

---

# Discourse and Sentiment:

---

Is there even anything interesting or relevant in their interaction at all or no?

---

# Example

---

We never feel anything for these characters, and as a result the film is basically just a curiosity.

---

# Questions

---

- 1) Are there any interesting correlations between discourse relations or specific discourse markers and sentiment?
  - 2) Can we leverage discourse and to provide polarity scores for values of attributes?
-

# Polarity shift

---

Rarely does the overall polarity of a sentence differ from the polarity of the second discourse segment.

Polarity of sentence and segment	# of sentences
Same	795
Different	137
(too little sentiment to tell)	1372

---



# Polarity of values for attributes

---

One minute, you think you're watching a **serious actioner**; the next, it's as though clips from *The Pink Panther Strikes Again* and/or *Sailor Moon* have been spliced in.

---

# Since

---

We haven't seen such **hilarity** since *Say it isn't so!*

It 's the **funniest** American comedy since *Graffiti Bridge*.

*Crush* could be the **worst film a man has made about women** since *Valley of the Dolls*.

---

# Because

---

The latest installment in the Pokemon canon, *Pokemon 4ever* is surprising **less moldy and trite** than the last two, likely because much of the Japanese anime is set in **a scenic forest where Pokemon graze in peace.**

---

# Future Work

---

- See if we can't automatically extract the attributes and values
  - Get an annotated corpus where discourse relations hold inter-sententially and learn the sentiment relationship between the first and second segments
-

# References

---

- Chomsky, N., & Halle, M. (1968). The sound pattern of English.
  - Chomsky, N. (1965). *Aspects of the Theory of Syntax* (No. 11). MIT press.
  - Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.
  - Chomsky, N. (2002). *Syntactic structures*. Walter de Gruyter.
  - Chomsky, N. (1995). *The minimalist program* (Vol. 28). Cambridge, MA: MIT press.
-

# References

---

- Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3), 175-204.
  - Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., & Joshi, A. K. (2008). Easily identifiable discourse relations. *Technical Reports (CIS)*, 884.
  - Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1631-1642).
  - Marcu, D. (1997, July). The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 96-103). Association for Computational Linguistics.
-

# BOOK REVIEWS & GENRES

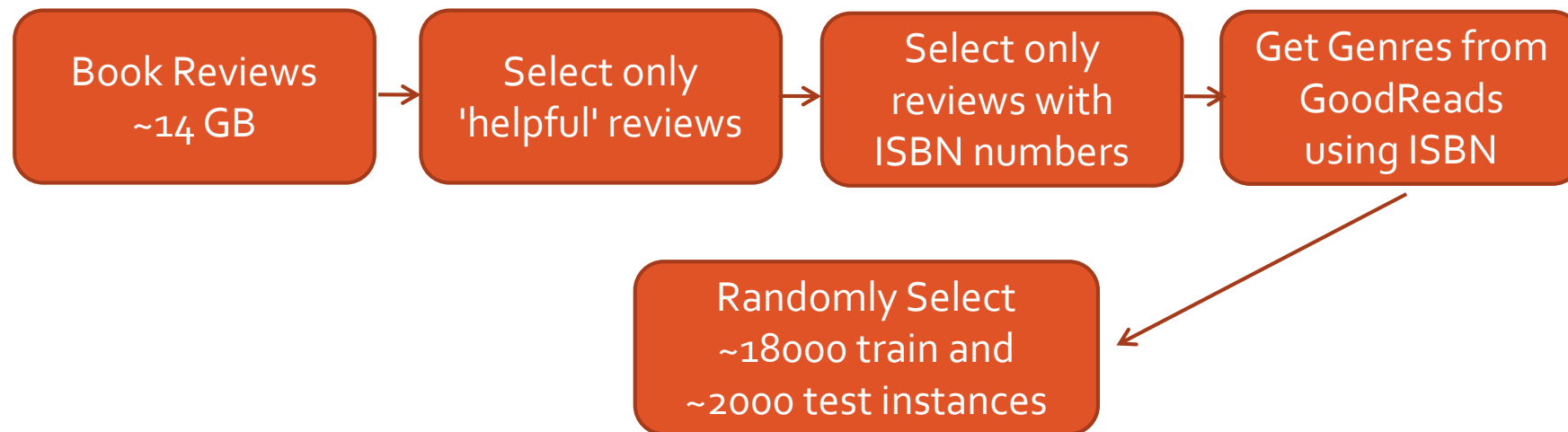
---

Antariksh Bothale and Maria Antoniak

LING 575 -- Spring 2014

# Corpus Collection

- Amazon Book Review Corpus





# Aspect Extraction

- We use MALLET's LDA model to extract topics for each sentence in each review.
- We would love to use seed words as in the Mukerjhee paper, but we could not find a package, and weren't sure if coding it from scratch in a short time line would be wise
- Maybe there is another method to get more specific results?

# Aspect Extraction

- story line told plot reader telling moving turns compelling slow tale interesting lines twists moves mystery bottom pace quickly
- series books left readers entire leave volume rest disappointed leaves trilogy find wanting happened set waiting direction pick fill
- characters character main story plot development developed interesting lead cast descriptions realistic drawn strong dialogue personality believable setting intriguing
- book excellent guide good reference advice practical complete introduction study resource purpose skills fast comprehensive essential title tool serve
- part parts chapters major authors variety close wide longer lead range discuss broken individual subjects contrast themes similar discusses
- man woman young states girl finds united beautiful named heart lady tells sees friend protect meets runs determined mysterious
- read book easy understand follow fun quick difficult enjoyable put easier helped helps pick entertaining full skip format fairly
- point view points starting position perspective views argument critical challenge fair support generally alternative arguments sides balanced ultimately offer
- writing style funny written prose humor entertaining engaging writer narrative author insightful wit brilliant voice makes witty tone clever

# Genre Merging

- We scraped book reviews from GoodReads
- User-defined and classified, lots of over-specific genres (Mermaids, Satanism, Sex Work and so on)
- Can't expect to discover them via plain LDA, and so manually merging them into 20-ish broad genres

# What next?

- Finish genre merging (too many genres)
- Use the aspects to cluster the book reviews into genres.
- Maybe reverse our task and use genres to better extract aspects.

# EmoViz

## Visualizing Emotional State

Shiri Azenkot

June 3, 2014







# **Realtime Emotion Detection & Visualization**



# Related Work

- Offline emotion detection
- Using static classifiers
- Detection of stress

# EmoViz Architecture

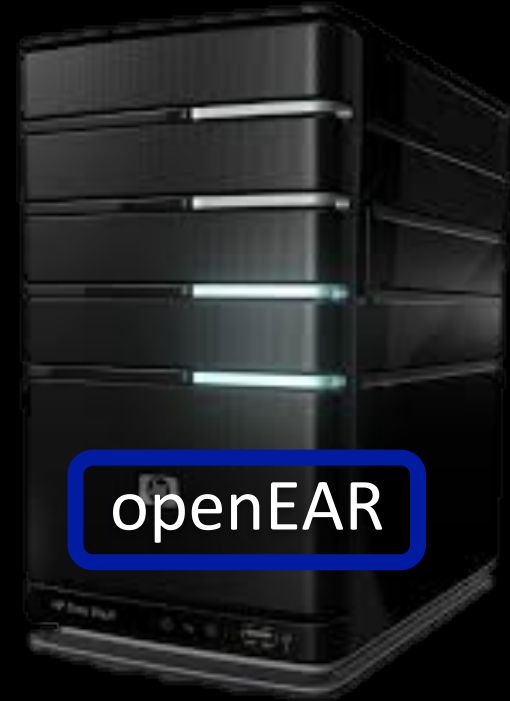


Mobile Device

wav file



'very frustrated'



Server

# Model

$$E_{actual} = E_{openEAR} * \alpha$$

$E$  = emotion

# Evaluation

- Informal
- ~ 5 users
- realtime

# Stance Classification Using Semantic Features in an Online Debate Corpus

C.J Hsu and Ryan Bielby

# Outline

- Motivation
- Reference Papers
- Corpus and Lexicon
- Methods
- Findings
  - Argument Features like appendix
  - Semantics as features vs Semantics as filters
- Future work (Linguistic Analysis)

# Motivation

- In this class, we studied almost 20 papers which address different domains, but all of them have some aspects in common.
- We are tired of this SOP; is there any other way to facilitate the information of semantics?

# Reference Papers

Somasundaran, Swapna and Janyce Wiebe.  
2010. "Recognizing Stances in Ideological On-Line Debates".

- Constructs an "Arguing Lexicon" from MPQA to predict the stances of the online debate posts

Wilson, et al. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis".

- Provides Subjectivity Lexicon built from news articles



# Corpus

- Somasundaran and Wiebe paper provides six categories of online debate posts: abortion, creation, guns, gay rights, god, and healthcare
- Total number of posts: 3167
- Although these posts belong to different categories, the authors have unified the class label (positive and negative argument) for different categories. We could merge all these six categories as single one.

# Building an Arguing Lexicon

- In the MPQA corpus, some text spans are marked with `attitude-type="arguing-pos"` or `attitude-type="arguing-neg"`
- Generate the unigrams, bigrams, and trigrams starting at the text spans which are marked `"arguing-pos"` or `"arguing-neg"`

# Building an Arguing Lexicon

- Remove these n-grams which are already presented in Subjective Lexicon
- Calculate the two conditional probabilities for each entry
  - i.e.,  $P(\text{type} = \text{"arguing-pos"} \mid \text{n-gram})$  and  $P(\text{type} = \text{"arguing-neg"} \mid \text{n-gram})$ .

# Subjectivity Lexicon

- 15,991 subjective expressions from 425 docs
  - devset: 66 docs, 2808 subjective expressions
  - 10-fold cross-validation: 359 docs, 7611 expressions
- Prior-polarity subjectivity lexicon (8,000 words)
  - Riloff and Wiebe, 2003
  - Hatzivassiloglou and McKeown, 1997
  - General Inquirer, 2000
  - Reliability tags: *strongsubj* and *weaksubj*
  - 33.1% *positive*, 59.7% *negative*, 0.3% *both*, 6.9% *neutral*

# Method: Unigram vs Arguing

- **Classifier:** Chose SVM over MaxEnt
- **Unigram Features:** unigram, non-stemmed; negate the unigram which appears after negator
- **Arguing Features:** Break each sentence from each post into trigrams, bigrams, unigrams; check if n-gram (starting with trigrams) is in arguing lexicon; find 'overall' sentiment (sentiment with greatest # arguing features); mark each word (sans stop words) in the sentence as such; e.g., nobody\_neg thinks\_neg

# Methods Results

- Results:
  - Unigrams: 10-fold, accuracy: 62.2198%
  - Arguing: 10-fold, accuracy: 58.236%

# Findings: Is an Arguing Lexicon Useful?

- Arguing Lexicon is like human appendix

Domain (#posts)	Distribution	Unigram	Sentiment	Arguing	Arg+Sent
<b>Overall</b> (2232)	50	62.50	55.02	62.59	63.93
Guns Rights (306)	50	66.67	58.82	69.28	70.59
Gay Rights (846)	50	61.70	52.84	62.05	63.71
Abortion (550)	50	59.1	54.73	59.46	60.55
Creationism (530)	50	64.91	56.60	62.83	63.96

Table 4: Accuracy of the different systems

- Almost 60% of the entries in "arguing-negative" have the token "not"
- Once we negate the word appearing after negator, the unigram feature could almost capture the essence just as the arguing feature does

# Findings

## Sentiment as Feature vs Sentiment as Filter

- Given a semantic lexicon, building the semantic features by counting and voting seems become a SOP in this field
- We think some online posts are suitable for this shallow processing based framework, however, some posts are not
- Could we identify those posts which are not suitable for this framework and perform additional analysis on them?



# Findings

## Sentiment as Feature vs Sentiment as Filter

- 817 posts have no any clue word of semantic lexicon and 2874 posts have at least one clue word of semantic lexicon
- The result of 10 fold C.V on the 2874 posts by unigram features is 59.53%
- The result of 10 fold C.V on those 817 posts by unigram features is 52.02%
- These posts have no pattern at all in unigram features! What causes this?

# Findings

Three categories for the 817 posts

- **Response:** this type of post does not propose any significant supporting points; just tries to deny others' points.
  - e.g., "You should spend more time thinking about what you say before you type ."
- **A/V response:** people are lazy and just post YouTube or other URL to argument their point.
  - e.g., "<http://americansfortruth.com/issues/the-agenda-glbtc-activist-groups/national-glbtc-activist-groups/sisters-of-perpetual-indulgence/page/2>"


# Findings

Three categories for the 817 posts

- **Negated Negatives:** author negates negative terms, but then alludes that they are true.
  - e.g., "Mark, you're not an asshole. You're just trying so hard to be!"

# Future Work

- For any semantics application, a two-stage framework deserves a try!
- Identify those sentences which are not suitable for shallow processing.
- Incorporate audio and video sentiment analysis to complement the text analysis.
- The semantic lexicons are mostly built on newspapers! They do not have slang words and other casual speech.



