Developing adjective scales from user-supplied textual metadata

Christopher Potts, Stanford Linguistics

NSF Workshop on Restructuring Adjectives in WordNet
September 30–Oct 1, 2011

Overview

Goals

Describe and evaluate a method for using naturally-occurring annotations to impose partial orderings on modifiers.

Plan

1. Data: user-supplied product and service reviews
2. Methods: hierarchical logistic regresssion
3. Evaluation: classification and scale induction against gold-standard lexicons
4. Looking ahead: alternative approaches and general issues

Overview

Limitations

My hope is that these techniques can *assist* WordNet annotators. They can't replace such annotators for serious lexicography.

Data and documentation

http://www.stanford.edu/~cgpotts/data/wordnetscales/

Related work

WordNet-based

Andreevskaia, Alina and Sabine Bergler. 2006. Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of EACL*, 209-216.

Blair-Goldensohn, Sasha; Kerry Hannan; Ryan McDonald; Tyler Neylon; George A. Reis; and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*.

Esuli, Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, 417-422.

Valitutti, Alessandro; Carlo Strapparava; and Oliviero Stock. 2004. Developing affective lexical resources. *PsychNology Journal* 2(1):61-83.

Related work

Open domains

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of ACL*, 172-182.

Täckström, Oscar and McDonald, Ryan. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of ACL*.

Turney, Peter D. and Michael L. Littman, 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems* 21.

Velikovich, Leonid; Sasha Blair-Goldensohn; Kerry Hannan; and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Proceedings of NAACL*.

Wiebe, Janyce; Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39(2–3): .

Scales in linguistics

General challenges

1. Lexical scales are fundamental to pragmatic inference, particularly for scalar conversational implicatures.
2. Scales involve gradable modifiers. Their vagueness and context-dependence affect the stability of scalar orderings.
3. Scales are affected by negation and intensional operators.

Foundational work

Fauconnier, Gilles. 1975. Pragmatic scales and logical structure. *Linguistic Inquiry* 6(3): 353375.

Hirschberg, Julia. 1985. *A Theory of Scalar Implicature*. PhD thesis, Penn.

Horn, Laurence R. 1972. *On the Semantic Properties of Logical Operators in English*. PhD thesis, UCLA.

Data

1. Data: user-supplied product and service reviews
2. Methods: hierarchical logistic regresssion
3. Evaluation: classification and scale induction against gold-standard lexicons
4. Discussion: alternative approaches and general issues

IMDB

User Reviews (Review this title)

294 out of 454 people found the following review useful.

WALL-E is one of the most cutest, lovable ch

Author: michael11391 from Augusta, Ga

Not only it's Pixar's best film of all-time but it's the b
animated films in years and surprisingly, one of the
mines. It's so beautiful, moving, hilarious & sad at tl
E, it's certainly one of his best right behind Finding I
WALL-E knocked off Ratatouille of the top spot in w
ever seen with Ratatouille right behind and Finding I
be remembered as one of the most lovable characte

Was the above review useful to you? **Yes** **No**

See more (855 total) »

IMDB

Rating	Reviews	Words	Vocabulary	Mean wrds/rev.
1	124,587 (9%)	28,962,201	172,346	203.84
2	51,390 (4%)	13,436,851	119,245	228.74
3	58,051 (4%)	15,987,151	132,002	241.10
4	59,781 (4%)	17,095,212	138,355	250.31
5	80,487 (6%)	23,293,790	164,476	253.34
6	106,145 (8%)	31,317,918	194,195	258.33
7	157,005 (12%)	45,913,948	240,876	255.99
8	195,378 (14%)	55,634,817	267,901	249.38
9	170,531 (13%)	45,941,763	236,249	236.19
10	358,441 (26%)	84,294,625	330,784	206.31
Total	1,361,796	361,878,276	800,743	232.83

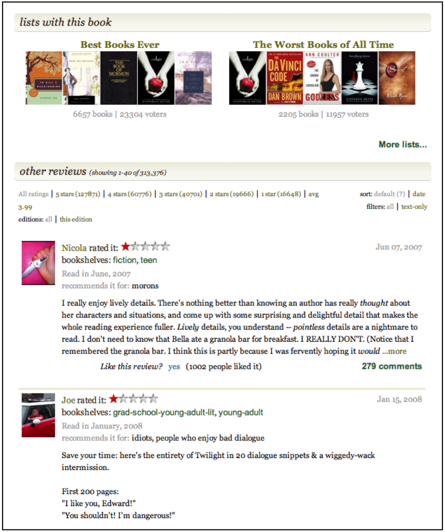
OpenTable


It was our third time at Firefly
OpenTable Diner Since 2008
Dined on 09/18/2011
It was our third time at Firefly and once again it was an incredibly memorable meal. The food preparation was imaginative and the quality of the food was outstanding. The desserts were over the top.
Special Features:
fit for foodies, neighborhood gem, notable wine list, special occasion
Food
Service
Ambiance
Noise Level
Moderate
SHARE
Report inappropriate content

OpenTable

Rating	Reviews	Words	Vocabulary	Mean wrds/rev.
1	9,352 (2%)	699,695	17,912	74.82
2	36,997 (8%)	2,507,147	34,818	67.77
3	73,064 (15%)	4,207,700	45,258	57.59
4	172,195 (35%)	7,789,649	64,143	45.24
5	197,757 (40%)	8,266,564	65,514	41.80
Total	489,365	23,470,755	116,406	47.96

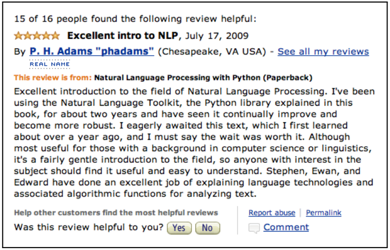
Goodreads


lists with this book
Best Books Ever
6657 books | 23304 voters
The Worst Books of All Time
2205 books | 11957 voters
More lists...
other reviews (showing 1-40 of 313,376)
All ratings | 5 stars (127871) | 4 stars (60776) | 3 stars (40701) | 2 stars (19666) | 1 star (16648) | avg 3.99
sort: default (?) | date
filters: all | text-only
editions: all | this edition
Nicola rated it:
Jun 07, 2007
bookshelves: fiction, teen
Read in June, 2007
recommends it for: morons
I really enjoy lively details. There's nothing better than knowing an author has really thought about her characters and situations, and come up with some surprising and delightful detail that makes the whole reading experience fuller. Lively details, you understand -- pointless details are a nightmare to read. I don't need to know that Bella ate a granola bar for breakfast. I REALLY DON'T. (Notice that I remembered the granola bar. I think this is partly because I was fervently hoping it would ...more
Like this review? yes (1002 people liked it)
279 comments
Joe rated it:
Jan 15, 2008
bookshelves: grad-school-young-adult-lit, young-adult
Read in January, 2008
recommends it for: idiots, people who enjoy bad dialogue
Save your time: here's the entirety of Twilight in 20 dialogue snippets & a wiggedy-wack intermission.
First 200 pages:
"I like you, Edward!"
"You shouldn't! I'm dangerous!"

Goodreads

Rating	Reviews	Words	Vocabulary	Mean wrds/rev.
1	32,057 (4%)	1,934,543	53,965	60.35
2	42,258 (5%)	2,576,214	59,501	60.96
3	81,530 (9%)	4,526,012	80,670	55.51
4	121,315 (13%)	6,037,719	95,341	49.77
5	178,225 (20%)	7,664,620	105,839	43.01
Total	455,385 (50%)	22,739,108	198,139	49.93

Amazon and Tripadvisor (pooled)



15 of 16 people found the following review helpful:

★★★★★ **Excellent intro to NLP**, July 17, 2009

By P. H. Adams "phadams" (Chesapeake, VA USA) - See all my reviews

REAL NAME

This review is from: **Natural Language Processing with Python (Paperback)**

Excellent introduction to the field of Natural Language Processing. I've been using the Natural Language Toolkit, the Python library explained in this book, for about two years and have seen it continually improve and become more robust. I eagerly awaited this text, which I first learned about over a year ago, and I must say the wait was worth it. Although most useful for those with a background in computer science or linguistics, it's a fairly gentle introduction to the field, so anyone with interest in the subject should find it useful and easy to understand. Stephen, Ewan, and Edward have done an excellent job of explaining language technologies and associated algorithmic functions for analyzing text.

