Web N-Grams as a Resource for Corpus Linguistics

Stefan Evert

English Computational Corpus Linguistics
Technische Universität Darmstadt, Germany
evert@linglit.tu-darmstadt.de

Heidelberg, 12.01.2012

TECHNISCHE UNIVERSITÄT DARMSTADT

Scaling up

because more data are better data for statistical NLP (Church and Mercer 1993)

- 1964: 1 million words (Brown Corpus)

Scaling up

because more data are better data for statistical NLP (Church and Mercer 1993)

- 1964: 1 million words (Brown Corpus)
- 1995: 100 million words (British National Corpus)

Scaling up

because more data are better data for statistical NLP (Church and Mercer 1993)

- 1964: 1 million words (Brown Corpus)
- 1995: 100 million words (British National Corpus)
- 2003: 1,000+ million words (English Gigaword, WaCky)

Scaling up

because more data are better data for statistical NLP (Church and Mercer 1993)

- 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)

The Google Web 1T 5-Gram database

Brants and Franz (2006)

- Not the full 1 trillion words of English Web text, but . . .
- Frequency counts for bigrams, trigrams, 4-grams and 5-grams extracted from this corpus
  - thresholds: $f \geq 200$ for terms, $f \geq 40$ for n-grams
- Multiple compressed text files with total size of 24.4 GiB
- No linguistic pre-processing (case-folding, lemmatization, POS tagging, parsing, word sense disambiguation, . . .)
- Little boilerplate cleanup ("*from collectibles to cars*")

The Google Web 1T 5-Gram database

Brants and Franz (2006)

word 1	word 2	word 3	f
supplement	depend	on	193
supplement	depending	on	174
supplement	depends	entirely	94
supplement	depends	on	338
supplement	derived	from	2668
supplement	des	coups	77
supplement	described	in	200

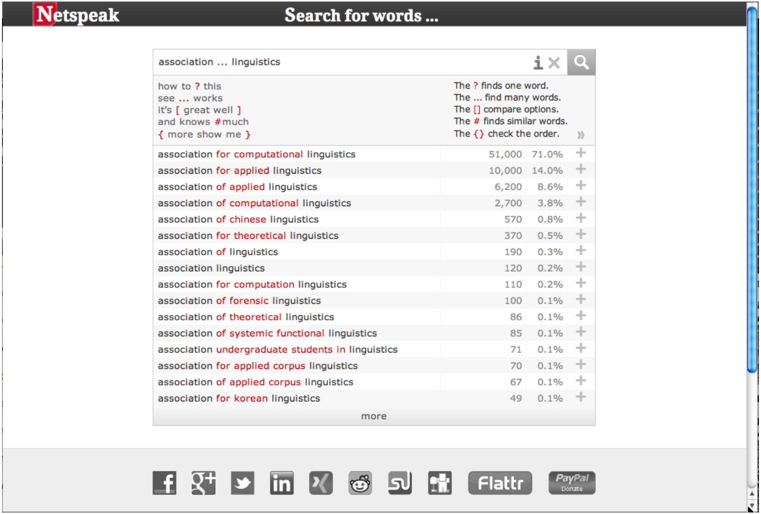
excerpt from file `3gm-0088.gz`

Some applications of Google Web1T5

- broad-coverage word-level n-gram models (of course ...)
  - machine translation
  - speech recognition
  - predictive typing
  - ...
- replacement for Google API in knowledge mining tools
- spelling correction (Bergsma *et al.* 2009)
- linguistic steganography (Chang and Clark 2010)
- near-synonym choice (Islam and Inkpen 2010)
- prediction of fMRI neural activation (Mitchell *et al.* 2008)
- testbed for n-gram search engines and analysis software (Stein *et al.* 2010; Sekine and Dalwani 2010; Lin *et al.* 2010)
  - e.g. `http://www.netspeak.org/` (University of Weimar)

Application example: Netspeak

http://www.netspeak.org/



Advantages of Google Web1T5

1. **Size**
   - more data are better data (Church and Mercer 1993)
   - three orders of magnitude larger than current Web corpora, four orders of magnitude larger than BNC
   - much better coverage of words and esp. phrases
   - data-driven NLP scales logarithmically (Banko and Brill 2001)
2. **Pre-compiled n-gram frequency data**
   - frequency counts for a trillion words of text need massive computing power and clever algorithms

Limitations of Google Web1T5

1. **Lack of linguistic annotation**
   - *part-time* is split into a trigram (*part*, –, *time*)
   - cannot search for *can*/N or verb-object combinations

Limitations of Google Web1T5

1. **Lack of linguistic annotation**
   - *part-time* is split into a trigram (*part*, –, *time*)
   - cannot search for *can*/N or verb-object combinations
2. **Frequency thresholds**
   - precise co-occurrence frequencies only for bigrams

Limitations of Google Web1T5

1. **Lack of linguistic annotation**
   - *part-time* is split into a trigram (*part*, –, *time*)
   - cannot search for *can*/N or verb-object combinations
2. **Frequency thresholds**
   - precise co-occurrence frequencies only for bigrams
3. **Lack of normalisation**
   - case-folding, deletion of non-words, numbers, URLs, ...

Limitations of Google Web1T5

1. **Lack of linguistic annotation**
   - *part-time* is split into a trigram (*part*, –, *time*)
   - cannot search for *can*/N or verb-object combinations
2. **Frequency thresholds**
   - precise co-occurrence frequencies only for bigrams
3. **Lack of normalisation**
   - case-folding, deletion of non-words, numbers, URLs, . . .
4. **No indexing for interactive search**
   - queries require linear scan of many GiB of compressed text
   - no suitable open-source indexing software available

Limitations of Google Web1T5

1. **Lack of linguistic annotation**
   - *part-time* is split into a trigram (*part*, –, *time*)
   - cannot search for *can*/N or verb-object combinations
2. **Frequency thresholds**
   - precise co-occurrence frequencies only for bigrams
3. **Lack of normalisation**
   - case-folding, deletion of non-words, numbers, URLs, . . .
4. **No indexing for interactive search**
   - queries require linear scan of many GiB of compressed text
   - no suitable open-source indexing software available
5. **Pre-compiled n-gram frequency data**
   - corpus linguists more interested in frequencies of patterns, association strength, collocations, distributional similarity
   - cannot use tagger, parser, . . . without original corpus data

Limitations of Google Web1T5

1. **Lack of linguistic annotation**
   - *part-time* is split into a trigram (*part*, –, *time*)
   - cannot search for *can*/N or verb-object combinations
2. **Frequency thresholds**
   - precise co-occurrence frequencies only for bigrams
3. **Lack of normalisation**
   - case-folding, deletion of non-words, numbers, URLs, ...
4. **No indexing for interactive search**
   - queries require linear scan of many GiB of compressed text
   - no suitable open-source indexing software available
5. **Pre-compiled n-gram frequency data**
   - corpus linguists more interested in frequencies of patterns, association strength, collocations, distributional similarity
   - cannot use tagger, parser, ... without original corpus data
6. **Web language** (sex, lolcats and advertising)

Web1T5 as a resource for corpus linguistics

Web1T5 as a resource for corpus linguistics

1. What can Web1T5 do for corpus linguists?
   - frequencies of words and phrases
   - collocation analysis
   - distributional semantics

Web1T5 as a resource for corpus linguistics

1. What can Web1T5 do for corpus linguists?
   - frequencies of words and phrases
   - collocation analysis
   - distributional semantics

2. Software implementation
   - how to compute lemmatised frequencies, association scores and distributional similarity
   - challenges: efficiency, limitations of Web1T5

Web1T5 as a resource for corpus linguistics

1. What can Web1T5 do for corpus linguists?
   - frequencies of words and phrases
   - collocation analysis
   - distributional semantics
2. Software implementation
   - how to compute lemmatised frequencies, association scores and distributional similarity
   - challenges: efficiency, limitations of Web1T5
3. Evaluating the quality of Web1T5
   - anecdotal evidence and pet peeves
   - direct comparison of frequencies and association scores
   - task-based evaluation: multiword extraction and distributional semantics

But first . . .

But first . . .

Introducing **Web1T5-Easy**

My solution to problems 3, 4 and 5

Requirements

A Web1T5 indexing software for corpus linguists should

- be flexible and powerful enough to support queries on multi-word patterns, collocation analysis, distributional semantics, word frequency distributions, . . .
- be open-source (or at least available free of charge)
- be easy to install and run (no GPU computing . . .)
- run on commodity hardware (e.g. a €5000 server)
- be fast enough for occasional interactive exploration
- connect to other analysis tools (Excel, R, . . .)

