus
  - surface co-occurrence (L0,R3) + POS filter (except Web1T5)
  - standard association measures: $G^2$, $t$, MI, Dice, $X^2$, $f$
- precision/recall graphs; overall quality: average precision (AP)

<table>
<thead>
<tr>
<th>verb</th>
<th>particle</th>
<th>$\chi^2$</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>talk</td>
<td>about</td>
<td>950906.9</td>
<td>-</td>
</tr>
<tr>
<td>lean</td>
<td>forward</td>
<td>510113.9</td>
<td>-</td>
</tr>
<tr>
<td>want</td>
<td>to</td>
<td>477739.8</td>
<td>-</td>
</tr>
<tr>
<td>sort</td>
<td>out</td>
<td>406072.7</td>
<td>+</td>
</tr>
<tr>
<td>base</td>
<td>on</td>
<td>398035.3</td>
<td>-</td>
</tr>
<tr>
<td>depend</td>
<td>on</td>
<td>330956.6</td>
<td>-</td>
</tr>
<tr>
<td>sit</td>
<td>down</td>
<td>329143.2</td>
<td>+</td>
</tr>
<tr>
<td>go</td>
<td>to</td>
<td>289818.9</td>
<td>-</td>
</tr>
<tr>
<td>slow</td>
<td>down</td>
<td>282418.3</td>
<td>+</td>
</tr>
<tr>
<td>lag</td>
<td>behind</td>
<td>257224.2</td>
<td>+</td>
</tr>
<tr>
<td>be</td>
<td>by</td>
<td>242827.7</td>
<td>-</td>
</tr>
<tr>
<td>set</td>
<td>aside</td>
<td>242238.1</td>
<td>+</td>
</tr>
</tbody>
</table>

$n$-best list ($n = 12$)

$P = \frac{5}{12} = 41.7$

$R = \frac{5}{440} = 1.1$
Evaluation on English VPC extraction task (Baldwin 2008)

Why does Web1T5 perform so badly in this task despite its size? Possible explanations include:

- Co-occurrence counts are underestimated for larger windows because of frequency threshold in n-gram database
  - quasi-collocations
- No part-of-speech annotation can be used to filter candidates

Identification of lexical collocations as habitual, recurrent word combinations (Firth 1957)

- essential for advanced learners (idiomatic English)
- different from lexicalised MWE
- semi-compositional or fully compositional
- no clear-cut linguistic criteria or tests available

Gold standard: BBI combinatory dict. (Benson et al. 1986)

- manually compiled based on lexicographer intuitions
- lexical collocations automatically extracted from BBI

\begin{itemize}
  \item injury \textit{n.} 1. to inflict (an) ~ on 2. to receive, suffer, sustain an ~ 3. a fatal; minor, slight; serious, severe ~ 4. bodily ~; an internal ~ 5. an ~ to (an ~ to the head) 6. (misc.) to add insult to ~
\end{itemize}
Evaluation on BBI collocation identification task
(Benson et al. 1986)

- Candidate extraction and ranking
  - extract all co-occurrences of BBI words within different spans (syntactic, L3/R3, L5/R5, L10/R10, sentence)
  - 1 million most frequent unordered word pairs as candidates in order to ensure fair comparison of corpora
  - ranked according to standard association measures
  - composite: AP50 = average precision up to 50% recall, selecting best measure for each data set
- Dictionary-based evaluation problematic (Evert 2004, 139f)
  - provides lower bound on n-best precision
  - coverage of native speaker intuitions by corpus data
  - may be biased against recent corpora, Web texts, etc.
- Manual validation of n-best lists
  - using custom Web-based annotation tool
  - work in progress

Why is Web1T5 so terrible?

Insufficient boilerplate removal & de-duplication:
- from collectibles to cars 9,443,572
- from collectables to cars 8,844,838
- from time to time 5,678,941
- from left to right 793,957
- from start to finish 749,705
- from a to z 572,917
- from year to year 486,669
- from top to bottom 372,935

"Traditional" Web corpora are much better:
- Google ≈ 121,000,000 hits
- Google.de ≈ 119,600,000 hits
- Web 1T 5-Grams 18,288,410 hits
- ukWaC 3 hits
- BNC 0 hits
Why is Web1T5 so terrible?