Help other customers find the most helpful reviews | Report abuse | Permalink

Was this review helpful to you? Yes No | Comment

Reviews you can trust

63% Do not recommend

26 reviews

Excellent	0
Very	5
Average	5
Poor	4
Terrible	12

By trip type

- All (26)
- Business (4)
- Couples (2)
- Family (5)
- Friends getaway (0)
- Solo travel (2)

3-7 of 26 « 1 2 3 ... 6 »

Sort by [Date] [Rating] English first

"Choose another hotel"

Penn Tower Hotel

Save Review

patrema 27 contributions
Lancaster County, PA

Jul 31, 2009 | Trip type: Business

1 person found this review helpful

The "service" is the worst I've ever experienced; the rooms have an old, tired, and dirty feel. The word is they are tearing the tower down in the near future, and not a minute too soon.

My son, daughter and I stayed here for a few days while visiting a family member in the excellent Hospital of the University of PA (HUP) which is across the street and attached by an enclosed walkway.

There are only two floors of the tower that are used as a hotel; the rest is an office building and owned, I believe, by HUP. This "hotel" really is a disgrace. I would only stay here again if I absolutely had to be going to and from the hospital at night. It probably is safer than walking the streets around the hospital.

However, after discovering how bad this place is, we checked out and stayed for about 5 days at the very nice Inn at Penn, a Hilton, which is just a few blocks from the hospital. I think they offer a hospital rate most of the time. I just made sure that my visitng hours were timed with lots of foot traffic on the streets and vehicle traffic on the roads around. The University of Penn's campus is right there, but they have had some crime problems in the past and now have a couple of campus guards on most corners. Still, even with this added safety factor, it's not the best place to be walking at night.

Bottom line: I wouldn't recommend this place to anyone unless safety is the ONLY concern.

My ratings for this hotel

Value
Rooms
Location
Cleanliness
Service

Date of stay July 2009

Visit was for Business

Traveled with Other

Member since April 10, 2008

Would you recommend this hotel to a friend? No

Amazon and Tripadvisor (pooled)

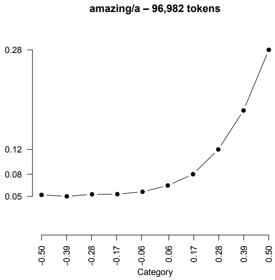
Rating	Reviews	Words	Vocabulary	Mean wrds/rev.
1	8,434 (4%)	4,756,322	53,704	563.95
2	7,545 (3%)	4,691,936	51,264	621.86
3	10,083 (4%)	5,883,625	58,396	583.52
4	28,186 (12%)	14,264,519	86,507	506.09
5	64,147 (27%)	28,135,240	124,389	438.61
Total	118,395 (50%)	57,731,642	184,487	487.62

Methods

1. Data: user-supplied product and service reviews
2. Methods: hierarchical logistic regresssion
3. Evaluation: classification and scale induction against gold-standard lexicons
4. Looking ahead: alternative approaches and general issues

Simple logistic regression

R	Cat.	Count	Total	$\Pr(w\mid c)$	$\Pr(c\mid w)$
1	−0.50	2,881	28,962,201	0.00010	0.05
2	−0.39	1,277	13,436,851	0.00010	0.05
3	−0.28	1,618	15,987,151	0.00010	0.05
4	−0.17	1,740	17,095,212	0.00010	0.05
5	−0.06	2,540	23,293,790	0.00011	0.06
6	+0.06	4,017	31,317,918	0.00013	0.07
7	+0.17	7,470	45,913,948	0.00016	0.08
8	+0.28	13,259	55,634,817	0.00024	0.12
9	+0.39	16,427	45,941,763	0.00036	0.18
10	+0.50	45,753	84,294,625	0.00054	0.28

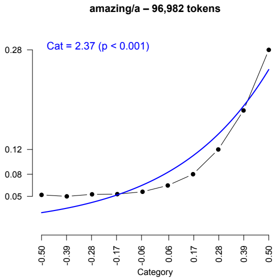


$$\Pr(w\mid c) \stackrel{def}{=} \frac{\mathrm{Count}(w,c)}{\mathrm{Total}(c)}$$

$$\Pr(c\mid w) \stackrel{def}{=} \frac{\Pr(w\mid c)}{\sum_{x\in \mathrm{Cat}} \Pr(w\mid x)}$$

Simple logistic regression

R	Cat.	Count	Total	Pr(w|c)	Pr(c|w)
1	−0.50	2,881	28,962,201	0.00010	0.05
2	−0.39	1,277	13,436,851	0.00010	0.05
3	−0.28	1,618	15,987,151	0.00010	0.05
4	−0.17	1,740	17,095,212	0.00010	0.05
5	−0.06	2,540	23,293,790	0.00011	0.06
6	+0.06	4,017	31,317,918	0.00013	0.07
7	+0.17	7,470	45,913,948	0.00016	0.08
8	+0.28	13,259	55,634,817	0.00024	0.12
9	+0.39	16,427	45,941,763	0.00036	0.18
10	+0.50	45,753	84,294,625	0.00054	0.28



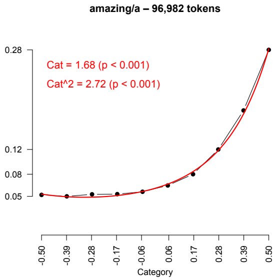
$$\Pr(w|c) \stackrel{def}{=} \frac{\text{Count}(w,c)}{\text{Total}(c)}$$

$$\Pr(c|w) \stackrel{def}{=} \frac{\Pr(w|c)}{\sum_{x\in\text{Cat}} \Pr(w|x)}$$

$$\text{logit}^{-1}\left(\begin{array}{c}\texttt{intercept} +\\ \texttt{category}\end{array}\right)$$

Simple logistic regression

R	Cat.	Count	Total	$\Pr(w\|c)$	$\Pr(c\|w)$
1	−0.50	2,881	28,962,201	0.00010	0.05
2	−0.39	1,277	13,436,851	0.00010	0.05
3	−0.28	1,618	15,987,151	0.00010	0.05
4	−0.17	1,740	17,095,212	0.00010	0.05
5	−0.06	2,540	23,293,790	0.00011	0.06
6	+0.06	4,017	31,317,918	0.00013	0.07
7	+0.17	7,470	45,913,948	0.00016	0.08
8	+0.28	13,259	55,634,817	0.00024	0.12
9	+0.39	16,427	45,941,763	0.00036	0.18
10	+0.50	45,753	84,294,625	0.00054	0.28



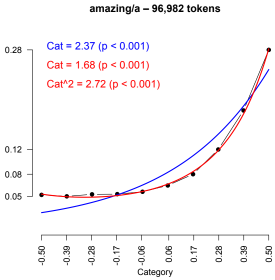
$$\Pr(w|c) \stackrel{def}{=} \frac{\text{Count}(w,c)}{\text{Total}(c)}$$

$$\Pr(c|w) \stackrel{def}{=} \frac{\Pr(w|c)}{\sum_{x\in\text{Cat}} \Pr(w|x)}$$

$$\text{logit}^{-1}\left(\begin{array}{r}\texttt{intercept} + \\ \texttt{category} + \\ \texttt{category}^2\end{array}\right)$$