Web1T5-Easy architecture

word 1	word 2	word 3	f
supplement	depend	on	193
supplement	depending	on	174
supplement	depends	entirely	94
supplement	depends	on	338
supplement	derived	from	2668
supplement	des	coups	77
supplement	described	in	200

- 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
- I'm not the only one to come up with this idea . . . (Evert 2010; Lam 2010)

Web1T5-Easy architecture

word	id
depend	6094
depending	3571
depends	3846
...	...
on	14
...	...
supplement	5095

id 1	id 2	id 3	f
5095	6094	14	193
5095	3571	14	174
5095	3846	4585	94
5095	3846	14	338
5095	4207	27	2668
5095	2298	62481	77
5095	1840	11	200

- 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

Which RDBMS?

Requirements

- Web1T5-Easy is more than an interactive end-user GUI
- Preferably on dedicated RDBMS not shared with other users
- Indexing expensive ➜ want to share pre-compiled database

My choice: **SQLite** [`www.sqlite.org`]

- Lightweight embedded SQL engine & RDBMS
- Database stored in single, platform-independent file
- Available for C, C++, Java, C#, Perl, Python, PHP, R, . . .

But it's all SQL & Perl, so you can substitute any other RDBMS!

(for all other technical details see Evert 2010)

Database encoding procedure

Pre-processing (normalisation, filtering, . . .)

Database encoding procedure

Pre-processing (normalisation, filtering, ...)

$\Downarrow$

Numeric ID coding & database insertion [1d 23h]

Database encoding procedure

Pre-processing (normalisation, filtering, . . .)

⇓

Numeric ID coding & database insertion [1d 23h]

⇓

Collapse duplicate rows (from normalisation) [6d 7h]

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]

Database encoding procedure

Pre-processing (normalisation, filtering, ...)

$\Downarrow$

Numeric ID coding & database insertion [1d 23h]

$\Downarrow$

Collapse duplicate rows (from normalisation) [6d 7h]

$\Downarrow$

Indexing of each n-gram position [3d 2h]

$\Downarrow$

Statistical analysis for query optimisation [not useful]

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]

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]

⇓

✌ 211 GiB

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

association **. . .** X*al* Y

Querying the database

It's easy to search the database for patterns like

association … X*al* Y

with a "simple" SQL query:

```
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;
```

Querying the database

It's easy to search the database for patterns like

association … X*al* Y

with a "simple" SQL query:

```
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:

`association ? %al *`

Web1T5-Easy demo

https://cogsci.uni-osnabrueck.de/~korpora/Web1T5/ (but currently broken)

Frequency list | Associations | Collocations

The Google Web 1T 5-Gram Database — SQLite Index & Web Interface

This is the Web interface of the Web1T5-Easy package, using a GOPHER page design.

Query Form

Search pattern: association ? %al *

Search | CSV | XML

- display first 50 N-grams with frequency ≥ 100
- variable elements are listed normally , constant elements are shown

Help | Debug | Optim.

Reset Form

Results

50 matches in 11.09 seconds

87979	association .. social workers
54756	association .. computational linguistics
54119	association .. trial lawyers
49715	association .. annual meeting
45917	association .. real estate
45703	association .. criminal defense
37246	association .. mental health
26721	association .. pharmaceutical scientists
26644	association .. professional engineers
26132	association .. artificial intelligence
24770	association .. annual conference
21821	association .. neurological surgeons

Web1T5-Easy query performance

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

(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.)

Corpus Linguistics with Web1T5

Approximating lemmatized frequency counts

- Web1T5 frequencies based on unnormalized word forms
- Web1T5-Easy can perform case-folding normalization during indexing (default)

Approximating lemmatized frequency counts

- Web1T5 frequencies based on unnormalized word forms
- Web1T5-Easy can perform case-folding normalization during indexing (default)
- Approximate lemmatized frequency counts by morphological query expansion

query	f
hear sound	36,304

Approximating lemmatized frequency counts

- Web1T5 frequencies based on unnormalized word forms
- Web1T5-Easy can perform case-folding normalization during indexing (default)
- Approximate lemmatized frequency counts by morphological query expansion

query	f
`hear sound`	36,304
`[hear,hears,heard,hearing] [sound,sounds]`	95,453

  - lazy approach: use TreeTagger lexicon, or extract from BNC
  - pooled frequency counts with SQL aggregates (`GROUP BY`)

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, . . .
  - collocation database is also a sparse representation of a distributional semantic model (term-term matrix)
- Web1T5 only provides co-occurrence frequencies for immediately adjacent bigrams (e.g. * day and day *)

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, ...
  - collocation database is also a sparse representation of a distributional semantic model (term-term matrix)
- Web1T5 only provides co-occurrence frequencies for immediately adjacent bigrams (e.g. * day and day *)
- Approximate counts for distance n from $n+1$-gram table

  day ? ? * and * ? ? day

➥ **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

Collocates of “corpus” (f=5137372)

50 matches in 0.20 seconds

collocate	t-score	frequency	expected	span distribution (left, right)					
christi	1582.37	2504283	198.3	00%	01%	01%	97%	01%	00%
tx	794.93	639346	3725.8	00%	14%	02%	00%	16%	67%
habeas	720.32	518962	52.8	00%	00%	99%	00%	00%	00%
texas	629.04	411495	7978.1	06%	09%	02%	00%	22%	61%
columbus	429.55	186575	1034.0	48%	16%	36%	00%	00%	00%
dallas	390.37	156254	1943.7	00%	00%	00%	00%	70%	30%
writ	372.46	138960	116.1	98%	00%	00%	01%	00%	00%
callosum	368.99	136174	8.8	01%	00%	00%	98%	01%	00%
m	327.51	146346	21058.1	45%	46%	08%	00%	00%	00%
hotels	287.67	114198	16985.0	11%	15%	16%	00%	52%	05%
luteum	275.98	76176	5.7	02%	00%	00%	97%	01%	00%
oh	265.20	80036	5009.5	03%	04%	93%	00%	00%	00%

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)

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.

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

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

Distributional semantics

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

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	**51**	**20**	**84**	**0**	**3**	**0**
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

sim(**???**, **knife**) = 0.770

Distributional semantics

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

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	**12**	**17**	**3**	**2**	**9**	**27**
banana	11	2	2	0	18	0

sim(**???**, **pig**) = 0.939

Distributional semantics

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

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	**52**	**58**	**4**	**4**	**6**	**26**
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

sim(**???**, **cat**) = 0.961

Distributional semantics

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

	get w_1	see w_2	use w_3	hear w_4	eat w_5	kill w_6
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
???	**115**	**83**	**10**	**42**	**33**	**17**
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

??? = dog

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 245.4 million cells ($\approx 1\%$) have nonzero values

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 245.4 million cells ($\approx 1\%$) have nonzero values
- We've already computed collocational profiles
  - 32 GiB collocations database = sparse co-occurrence matrix
  - export for further processing with 250k most frequent word forms as target terms (rows) and 100k mid-frequency word forms as feature terms (columns)

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 245.4 million cells ($\approx 1\%$) have nonzero values
- We've already computed collocational profiles
  - 32 GiB collocations database = sparse co-occurrence matrix
  - export for further processing with 250k most frequent word forms as target terms (rows) and 100k mid-frequency word forms as feature terms (columns)
- DSM implemented in **R** (experimental wordspace package)
  - column-compressed sparse matrix
  - t-score feature weights with sqrt transformation
  - cosine similarity measure (converted to angle = distance)
  - dim. reduction with randomized SVD (Halko *et al.* 2009)
  - needs 20 GiB RAM and half a day (or else a weekend)

DSM with Web1T5: nearest neighbours

Neighbours of **linguistics** (cosine angle):

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

DSM with Web1T5: nearest neighbours

Neighbours of **linguistics** (cosine angle):

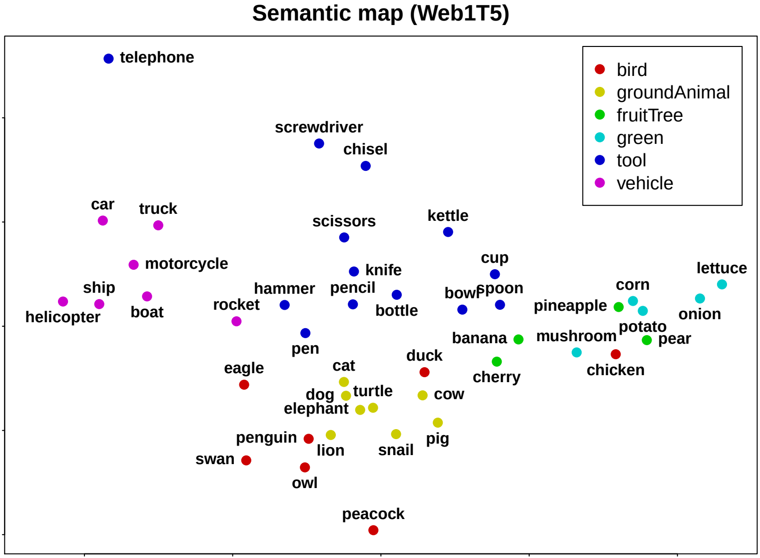
- ☞ sociology (24.6), sociolinguistics (24.6), criminology (29.5), anthropology (30.8), mathematics (31.2), phonetics (33.1), phonology (33.2), philology (33.2), literatures (33.5), gerontology (35.3), proseminar (35.5), geography (35.8), humanities (35.9), archaeology (35.9), science (36.5), ...

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), dandie (34.7), chihuahua (34.9), dinmont (35.0), elkhound (35.0), ...

DSM with Web1T5: semantic map

(data from ESSLLI 2008 shared task on concrete noun categorization)



Evaluating the quality of Web1T5

Anecdotal Evidence

Insufficient boilerplate removal & de-duplication:

```
from * to *
```

Anecdotal Evidence

Insufficient boilerplate removal & de-duplication:

`from * to *`

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

Anecdotal Evidence

Insufficient boilerplate removal & de-duplication:

`from * to *`

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 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

Anecdotal Evidence

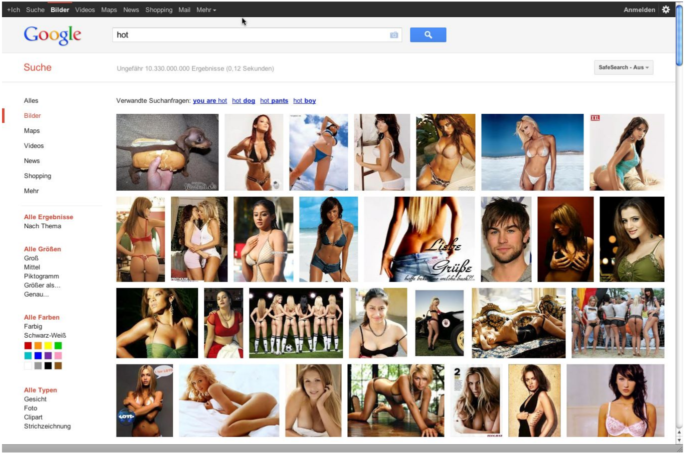
Which words are semantically similar to **hot** (in DSM)?

- I hope there are no minors in the room!

Anecdotal Evidence

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

- I hope there are no minors in the room!



Anecdotal Evidence

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 . . .

Linguistic Evaluation of Web 1T 5-Grams

- Compare Web1T5 with British National Corpus (Aston and Burnard 1998) and ukWaC Web corpus (Baroni *et al.* 2009)

Linguistic Evaluation of Web 1T 5-Grams

- Compare Web1T5 with British National Corpus (Aston and Burnard 1998) and ukWaC Web corpus (Baroni *et al.* 2009)

- Method 1: Direct comparison of frequency counts
  - expect good correlation, but better coverage from Web1T5
  - Baroni *et al.* (2009) use a similar approach to compare their ukWaC Web corpus against the BNC
  - same for association scores (bigrams, collocations)

Linguistic Evaluation of Web 1T 5-Grams

- Compare Web1T5 with British National Corpus (Aston and Burnard 1998) and ukWaC Web corpus (Baroni *et al.* 2009)

- Method 1: Direct comparison of frequency counts
  - expect good correlation, but better coverage from Web1T5
  - Baroni *et al.* (2009) use a similar approach to compare their ukWaC Web corpus against the BNC
  - same for association scores (bigrams, collocations)
- Method 2: Task-based evaluation
  - do applications benefit from the Web1T5 data?
  - multiword extraction: English particle verbs (VPC, Baldwin 2008) and light verb constructions (LVC, Tu and Roth 2011)
  - standard shared tasks for distributional models, such as TOEFL synonyms and WordSim-353 (Finkelstein *et al.* 2002)

Comparison of frequency counts

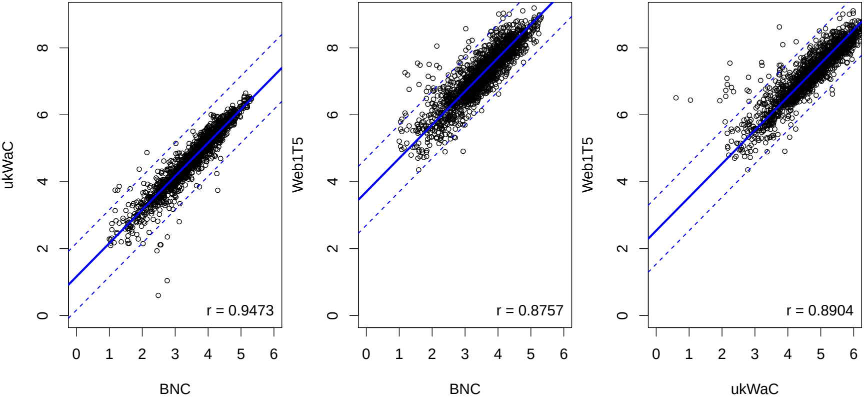
- Scatterplots of (log) frequencies in different corpora
  - BNC vs. ukWaC vs. Web 1T 5-Grams
  - only include items that occur in all three corpora (➜ not interested in coverage / idiosyncrasies)
  - correlation r from regression model $f_{\text{ukWaC}} \sim \beta \cdot f_{\text{BNC}}$ etc.

Comparison of frequency counts