Which words are semantically similar to hot (in DSM)?

- I hope there are no minors in the room!

big (29.5), butt (31.1), ass (31.1), wet (31.2), naughty (31.6), pussy (31.6), sexy (31.6), chicks (32.0), cock (32.2), ebony (32.3), fat (32.4), girls (32.4), asian (32.7), cum (33.1), babes (33.2), dirty (33.2), bikini (33.3), granny (33.4), teen (33.8), pics (33.8), gras (34.1), fucking (34.1), galleries (34.2), fetish (34.3), babe (34.3), blonde (34.5), pussies (34.5), whores (34.6), fuck (34.6), horny (34.7)

Please don’t ask about cats and dogs . . .

Distributional Semantics

Distributional hypothesis (Harris 1954): meaning of a word can be inferred from its distribution across contexts

“You shall know a word by the company it keeps!” — (Firth 1957)

Reality check: What is the mystery word?

- He handed her her glass of XXXXX.
- Nigel staggered to his feet, face flushed from too much XXXXX.
- Malbec, one of the lesser-known XXXXX grapes, responds well to Australia’s sunshine.
- I dined off bread and cheese and this excellent XXXXX.
- The drinks were delicious: blood-red XXXXX as well as light, sweet Rhenish.

XXXXX = claret

- all examples from BNC (carefully selected & slightly edited)
Distributional semantics

- A computer can (sometimes) do the same, with sufficient amounts of corpus data and full collocational profiles

<table>
<thead>
<tr>
<th></th>
<th>get</th>
<th>see</th>
<th>use</th>
<th>hear</th>
<th>eat</th>
<th>kill</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>51</td>
<td>20</td>
<td>84</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>cat</td>
<td>52</td>
<td>58</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>???</td>
<td>115</td>
<td>83</td>
<td>10</td>
<td>42</td>
<td>33</td>
<td>17</td>
</tr>
<tr>
<td>boat</td>
<td>59</td>
<td>39</td>
<td>23</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cup</td>
<td>98</td>
<td>14</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pig</td>
<td>12</td>
<td>17</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>banana</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>

$\text{sim}(???, \text{knife}) = 0.770$

Distributional semantics with Web1T5

- Basis of distributional semantic model (DSM): term-term co-occurrence matrix of collocational profiles
  - very sparse: e.g. 250k $\times$ 100k matrix with 24.2 billion cells, but only 249.4 million cells ($\approx$ 1%) have nonzero values
  - We've already computed collocational profiles
    - 32 GiB collocations database = sparse co-occurrence matrix
    - export matrix with 25k target words (rows) and 50k high-frequency word forms as features (columns)
- DSM implemented in R (experimental wordspace package)
  - column-compressed sparse matrix
  - log $G^2$ weights, $L_2$-normalized, angular distance (= cosine), 500 latent dimensions + 50 skipped (randomized SVD)
  - parameter settings according to Lapesa & Evert (submitted)
  - needs 10 GiB RAM and less than an hour
DSM with Web1T5: nearest neighbours

Neighbours of linguistics (cosine angle):

- sociology (24.6), sociolinguistics (24.6), criminoology (29.5),
  anthropology (30.8), mathematics (31.2), phonetics (33.1),
  phonology (33.2), philology (33.2), literatures (33.5),
  gerontology (35.3), prosemear (35.5), geography (35.8),
  humanities (35.9), archaeology (35.9), science (36.5), ...

Neighbours of spaniel (cosine angle):

- terrier (23.0), schnauzer (26.5), pinscher (27.0), weimaraner
  (28.3), keeshond (29.1), pomeranian (29.4), pekingese (29.6),
  bichon (30.1), vizsla (30.5), labradoodle (30.6), apso (31.1),
  spaniels (32.0), frise (32.0), yorkie (32.1), sheepdog (32.3),
  dachshund (32.4), retriever (32.7), whippet (32.9), havanese
  (33.1), westie (34.5), mastiff (34.6), dandle (34.7), chihuahua
  (34.9), dinkmot (35.0), elkhound (35.0), ...