Simple logistic regression

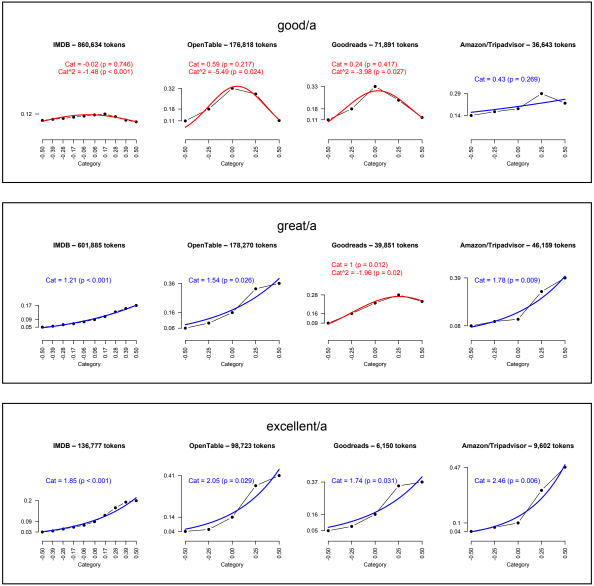
R	Cat.	Count	Total	Pr(w|c)	Pr(c|w)
1	−0.50	2,881	28,962,201	0.00010	0.05
2	−0.39	1,277	13,436,851	0.00010	0.05
3	−0.28	1,618	15,987,151	0.00010	0.05
4	−0.17	1,740	17,095,212	0.00010	0.05
5	−0.06	2,540	23,293,790	0.00011	0.06
6	+0.06	4,017	31,317,918	0.00013	0.07
7	+0.17	7,470	45,913,948	0.00016	0.08
8	+0.28	13,259	55,634,817	0.00024	0.12
9	+0.39	16,427	45,941,763	0.00036	0.18
10	+0.50	45,753	84,294,625	0.00054	0.28



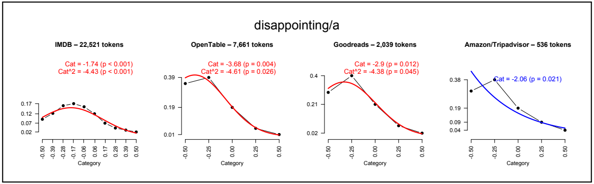
$$\Pr(w|c) \overset{def}{=} \frac{\mathrm{Count}(w,c)}{\mathrm{Total}(c)}$$

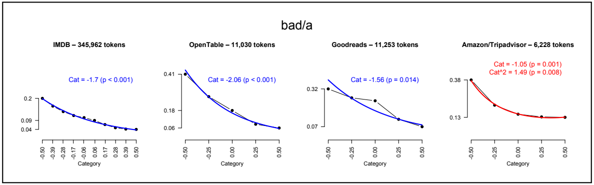
$$\Pr(c|w) \overset{def}{=} \frac{\Pr(w|c)}{\sum_{x \in \mathrm{Cat}} \Pr(w|x)}$$

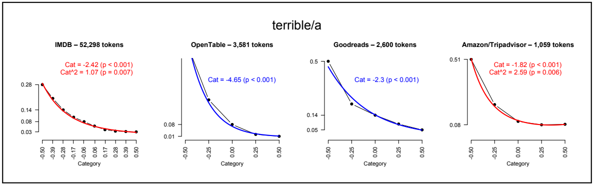
Example: positive scalars



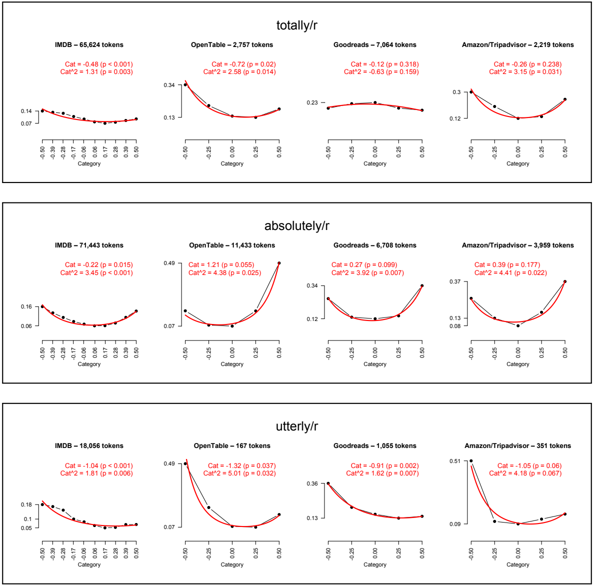
Example: negative scalars



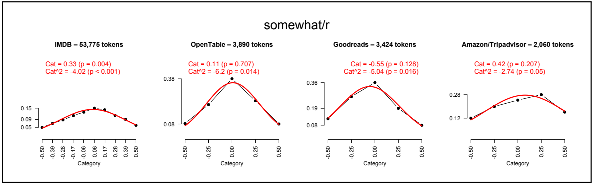


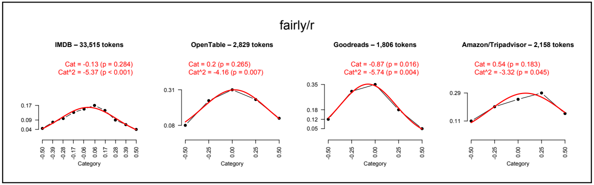


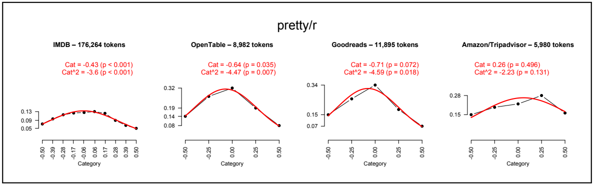
Example: emphatics



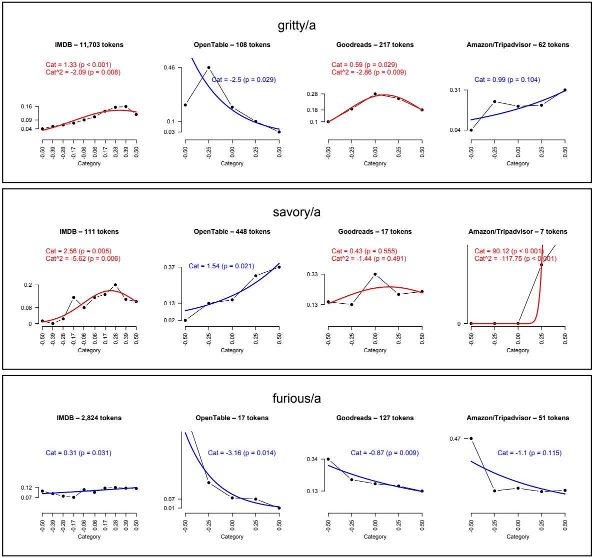
Example: attenuators







Corpus-level effects (and lack thereof)



(true domain effect)

(frequency effect)

(movies: *The Fast and the Furious 1-5*)

Hierarchical logistic regression

Cat.	Count	Total	Corpus
−0.50	4	699,695	OpenTable
−0.25	4	2,507,147	OpenTable
0.00	3	4,207,700	OpenTable
0.25	5	7,789,649	OpenTable
0.50	1	8,266,564	OpenTable
−0.50	13	3,419,923	Amazon
−0.25	4	3,912,625	Amazon
0.00	7	6,011,388	Amazon
0.25	10	10,187,257	Amazon
0.50	17	16,202,230	Amazon
−0.50	22	3,419,923	Goodreads
−0.25	15	3,912,625	Goodreads
0.00	20	6,011,388	Goodreads
0.25	31	10,187,257	Goodreads
0.50	39	16,202,230	Goodreads
−0.50	212	28,962,201	IMDB
−0.39	85	13,436,851	IMDB
−0.28	86	15,987,151	IMDB
−0.17	84	17,095,212	IMDB
−0.06	183	23,293,790	IMDB
0.06	214	31,317,918	IMDB
0.17	388	45,913,948	IMDB
0.28	483	55,634,817	IMDB
0.39	387	45,941,763	IMDB
0.50	702	84,294,625	IMDB

Table: furious/a

Linear model allowing intercept and slope to vary by corpus:

$$\text{logit}^{-1}\left(\begin{array}{r}\texttt{intercept} + \\ \texttt{category} + \\ (1+\texttt{category}|\texttt{corpus})\end{array}\right)$$

Quadratic model allowing intercept, slope, and curve to vary by corpus:

$$\text{logit}^{-1}\left(\begin{array}{r}\texttt{intercept} + \\ \texttt{category} + \\ (1+\texttt{category}|\texttt{corpus}) + \\ (0+\texttt{category}^2|\texttt{corpus})\end{array}\right)$$