# Studying the Impact of Multimodality in Sentiment Analysis

Ahmad Elshenawy  
Steele Carter



# Goals/Motivation

- How are judgments influenced by different modalities?
- Compare sentiment contributions of different modalities
- Use Interannotator agreement to measure objectivity of sentiment and ease of judgment
- Observe how results change for fine grained judgments of review chunks

# Background/prior work

- [Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web](#) (Morency et al)
  - Built sentiment classifiers using features from 3 different modalities:
    - Text
    - Audio
    - Video
  - Created YouTube corpus of video reviews
  - Found that integrating all 3 modalities yields best performance

# Corpus

- We created our own corpus of Youtube video reviews, consisting of 3-5 minute long book reviews.
- Originally 35 videos were found and analyzed, but the experiment uses only 20 videos.
  - corpus reduced primarily due to cost concerns
  - 6 positive, 6 negative, 8 neutral
- Originally video transcriptions were obtained via crowdsourcing
  - was way too slow, and way too expensive



# Annotation

- Transcribed each video by hand
  - Labeled disfluencies (um, er, etc.)
- Also labeled our own evaluations of sentiment for comparison and spam filtering
- Added timestamps dividing transcriptions into chunks

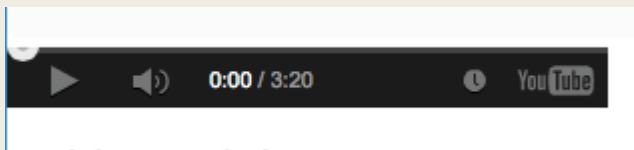
# Modalities

We experiment on four different modalities here:

- **Text only:** typical in sentiment analysis, workers are given only a piece of text.

'You wrote House of Leaves, remember.' 'Yes.' 'House of Leaves, amazing book, that simultaneously destroyed and reconstructed modern storytelling, that house of leaves?' 'mm hmmm.' 'and you think a fitting addition to your cannon, is a fairy tale, told quite unnecessarily as a long-form poem?' 'Absolutely.'

- **Audio only:** workers are given an audio-only piece of the review.



# Modalities - cont'd

- **Video only:** workers are given a video piece of the review where the video is muted, and they are given no option to increase the volume.



- **Audio/Video:** a complete piece of a video, with sound and video intact.

# Video Chunks

- Videos were annotated with timestamps, breaking up videos into ~20-30 second chunks, typically also demarcating new topics within the review.
- A HIT was designed where workers are presented with 5 of these chunks, and asked to judge the sentiment of that chunk.

# HIT Design

- Experiment ended up needing 8 Mechanical Turk HITs.
  - One set of HITs for each modality.
    - Text only, audio only, video only, audio/video
  - One set of HITs for chunks vs whole reviews
- Required **a lot** of javascript and HTML coding
- Collected 10 judgments per video/fragment, paying about \$0.15 per task.
  - 20 video HITs per modality
  - 21 5-chunk HITs per modality

## Preview HITS

1 Select HIT Template 2 Upload Input Data 3 Preview 4 Confirm and Publish

This is how your HIT will look to Workers. Make sure that any variables in the HIT are correctly replaced by your input data, then click "Next".

### Analyze the sentiment in text fragments of a review

Analyze the sentiment in text fragments of a review

**Requester:** Ahmad Elshenawy

**Reward:** \$0.15 per HIT

**HITs available:** 21

**Duration:** 1 Hours

**Qualifications Required:** HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 ,  
Number of HITs Approved greater than or equal to 500

#### HIT Preview

##### Instructions

- Instruction #1: First please fill out a brief survey, providing information about gender, age group and country,
- Instruction #2: Please read five fragments from a review, and tell us whether you feel that fragment was positive, negative or neutral in tone in accordance with the table immediately below.

5	Strongly Positive	Select this if the item embodies emotion that was extremely happy or excited toward the topic. For example, "Their customer service is the best that I've seen!!!!"
4	Positive	Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."
3	Neutral	Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"
2	Negative	Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."
1	Strongly Negative	Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are terrific... NOTTTTT!!!!!!" or "I will NEVER shop there again!!!"

## Instructions

Number of HITs Approved greater than or equal to 500

### HIT Preview

4	Positive	Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."
3	Neutral	Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"
2	Negative	Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."
1	Strongly Negative	Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are terrific... NOTTTTT!!!!!" or "I will NEVER shop there again!!!"

#### Pre-Survey

**1. What is your gender?**

- Male
- Female

**2. What is your age group?**

- select one -

**3. Country of Residence**

### Fragment 1

Fragment 1: 169seconds - 197seconds



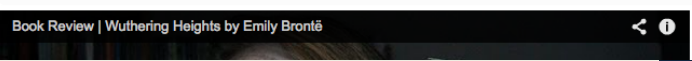
**Fragment 1 Partial Transcription**  
Please transcribe the first 10 words of this fragment. Include utterances like 'umm', 'like' and so on:

**Fragment 1 Sentiment**

- select one -

### Fragment 2

Fragment 2: 140seconds - 155seconds





Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 ,  
Number of HITs Approved greater than or equal to 500

### HIT Preview

#### Fragment 1

There's the bit where the pastor has to deal with all the crazy ladies in his congregation and one woman didn't show up to the meeting and all the other ladies are pissed that she isn't there, and then the pastor says, 'Isn't it funny that the rapture finally happens and the only person to be taken away is Cynthia?' This is the same lady again who says that, 'She once heard that the best thing for the planet would be for everyone to stay in one place for five years. No more transience, no more geographical cures, no more petro holidays. Just a simple commitment to one spot.'

#### Fragment 1 Sentiment

- select one -

#### Fragment 2

I had to take it segmented. I had to read a chapter and then I watched clueless, and then I read a chapter, and then I watched Romy and Michele's High School Reunion, and then I read a chapter, and I watched Heather's -- Heather's probably wasn't the best decision in hindsight.

#### Fragment 2 Sentiment

- select one -

#### Fragment 3

Anyway there's no point in me telling you the story of Pride and Prejudice, everyone knows the story of Pride and Prejudice. You know how it ends. I'm not gonna tell you how it ends, just in case you haven't read this novel, I mean you have had 200 years, so you know, it's kinda your own fault if you read spoilers anywhere.

#### Fragment 3 Sentiment

- select one -

#### Fragment 4

Example of a Text Chunk HIT

# Spam detection/prevention

- HITs with audio, ask workers to transcribe first 10 words
- Label Gold sentiment chunks
  - Discard HITs that disagree with Gold polarity (eg if Gold is 5, discard 3 but keep 5)
  - Issue: can't label video only modality
- Compare submissions to average MTurk worker judgments
- Currently, spam filtration has caught 175+ spam submissions

# Results

- In progress
- Results so far...

experiment	Audio Fragments	Audio Full	AV Fragments	AV Full	Text Fragments	Text Full	Video Fragments	Video Full
kappa	0.7704488	0.4029066	XXXXXXXX	0.3512912	0.4193037	0.3348412	0.2079012	0.1747049

# Potential Analysis

- Interannotator Agreement
- Agreement between modalities
- Compare to Gold
- Compare Chunk deviation from full video sentiment judgment

# Reference

- Morency, Louis-Phillipe and Mihalcea, Rada and Doshi, Payal. [Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web](#), Proceedings of ICMI '11 Proceedings of the 13th international conference on multimodal interfaces, p. 169-176.

# Using Author Types to Predict Review Ratings

*Julian Chan, Laurel Hart, and Ruth Morrison*

# Goal

- Predict rating of review based on review text
- Intuition: “dogs of the same street bark alike” -- authors with similar styles will rate similarly
- Amazon review corpus (Bing Liu et. al)
- Mallet for classification (MaxEnt classifier)

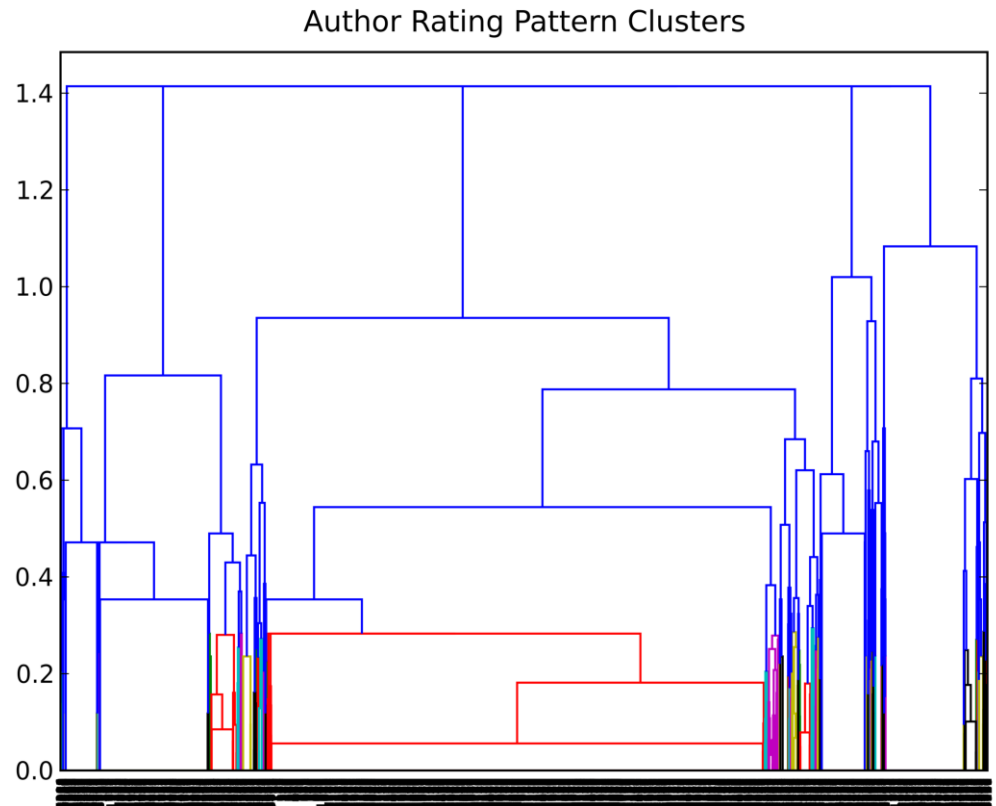
# Features

- N-grams
  - unigrams, bigrams, trigrams, 4-grams, and 5-grams
  - top discriminating n-grams
- Author profile
  - Previous rating behaviors
- Stylistic features
  - Review length, negation, readability
- Miscellaneous
  - product type/genre path



# Author Rating Pattern Clustering

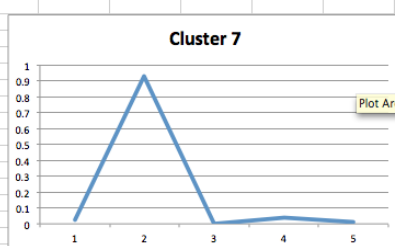
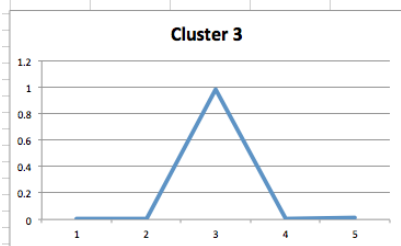
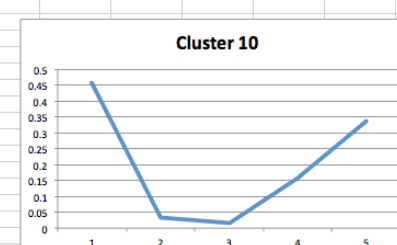
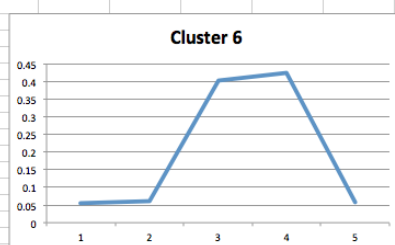
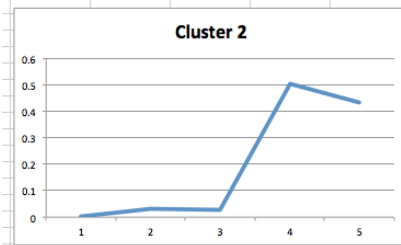
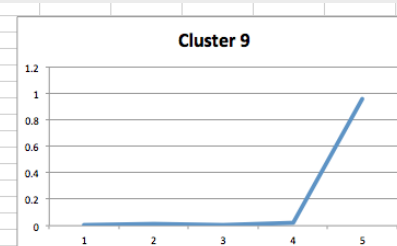
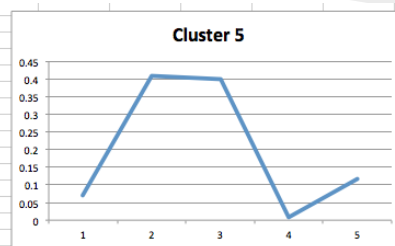
- Each author represented by a 5-dimensional vector.
- Hierarchical clustering from 10000 author samples.
- Cosine distance between author vectors



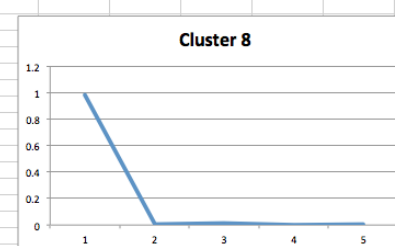
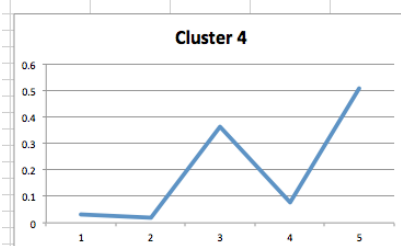
# Five Clusters



# Ten Clusters



Plot Area



# Evaluation

- Strict accuracy is not that informative.
- Credit should be given to a close guess.
- Wildly inaccurate guesses should be penalized more harshly.
- Solution: Mean Squared Error

# Using Five-Cluster Author Type as Feature

AllBigrams

	1	2	3	4	5	Total Squared Error	Instances	MSE
1	39647	2613	2715	2834	48005	807059	95814	8.423184503
2	11912	4569	7976	6798	31807	333343	63062	5.285956678
3	5881	3132	14731	21955	55344	269987	101043	2.672001029
4	3828	1201	8532	44456	173848	221636	231865	0.955883812
5	5831	857	3372	25533	631164	140030	666757	0.210016543
						1772055	1158541	
							Overall MSE	<b>1.529557435</b>
							Normalized MSE	3.509408513

AllBigrams and 5-cluster Author-Type

	1	2	3	4	5	Total Squared Error	Instances	MSE
1	40280	2850	3975	3688	45021	772278	95814	8.0601791
2	11663	3925	8943	7862	30669	328075	63062	5.20241984
3	6018	2533	14914	23721	53857	265754	101043	2.63010797
4	4367	1133	9221	47582	169562	222618	231865	0.96011903
5	7520	1007	4663	29703	623864	177738	666757	0.26657088
						1766463	1158541	
							Overall MSE	<b>1.52473067</b>
							Normalized MSE	3.42387937

It helped \*a little bit\*...