Evaluating distributional similarity

- Correlation with human similarity ratings
  - RG65: Rubenstein and Goodenough (1965)
- Multiple-choice tasks
  - TOEFL synonym questions
  - SAT analogy questions
- Decision tasks for consistent/inconsistent prime
  - SPP: Semantic Priming Project (Hutchison et al. 2013)
  - GEK: Generalized Event Knowledge (Ferretti et al. 2001;
    McRae et al. 2005; Hare et al. 2009)
- Noun clustering vs. semantic classification
  - AP: Almuhareb/Poesio (Almuhareb 2006)
  - Battig: Battig/Montague norms (Van Overschelde et al. 2004)
  - ESSLLI: basic-level concrete nouns (ESSLLI '08 shared task)
- Psycholinguistic norms & experiments
  - free association norms
  - property norms
  - priming effects (ΔRT)
Correlation with human similarity ratings

- Direct comparison with semantic similarity ratings
  (WordSim-353, Finkelstein et al. 2002)
  - 353 noun-noun pairs with “relatedness” ratings
  - rated on scale 0–10 by 16 test subjects
  - closely related: money/cash, soccer/football, type/kind, …
  - unrelated: king/cabbage, noon/string, sugar/approach, …
  - NB: not all "nouns" are nouns in a traditional sense
    (five, live, eat, stupid, …)

- Correlation with DSM distances for different corpora
  - DSM parameters: term-term matrix, L4/R4 surface window, 30k feature terms, log G^2 weighting, cosine similarity, SVD to 500 dimensions / 50 skipped (Lapesa & Evert submitted)
  - quantitative measure: Spearman’s rank correlation ρ
    (robust against non-linearities)

(Finkelstein et al. 2002)

Correlation with human relatedness ratings

\[
\rho = 0.644, \quad p < 0.000, \quad |r| = 0.500 .. 0.640 \text{ (1 pairs not found)}
\]

Overview & Discussion

Result overview: Frequency comparison

Frequency Comparison: Basic English (lemmatized)
Result overview: MWE identification

**MWE Extraction: English VPC (L0/R3 span)**

![Chart showing average precision across different corpora for MWE extraction.](chart1)

**MWE Extraction: BBI collocations (L3/R3)**

![Chart showing R50-averaged precision across different corpora for MWE extraction.](chart2)

**DSM Evaluation: WordSim-353**

![Chart showing rank correlation (ρ) across different corpora for DSM evaluation.](chart3)

**DSM Evaluation: multiple choice (SPP)**

![Chart showing accuracy across different corpora for DSM evaluation.](chart4)
Result overview: MWE identification

DSM Evaluation: noun clustering (AP)

Cluster purity (%)

BNC
DESC
Gigaword
WP500
Wackypedia
ukWaC
WebBase
UKCOW
Joint Web
LCC
LCC \[f >= 5\]
LCC \[f >= 10\]
Web1T5

40 50 60 70 80

●
●
●
●
●
●
● ●
● ●
●
●
●

S. Evert (stefan.evert@fau.de)

References I


References II


Evert, Stefan (2010). Google Web 1T5 n-grams made easy (but not for the computer). In *Proceedings of the 6th Web as Corpus Workshop (WAC-6)*, Los Angeles, CA.


Hare, Mary; Jones, Michael; Thomson, Caroline; Kelly, Sarah; McRae, Ken (2009). Activating event knowledge. *Cognition*, 111(2), 151–167.

References III

Hutchison, Keith A.; Balota, David A.; Neely, James H.; Cortese, Michael J.; Cohen-Shikora, Emily R.; Tse, Chi-Shing; Yap, Melvin J.; Bengson, Jesse J.; Niemeyer, Dale; Buchanan, Erin (2013). The semantic priming project. Behavior Research Methods, 45(4), 1099–1114.

McRae, Ken; Hare, Mary; Elman, Jeffrey L.; Ferretti, Todd (2005). A basis for generating expectancies for verbs from nouns. Memory & Cognition, 33(7), 1174–1184.