Hierarchical logistic regression

Cat.	Count	Total	Corpus
−0.50	4	699,695	OpenTable
−0.25	4	2,507,147	OpenTable
0.00	3	4,207,700	OpenTable
0.25	5	7,789,649	OpenTable
0.50	1	8,266,564	OpenTable
−0.50	13	3,419,923	Amazon
−0.25	4	3,912,625	Amazon
0.00	7	6,011,388	Amazon
0.25	10	10,187,257	Amazon
0.50	17	16,202,230	Amazon
−0.50	22	3,419,923	Goodreads
−0.25	15	3,912,625	Goodreads
0.00	20	6,011,388	Goodreads
0.25	31	10,187,257	Goodreads
0.50	39	16,202,230	Goodreads
−0.50	212	28,962,201	IMDB
−0.39	85	13,436,851	IMDB
−0.28	86	15,987,151	IMDB
−0.17	84	17,095,212	IMDB
−0.06	183	23,293,790	IMDB
0.06	214	31,317,918	IMDB
0.17	388	45,913,948	IMDB
0.28	483	55,634,817	IMDB
0.39	387	45,941,763	IMDB
0.50	702	84,294,625	IMDB

Table: furious/a

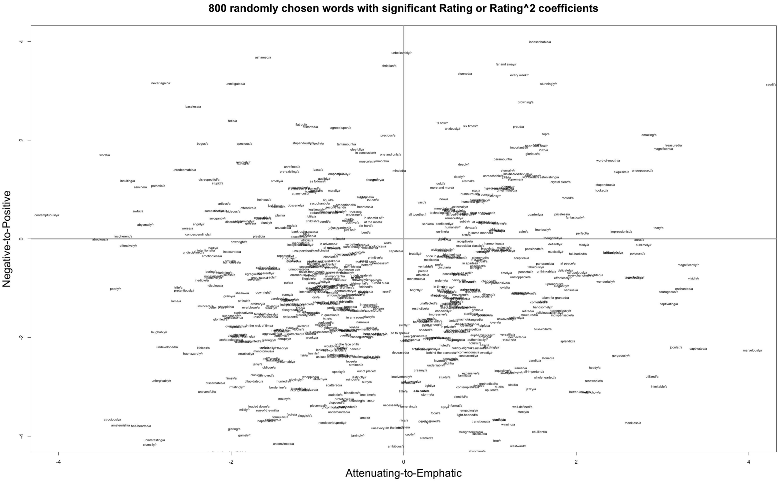
Linear model allowing intercept and slope to vary by corpus:

$$\text{logit}^{-1}\left(\begin{array}{r} \texttt{intercept} + \\ \texttt{category} + \\ (1+\texttt{category}|\texttt{corpus}) \end{array}\right)$$

Fixed	Coef. est.	*p*-value
intercept	−12.91	< 0.001
category	−1.06	0.037

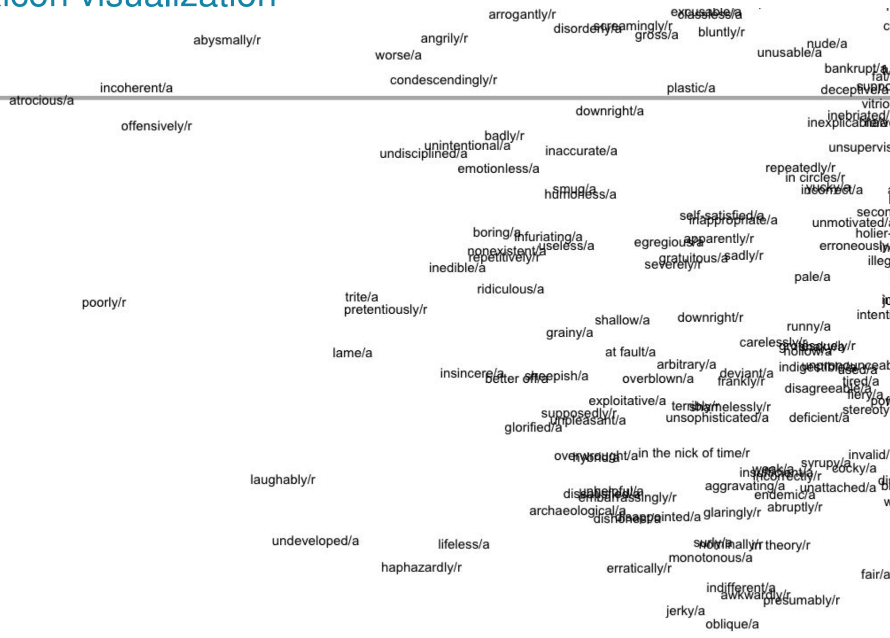
Random	Intercept	Category
Amazon	−13.31	−1.57
Goodreads	−12.58	−0.66
IMDB	−11.81	0.31
OpenTable	−13.88	−2.29

Lexicon visualization



Lexicon visualization

Lexicon visualization



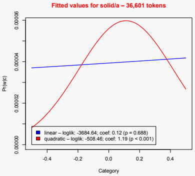
Categorization experiments

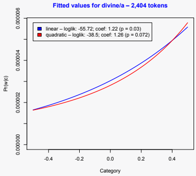
1. Data: user-supplied product and service reviews
2. Methods: hierarchical logistic regresssion
3. Evaluation: classification and scale induction against gold-standard lexicons
4. Looking ahead: alternative approaches and general issues

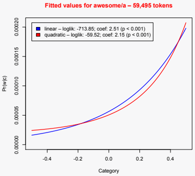
Polarity categorization

Choose the significant model ($p < 0.05$) with the greater log-likelihood. (If there is no such model, the word is neutral.)

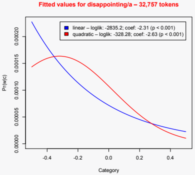
Positive: positive linear coef. or turning point coef.

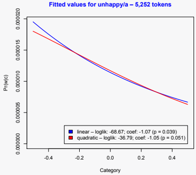


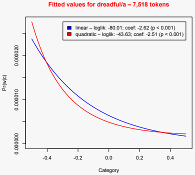




Negative: negative linear coef. or turning point coef.







Adjective category random samples

Positive	
absolute	intangible
acrobatic	invaluable
adventurous	legendary
alluring	little-known
amazing	nice
belgian	outside
brisk	personable
ceremonial	quiet
colourful	romantic
controversial	sociable
delectable	soulful
dignified	stinging
diligent	sublime
earthly	twenty-eight
finicky	unrivaled
first-class	unyielding
glorious	weathered
high-energy	well-made
idyllic	worldly
injured	youthful

Negative	
accented	one and only
angered	paper thin
copious	passable
defensible	perky
discernible	presumptuous
disjointed	pretty
disorganized	problematic
downright	psychedelic
entire	ripe
faux	shabby
flimsy	stiff
glowing	tame
gratuitous	thin
inauthentic	torturous
inconceivable	unhappy
low	unneeded
lucrative	unrefined
melted	wet
mercenary	whiney
merry	would-be

Adverb category random samples

Positive	
all in all	instantly
and how	intensively
authentically	joyously
closer	lately
complexly	now and then
daringly	out of sight
darkly	refreshingly
elegantly	remarkably
enjoyably	six times
far and wide	soon
for keeps	sublimely
harmoniously	the right way
heart and soul	thus far
high up	to perfection
in a flash	undeniably
in good time	uniquely
in hiding	unknowingly
in particular	unusually
in the lead	very fast
inside	yet

Negative	
abysmally	haphazardly
ad nauseam	harshly
also known as	in the nick of time
angrily	intolerably
arbitrarily	knowingly
arrogantly	lazily
asleep	liberally
at the worst	massively
clear	maybe
decidedly	more often than not
downright	ostensibly
earlier	other than
en masse	piecemeal
even	profusely
first and last	properly
flat out	repetitively
flatly	satisfactorily
frightfully	under that
gravely	unforgivably
grossly	upstairs

Experimental set-up

1. Underlying data: The IMDB, OpenTable, Goodreads, and Amazon/Tripadvisor corpora described earlier
2. Vocabulary: the 5,801 adjectives and adverbs derived from WordNet lemmas and appearing in all four corpora
3. Scores: derived using the categorization scheme just described, with threshold $p < 0.05$
4. Assessment: Comparisons with gold-standard or near-gold-standard sentiment lexicons *always over the intersection of the sentiment lexicon in question with the vocabulary in* 2.