- Scatterplots of (log) frequencies in different corpora
  - BNC vs. ukWaC vs. Web 1T 5-Grams
  - only include items that occur in all three corpora (➜ not interested in coverage / idiosyncrasies)
  - correlation r from regression model $f_{\text{ukWaC}} \sim \beta \cdot f_{\text{BNC}}$ etc.

- Test data sets
  - Basic English words (lemmatised vs. word form in Web1T5)
  - inflected forms of Basic English words
  - binary compound nouns extracted from WordNet 3.0
  - English particle verbs from VPC task (adjacent bigrams)
  - English particle verbs (co-occurrence in L0/R3 window)

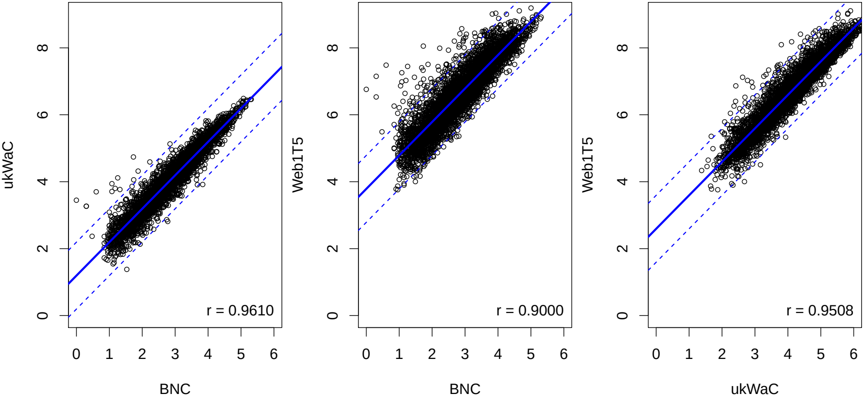
Comparison of frequency counts



Basic English (lemmatised vs. word forms)

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

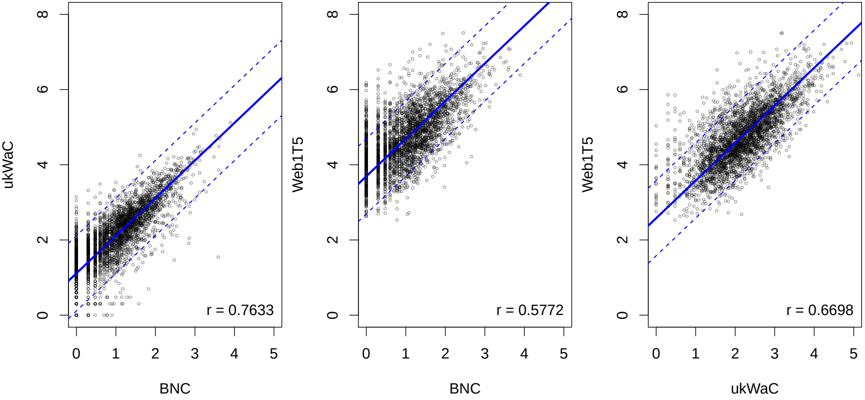
Comparison of frequency counts



Basic English (inflected forms)

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

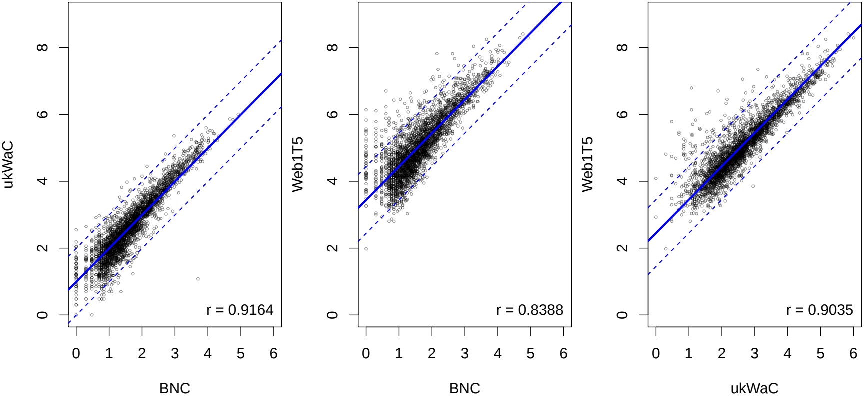
Comparison of frequency counts



Binary compound nouns (WordNet)

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

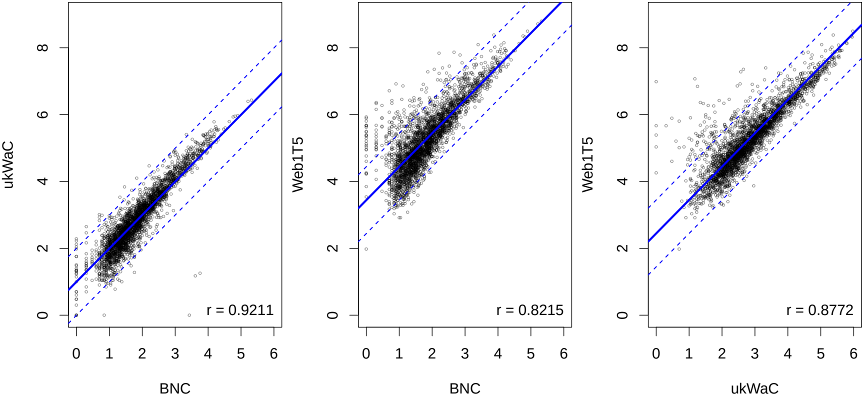
Comparison of frequency counts



Particle verbs (adjacent bigrams)

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

Comparison of frequency counts



Particle verbs (L0/R3 quasi-collocations)

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

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*, . . .

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

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

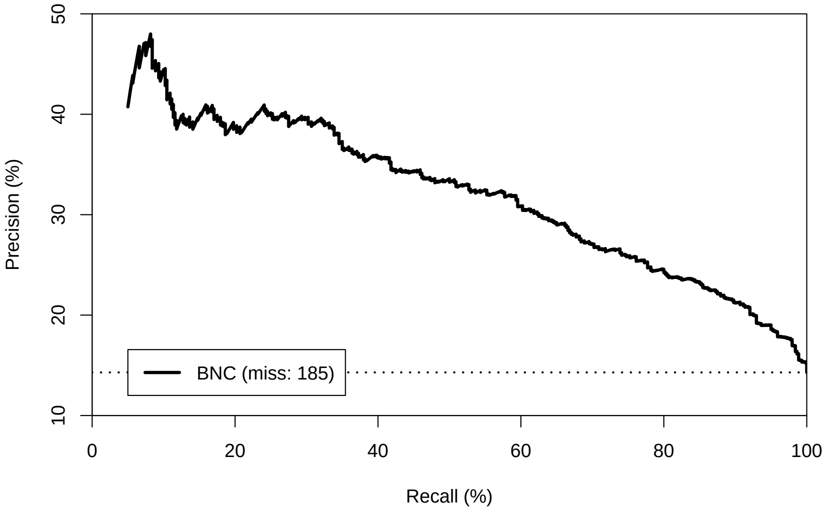
Evaluation on English VPC extraction task

(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 from BNC/ukWaC/Web1T5
  - surface co-occcurrence (L0,R3) + POS filter (except Web1T5)
  - Web1T5 without/with morphological expansion
  - using best association measure for each corpus (X^2, X^2, t, G^2, Dice)

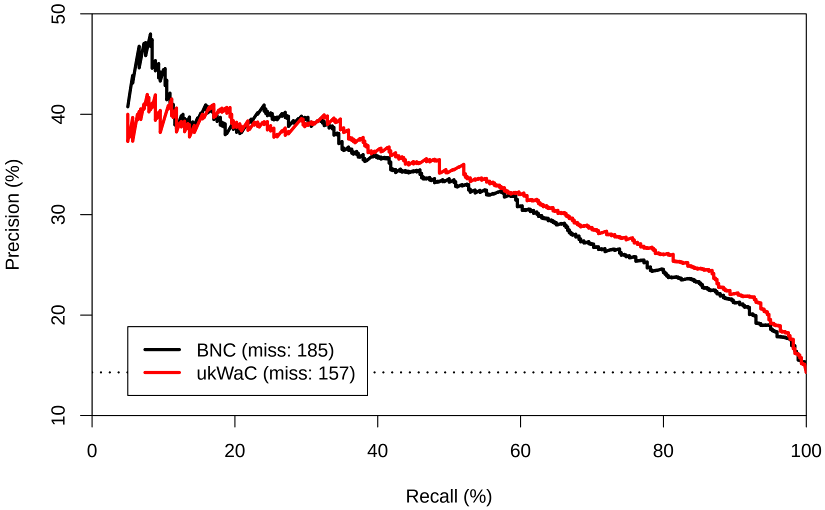
Evaluation on English VPC extraction task

(Baldwin 2008)



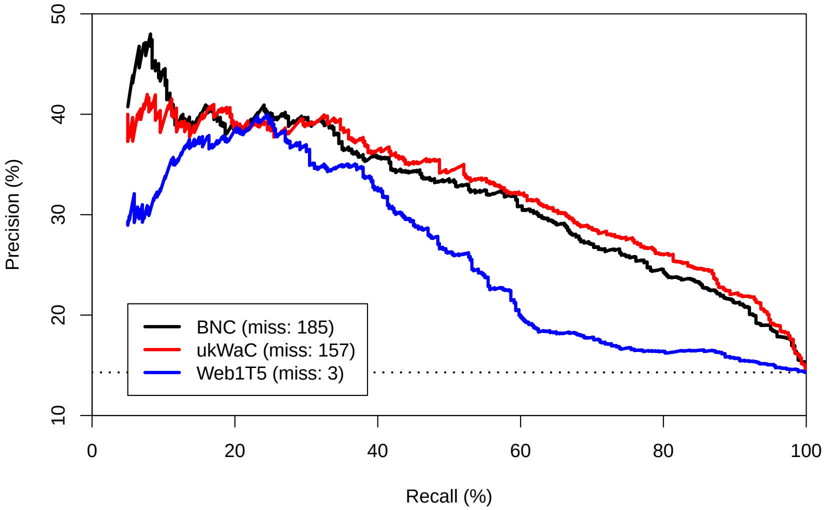
Evaluation on English VPC extraction task

(Baldwin 2008)



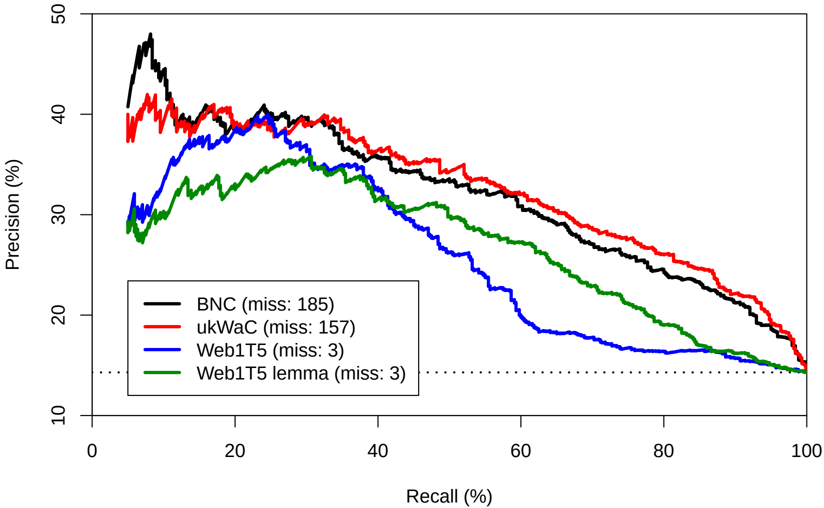
Evaluation on English VPC extraction task

(Baldwin 2008)



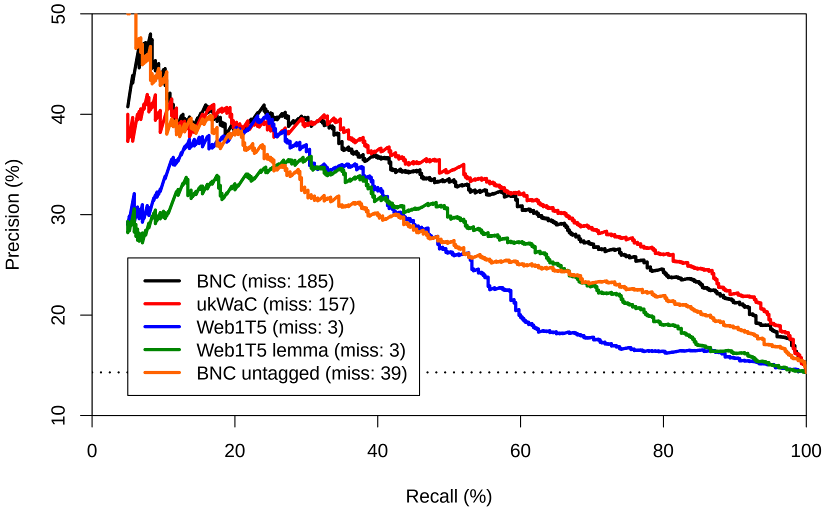
Evaluation on English VPC extraction task

(Baldwin 2008)



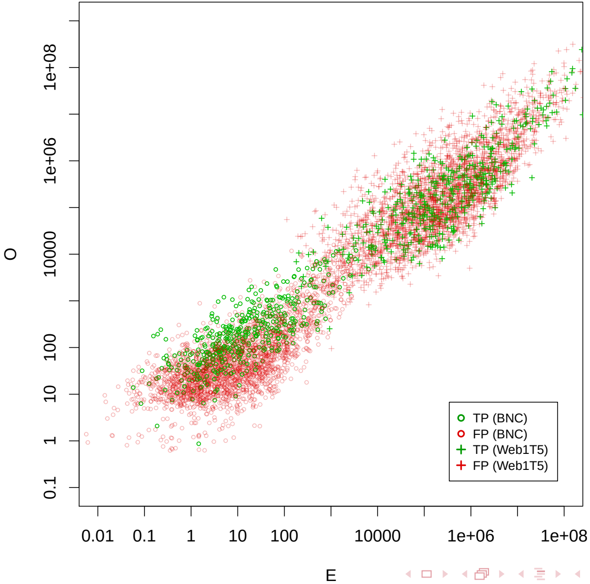
Evaluation on English VPC extraction task

(Baldwin 2008)



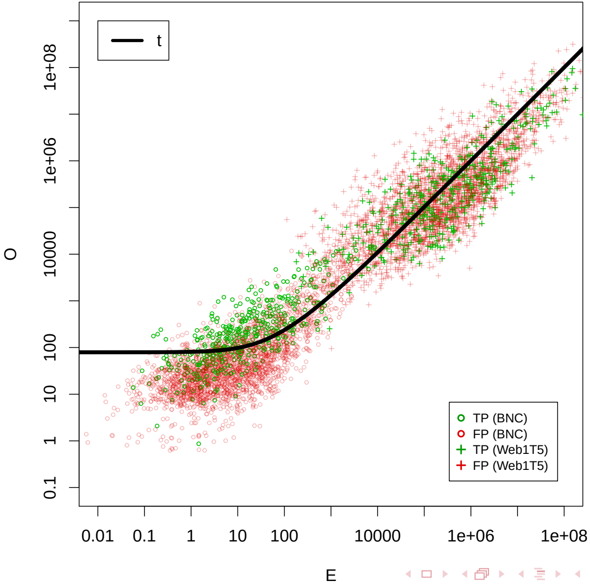
Do association measures scale badly?