# Our best results so far

AllCaseInsensitiveBigramsBalanced

	1	2	3	4	5	Total Squared Error	Instances	MSE
1	67172	16111	4549	2255	5727	146234	95814	1.5262279
2	18318	23840	12458	4144	4302	86070	63062	1.364847293
3	12514	20282	37062	20061	11124	134895	101043	1.335025682
4	16291	13824	42706	85784	73260	317881	231865	1.370974489
5	51675	16602	32257	111473	454750	1216719	666757	1.824831235
						1901799	1158541	
							Overall MSE	<b>1.641546566</b>
							Normalized	
							MSE	1.48438132

- Rebalanced training data by down-sampling
- Using case-insensitive bigrams results in error reduction
- Incorporating author-profile actually resulted in performance degradation.
- We tried trigrams, tetragrams, and fivegrams. Nothing beat good ol' bigrams.
- A disproportionate number of 5s got classified as 1s. Perhaps some negation resolution could help here.

# Human Performance

- We set up a website showing ten reviews to viewers and asked them to guess the ratings.
- Accuracy of 57.78%
- Mean Squared Error of **0.7889**
- Humans have much better MSE.
- MaxEnt had better accuracy on unbalanced training data, simply because it guessed 5-star more often.
- MaxEnt has similar accuracy as human when trained on balanced data.

# What influences author-type?

We found more than 50% of the data are 5-star reviews.

Most authors also only give 5-star reviews.

Could that be influenced by things like location, time, day of week, etc?

For example, do Americans generally give more positive reviews than people in the UK?



# In Summary...

Nothing beats balanced case-insensitive bigrams (so far), but we're still investigating certain style features (negation, length, readability).

We could explore giving author-type features more weight instead of just throwing everything into MaxEnt

# Learning Sentiment Polarity of Multiword Expressions



MAX KAUFMANN, NICK CHEN, JEREMY  
MCLAIN

# What?



- Previous work
  - Contextual polarity of single words
- Our work
  - Contextual polarity of multiword expressions
- MWE = multiple words that are one single lexical item.
  - throw up, make out, kick the bucket
- Train a classifier that can find sentiment of MWEs

# Why?



- **Noncompositional semantics == noncompositional polarity**
  - Problem:  $\text{sentiment}(\text{playing with fire}) \neq \text{sentiment}(\text{play}) + \text{sentiment}(\text{with}) + \text{sentiment}(\text{fire})$
  - Solution: special classifier
- **Noncompositional semantics == hard to detect**
  - Kick the bucket vs Kick the ball
  - One approach is to use semantic context (a la lesk)
  - Maybe “polarity context” will help us detect them?

# How?



- Based off of the paper *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis* by Wilson et al.
- Create a list of MWEs from the figurative language category of Wiktionary.
- Treat the sentiment of these expressions from the Stanford Sentiment Treebank as the gold standard.
- Using the same corpus used to build the Treebank, create a list of contextual features for each MWE.
- Use these features and the gold standard to train a classifier.

# Features

- POS
- Prior polarity (General Inquirer)
- Previous/next 1 and 2 words
- Previous/next POS
- Contains intensifier?
- Sentence has pronoun?
- Sentence has modal?
- Adjective count
- Adverb count
- Weak/strong subjectivity clue count (MPQA)
- Subjective modifier count

# Progress



	MWEs Count	Accuracy
Training	1478	83%
Testing	987	53%

	Negative	Very Negative	Neutral	Positive	Very Positive	Total
Negative	173	0	60	87	0	320
Very Negative	3	0	1	2	0	6
Neutral	72	0	187	87	0	348
Positive	78	0	68	162	0	308
Very Positive	1	0	3	2	0	5

# This week



- Things we will do
  - 2 classifiers
    - ✦ Binary: Neutral vs polar
    - ✦ Positive vs Very Positive vs Very Negative vs Negative
  - Feature Ablation
  - Use definitions from Wiktionary
    - ✦ Playing with fire -> in a dangerous situation
- Things we wont do
  - Incorporate sense information
    - ✦ Kick the bucket (fig.) vs Kick the bucket (lit.)



# Multilingual Sentiment Analysis for Twitter



George Cooper and Scott Simpson

# Goals

- Create methodology for building sentiment analysis tools for twitter for languages that don't have sentiment-specific resources but do have other resources (bilingual dictionary, machine translation software, POS tagger). Test methodology using Spanish tweets.

# Methodology

- Compare two approaches:
  - Use machine translation to translate input documents into English and use an English sentiment analysis system (NRC-Canada)
  - Adapt English resources to build a sentiment analysis system in the target language



# Translating Input Documents

- Use Google Translate
- Clean tweets in preparation for translation:
  - Remove repeated letters (e.g. “aaawesooooome” -> “awesome”)
  - Split multi-word hashtags (e.g. “#thebestever” -> “#the best ever”)
  - Correct transposed letters (e.g. “hte” -> “the”)

# Adapting English Resources

- Translate training data from English into target language using Google Translate (with cleanup)
- Generate new training data in target language by searching for positive and negative emoticons and hashtags
- Translate sentiment lexicons using a bilingual dictionary
- Translate negation word list using a bilingual dictionary

# Completed Tasks

- English Sentiment Analysis tool built
- List of positive and negative Spanish hashtags built
- All resources acquired (except for a few that could be useful for tweet cleanup)
- Translated negation word list

# In-Progress Tasks

- Queries currently running for Spanish tweets with positive and negative emoticons and hashtags
- Automatically translating sentiment lexicons from English into Spanish
- Writing code to clean tweets in preparation for machine translation



# Remaining Tasks

- Use machine translation to translate training data from English into Spanish
- Run Spanish experiments using translated training data, translated inputs, and training data from tweets with specific emoticons and hashtags.



# Preliminary Results

- English twitter sentiment analysis tool achieves 65.12 averaged F-score (compared to 69.02 reported in paper)

# Predicted Results

- We predict that there will be a modest drop in accuracy for Spanish sentiment analysis tools
- Don't know whether translating input tweets into English or adapting English resources will yield better accuracy

# Aspect Based Sentiment Analysis

*Jared Kramer and Clara Gordon*

# Overview

- Background
- Our Task
- Our Approach
- Results!

# Background

- Entity: The thing being described
- Aspect: A part of the thing being described

The screen is too small.

- Entity = laptop
- Aspect = screen
  
- Aspect detection and sentiment analysis has many downstream applications in automatic review summarization and aggregation

# The Whole Task

## Dataset

- 2 sets of sentences extracted from reviews, ~3K apiece
- Domains: laptop and restaurant
- Labeled for aspect, aspect polarity, and aspect category

## Task breakdown

- Subtask 1: Extract aspects
- **Subtask 2: Classify polarity of aspects**
- Subtask 3: Group aspects into categories
- Subtask 4: Classify polarity of categories

# Subtask 2

- Given a sentence with a list of aspects, classify the polarity of each aspect.
  - Not all sentences have aspects
- Two kinds of data: Laptops and Restaurants
- Polarity labels:
  - positive, negative, neutral, conflict

# Baseline

- From SemEval-provided script, using random 20% of data as test:
  - 0.4705
  - Pretty easy to beat
  - Based on <aspect term, polarity> tuple frequencies gathered from the training corpus
  - Given 4 different categories, indicates that there are some correlations between aspect and polarity



# Our Approach

- Throw tons of features at Mallet!
- Use multiple classifiers
  - Naive Bayes, Max Ent, Decision Tree
- Start with shallow features and move deeper

# Shallow Features

- N-grams
  - sentiment backoff using Sentistrength
    - Screen size is POS for portable use
  - POS labeling
  - Aspect labeling
    - ASPECT is perfect for portable use
  - Punctuation stripping
  - Stopword removal
  - Proximity labeling
  - “Window” around aspect span
  - Wordnet expansion for adjectives
- Metadata
  - Punc, token, POS counts

# Preliminary Results (laptops)

Features	Naive Bayes	MaxEnt	Decision Tree
All Unigrams	.6348	.6348	.5132
5 - Window unigrams	.6045	.6045	.4158
All uni+bi-grams	.5943	.6531	.5131
All uni+bi+tri-grams	.5598	<b>.6551</b>	.5132
Uni + POS tags	<b>.6511</b>	.6409	<b>.5476</b>
Bi + Aspect Backoff	.5923	.6227	.5416
Uni + Positions	.6206	.5963	.4787
Bi + Sentiment Backoff	.5930	.6227	.5416
Uni + WordNet	.5223	.5355	.4604

\*\* Official results range between 0.3654 and 0.7049 -- not bad!

# Conclusions so far

- Bag-of-words is hard to beat :(
- Similarity of aspect and sentence polarity
  - Sentence level features generally outperform “window”-focused features
  - The more data gathered from the sentence, the better
- Aspect backoff hurts performance
  - There might be trends in which types of aspects are discussed negatively and positively
- Revised focus: focus on identifying and analyzing sentences where aspect polarities differ from overall

# Back of the envelope...

- Of 100 manually-examined sentences, 69% had the matching sentence and aspect polarities
- Of those with different aspect polarities, an overwhelming number of the differing aspects were neutral
- Single-aspect sentences more likely to match

# Polarity Differences

## Negative-Positive:

It's like 9 **punds**, but if you can look past it, it's GREAT!

Still testing the **battery life** as i thought it would be better, but am very happy with the upgrade

Everything is so easy to use, Mac software is just so much simpler than **Microsoft software**.

I love WIndows 7 which is a vast improvment over **Vista**.

## Neutral-Polar (far more common)

I charge it at night and skip taking the **cord** with me because of the good battery life

I took it back for an Asus and same thing- blue screen which required me to remove the **battery** to reset.

# Data Issues

In the shop, these MacBooks are encased in a soft **rubber enclosure** - so you will never know about the razor edge until you buy it, get it home, break the seal and use it (very clever con).

I was looking for a mac which is portable and has all the **features** that I was looking for.

- Are these aspects really positive?

# In progress...

- More systematic examination of all possible shallow feature combinations
- Dependendency triples
- Other types of expansion
  - Lin thesaurus, distributional similarity
- Two-part identification: different procedures for single and multiple aspects



Thanks for listening!



# wPod

Weibo Public Opinion (Polarity) Detection

Haotian He & Sanae Sato

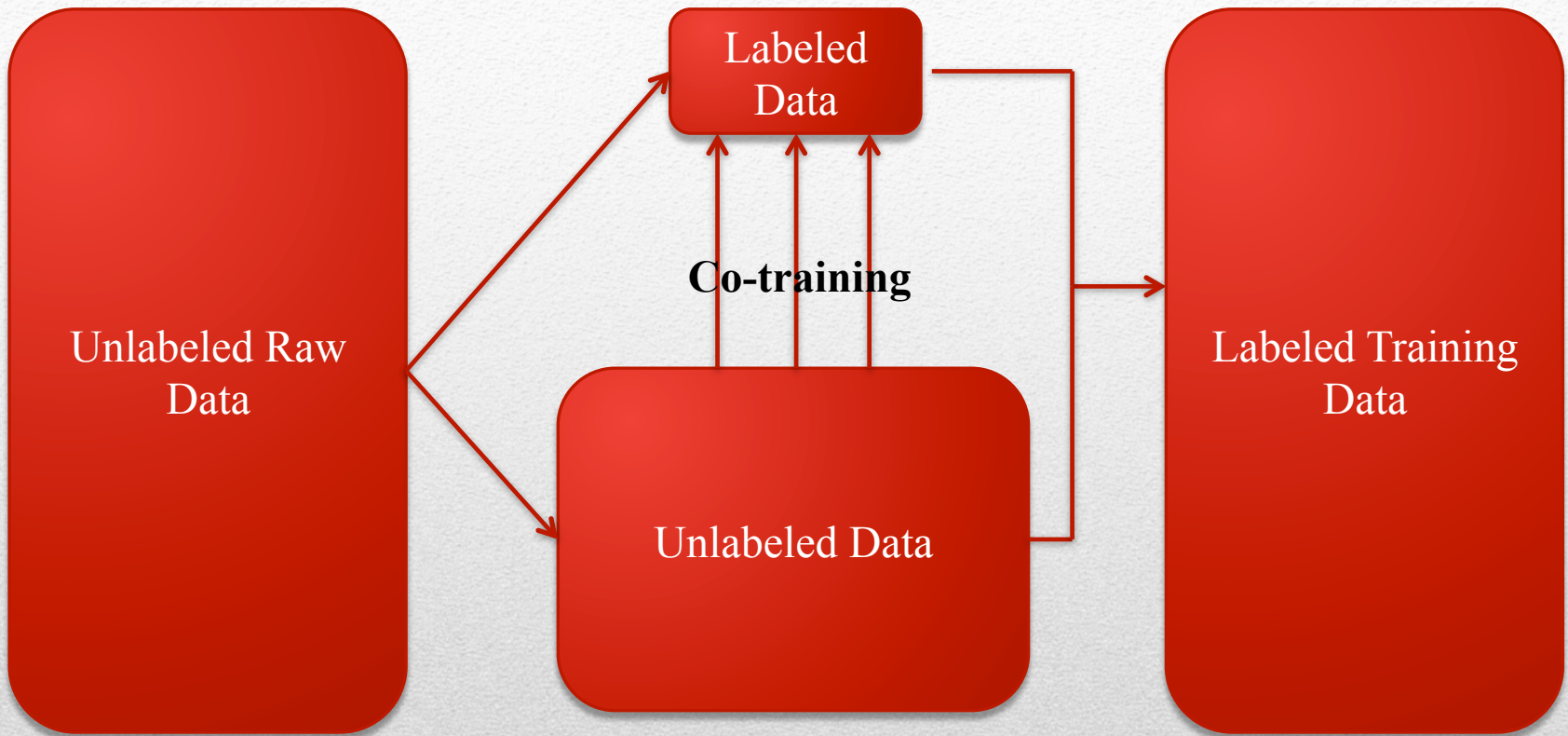
---

- Microblog is always updating, and new issues and terms come out. There are not many valuable handy labeled data.
- Co-training is a semi-supervised learning technique that provides a good way for such a case.
- Co-training employs two classifiers, with two sets of features separately, in a loop to label all the unlabeled examples. Each classifier takes turns to select the most confidently predicted examples and add these into the training set. Both classifiers then re-learn on the enlarged training set so that they take into account the newly added data.