Harvard Inquirer

	Entry	Positiv	Negativ	Hostile	. . . (184 classes)	Othtags	Defined
1	A					DET ART	. . .
2	ABANDON		Negativ			SUPV	
3	ABANDONMENT		Negativ			Noun	
4	ABATE		Negativ			SUPV	
5	ABATEMENT					Noun	
⋮							
35	ABSENT#1		Negativ			Modif	
36	ABSENT#2					SUPV	
⋮							
11788	ZONE					Noun	

Table: '#*n*' differentiates senses. Binary category values: 'Yes' = category name; 'No' = blank. Heuristic mapping from Othtags into {a,n,r,v}.

Positiv: 1,915 words   Negativ: 2,291   (classes disjoint)

Harvard Inquirer

Absence of Positiv/Negativ is arguably not absence of polarity. For example, the following words all fall outside both classes:

accidental/a	greatest/a	rugged/a
afraid/a	idle/a	sharp/a
ambitious/a	implausible/a	strenuous/a
beautiful/a	inexcusable/a	unchecked/a
brutal/a	intense/a	unprecedented/a
convenient/a	jealous/a	unrealistic/a
drastic/a	joyous/a	utterly/r
embarrassed/a	massive/a	vigorous/a
exceptional/a	persistent/a	weak/a
favourable/a	preferable/a	well-informed/a
futile/a	ragged/a	zealous/a

Thus, I assess only against words that are in Positiv/Negativ.

Harvard Inquirer categorization

	Predicted	
Inquirer	positive	negative
Positiv	241	49
Negativ	58	241

Table: Confusion matrix. Accuracy: 82%.

	Precision	Recall
Positive	0.81	0.83
Negative	0.83	0.81

Table: Effectiveness

20 randomly selected errors

Word	Inquirer	Predicted
actual/a	Positiv	negative
coherent/a	Positiv	negative
benign/a	Positiv	negative
competent/a	Positiv	negative
capable/a	Positiv	negative
colossal/a	Positiv	negative
complete/a	Positiv	negative
better/r	Positiv	negative
clear/r	Positiv	negative
charitable/a	Positiv	negative
constructive/a	Positiv	negative
credible/a	Positiv	negative
acceptable/a	Positiv	negative
austere/a	Negativ	positive
complex/a	Negativ	positive
audacious/a	Negativ	positive
competitive/a	Negativ	positive
bleak/a	Negativ	positive
apprehensive/a	Negativ	positive
brittle/a	Negativ	positive

MPQA subjectivity lexicon

`http://www.cs.pitt.edu/mpqa/`

1.	type=weaksubj	len=1	word1=abandoned	pos1=adj	stemmed1=n	priorpolarity=negative
2.	type=weaksubj	len=1	word1=abandonment	pos1=noun	stemmed1=n	priorpolarity=negative
3.	type=weaksubj	len=1	word1=abandon	pos1=verb	stemmed1=y	priorpolarity=negative
4.	type=strongsubj	len=1	word1=abase	pos1=verb	stemmed1=y	priorpolarity=negative
5.	type=strongsubj	len=1	word1=abasement	pos1=anypos	stemmed1=y	priorpolarity=negative
6.	type=strongsubj	len=1	word1=abash	pos1=verb	stemmed1=y	priorpolarity=negative
7.	type=weaksubj	len=1	word1=abate	pos1=verb	stemmed1=y	priorpolarity=negative
8.	type=weaksubj	len=1	word1=abdicate	pos1=verb	stemmed1=y	priorpolarity=negative
9.	type=strongsubj	len=1	word1=aberration	pos1=adj	stemmed1=n	priorpolarity=negative
10.	type=strongsubj	len=1	word1=aberration	pos1=noun	stemmed1=n	priorpolarity=negative
11.	type=strongsubj	len=1	word1=abhor	pos1=anypos	stemmed1=y	priorpolarity=negative
12.	type=strongsubj	len=1	word1=abhor	pos1=verb	stemmed1=y	priorpolarity=negative
13.	type=strongsubj	len=1	word1=abhorred	pos1=adj	stemmed1=n	priorpolarity=negative
14.	type=strongsubj	len=1	word1=abhorrence	pos1=noun	stemmed1=n	priorpolarity=negative
15.	type=strongsubj	len=1	word1=abhorrent	pos1=adj	stemmed1=n	priorpolarity=negative
16.	type=strongsubj	len=1	word1=abhorrently	pos1=anypos	stemmed1=n	priorpolarity=negative
17.	type=strongsubj	len=1	word1=abhors	pos1=adj	stemmed1=n	priorpolarity=negative
18.	type=strongsubj	len=1	word1=abhors	pos1=noun	stemmed1=n	priorpolarity=negative
19.	type=strongsubj	len=1	word1=abidance	pos1=adj	stemmed1=n	priorpolarity=positive
20.	type=strongsubj	len=1	word1=abidance	pos1=noun	stemmed1=n	priorpolarity=positive
⋮						
8221.	type=strongsubj	len=1	word1=zest	pos1=noun	stemmed1=n	priorpolarity=positive

MPQA classification

	Predicted		
MQAP	positive	negative	neutral
positive	456	96	394
negative	140	540	394
neutral	35	36	98

Table: Confusion matrix. Accuracy: 50%. Pos/neg accuracy: 81%

	Precision	Recall
positive	0.72	0.48
negative	0.80	0.50
neutral	0.11	0.58

Table: Effectiveness

20 randomly selected errors		
Word	MPQA	Predicted
advantageous/a	positive	neutral
advanced/a	positive	neutral
acceptable/a	positive	negative
above/r	positive	neutral
active/a	positive	neutral
admittedly/r	positive	negative
above/a	positive	negative
accurate/a	positive	neutral
admirable/a	positive	neutral
adequate/a	positive	neutral
abundant/a	positive	neutral
affected/a	neutral	neutral
actually/r	neutral	negative
actual/a	neutral	negative
absolute/a	neutral	positive
agonizing/a	negative	neutral
aggressive/a	negative	neutral
adamantly/r	negative	neutral
accidental/a	negative	positive
abnormal/a	negative	neutral

Micro-WNOp

Documentation and download: `http://www.unipv.it/wnop`

	Pos	Neg	Synset
1	1	0	true·a·2 real·a·4
2	1	0	illustrious·a·1 famous·a·1 . . .
3	0.5	0	real·a·6 tangible·a·2
4	0.25	0	existent·a·2 real·a·1
5	0.125	0.125	real·a·2
⋮			
110	0	0	demand·v·6

Table: ‘Common’: Five evaluators working together, 110 synsets.

Micro-WNOp

Documentation and download: `http://www.unipv.it/wnop`

	Evaluator 1		Evaluator 2		Evaluator 3		
	Pos1	Neg1	Pos2	Neg2	Pos3	Neg3	Synset
1	1	0	1	0	1	0	good·a·15 well·a·2
2	1	0	1	0	0.75	0	sweet-smelling·a·1 perfumed·a·2 . . .
3	1	0	1	0	1	0	good·a·23 unspoilt·a·1 unspoiled·a·1
4	0.5	0	0.25	0	0.25	0	hot·a·16
⋮							
496	0.5	0	1	0	0.5	0	heal·v·3 bring_around·v·2 cure·v·1

Table: ‘Group 1’: Three evaluators working separately, 496 synsets. Complete agreement on 197 (40%). Polarity agreement (sign(Pos1−Neg1) = sign(Pos2−Neg2) = sign(Pos3−Neg3)) on 387 (78%).

Micro-WNOp

Documentation and download: `http://www.unipv.it/wnop`

	Evaluator 1		Evaluator 2		
	Pos1	Neg1	Pos2	Neg2	Synset
1	0	1	0	1	forlorn·a·1 godforsaken·a·2 lorn·a·1 desolate·a·2
2	0	1	0	1	rotten·a·2
3	1	0	1	0	intimate·a·2 cozy·a·2 informal·a·4
4	0	0	0	0	federal·a·1
⋮					
499	0	0	0	0	term·v·1

Table: 'Group 2': Two evaluators working separately, 499 synsets. Complete agreement on 395 (79%). Polarity agreement (sign(Pos1−Neg1) = sign(Pos2−Neg2)) on 471 (90.4%).