fitted to BNC data



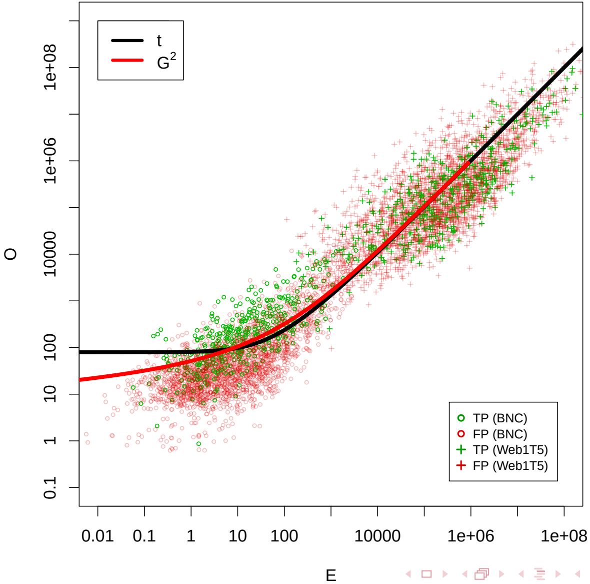
Do association measures scale badly?

fitted to BNC data



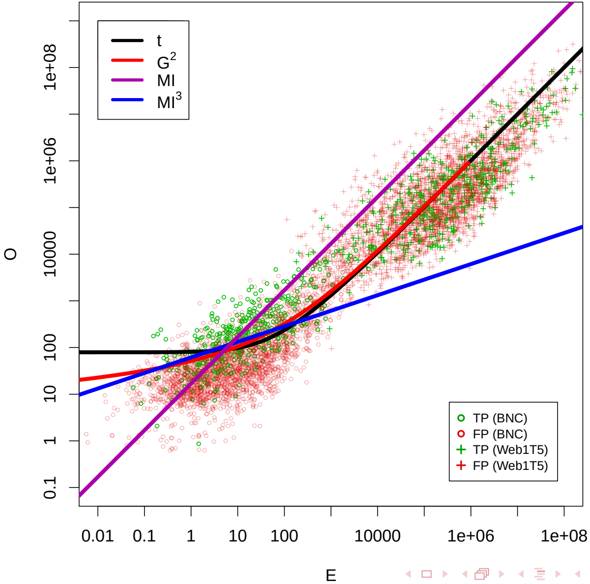
Do association measures scale badly?

fitted to BNC data



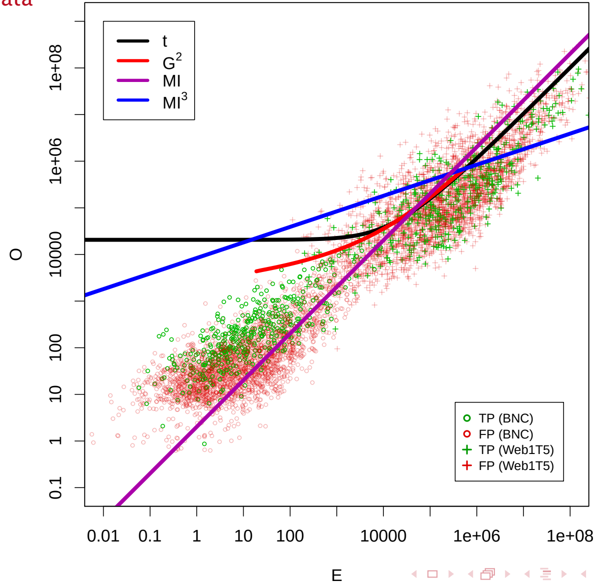
Do association measures scale badly?

fitted to BNC data

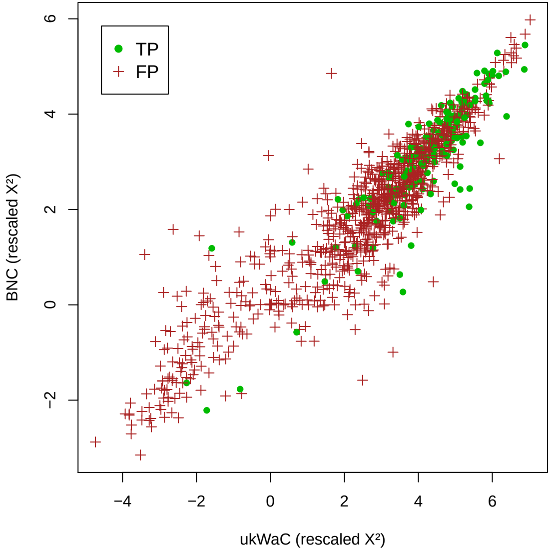


Do association measures scale badly?

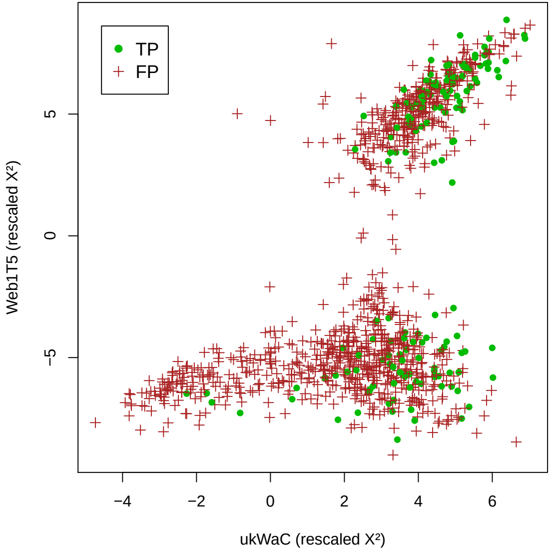
fitted to Web1T5 data



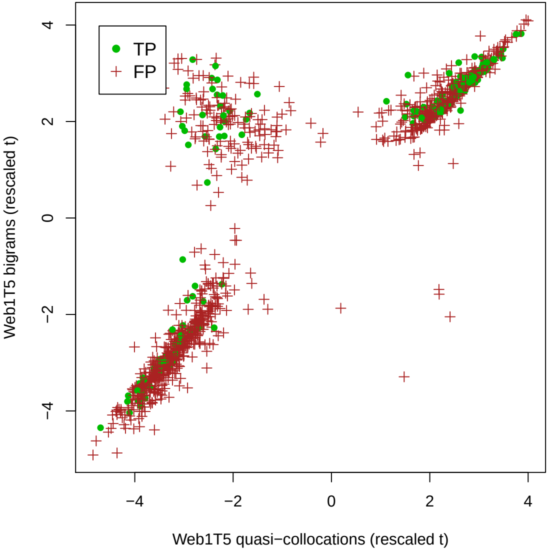
What's wrong with Web1T5 quasi-collocations?



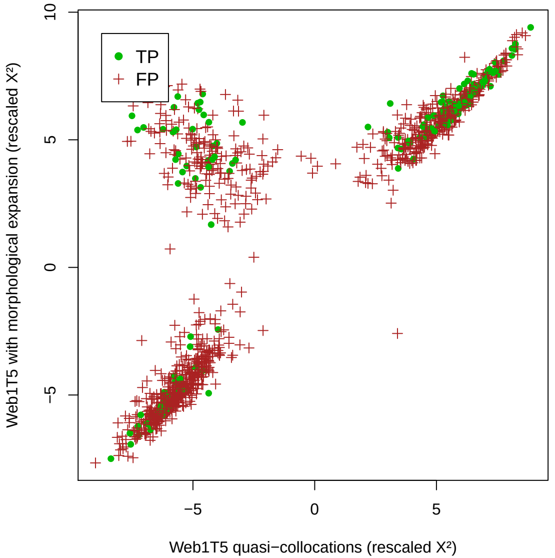
What's wrong with Web1T5 quasi-collocations?



What's wrong with Web1T5 ~~quasi-collocations~~?



What's wrong with Web1T5 ~~quasi-collocations~~?



Evaluation on English LVC extraction task

(Tu and Roth 2011)

- English **light verb constructions** (LVC) consisting of verb (semantically bleached) + object noun (often deverbal)
  - *take a walk*, *give a speech*, *have a look*, *make a call*, ...

Evaluation on English LVC extraction task

(Tu and Roth 2011)

- English **light verb constructions** (LVC) consisting of verb (semantically bleached) + object noun (often deverbal)
  - *take a walk*, *give a speech*, *have a look*, *make a call*, . . .
- Data set of 2,162 candidate LVC tokens
  - extracted from BNC with parser and various heuristics (e.g. object NP must have deverbal head noun)
  - only for verbs *do*, *get*, *give*, *have*, *make* and *take*

Evaluation on English LVC extraction task

(Tu and Roth 2011)

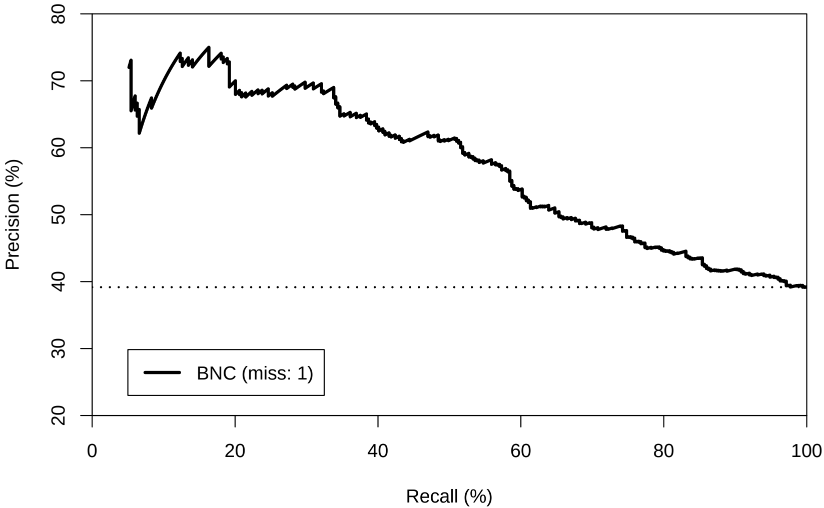
- English **light verb constructions** (LVC) consisting of verb (semantically bleached) + object noun (often deverbal)
  - *take a walk*, *give a speech*, *have a look*, *make a call*, ...
- Data set of 2,162 candidate LVC tokens
  - extracted from BNC with parser and various heuristics (e.g. object NP must have deverbal head noun)
  - only for verbs *do*, *get*, *give*, *have*, *make* and *take*
- Manually annotated as LVC / non-LVC in sentence context
  - reduced to 891 verb + head noun types for this experiment
  - type considered a LVC if at least 50% of its tokens are LVC
  - baseline: **39.2%** LVC (349 / 891 candidate types)

Evaluation on English LVC extraction task
(Tu and Roth 2011)

- English **light verb constructions** (LVC) consisting of verb (semantically bleached) + object noun (often deverbal)
  - *take a walk*, *give a speech*, *have a look*, *make a call*, ...