# Algorithm

---





# Co-training Architecture

---

- Microblogs of 2565 users from the signed-up day of each user till March 15, 2014, more than 3 million in total.
- Randomly picked up 330 microblogs to manually label the polarity as the initial training dataset.
- Co-trained for loops to reach a training dataset as 65,000 microblogs.
- Microblogs for test dataset are extracted from March 3 to 5 during the NPC (The House) and CPPCC (Senate) annual joint conferences in 2014, and divided to 12 different categories according to the key words.
- Manually labeled two test datasets for evaluation.

# Dataset

---



Category	Key Words	Count
Report of the Government (pre)	Premier, State Council, Report	1762
Report of the Supreme Court (spc)	Supreme Court, Report	315
Report of the Supreme Procuratorate (spp)	Supreme Procuratorate, Report	158
Education Equality (edu)	Education Equality	123
Second Child (sch)	Second Child	210
Environment, Air pollution (env)	Environment, pollution, PM2.5	4439
Anti-corruption (cor)	Anti-corruption, ...	1062
Medical System Reform (med)	Medical Reform	456
Public Funding Usage (pfu)	Public Funding Usage	168
State-Owned Enterprise Reform (ser)	State-Owned Enterprise Reform	922
Real Estate (hou)	Real Estate, Price, ...	7444
Food Safety (fst)	Food Safety	2333

# Dataset

---

- Features:
  - Set 1: Unigram + Bigram + Polarity Words
  - Set 2: Trigram + Emoticon
- Classifier:
  - Naïve Bayes
  - Maximum Entropy (better performance / final choice)

# Features and Classifier

---



	<b>Accuracy</b>
Education Equality	0.8048780487804879
Second Child Policy	0.8333333333333334

# Evaluation

---



	Accuracy
Education Equality	0.8048780487804879
Second Child Policy	0.8333333333333334

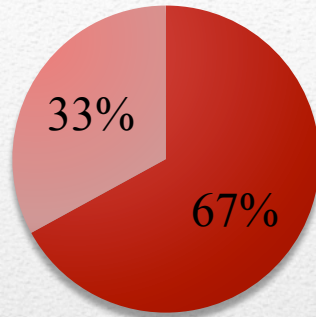
### Top 3 predicted negative microblogs for Second Child Policy:

Microblog text	Translation
2013年两会也是马旭这只计生狗大言不惭的讲到单独二胎不会一下子放开。	Ma Xu, this dog official, shamelessly said the second child policy would not be released in 2013 NPC and CPPCC.
计生委早死。老百姓就不会死。计生委不死。老百姓就会断子绝孙	If Family Planning Office is shut down, people would survive. If it is not, people would die without sons.
反人类废除计划生育2013:请记住这计划生育利益集团代言人的丑陋嘴脸。	Repeal the antihuman family plan 2013: Please remember the ugly face of the Family Plan interest group.

# Evaluation

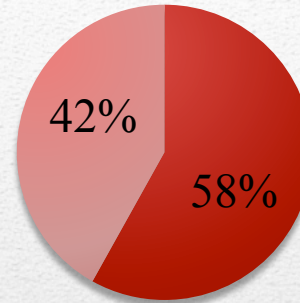
---

## Report of the Government Work



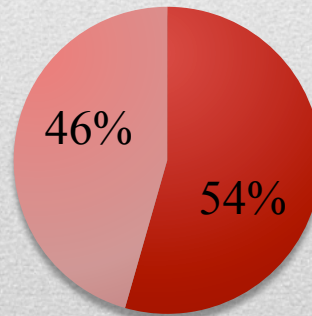
■ Positive  
■ Negative

## Report of the Supreme People's Court



■ Positive  
■ Negative

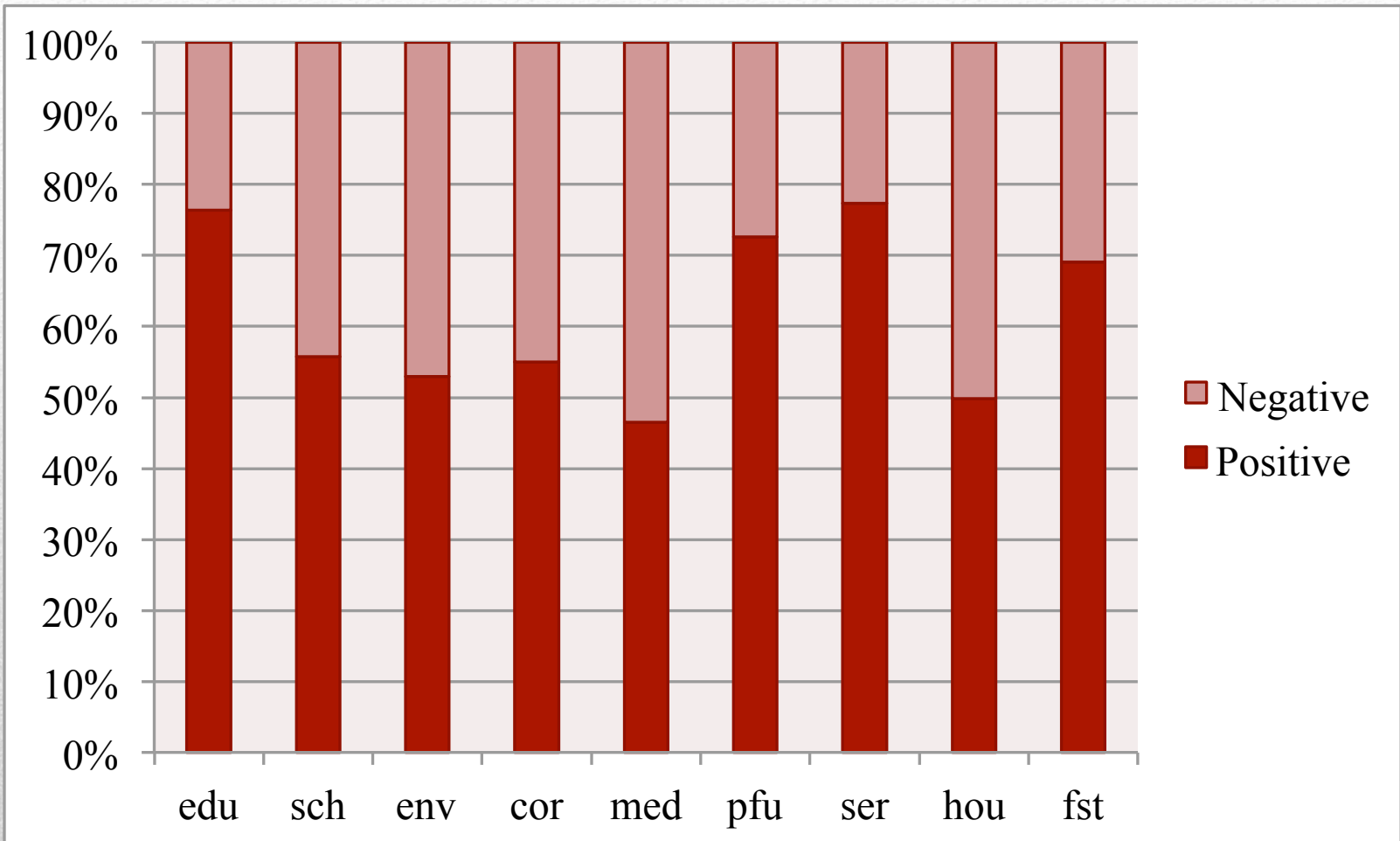
## Report of the Superme People's Procuratorate



■ Positive  
■ Negative

# Results – Three Reports

---



# Results – Popular Issues

---



- Want to explore diachronic changes on the same issues. But currently did not have time to extract the previous years' Weibo data.
- Not only add more features, but add more sentiments as happy, sad, or angry to the system.
- If in the future only do the political analysis, should specify the training data to be related topics, instead of all the general data.
- Try SVM classifier.

# Future Work

---

# Multilingual Sentiment Analysis

**Comparing techniques in sentiment  
analysis on different languages**

LING 575

Claire Jaja, Andrea Kahn

# Problem Definition

Sentiment analysis techniques are typically developed on English.

Current approaches to other languages often involve automatic translation or use of “language agnostic” techniques like machine learning.

This raises two research questions:

1. Are machine learning techniques really language agnostic?
2. How do the results obtained when using resources translated/pivoted from English compare to those with resources developed in the test language?

# Datasets

## IMDb movie reviews (Pang and Lee, 2004)

- English
- 1000 positive, 1000 negative reviews, pre-processed

## CorpusCine movie reviews (Cruz Mata, 2011)

- Spanish
- 3878 reviews with 1 - 5 star ratings
- processed by us, discarding 3 star reviews, then choosing 1000 positive and 1000 negative

## quotations from newspaper articles, annotated for polarity

- English (Balahur-Dobrescu and Ralf, 2009)
  - 1590 total, where annotators agree: 863 obj, 193 pos, 234 neg
- German (Balahur-Dobrescu, 2011)
  - 2387 total, where annotators agree: 591 obj, 514 pos, 379 neg

# Methodology

classifiers: MaxEnt, Naive Bayes

features: unigram (with and without frequency cut-off),  
bigram, trigram, unigrams from General Inquirer sentiment  
lexicon

use 10-fold cross validation



# Results: MaxEnt

features	IMDb			CorpusCine		
	average	min	max	average	min	max
unigram	86.00%	81.00%	91.50%	83.40%	81.50%	86.00%
unigram > 4	68.20%	62.00%	71.50%	67.45%	60.50%	73.00%
bigram	84.65%	80.50%	88.00%	83.10%	78.00%	87.00%
trigram	50.05%	49.50%	51.00%	81.00%	76.00%	87.00%
unigram + bigram	85.35%	82.50%	89.00%	82.70%	79.50%	86.50%
sentiment lexicon	78.70%	74.00%	83.00%	?	?	?

# Results: Naive Bayes

features	IMDb			CorpusCine		
	average	min	max	average	min	max
unigram	81.65%	75.00%	87.00%	82.70%	79.00%	86.50%
unigram > 4	69.20%	62.50%	74.50%	64.75%	59.50%	69.50%
bigram	81.15%	73.50%	85.50%	81.80%	78.50%	85.00%
trigram	80.95%	73.00%	86.00%	81.55%	78.50%	85.00%
unigram + bigram	81.45%	74.50%	85.50%	81.85%	78.00%	85.00%
sentiment lexicon	78.50%	75.00%	82.50%	?	?	?

# Results: Discussion

- using a unigram frequency cut off of 4 drastically drops results
- MaxEnt is better than Naive Bayes on IMDb using unigram and/or bigram features
- MaxEnt is weirdly bad using trigram features on IMDb
- CorpusCine results are worse than IMDb results using MaxEnt and unigram and/or bigram features
- IMDb and CorpusCine results are comparable using Naive Bayes - NB is more language agnostic? (when it comes to two similar languages like English and Spanish...)

# Future Work

- translate sentiment lexicon into Spanish, use for CorpusCine
- find Spanish sentiment lexicon, use for CorpusCine
- translate CorpusCine test set(s) into English, use IMDb trained classifiers
- address negation in the text
- lemmatize text
- try subjectivity classification for English and German newspaper quotes

**Thanks for listening!**

# Project Presentation

*Veljko Miljanic*

# Task

Analyze sentiment in subordinating and coordinating conjunctions with respect to cue phrase.