Micro-WNOp

Documentation and download: `http://www.unipv.it/wnop`

- Limit attention to the 702 items on which all annotators agree on all values (intersection with our vocab: 175)
- Positive: positive score was higher
- Negative: negative score was higher
- Objective: the two scores were the same (even if > 0)

Micro-WNOp classification

MQAP	Predicted positive	Predicted negative	Predicted neutral
positive	32	6	31
negative	8	29	19
neutral	15	9	26

Table: Confusion matrix. Accuracy: 50%. Pos/neg accuracy: 81%

	Precision	Recall
positive	0.58	0.46
negative	0.66	0.52
neutral	0.34	0.52

Table: Effectiveness

20 randomly selected errors

Word	Micro-WNOp	Predicted
distinctive/a	neutral	positive
desolate/a	negative	positive
criminal/a	negative	neutral
beneficial/a	positive	neutral
cynical/a	negative	neutral
elastic/a	positive	neutral
adverse/a	negative	neutral
actual/a	neutral	negative
capable/a	positive	negative
diplomatic/a	positive	neutral
collectively/r	neutral	positive
desirable/a	positive	neutral
able/a	neutral	positive
angelic/a	positive	neutral
amiable/a	positive	neutral
celebrated/a	positive	neutral
despicable/a	negative	neutral
able-bodied/a	positive	neutral
benevolent/a	positive	neutral
chiefly/r	neutral	negative

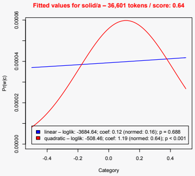
Scale induction experiments

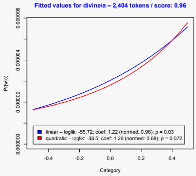
1. Data: user-supplied product and service reviews
2. Methods: hierarchical logistic regresssion
3. Evaluation: classification and scale induction against gold-standard lexicons
4. Looking ahead: alternative approaches and general issues

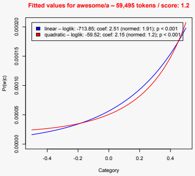
Scale induction, informal version

Challenge: linear and quadratic coefficients are not on the same scale. Solution: z-score normalize them against their populations

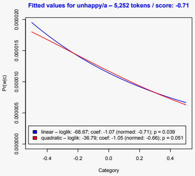
Positive: positive linear coef. or turning point coef.

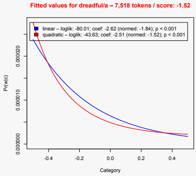


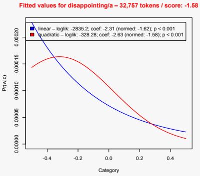




Negative: negative linear coef. or turning point coef.







Adjective scales induced by the informal version

Positive	
unturned/a	6.90
unhurried/a	4.72
melodious/a	4.28
peerless/a	3.80
fledged/a	3.75
small-scale/a	3.20
saudi/a	3.18
geometric/a	3.11
symphonic/a	3.03
sterling/a	2.96
unforgettable/a	2.81
indefatigable/a	2.75
stately/a	2.66
skittish/a	2.66
captivated/a	2.57
unobtrusive/a	2.53
unsung/a	2.50
thankless/a	2.47
jocular/a	2.46
enchanted/a	2.45
⋮	

Negative	
unrewarding/a	-3.47
flowery/a	-3.34
god-awful/a	-2.71
redeeming/a	-2.63
insipid/a	-2.58
atrocious/a	-2.50
slapdash/a	-2.38
amateurish/a	-2.33
uninspiring/a	-2.33
incoherent/a	-2.30
lowbrow/a	-2.28
incompetent/a	-2.22
unprofessional/a	-2.20
unimaginative/a	-2.19
awful/a	-2.18
uninspired/a	-2.17
inane/a	-2.16
half-hearted/a	-2.13
acting/a	-2.12
laughable/a	-2.10
⋮	

Adverb scales induced by the informal version

Positive	
daringly/r	5.44
skilfully/r	4.24
craftily/r	4.18
marvelously/r	3.00
splendidly/r	2.65
capably/r	2.63
loyally/r	2.55
magnificently/r	2.48
pleasantly/r	2.41
expertly/r	2.40
comprehensively/r	2.23
for keeps/r	2.23
steadfastly/r	2.22
breezily/r	2.20
sublimely/r	2.10
flawlessly/r	2.09
for all practical purposes/r	2.08
par excellence/r	2.07
vividly/r	2.06
best of all/r	2.05
⋮	

Positive	
promisingly/r	-4.17
incompetently/r	-3.50
dismally/r	-3.17
contemptuously/r	-2.95
appallingly/r	-2.39
atrociously/r	-2.39
poorly/r	-2.37
in name only/r	-2.30
offensively/r	-2.26
blankly/r	-2.21
lamely/r	-2.21
interminably/r	-2.21
clumsily/r	-2.12
abysmally/r	-2.11
unceremoniously/r	-2.11
insultingly/r	-2.09
inexcusably/r	-2.03
sloppily/r	-1.94
unforgivably/r	-1.90
at best/r	-1.90
⋮	

Systematic comparison

	Cat.	Count	Total	Corpus	Stronger
	−0.50	1,191	699,695	OpenTable	0
	−0.25	6,601	2,507,147	OpenTable	0
	0.00	19,151	4,207,700	OpenTable	0
	0.25	69,395	7,789,649	OpenTable	0
great/a	0.50	81,932	8,266,564	OpenTable	0
	−0.50	1,118	3,419,923	Amazon	0
	−0.25	1,758	3,912,625	Amazon	0
			⋮		
	0.50	197,461	84,294,625	IMDB	0
	−0.50	445	699,695	OpenTable	1
	−0.25	2,064	2,507,147	OpenTable	1
	0.00	8,362	4,207,700	OpenTable	1
	0.25	38,771	7,789,649	OpenTable	1
excellent/a	0.50	49,081	8,266,564	OpenTable	1
	−0.50	122	3,419,923	Amazon	1
	−0.25	226	3,912,625	Amazon	1
			⋮		
	0.50	48,160	84,294,625	IMDB	1

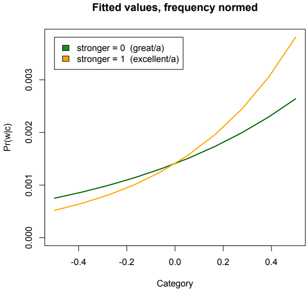
Table: great/a and excellent/a

$$\text{logit}^{-1}\left(\begin{array}{r}\texttt{intercept}\ +\\ \texttt{rating} * \texttt{stronger}\ +\\ (1{+}\texttt{rating}|\texttt{corpus})\end{array}\right)$$

Systematic comparison

Fixed	Coef. est.	*p*-value	Gloss
`intercept`	−6.56	< 0.001	
`category`	1.26	< 0.001	*positivity*
`stronger`	−1.46	< 0.001	*'excellent' less frequent*
`category*stronger`	0.74	< 0.001	*'excellent' pos. than 'great'*

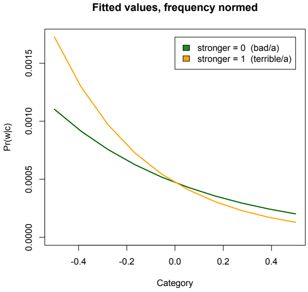
Table: great/a and excellent/a



Systematic comparison

Fixed	Coef. est.	p-value	Gloss
`intercept`	7.66	< 0.001	
`category`	-1.70	< 0.001	*positivity*
`stronger`	-1.92	< 0.001	*'terrible' less frequent*
`category*stronger`	-0.90	< 0.001	*'terrible' more neg. than 'bad'*

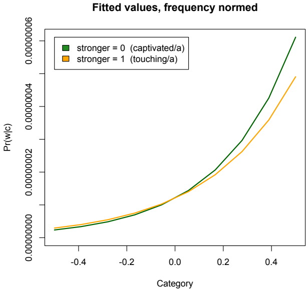
Table: bad/a and terrible/a



Systematic comparison

Fixed	Coef. est.	*p*-value	Gloss
`intercept`	−18.24	< 0.001	
`category`	3.26	< 0.001	*positivity*
`stronger`	4.85	< 0.001	*'touching' more frequent*
`category*stronger`	0.44	0.385	*no solid ordering inferrable*

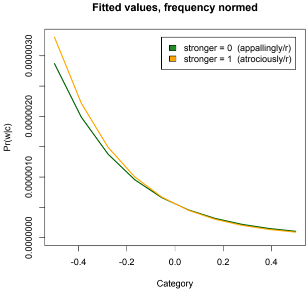
Table: captivated/a and touching/a



Systematic comparison

Fixed	Coef. est.	p-value	Gloss
`intercept`	−14.41	< 0.001	
`category`	−3.30	< 0.001	*negativity*
`stronger`	−0.59	< 0.001	*'atrociously' less frequent*
`category*stronger`	−0.28	0.156	*no solid ordering inferrable*

Table: appallingly/r and atrociously/r



Scale induction, formal version

To compare two words w_1 and w_2:

Step 1: Fit the model

Randomly select one of the words w_i to call 'stronger'.