- Data set of 2,162 candidate LVC tokens
  - extracted from BNC with parser and various heuristics (e.g. object NP must have deverbal head noun)
  - only for verbs *do*, *get*, *give*, *have*, *make* and *take*
- Manually annotated as LVC / non-LVC in sentence context
  - reduced to 891 verb + head noun types for this experiment
  - type considered a LVC if at least 50% of its tokens are LVC
  - baseline: **39.2%** LVC (349 / 891 candidate types)
- Evaluation: candidate ranking from BNC/ukWaC/Web1T5
  - surface co-occcurrence (L3,R3) + POS filter (except Web1T5)
  - association measure: G^2 with POS filter, MI without

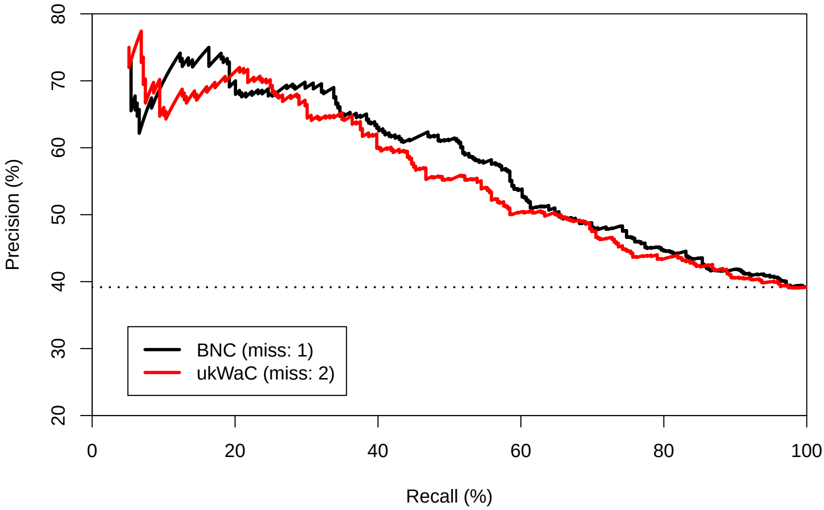
Evaluation on English LVC extraction task

(Tu and Roth 2011)



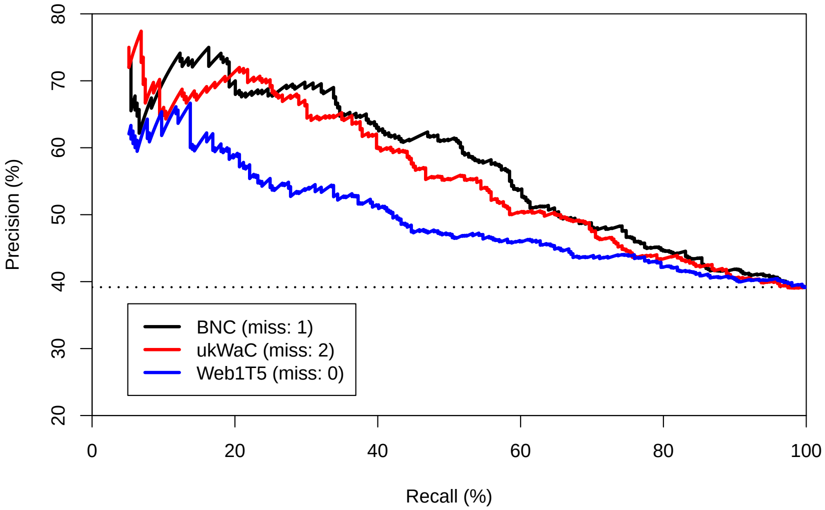
Evaluation on English LVC extraction task

(Tu and Roth 2011)



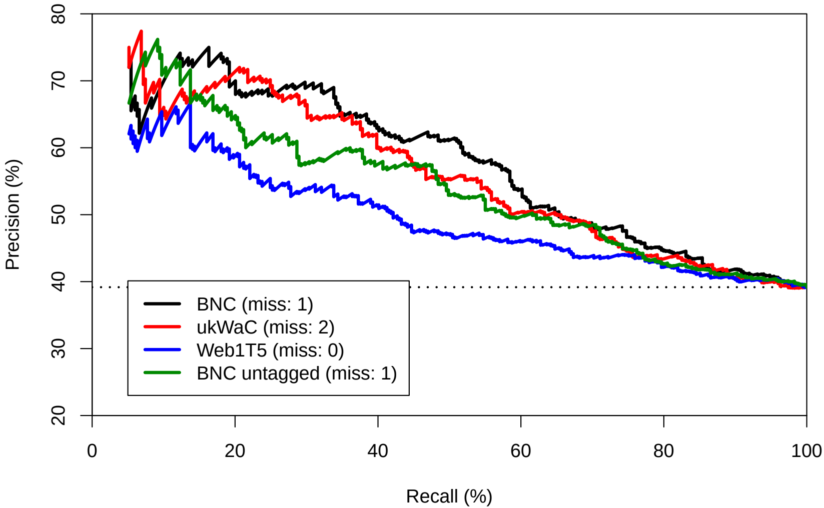
Evaluation on English LVC extraction task

(Tu and Roth 2011)



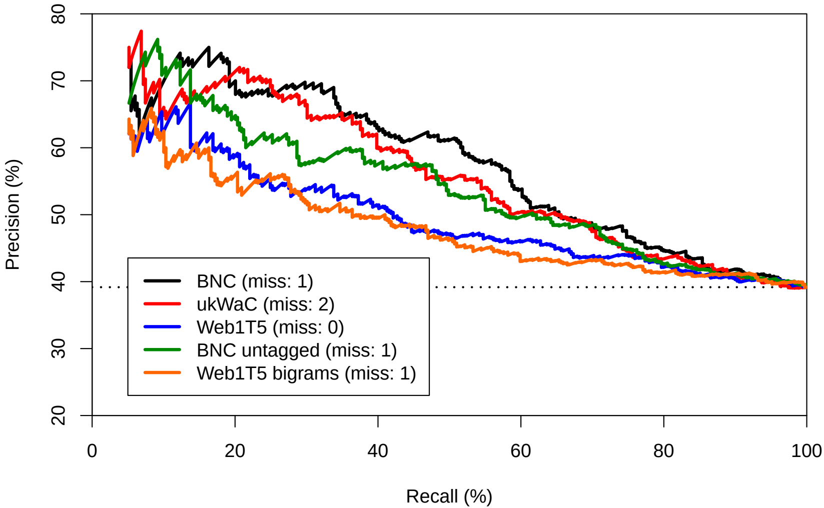
Evaluation on English LVC extraction task

(Tu and Roth 2011)



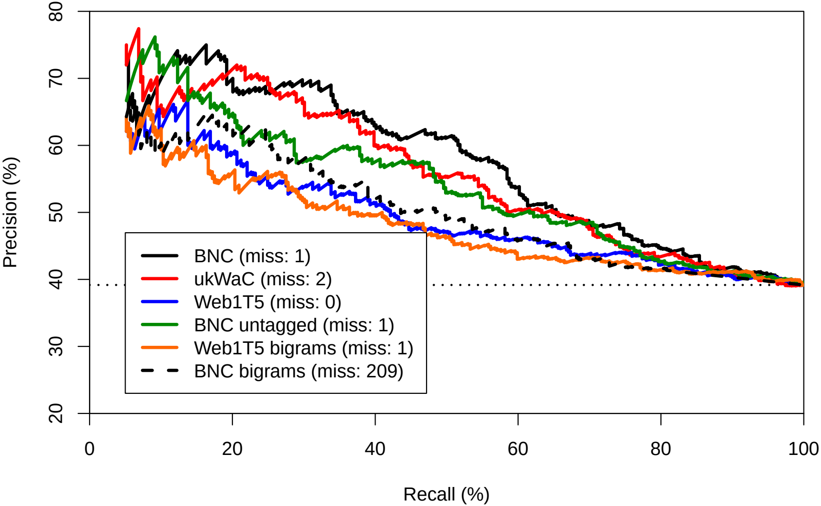
Evaluation on English LVC extraction task

(Tu and Roth 2011)

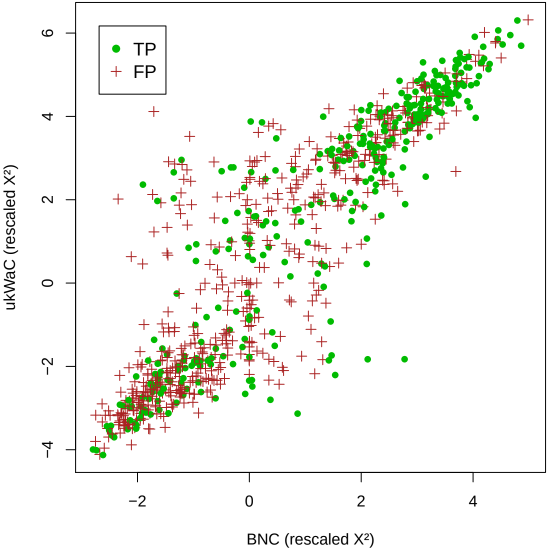


Evaluation on English LVC extraction task

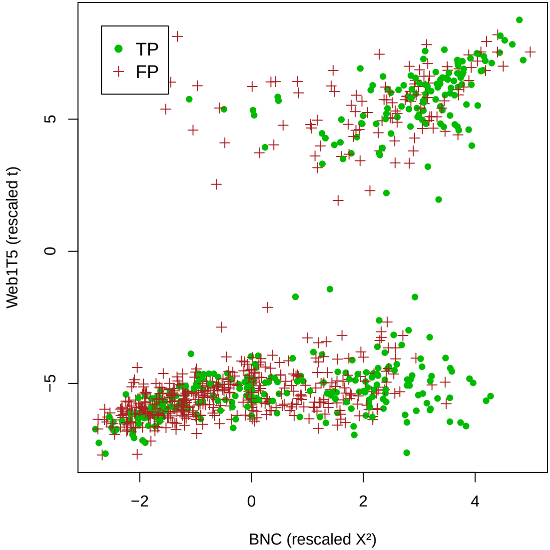
(Tu and Roth 2011)



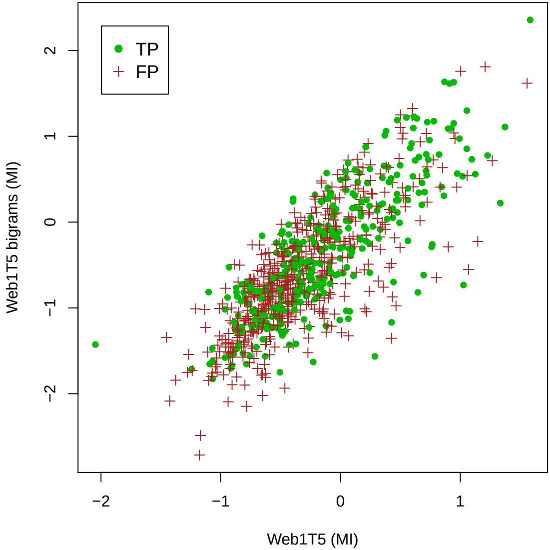
Comparison of association scores for English LVC



Comparison of association scores for English LVC



Comparison of association scores for English LVC



Evaluating distributional similarity in Web1T5

- Distributional semantic model built from Web1T5 can be evaluated in various shared tasks (e.g. ESSLLI 2008)

Evaluating distributional similarity in Web1T5

- Distributional semantic model built from Web1T5 can be evaluated in various shared tasks (e.g. ESSLLI 2008)

- Here: 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*, ...

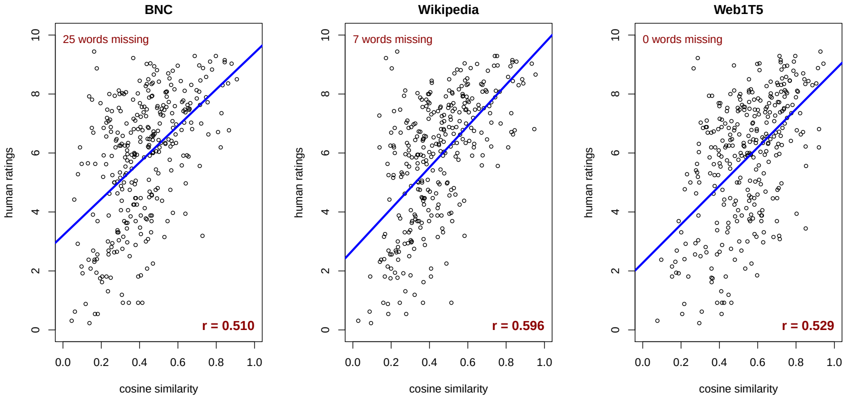
Evaluating distributional similarity in Web1T5

- Distributional semantic model built from Web1T5 can be evaluated in various shared tasks (e.g. ESSLLI 2008)

- Here: 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*, ...
- Correlation with DSM similarity in BNC/Wikipedia/Web1T5
  - DSM parameters: term-term matrix, (L2,R2) surface context, $\sqrt{t}$ weighting, cosine similarity, SVD to 300 dimensions
  - lemma vs. POS-disambiguated lemma on BNC and Wikipedia
  - word forms on Web1T5

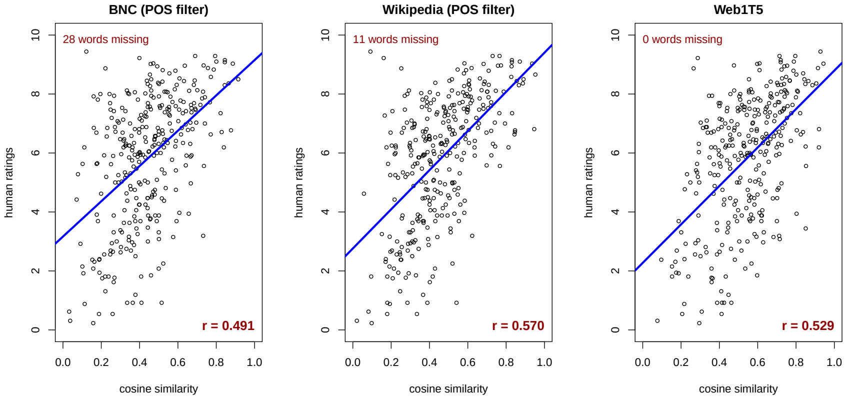
Evaluating distributional similarity in Web1T5