Hypothesis:

- Cue phrases are strong signal of sentiment relationship between clauses
- Because cue phrases are strong indicators RST relationships (e.g. contrast [BOS ... ][but ... EOS])

# Example

(S  
    (SBAR Although (S What Time offers Tsai 's usual style (CC and) themes))  
    , it has a more colorful , more playful tone than his other films .  
)

Sentiment:

SBAR = 0.56

SBAR clause = 0.72

Whole sentence = 0.81



# Method

I am using Socher at al dataset because it has sentiment annotations on phrase level

However I still have to identify cue phrases and clauses:

1. Run parser and recover S, SBAR and CC labels
2. Identify clauses and cue phrases by matching simple patterns like: [S ... [SBAR cue phrase [S ...] ]

# Metrics

Metrics should:

- Quantify relationship strength
- Qualify kind of relationship

Metrics

1. Mutual information
2. Sentiment delta (mean, std)
3. Count of + to - and - to +
4. Count of + to ++ and - to --

# Progress

## Done so far:

- Extraction of subordinate / main clause and extraction of coordinate clauses
- Merging of extractions with sentiments
- Mutual information calculation

## To do:

- Rest of the metrics
- Analysis

The End

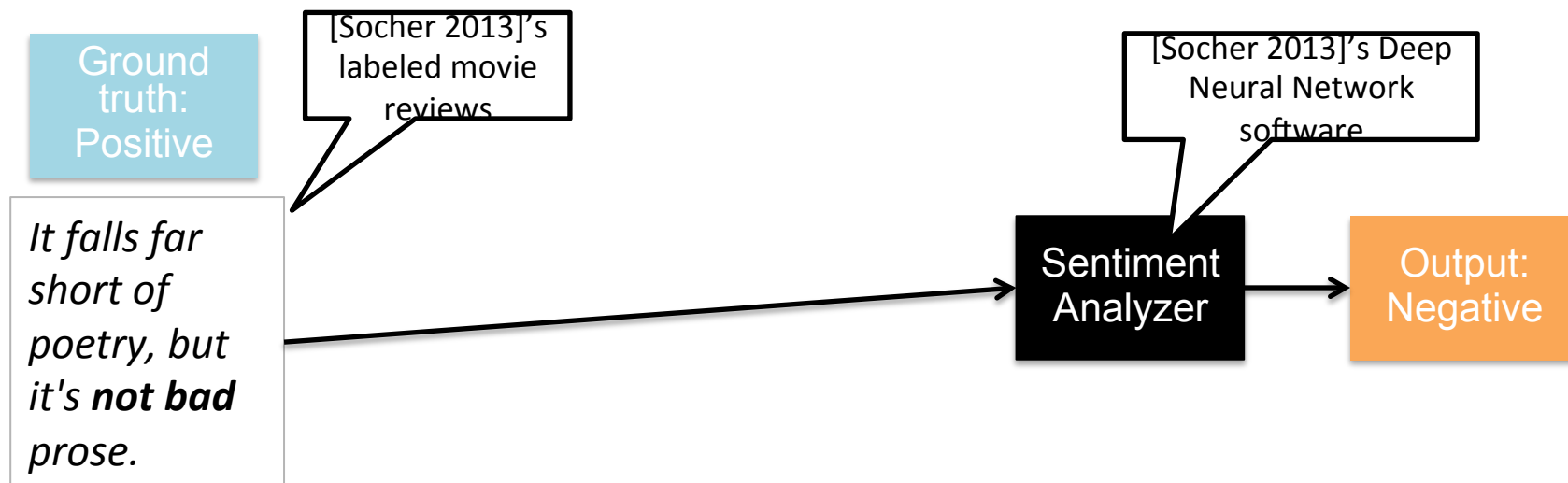
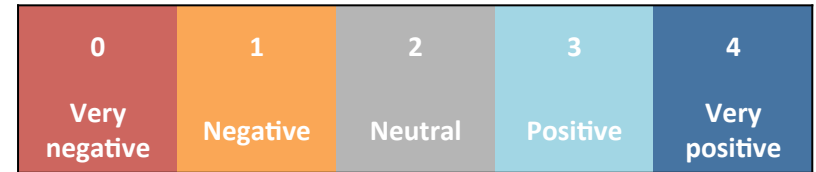
# Paraphrasing Negation Structures for Sentiment Analysis

# Overview

---

- Problem:
  - Negation structures (e.g. “not”) may reverse or modify sentiment polarity
  - Can cause sentiment analyzers to misclassify the polarity
- Our approach:
  - **Remove** the negation by restructuring and then resurfacing the sentence
- Hypothesized benefits
  - **Improves** sentiment analysis accuracy
  - **Reduces** work for sentiment analysis implementers
- Results (so far!):
  - Implemented the paraphraser using Java, Stanford Parser, and Wordnet
  - Used data set and black-box classifier from [Socher 2013]
  - Reduction of 1.4% RMSE between ground-truth and classification on paraphrases

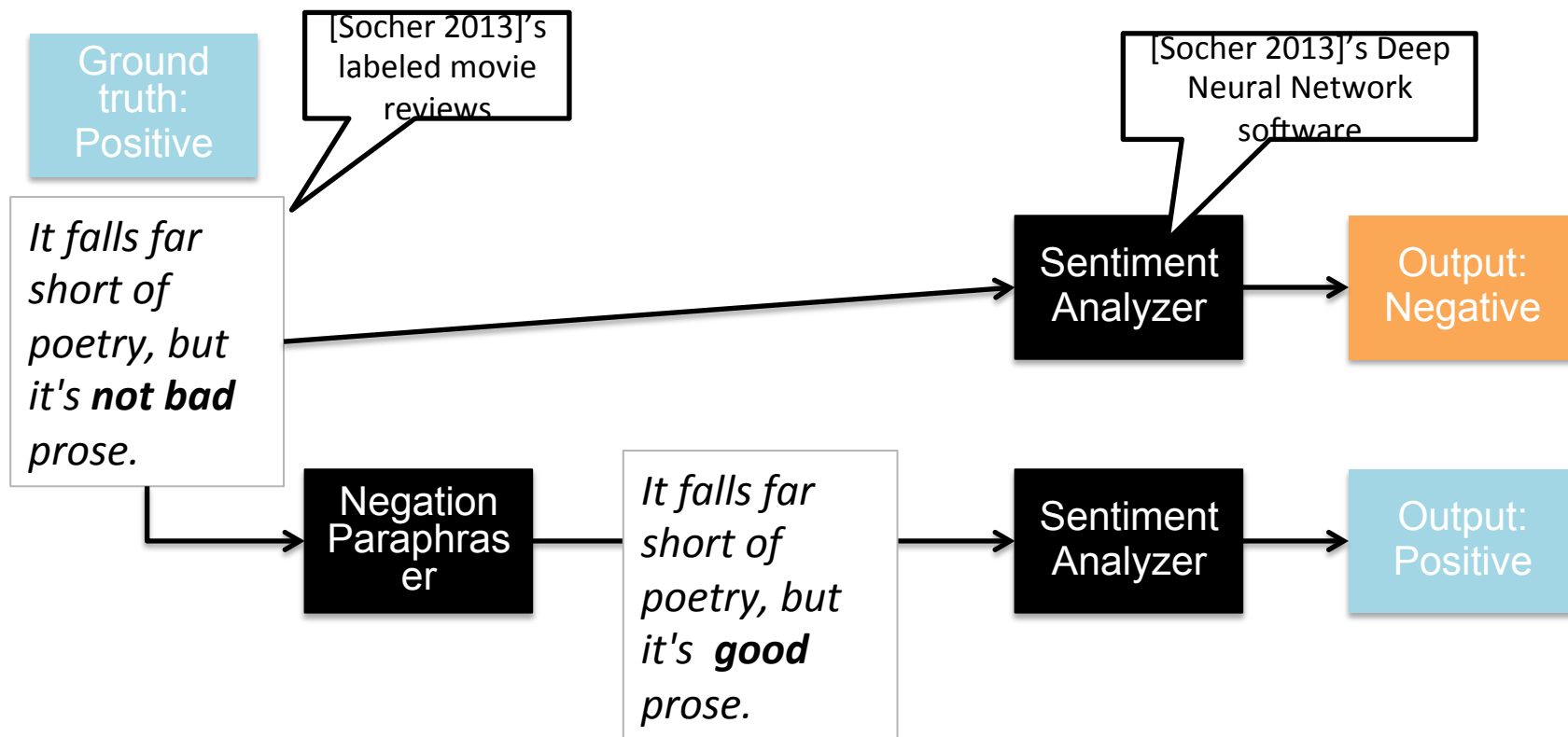
# Example



R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," In Proceedings of EMNLP. 2013.

# Example

0	1	2	3	4
Very negative	Negative	Neutral	Positive	Very positive

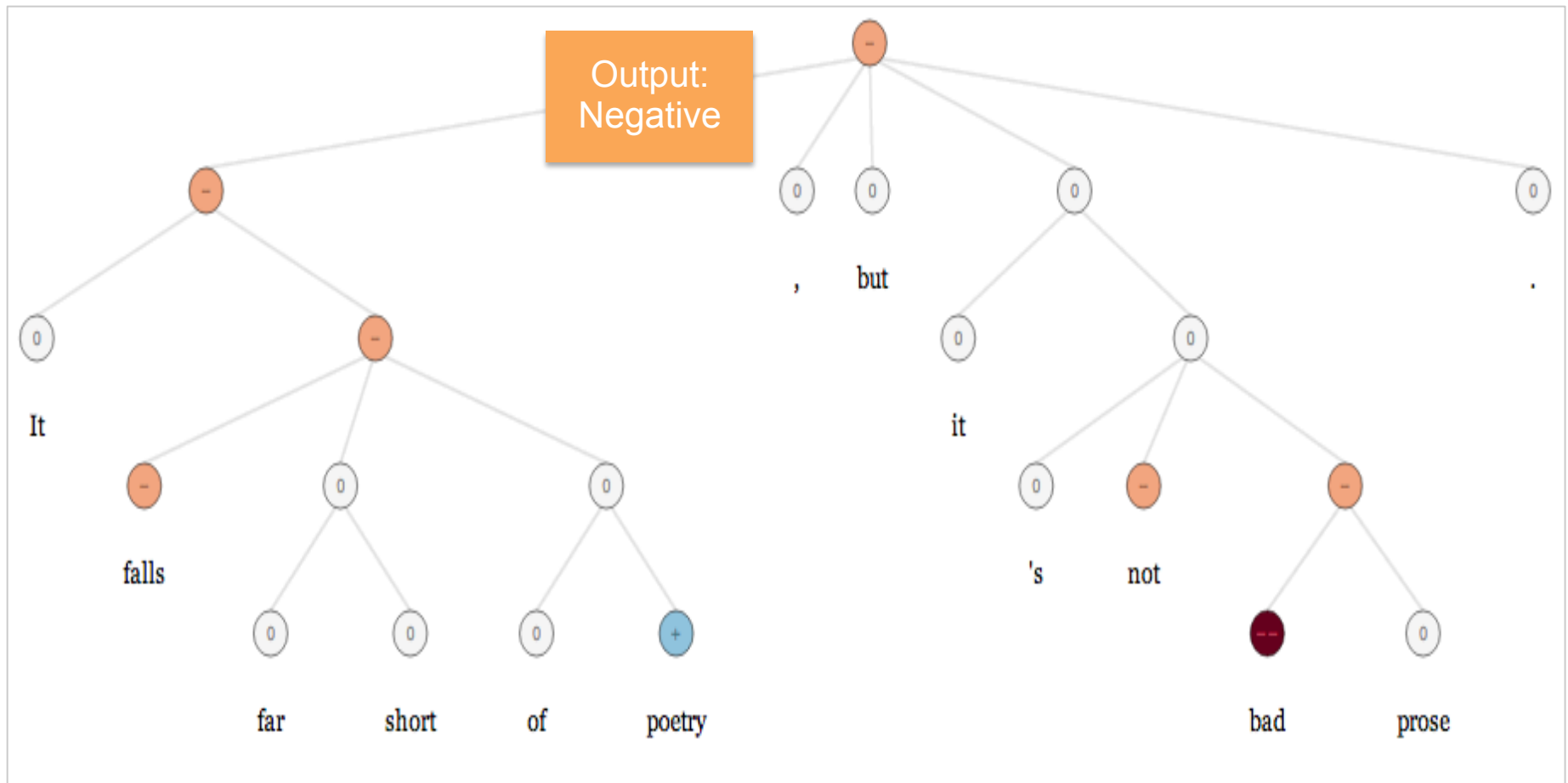


R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," In Proceedings of EMNLP. 2013.