Step 2: Inspect the `stronger` coefficient

- If its $p >$ threshold, the words are of equal strength;
- else if sign(`stronger`) = sign(`category`), then w_i is the stronger of the two;
- else w_i is the weaker of the two.

Drawback

Computationally intensive. However, the orderings are transitive and asymmetric, greatly reducing the space of pairs to test.

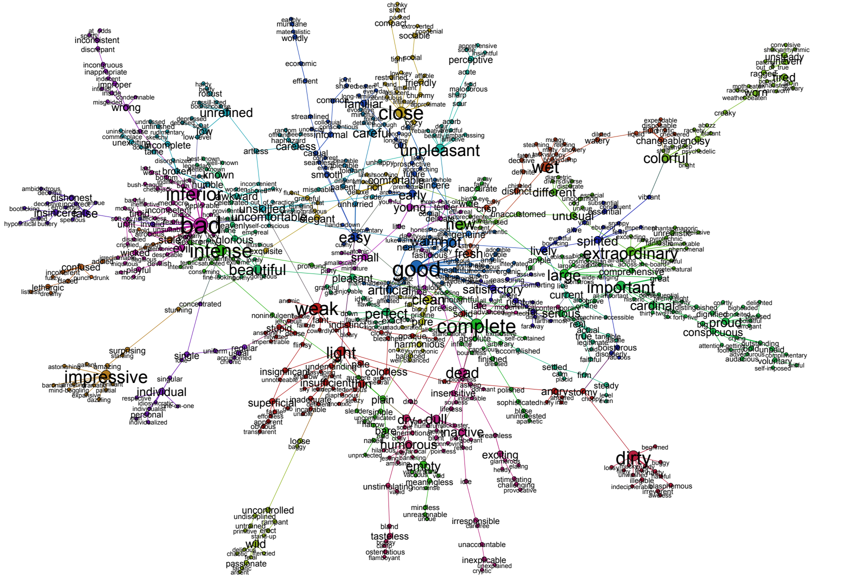
Incomparables

It's a mistake to assume that we can partially order all modifiers along a single dimension; asking whether the following orderings are correct or not seems mistaken:

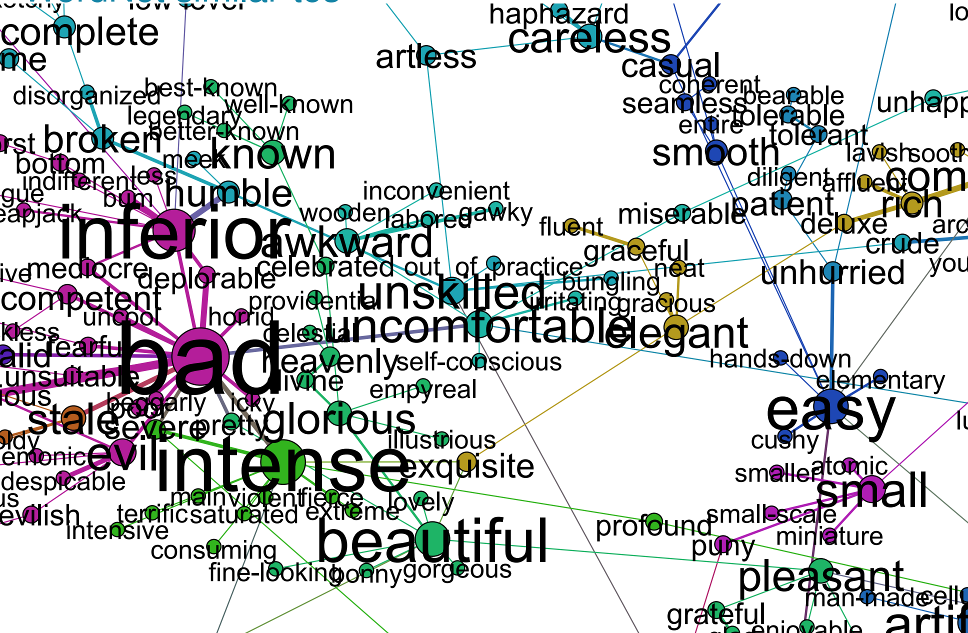
symphonic/a	3.03	amazing/a	1.22	poorly/r	−2.37
delicious/a	1.25	funny/a	0.02	offensively/a	−2.26

Scales are meaningful only internal to semantically coherent groups of words.

WordNet similar-tos



WordNet similar-tos



WordNet similar-tos

Experimental set-up

1. From (word, tag) pairs to synsets:

	synsets	similar-tos	(word, tag) pairs
		severe.s.06	(severe, a)
		naughty.s.02	(naughty, a)
	bad.a.01	corked.s.01	(corked, a)
	bad.s.02	poor.s.06	(poor, a)
(bad, a)	bad.s.03	icky.s.01	(icky, a)
	⋮	intense.a.01	(intense, a)
		uncomfortable.a.02	(uncomfortable, a)
		⋮	⋮

Experimental set-up

1. From (word, tag) pairs to synsets:

	synsets	similar-tos	(word, tag) pairs
		severe.s.06	(severe, a)
		naughty.s.02	(naughty, a)
	bad.a.01	corked.s.01	(corked, a)
	bad.s.02	poor.s.06	(poor, a)
(bad, a)	bad.s.03	icky.s.01	(icky, a)
	⋮	intense.a.01	(intense, a)
		uncomfortable.a.02	(uncomfortable, a)
		⋮	⋮

2. Restrict attention to words in the gold-standard lexicon and induce pairwise orderings:

Gold		
bad/a	$=$	severe/a
bad/a	$>$	poor/a
bad/a	$=$	deplorable/a
bad/a	$=$	atrocious/a
	⋮	

Experimental set-up

1. From (word, tag) pairs to synsets:

	synsets	similar-tos	(word, tag) pairs
		severe.s.06	(severe, a)
		naughty.s.02	(naughty, a)
	bad.a.01	corked.s.01	(corked, a)
	bad.s.02	poor.s.06	(poor, a)
(bad, a)	bad.s.03	icky.s.01	(icky, a)
	⋮	intense.a.01	(intense, a)
		uncomfortable.a.02	(uncomfortable, a)
		⋮	⋮

2. Restrict attention to words in the gold-standard lexicon and induce pairwise orderings:

Gold			Normed score cmp	Cat. coef	Cmp coef.
bad/a	$=$	severe/a	*abs*$(-1.08) >$ *abs*(-0.32)	-1.57	1.18 $(>)$
bad/a	$>$	poor/a	*abs*$(-1.08) <$ *abs*(-1.53)	-1.75	-0.11 $(<)$
bad/a	$=$	deplorable/a	*abs*$(-1.08) <$ *abs*(-1.93)	-1.58	-0.40 $(<)$
bad/a	$=$	atrocious/a	*abs*$(-1.08) <$ *abs*(-2.50)	-1.57	-1.48 $(<)$
	⋮		⋮		

