Sentiment Analysis

Learning Sentiment Lexicons

Semi-supervised learning of lexicons

- Use a small amount of information
  - A few labeled examples
  - A few hand-built patterns
- To bootstrap a lexicon

Hatzivassiloglou and McKeown intuition for identifying word polarity

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181

- Adjectives conjoined by "*and*" have same polarity
  - Fair **and** legitimate, corrupt **and** brutal
  - *fair **and** brutal, *corrupt **and** legitimate
- Adjectives conjoined by "*but*" do not
  - fair **but** brutal

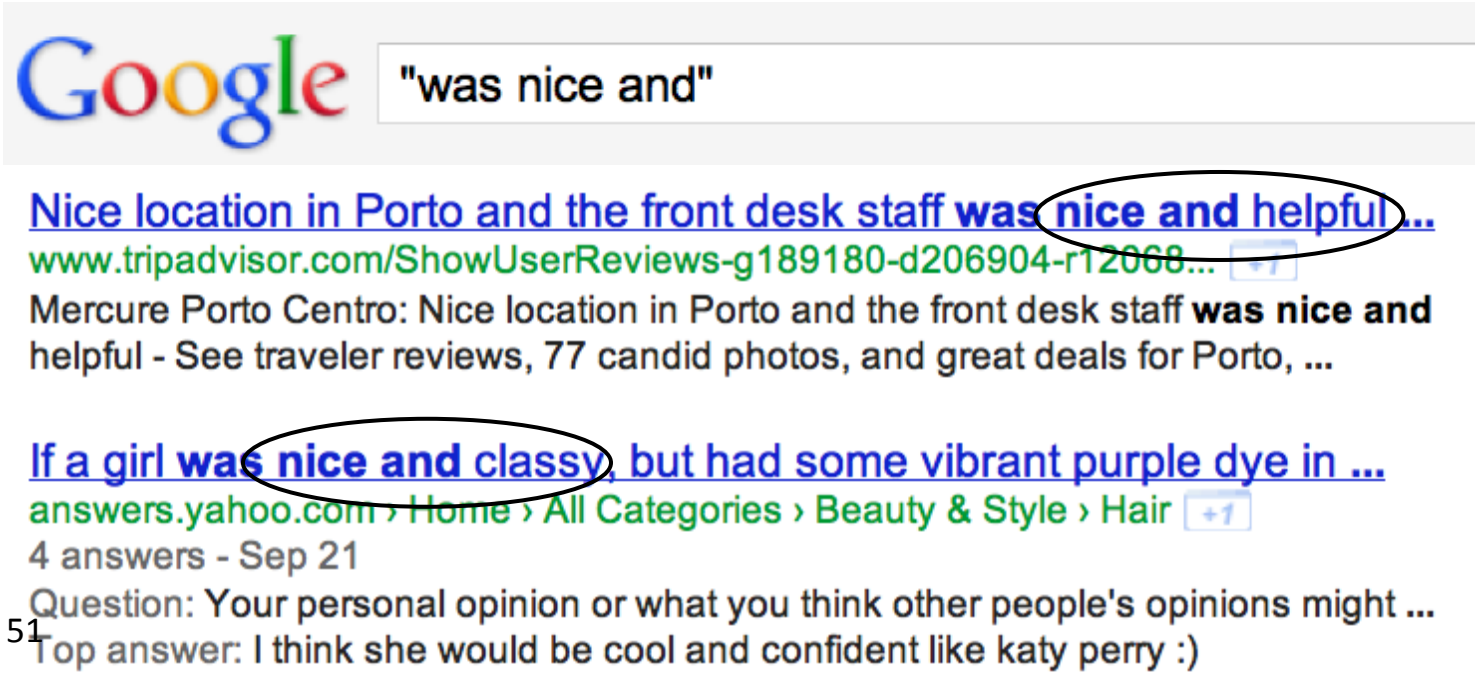
Hatzivassiloglou & McKeown 1997 Step 1

- Label **seed set** of 1336 adjectives (all >20 in 21 million word WSJ corpus)
  - 657 positive
    - adequate central clever famous intelligent remarkable reputed sensitive slender thriving…
  - 679 negative
    - contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting…