# The (observed) effect of negation on polarity classification

0	1	2	3	4
Very negative	Negative	Neutral	Positive	Very positive



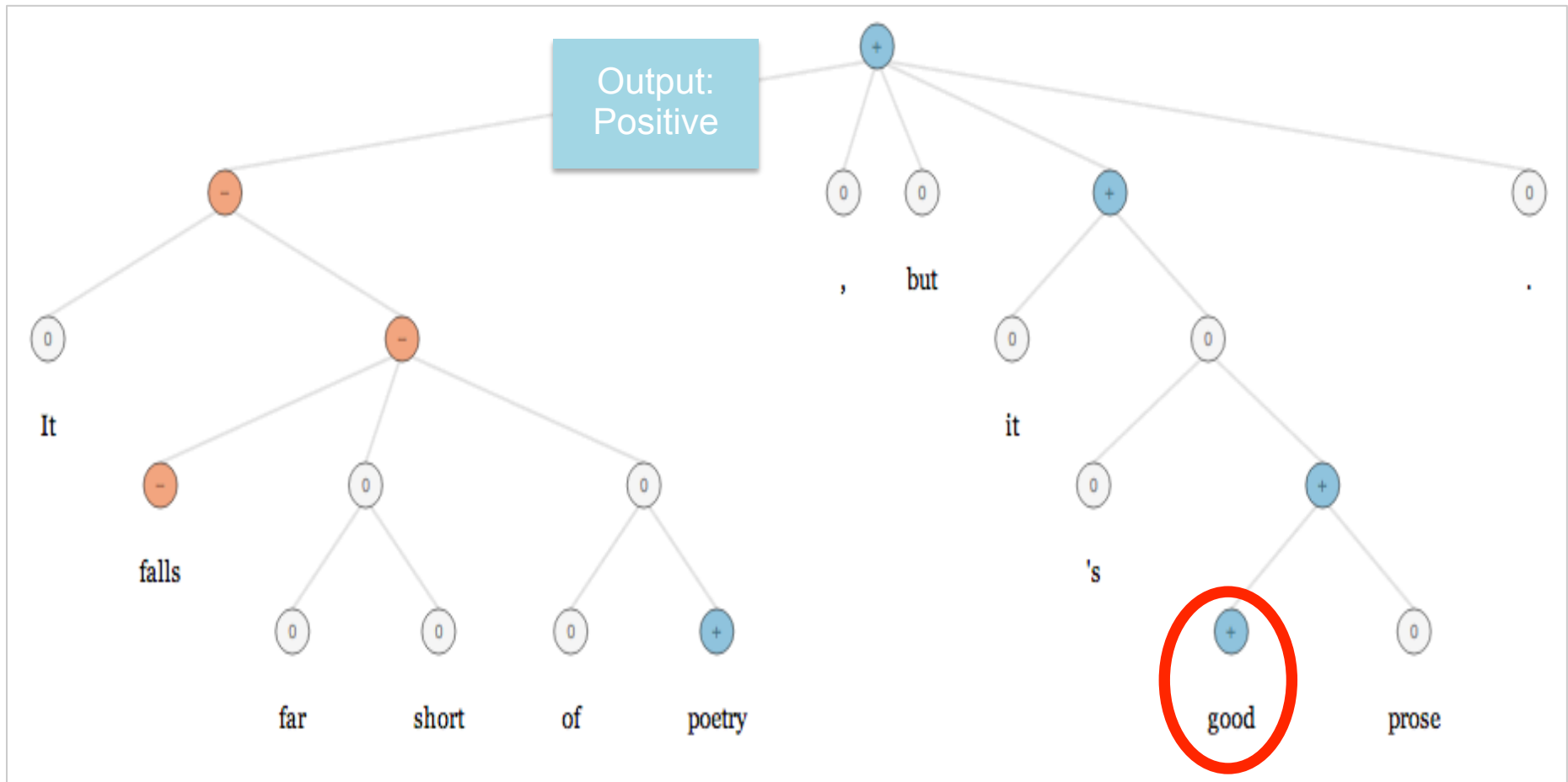
Ground truth: Positive

*It falls far short of poetry, but it's **not bad** prose.*



# The (observed) effect of negation on polarity classification

0	1	2	3	4
Very negative	Negative	Neutral	Positive	Very positive



Ground truth: Positive

*It falls far short of poetry, but it's **good** prose.*

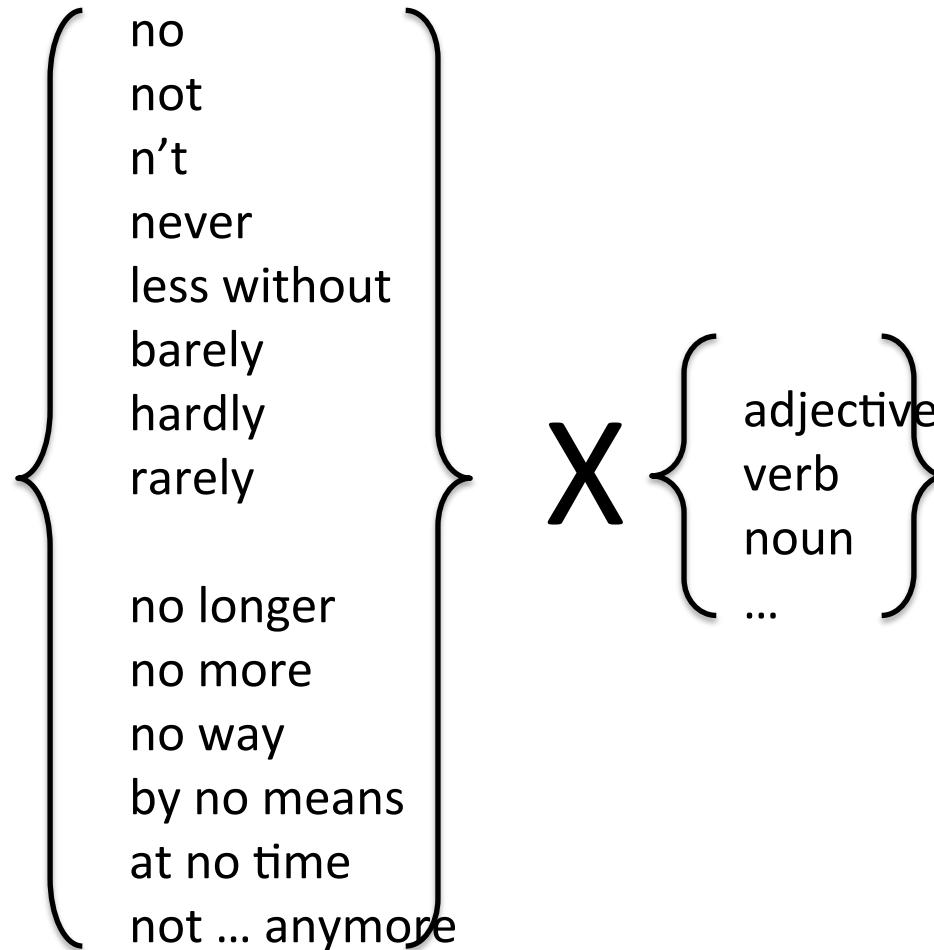
# Related work on the treatment of negation

---

- Heuristic rules
  - A. Hogenboom, P. van Iterson, B. Heerschop, F. Frasinca, and U. Kaymak. “Determining Negation Scope and Strength in Sentiment Analysis,” In Proceedings of IEEE SMC, 2011.
  - M. Hu and B. Liu. “Mining and Summarizing Customer Reviews,” In Proceedings of ACM KDD, 2004.
  - L. Jia, C. Yu, and W. Meng. “The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness,” In Proceedings of ACM CIKM, 2009.
- Supervised machine learning
  - E. Lapponi, J. Read, and L. Ovreliid. “Representing and Resolving Negation for Sentiment Analysis,” In Proceedings of IEEE ICDMW, 2012.
  - T. Wilson, J. Wiebe, and P. Hoffman. “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,” In Proceedings of EMNLP, 2005.

# Design & Implementation: Negation structures as polarity shifters

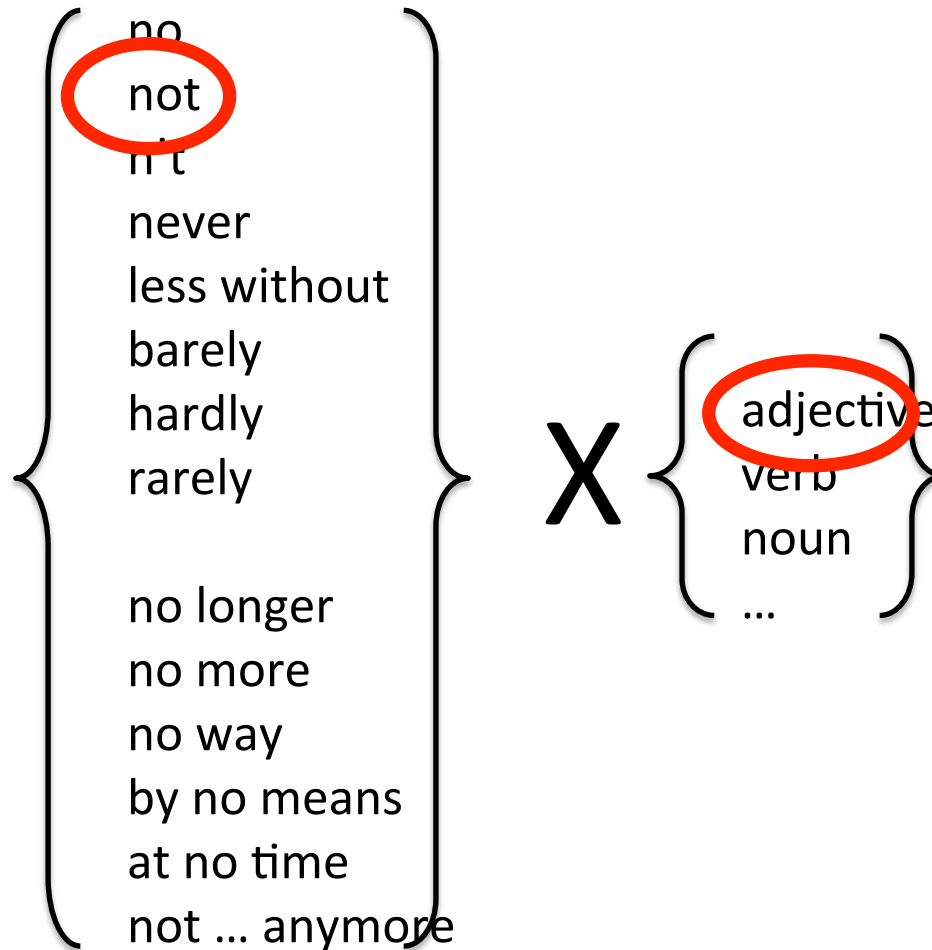
---



List from [Jia 2009]

# Design & Implementation: Negation structures as polarity shifters

---



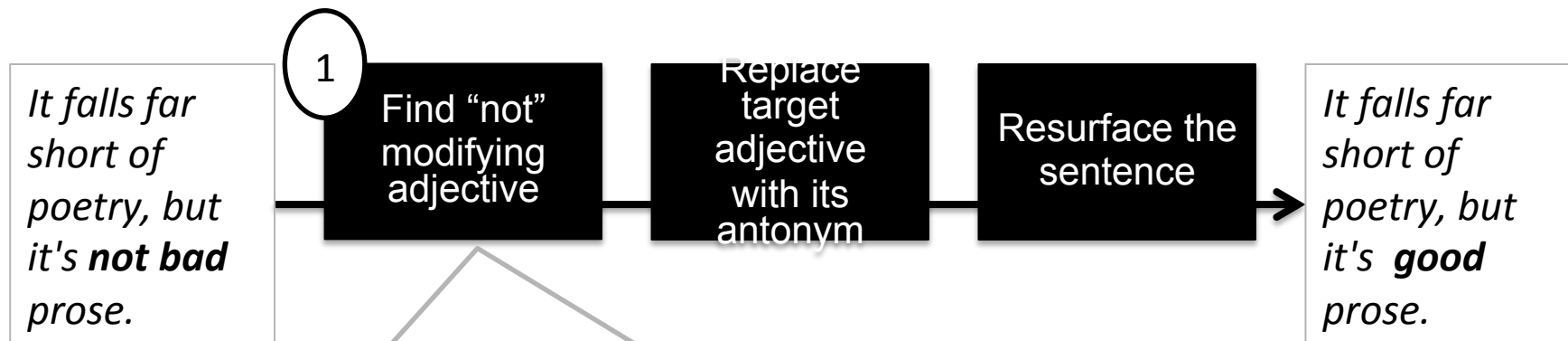
List from [Jia 2009]

# Design & Implementation: Negation Paraphraser Pipeline

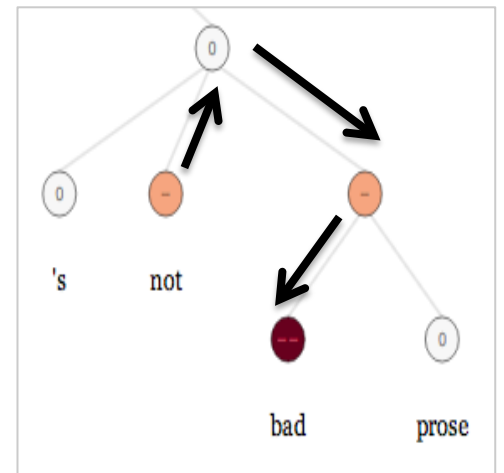
---



# Design & Implementation: Negation Paraphraser Pipeline

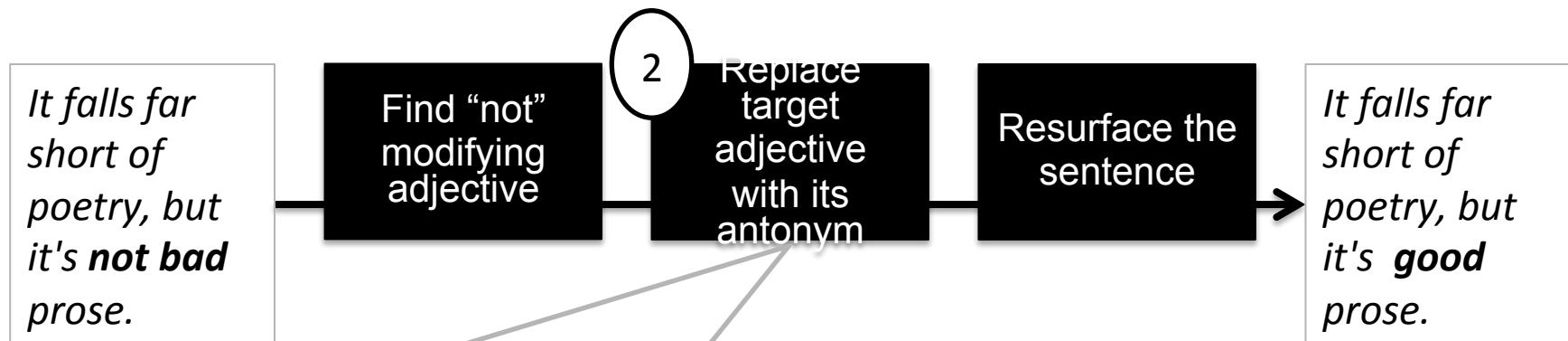


- Used the Stanford Parser software to build a parse tree
- Find "not"
- Find the first adjective who is a right descendent of