correlation with human relatedness ratings (Finkelstein *et al.* 2002)



Evaluating distributional similarity in Web1T5

correlation with human relatedness ratings (Finkelstein *et al.* 2002)



Work in progress

- Find out what's really wrong with the Web 1T 5-grams
  - qualitative error analysis: which words and pairs are off?
  - further experiments on scaling of association measures, direct comparison of frequencies and association score, etc.
  - esp. usefulness of morphological expansion
  - linguistic quality of Web data (topics, slang, . . .)
- Software improvements (Web1T5-Easy 2.0)
  - adapt to Web1T5 European edition (Brants and Franz 2009)
  - better customisation (e.g. normalisation, tagged data)
  - consistent Unicode support, more flexible Web GUI
  - include distributional model in open-source code
- Partial POS-tagging and lemmatisation of n-grams possible?

That's all folks!

http://webascorpus.sf.net/Web1T5-Easy/

Try the online demo at
http://cogsci.uos.de/~korpora/Web1T5/
— currently offline —

Thanks for listening!

References I

Aston, Guy and Burnard, Lou (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh. See also the BNC homepage at http://www.natcorp.ox.ac.uk/.

Baldwin, Timothy (2008). A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 1–2, Marrakech, Morocco.

Banko, Michele and Brill, Eric (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano; Zanchetta, Eros (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, **43**(3), 209–226.

Bergsma, Shane; Lin, Dekang; Goebel, Randy (2009). Web-scale N-gram models for lexical disambiguation. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI '09)*, pages 1507–1512, Pasadena, CA. Morgan Kaufmann Publishers Inc.

Brants, Thorsten and Franz, Alex (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, PA. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13.

References II

Brants, Thorsten and Franz, Alex (2009). *Web 1T 5-gram, 10 European Languages Version 1*. Linguistic Data Consortium, Philadelphia, PA. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T25`.

Chang, Ching-Yun and Clark, Stephen (2010). Linguistic steganography using automatically generated paraphrases. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 591–599, Los Angeles, CA.

Church, Kenneth W. and Mercer, Robert L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, **19**(1), 1–24.

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.