Dan Jurafsky

Hatzivassiloglou & McKeown 1997 Step 2

- Expand seed set to conjoined adjectives

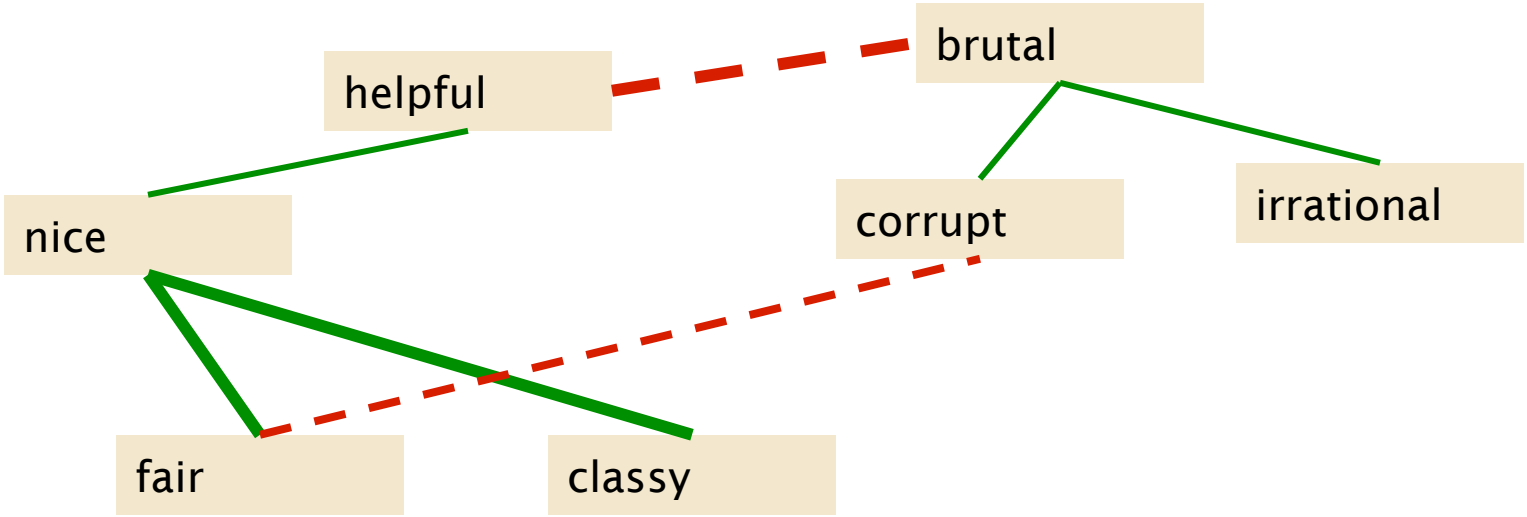


nice, helpful

nice, classy

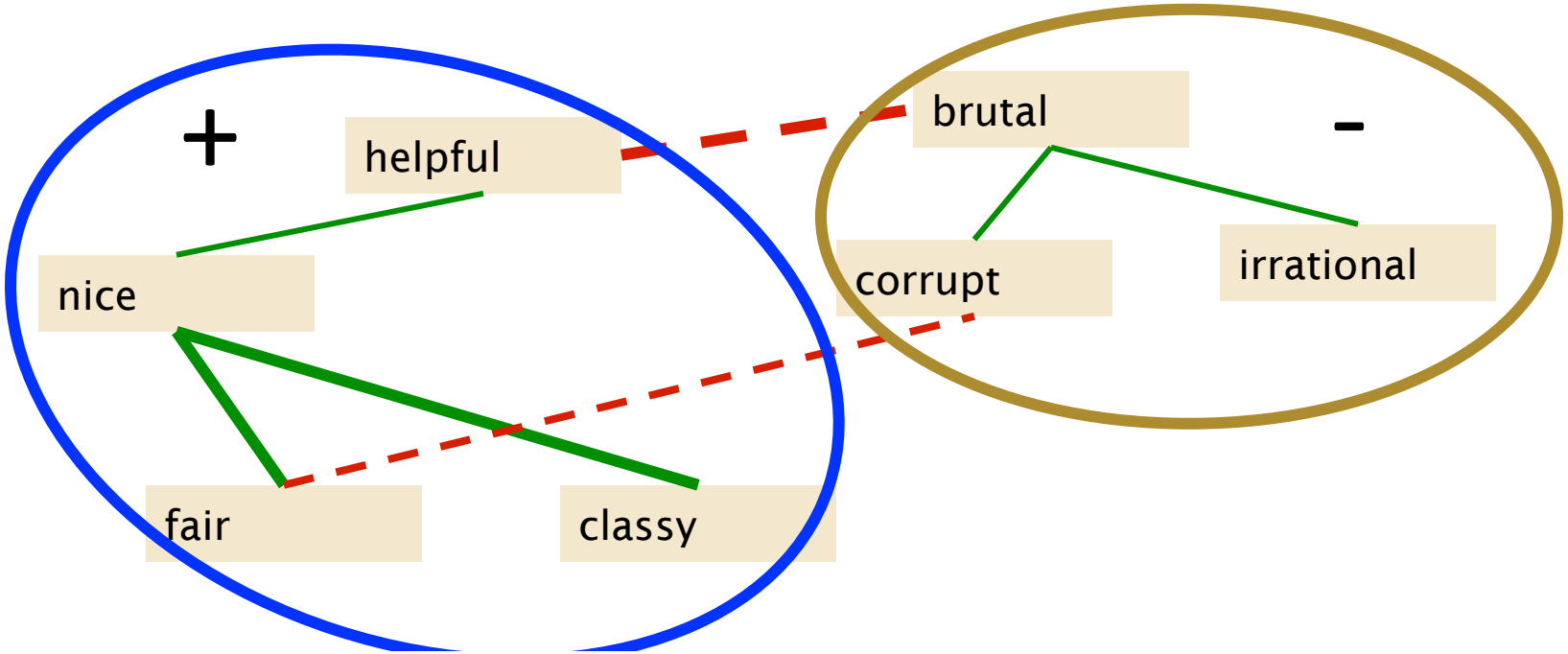
Hatzivassiloglou & McKeown 1997 Step 3

- Supervised classifier assigns "polarity similarity" to each word pair, resulting in graph:



Hatzivassiloglou & McKeown 1997 Step 4

- Clustering for partitioning the graph into two



Output polarity lexicon

- Positive
  - bold decisive disturbing generous good honest important large mature patient peaceful positive proud sound stimulating straightforward strange talented vigorous witty…
- Negative
  - ambiguous cautious cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor outspoken pleasant reckless risky selfish tedious unsupported vulnerable wasteful…

Output polarity lexicon

- Positive
  - bold decisive **disturbing** generous good honest important large mature patient peaceful positive proud sound stimulating straightforward **strange** talented vigorous witty…
- Negative
  - ambiguous **cautious** cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor **outspoken pleasant** reckless risky selfish tedious unsupported vulnerable wasteful…

Turney Algorithm

Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

1. Extract a *phrasal lexicon* from reviews
2. Learn polarity of each phrase
3. Rate a review by the average polarity of its phrases

Extract two-word phrases with adjectives

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

How to measure polarity of a phrase?

- Positive phrases co-occur more with *"excellent"*
- Negative phrases co-occur more with *"poor"*
- But how to measure co-occurrence?

Pointwise Mutual Information

- **Mutual information** between 2 random variables X and Y

$$\mathrm{I}(X,Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **Pointwise mutual information**:
  - How much more do events x and y co-occur than if they were independent?

$$\mathrm{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Pointwise Mutual Information

- **Pointwise mutual information**:
  - How much more do events x and y co-occur than if they were independent?

$$\text{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **PMI between two words**:
  - How much more do two words co-occur than if they were independent?

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

How to Estimate Pointwise Mutual Information

- Query search engine (Altavista)
  - P(word) estimated by `hits(word)/N`
  - P(word$_1$,word$_2$) by `hits(word1 NEAR word2)/N`^{2}

$$\text{PMI}(word_1, word_2) = \log_2 \frac{hits(word_1 \text{ NEAR } word_2)}{hits(word_1)hits(word_2)}$$

Does phrase appear more with "poor" or "excellent"?

$$\text{Polarity}(\textit{phrase}) = \text{PMI}(\textit{phrase},\text{"excellent"}) - \text{PMI}(\textit{phrase},\text{"poor"})$$

$$= \log_2 \frac{\text{hits}(\textit{phrase}\ \text{NEAR "excellent"})}{\text{hits}(\textit{phrase})\text{hits}(\text{"excellent"})} - \log_2 \frac{\text{hits}(\textit{phrase}\ \text{NEAR "poor"})}{\text{hits}(\textit{phrase})\text{hits}(\text{"poor"})}$$

$$= \log_2 \frac{\text{hits}(\textit{phrase}\ \text{NEAR "excellent"})}{\text{hits}(\textit{phrase})\text{hits}(\text{"excellent"})} \frac{\text{hits}(\textit{phrase})\text{hits}(\text{"poor"})}{\text{hits}(\textit{phrase}\ \text{NEAR "poor"})}$$

$$= \log_2 \left(\frac{\text{hits}(\textit{phrase}\ \text{NEAR "excellent"})\text{hits}(\text{"poor"})}{\text{hits}(\textit{phrase}\ \text{NEAR "poor"})\text{hits}(\text{"excellent"})} \right)$$

Phrases from a thumbs-up review

Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
…		
low fees	JJ NNS	0.33
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5
Average		0.32

Phrases from a thumbs-down review

Phrase	POS tags	Polarity
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
…		
virtual monopoly	JJ NN	-2.0
lesser evil	RBR JJ	-2.3
other problems	JJ NNS	-2.8
low funds	JJ NNS	-6.8
unethical practices	JJ NNS	-8.5
Average		-1.2

Results of Turney algorithm

- 410 reviews from Epinions
  - 170 (41%) negative
  - 240 (59%) positive
- Majority class baseline: 59%
- Turney algorithm: 74%

- Phrases rather than words
- Learns domain-specific information

Using WordNet to learn polarity

S.M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. COLING 2004
M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of KDD, 2004

- WordNet: online thesaurus (covered in later lecture).
- Create positive ("good") and negative seed-words ("terrible")
- Find Synonyms and Antonyms
  - Positive Set: Add synonyms of positive words ("well") and antonyms of negative words
  - Negative Set: Add synonyms of negative words ("awful") and antonyms of positive words ("evil")
- Repeat, following chains of synonyms
- Filter

Summary on Learning Lexicons

- Advantages:
  - Can be domain-specific
  - Can be more robust (more words)
- Intuition
  - Start with a seed set of words ('good', 'poor')
  - Find other words that have similar polarity:
    - Using "and" and "but"
    - Using words that occur nearby in the same document
    - Using WordNet synonyms and antonyms

Sentiment Analysis

Learning Sentiment Lexicons