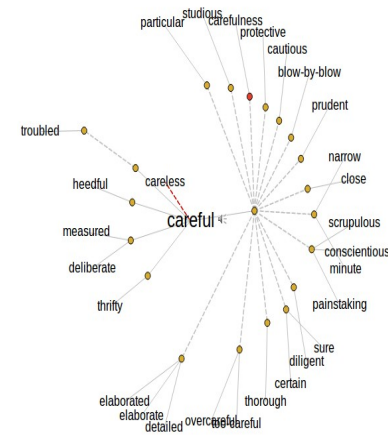




# Design & Implementation: Negation Paraphraser Pipeline

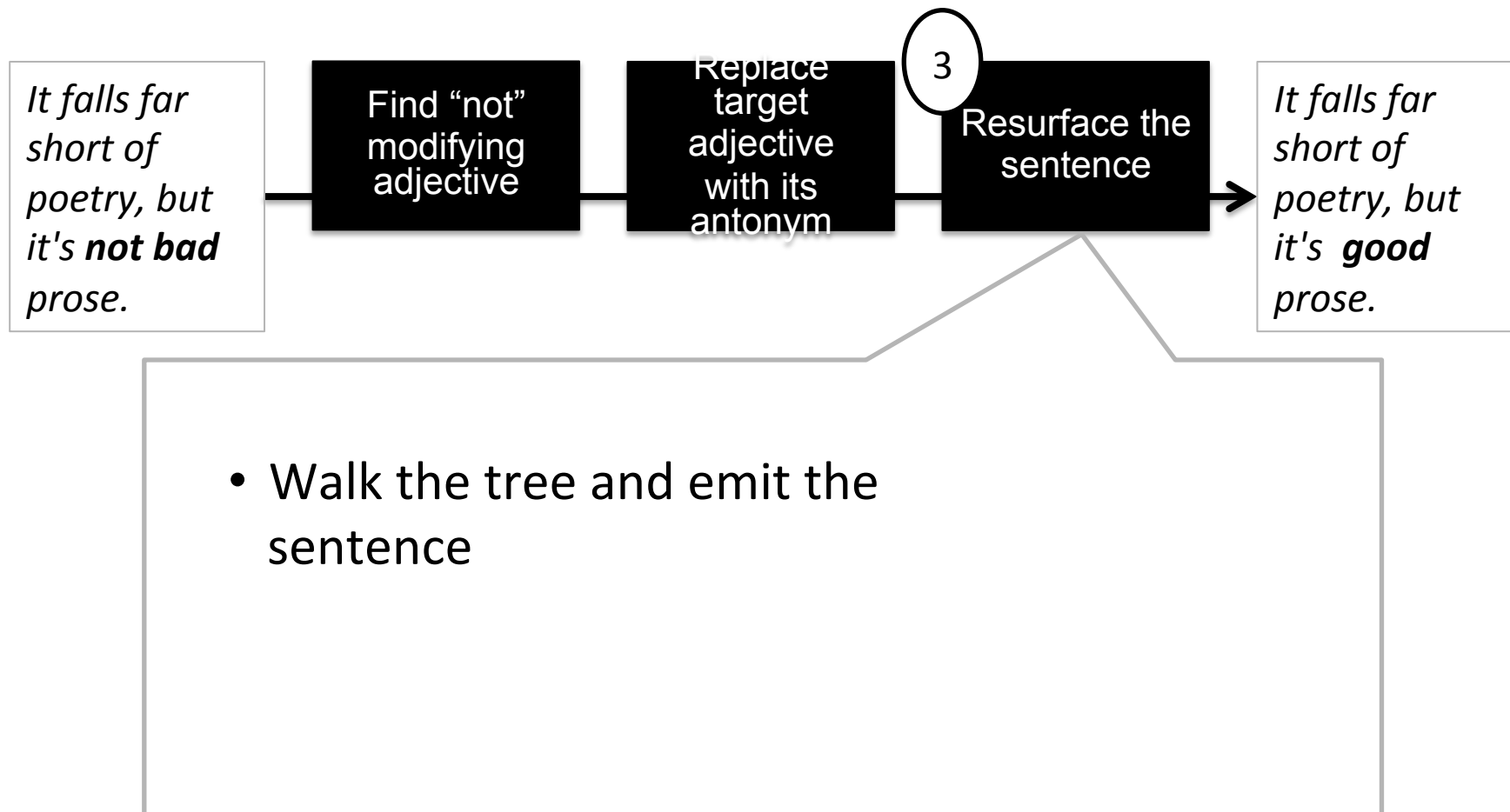


- Used Wordnet
- Find the adjective synset
- Find head synset
- Find antonym
- Replace adjective with antonym in tree



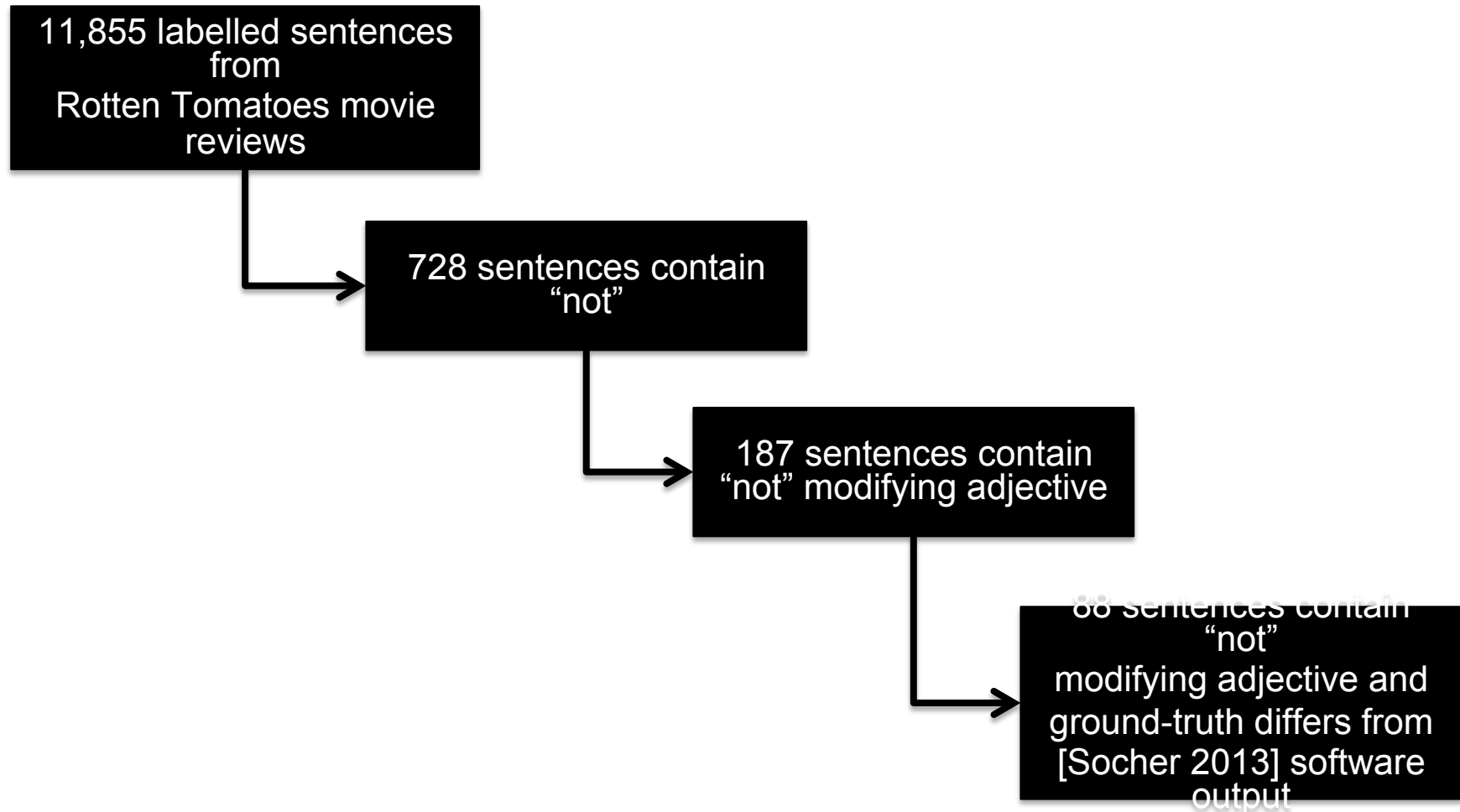
# Design & Implementation: Negation Paraphraser Pipeline

---



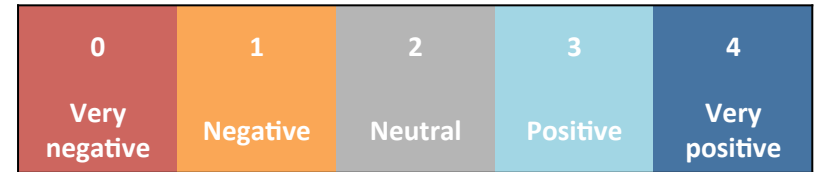
# Experiments: Data set from [Socher 2013]

---



# Results:

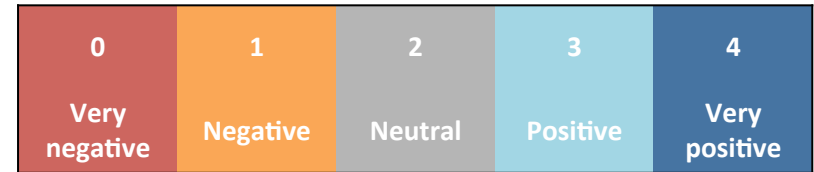
## Good examples



Input sentence	Ground-truth polarity	Output of [Socher 2013] classifier	Paraphrased sentence	Output of [Socher 2013] classifier on paraphrased
<i>S1MONE 's satire is <b>not subtle</b> , but it is effective .</i>	Positive	Negative	<i>S1MONE 's satire is <b>palpable</b> , but it is effective .</i>	Positive
<i>Certainly not a good movie , but it was <b>not horrible</b> either .</i>	Negative	Neutral	<i>Certainly a bad movie , but it was <b>innocuous</b> either .</i>	Negative
<i>At times a bit melodramatic and even a little dated (depending upon where you live), Ignorant Fairies is still quite good-natured and <b>not a bad</b> way to spend an hour or two .</i>	Positive	Negative	<i>At times a bit melodramatic and even a little dated (depending upon where you live), Ignorant Fairies is still quite good-natured and a <b>good</b> way to spend an hour or two .</i>	Positive

# Results:

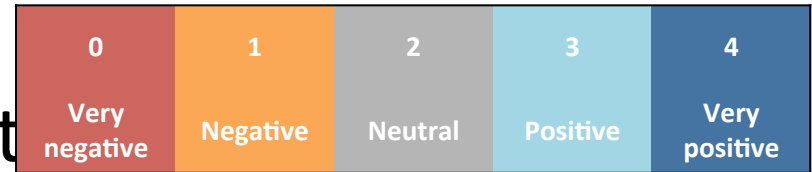
## Not good examples



Input sentence	Ground-truth polarity	Output of [Socher 2013] classifier	Paraphrased sentence	Output of [Socher 2013] classifier on paraphrased
<i>It 's one of the saddest films I have ever seen that still manages to be uplifting but <b>not</b> overly <b>sentimental</b> .</i>	<b>Very Positive</b>	<b>Negative</b>	<i>It 's one of the saddest films I have ever seen that still manages to be uplifting but overly <b>tough</b> .</i>	<b>Negative</b>
<i>It uses an old-time formula , it 's <b>not</b> terribly <b>original</b> and it 's rather messy -- but you just have to love the big , dumb , happy movie My Big Fat Greek Wedding .</i>	<b>Positive</b>	<b>Negative</b>	<i>It uses an old-time formula , it 's terribly <b>unoriginal</b> and it 's rather messy -- but you just have to love the big , dumb , happy movie My Big Fat Greek Wedding .</i>	<b>Very Negative</b>

# Results:

## Overall evaluation of 88 sent



		Predicted									
		0	1	2	3	4	0	1	2	3	4
Ground Truth	0:	0	20	1	0	0	3	14	2	2	0
	1:	3	0	2	2	0	3	2	0	2	0
	2:	2	27	0	6	0	3	18	2	12	0
	3:	0	14	1	0	0	1	9	1	4	2
	4:	0	4	0	4	0	0	3	0	4	1
<b>RMSE =</b>		<b>1.418</b>					<b>RMSE =</b>		<b>1.398</b>		

Without Negation Paraphraser

With Negation Paraphraser

# Conclusion

---

- What's right
  - Some examples demonstrate improvement
  - Overall 1.4% improvement with “not” modifying adjectives
- What's wrong
  - Generated antonyms may have wrong sense – need some disambiguation
  - Generated antonyms affected by other modifiers
  - Generated antonyms were not in training set
  - Generated antonyms simply do not affect the classifier
- What's next
  - Try out different negation structures

# Looking for Subjectivity in Medical Discharge Summaries

The Obesity NLP i2b2 Challenge (2008)

Michael Roylance and Nicholas Waltner

Tuesday 3<sup>rd</sup> June, 2014



# Paper

Journal of the American Medical Informatics Association Volume 16 Number 4 July / August 2009

561

## Focus on i2b2 **Obesity NLP Challenge**



*Viewpoint Paper* ■

### Recognizing Obesity and Comorbidities in Sparse Data

---

ÖZLEM UZUNER, PhD

# General Factoids

- The BioMedical field is awash in data.
- It is argued that up to 70% of important data about a patient is stored in *largely unstructured free text fields*<sup>1</sup>
- Although local hospitals like Swedish have heads of Informatics, there is still an active debate over how much machine learning can do to accurately diagnose patient using textual approaches.
- In spite of its enormous success in *Jeopardy!*, IBM's Watson has yet to make expected inroads in field medicine, although may well as Watson is distributed to mobile devices.
- Maybe the human doctors are the obstacle or maybe not?

---

<sup>1</sup>Please see: Shah, Stanford University.

<http://med.stanford.edu/ism/2013/april/clinical-notes.html#sthash.Gb42nykc.dpuf>.

# Task

- We worked on a medical dataset consisting of 1,237 patient discharge summaries used in the Obesity Challenge.
- Along with Obesity each patient was evaluated for an additional 15 co-morbidities such as Hypertension, Diabetes, Heart Disease, etc.
- Each patient's record was annotated using *textual* and *intuitive* classifications.
- The diseases were judged to be either Present, Absent, Questionable or Unmentioned for each patient.
- This led to a training corpus with 22,285 cases and a test one with 15,443.