Finkelstein, Lev; Gabrilovich, Evgeniy; Matias, Yossi; Rivlin, Ehud; Solan, Zach; Wolfman, Gadi; Ruppin, Eytan (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20**(1), 116–131.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford. Reprinted in Palmer (1968), pages 168–205.

References III

Halko, N.; Martinsson, P. G.; Tropp, J[oel] A. (2009). Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. Technical Report 2009-05, ACM, California Institute of Technology.

Harris, Zellig (1954). Distributional structure. *Word*, **10**(23), 146–162. Reprinted in Harris (1970, 775–794).

Islam, Aminul and Inkpen, Diana (2010). Near-synonym choice using a 5-gram language model. *Research in Computing Science: Special issue on Natural Language Processing and its Applications*, **46**, 41–52.

Lam, Yan Chi (2010). Managing the Google Web 1T 5-gram with relational database. *Journal of Education, Informatics, and Cybernetics*, **2**(2).

Lin, Dekang; Church, Kenneth; Ji, Heng; Sekine, Satoshi; Yarowsky, David; Bergsma, Shane; Patil, Kailash; Pitler, Emily; Lathbury, Rachel; Rao, Vikram; Dalwani, Kapil; Narsale, Sushant (2010). New tools for web-scale n-grams. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association (ELRA).

Mitchell, Tom M.; Shinkareva, Svetlana V.; Carlson, Andrew; Chang, Kai-Min; Malave, Vicente L.; Mason, Robert A.; Just, Marcel Adam (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, **320**, 1191–1195.

References IV

Sekine, Satoshi and Dalwani, Kapil (2010). Ngram search engine with patterns combining token, pos, chunk and ne information. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association (ELRA).

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Sinclair, John McH. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins (eds.), *In Memory of J. R. Firth*, pages 410–430. Longmans, London.

Stein, Benno; Potthast, Martin; Trenkmann, Martin (2010). Retrieving customary Web language to assist writers. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. M. Rüger, and K. van Rijsbergen (eds.), *Advances in Information Retrieval: 32nd European Conference on Information Retrieval (ECIR '10)*, volume 5993 of *Lecture Notes in Computer Science*, pages 631–635. Springer, Berlin, Heidelberg, New York.

Tu, Yuancheng and Roth, Dan (2011). Learning English light verb constructions: Contextual or statistical. In *Proceedings of the ACL 2011 Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, Portland, OR.