3. Predictions.

MPQA scales

`http://www.cs.pitt.edu/mpqa/`

1.	type=weaksubj	len=1	word1=abandoned	pos1=adj	stemmed1=n	priorpolarity=negative
2.	type=weaksubj	len=1	word1=abandonment	pos1=noun	stemmed1=n	priorpolarity=negative
3.	type=weaksubj	len=1	word1=abandon	pos1=verb	stemmed1=y	priorpolarity=negative
4.	type=strongsubj	len=1	word1=abase	pos1=verb	stemmed1=y	priorpolarity=negative
5.	type=strongsubj	len=1	word1=abasement	pos1=anypos	stemmed1=y	priorpolarity=negative
6.	type=strongsubj	len=1	word1=abash	pos1=verb	stemmed1=y	priorpolarity=negative
7.	type=weaksubj	len=1	word1=abate	pos1=verb	stemmed1=y	priorpolarity=negative
8.	type=weaksubj	len=1	word1=abdicate	pos1=verb	stemmed1=y	priorpolarity=negative
9.	type=strongsubj	len=1	word1=aberration	pos1=adj	stemmed1=n	priorpolarity=negative
10.	type=strongsubj	len=1	word1=aberration	pos1=noun	stemmed1=n	priorpolarity=negative
11.	type=strongsubj	len=1	word1=abhor	pos1=anypos	stemmed1=y	priorpolarity=negative
12.	type=strongsubj	len=1	word1=abhor	pos1=verb	stemmed1=y	priorpolarity=negative
13.	type=strongsubj	len=1	word1=abhorred	pos1=adj	stemmed1=n	priorpolarity=negative
14.	type=strongsubj	len=1	word1=abhorrence	pos1=noun	stemmed1=n	priorpolarity=negative
15.	type=strongsubj	len=1	word1=abhorrent	pos1=adj	stemmed1=n	priorpolarity=negative
16.	type=strongsubj	len=1	word1=abhorrently	pos1=anypos	stemmed1=n	priorpolarity=negative
17.	type=strongsubj	len=1	word1=abhors	pos1=adj	stemmed1=n	priorpolarity=negative
18.	type=strongsubj	len=1	word1=abhors	pos1=noun	stemmed1=n	priorpolarity=negative
19.	type=strongsubj	len=1	word1=abidance	pos1=adj	stemmed1=n	priorpolarity=positive
20.	type=strongsubj	len=1	word1=abidance	pos1=noun	stemmed1=n	priorpolarity=positive
⋮						
8221.	type=strongsubj	len=1	word1=zest	pos1=noun	stemmed1=n	priorpolarity=positive

MPQA scale prediction: informal method

	Predicted		
MPQA	stronger	weaker	same
stronger	98	64	2
weaker	43	103	2
same	277	262	27

Table: Overall accuracy: 26%. Stronger/weaker accuracy: 65%.

Criticism

Two non-0 scores are almost never the same, and this method cannot assess whether two scores are genuinely different, so it flounders on ‘same’. The problem is especially serious since ‘same’ is the largest MPQA category.

MPQA scale prediction: formal method

	Predicted		
MPQA	stronger	weaker	same
stronger	27	67	58
weaker	47	33	55
same	145	216	163

Table: Overall accuracy: 27%. Stronger/weaker accuracy: 34%. The two-step MPQA scale is much coarser than our predictions.

	Precision	Recall
stronger	0.12	0.18
weaker	0.10	0.24
same	0.59	0.31

Table: Effectiveness

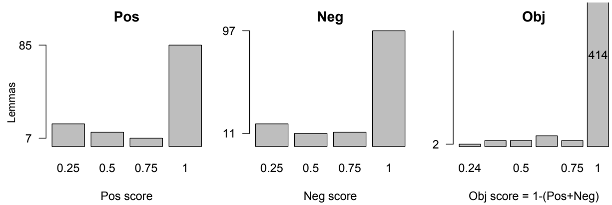
20 randomly selected errors			
Word	SimWord	MQAP	Pred.
foolish/a	asinine/a	$=$	$<$
unnatural/a	artificial/a	$=$	$>$
senseless/a	stupid/a	$=$	$<$
flimsy/a	weak/a	$=$	$>$
feeble/a	weak/a	$=$	$>$
low/a	insufficient/a	$=$	$<$
stupid/a	confused/a	$=$	$>$
marginal/a	narrow/a	$=$	$>$
mean/a	nasty/a	$=$	$<$
primary/a	particular/a	$=$	$>$
dreadful/a	unpleasant/a	$=$	$>$
pathetic/a	unfortunate/a	$=$	$>$
harmonious/a	balanced/a	$>$	$=$
clumsy/a	unskilled/a	$>$	$=$
false/a	incorrect/a	$>$	$=$
incorrect/a	erroneous/a	$<$	$=$
colorful/a	vibrant/a	$<$	$=$
careful/a	mindful/a	$<$	$=$
refined/a	gracious/a	$<$	$=$
sunny/a	cheerful/a	$<$	$=$

Micro-WNOp scales

Unfortunately, Micro-WNOp contains data on only 5 pairs that are related by similar-to relations in the way described above:

Word	SimWord	Polarity	WordStrength	SimStrength
cordial/a	sincere/a	positive	1	1
good/a	solid/a	positive	1	1
sincere/a	cordial/a	positive	1	1
solid/a	good/a	positive	1	1
true/a	sincere/a	positive	1	1

This score distribution if not atypical; the majority of Micro-WNOp scores are either 0 or 1:



Looking ahead

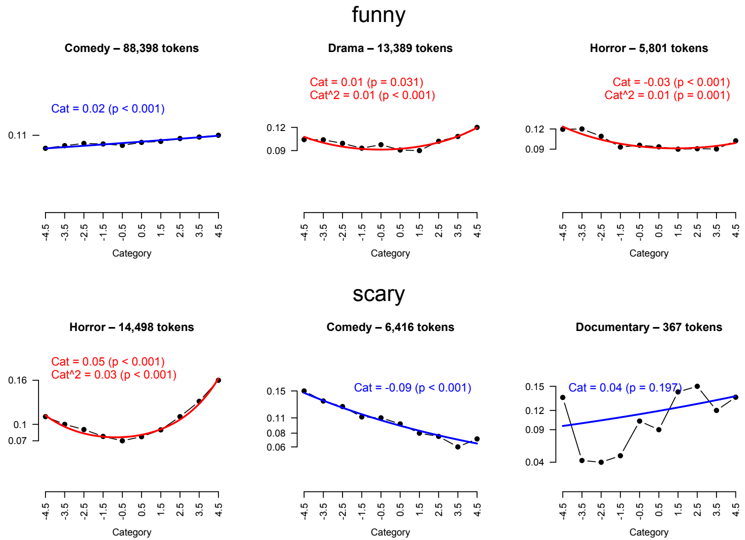
1. Data: user-supplied product and service reviews
2. Methods: hierarchical logistic regresssion
3. Evaluation: classification and scale induction against gold-standard lexicons
4. Looking ahead: alternative approaches and general issues

Discussion

- The categorization experiments provide solid assessment information.
- This is less clear for the scale induction experiments, where the gold standard resources are far more coarse-grained than the current approach.

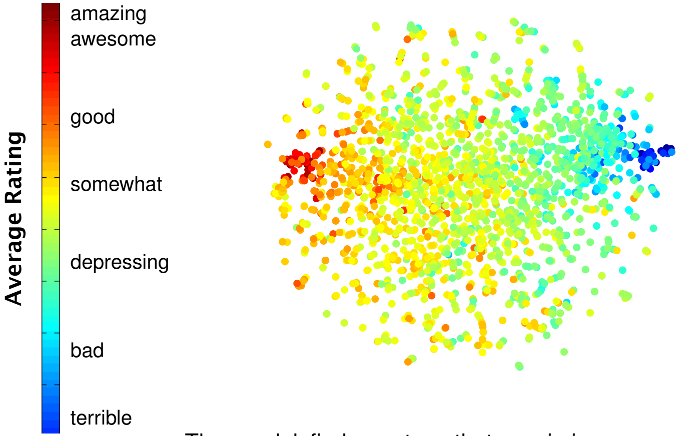
Context-dependence

Profiles varying by genre in the IMDB:



Polarity-rich word vectors

Simultaneously learning coherent adjective groups and scalar orderings:



The model finds vectors that maximize unsupervised log-likelihood and supervised prediction accuracy of star ratings.

Joint work with Andrew Maas and Andrew Ng.

In closing

Summary

1. Data: user-supplied product and service reviews
2. Methods: hierarchical logistic regresssion
3. Evaluation: classification and scale induction against gold-standard lexicons

http://www.stanford.edu/~cgpotts/data/wordnetscales/

Next steps

- Contextual variability
- Incomparability (semantically coherent word clusters)
- Improved assessment metrics