# Data Set - Textual Judgements

**Table :** Distribution of Textual Judgements into Training and Test Sets

Diseases	Present		Absent		Questionable		Unmentioned		Total	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Asthma	93	68	3	2	2	2	630	432	728	504
CAD	399	277	23	22	7	2	292	196	721	497
CHF	310	205	11	11	0	0	399	280	720	496
Depression	104	72	0	0	0	0	624	434	728	506
Diabetes	485	338	15	12	7	3	219	150	726	503
GERD	118	69	1	1	5	1	599	433	723	504
Gallstones	109	87	4	2	1	0	615	418	729	507
Gout	90	52	0	0	4	0	634	453	728	505
Hypercholesterolemia	304	213	13	6	1	4	408	279	726	502
Hypertension	537	374	12	6	0	0	180	121	729	501
Hypertriglyceridemia	18	10	0	0	0	0	711	497	729	507
OA	115	86	0	0	0	0	613	416	728	502
OSA	105	69	1	0	8	2	614	432	728	503
Obesity	298	198	4	3	4	3	424	289	730	493
PVD	102	64	0	0	0	0	627	443	729	507
Venous.Insufficiency	21	10	0	0	0	0	707	497	728	507
<b>Total</b>	<b>3,208</b>	<b>2,192</b>	<b>87</b>	<b>65</b>	<b>39</b>	<b>17</b>	<b>8,296</b>	<b>5,770</b>	<b>11,630</b>	<b>8,044</b>

Notes: CAD = coronary artery disease; CHF = congestive heart failure; DM = diabetes mellitus;  
GERD = gastroesophageal reflux disease; HTN = hypertension; OSA = obstructive sleep apnea;  
OA = osteo arthritis; PVD = peripheral vascular disease.

# Data Set - Intuitive Judgements

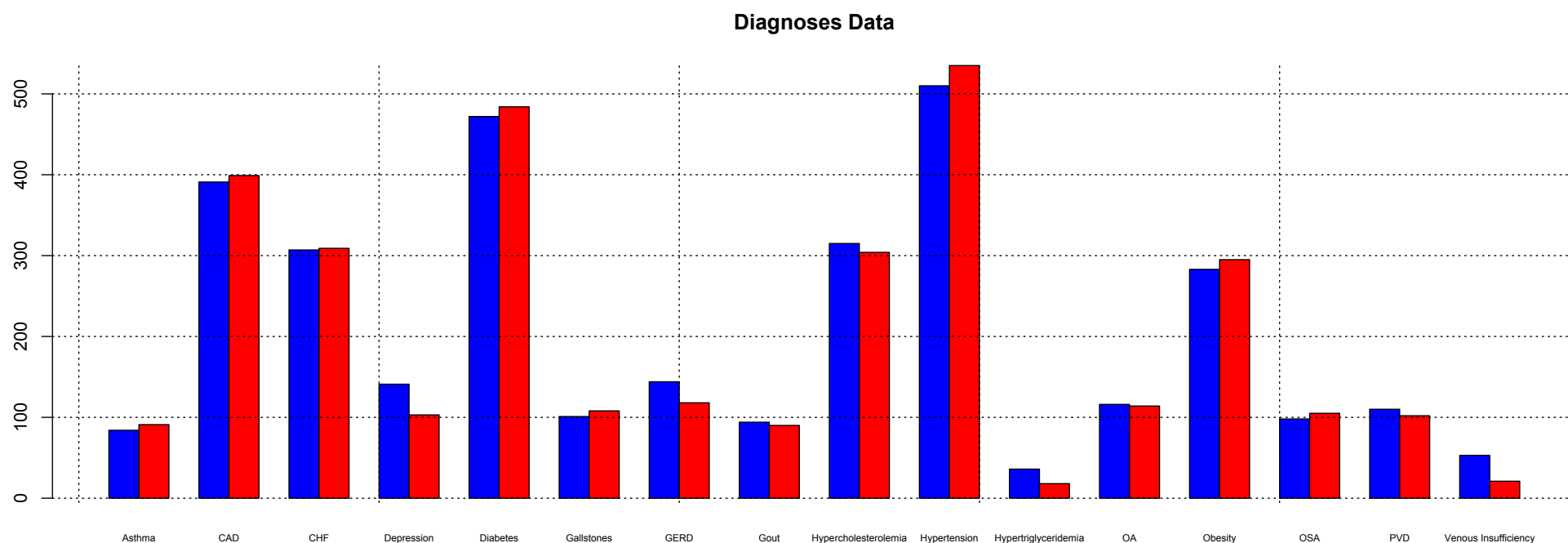
**Table :** Distribution of Intuitive Judgements into Training and Test Sets

Diseases	Present		Absent		Questionable		Unmentioned		Total	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Asthma	86	68	596	403	0	0	0	0	682	471
CAD	391	272	265	185	5	1	0	0	661	458
CHF	308	205	318	229	1	4	0	0	627	438
Depression	142	105	555	372	0	0	0	0	697	477
Diabetes	473	333	205	146	5	0	0	0	683	479
GERD	144	93	447	331	1	2	0	0	592	426
Gallstones	101	80	609	411	0	0	0	0	710	491
Gout	94	61	616	439	2	0	0	0	712	500
Hypercholesterolemia	315	242	287	189	1	0	0	0	603	431
Hypertension	511	358	127	88	0	0	0	0	638	446
Hypertriglyceridemia	37	25	665	461	0	0	0	0	702	486
OA	117	91	554	367	1	4	0	0	672	462
OSA	99	66	606	427	8	2	0	0	713	495
Obesity	285	192	379	255	1	0	0	0	665	447
PVD	110	65	556	399	1	1	0	0	667	465
Venous.Insufficiency	54	29	577	398	0	0	0	0	631	427
<b>Total</b>	<b>3,267</b>	<b>2,285</b>	<b>7,362</b>	<b>5,100</b>	<b>26</b>	<b>14</b>	<b>0</b>	<b>0</b>	<b>10,655</b>	<b>7,399</b>

Notes: CAD = coronary artery disease; CHF = congestive heart failure; DM = diabetes mellitus;  
GERD = gastroesophageal reflux disease; HTN = hypertension; OSA = obstructive sleep apnea;  
OA = osteo arthritis; PVD = peripheral vascular disease.

# Textual and Intuitive Counts

- The textual data is lumpy with the top four diseases (Hypertension, Diabetes, CAD (Coronary-Arterial) and Hypercholesterolemia) account for more than 50% of the data.
- Low frequency cases could cause classification confusion.



## Data Set - A Quick Look

- Uzner reports high agreement kappa ( $\kappa$ ) levels between annotators.
- The textual and intuitive diagnoses generally agreed quite well except for Depression, GERD, Hypertriglyceridemia and Venous Insufficiency.

Table : Agreement and Correlation between Textual and Intuitive Datasets

Diseases	Textual $\kappa$	Intuitive $\kappa$	Correlation
Asthma	0.90	0.76	0.919
CAD	0.78	0.81	0.928
CHF	0.91	0.74	0.858
Depression	0.92	0.86	<b>0.748</b>
Diabetes	0.91	0.87	0.926
GERD	0.92	0.90	<b>0.763</b>
Gallstones	0.89	0.59	0.956
Gout	0.93	0.92	0.885
Hypercholesterolemia	0.87	0.68	0.851
Hypertension	0.82	0.67	0.808
Hypertriglyceridemia	0.71	0.72	<b>0.523</b>
OA	0.91	0.86	0.815
OSA	0.92	0.92	0.933
Obesity	0.76	0.76	0.872
PVD	0.94	0.73	0.907
VenousInsufficiency	0.79	0.44	<b>0.473</b>
<b>Averages</b>	0.87	0.76	0.820

# Competition Results

30 teams submitted results...textual macro-average F-scores were between 0.61 and 0.80 for the top ten teams.

Table 7 Micro- and Macro-averaged Results on Textual Judgments, Sorted by Macro-averaged F-Measure

Systems	Macro-Averaged			Micro-Averaged		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Yang et al.	0.8482	0.7737	<b>0.8052</b>	0.9723	0.9723	0.9723
Solt et al.	0.8318	0.7776	0.8000	0.9756	0.9756	0.9756
Ware et al.	0.8314	0.7542	0.7821	0.9718	0.9718	0.9718
Childs et al.	0.8169	0.7454	0.7762	0.9773	0.9773	<b>0.9773</b>
Mishra et al.	0.7485	0.8050	0.7718	0.9704	0.9704	0.9704
Szarvas et al.	0.7644	0.7600	0.7622	0.9729	0.9729	0.9729
Savova et al.	0.7701	0.7147	0.7377	0.9668	0.9668	0.9668
Patrick et al.	0.7971	0.6219	0.6737	0.9693	0.9693	0.9693
* Jazayeri et al.	0.7849	0.5779	0.6205	0.9514	0.9514	0.9514
†DeShazo et al.	0.8552	0.6240	0.6140	0.9639	0.9639	0.9639



# Competition Results

30 teams submitted results...intuitive results were lower at 0.63 to 0.67, as one might expect.

Table 9 Micro- and Macro-averaged Results on Intuitive Judgments, Sorted by Macro-averaged F-Measure

Systems	Macro-Averaged			Micro-Averaged		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Solt et al.	0.7485	0.6571	<b>0.6745</b>	0.9590	0.9590	0.9590
Szarvas et al.	0.6999	0.6588	0.6727	0.9642	0.9642	0.9642
Childs et al.	0.7061	0.6540	0.6696	0.9582	0.9582	0.9582
Ware et al.	0.6410	0.6399	0.6404	0.9654	0.9654	<b>0.9654</b>
Ambert et al.	0.6383	0.6307	0.6344	0.9558	0.9558	0.9558
Meystre	0.6304	0.6387	0.6343	0.9566	0.9566	0.9566
Yang et al.	0.6383	0.6294	0.6336	0.9572	0.9572	0.9572
†DeShazo et al.	0.9722	0.6216	0.6292	0.9524	0.9523	0.9524
Matthews	0.6325	0.6256	0.6288	0.9509	0.9509	0.9509
Jazayeri et al.	0.6320	0.6257	0.6287	0.9508	0.9508	0.9508

# Take Aways

What did we learn from the paper:

- Most of the team did not rely super-heavily on pure ML, rather rule building on “standard language” seem to dominate the systems along with a lot of work on the naming of various diseases, etc.
- Intuitive judgements seem to be harder to machine learning (not so surprising).
- Each patient was diagnosed with 4.36 diseases - are the diseases similar or is there confusion?
- Possibly, sentiment measures could improve over a baseline, especially in areas where there was not strong agreement between textual and intuitive annotation, i.e. the human knew something that was not obvious in the text or vice versa.

# Methodology

We obtained the dataset from i2b2 organization in XML format.

- Built a MySql database to house the data and build various tables around the data.
- Basic scrubbing and ETL (Extract, Transform and Load) was performed in Python and Perl.
- Used the Stanford Parser for POS tagging.
- Classification was done using Mallet andSKLearn (very handy especially with micro- and macro-averaging).
- Established a two class baseline (Present and Absent) and then added sentiment/subjectivity features.

# Comp Ling Issues

As Gina pointed out in Week 6, “biomedical texts are not really English” !!!!

- POS X comes up nearly 30% of the time.
- Punctuation is very heavy owing to abbreviations.

Table : Part of Speech Counts

POS	Count	Percentage	POS	Count	Percentage
X	354,165	28.4	CC	28,902	2.3
NN	198,815	15.9	VBN	28,441	2.3
PUNC	147,095	11.8	RB	28,031	2.2
NNP	124,185	9.9	VB	20,515	1.6
JJ	93,352	7.5	PRP	18,060	1.4
IN	91,270	7.3	TO	17,915	1.4
CD	66,893	5.4	VBZ	16,474	1.3
DT	54,860	4.4	PRP\$	12,653	1.0
VBD	46,635	3.7	VBP	10,895	0.9
NNS	46,234	3.7	VBG	9,972	0.8

# Results

Sentiment and subjectivity features in many cases lowered classification accuracy. However, notable gains were found in the intuitive categories.

Table : Classification Results

Category	Sub-Task	Micro/Macro Intuitive	Micro/Macro Textual	Comment
Base Line	Uni-gram without StopWords	47.6 / 83.1	51.6 / 87.1	
	without X POS	39.1 / 72.5	40.9 / 73.1	
	without X -LBR- -RRB . , etc	39.1 / 72.5	40.9 / 73.1	
POS Tags	Pronouns-only	47.4 / 82.4	51.3 / 87.0	
	Nouns-only	47.4 / 82.1	50.2 / 84.7	
	Verbs-only	45.0 / 76.6	48.5 / 84.4	
	Adjectives-only	46.6 / 80.5	49.6 / 85.0	
	Adverbs-only	47.2 / 78.9	50.5 / 85.7	
	Adjectives and Adverbs-only	45.6 / 75.9	49.3 / 83.0	
	All Tags	47.9 / 80.2	51.0 / 86.0	
Polarity	Simple (positive/negative counts)	48.0 / 80.2	51.0 / 56.0	
	Complex (positive weak, positive strong)	47.2 / 82.6	51.3 / 86.8	
Combinations	Simple Polarity without X	39.2 / 73.2	40.8 / 72.3	
	Complex Polarity without X	39.5 / 71.8	40.8 / 71.6	
Other	Unique Words per Diagnosis	46.4 / 65.1	46.6 / 69.9	
	Highest Probability Words per Diagnosis	46.1 / 76.5	48.2 / 74.7	

# Initial Conclusions

## Did we fail or is something else going on?

- It may simply be the case that medical literature is largely absent emotive descriptions of patient discharge summaries.
- Alternatively, it may simply be the case that standard lexicons of subjectivity are insufficient for the medical domain.
- However, it is clear that there is a high degree of correlation between the various diseases.
- Hence, a more interesting question might be to ask whether there are fundamental drivers underneath these 16 diseases?
- Perhaps, unsupervised machine learning techniques can shed further light on what we already know?

# An Unsupervised Approach

Both cluster and principal component analysis indicate that there is a higher structure to the co-morbidity data. PCA indicates that five-factors explain 50% of the variance in patient diagnoses...

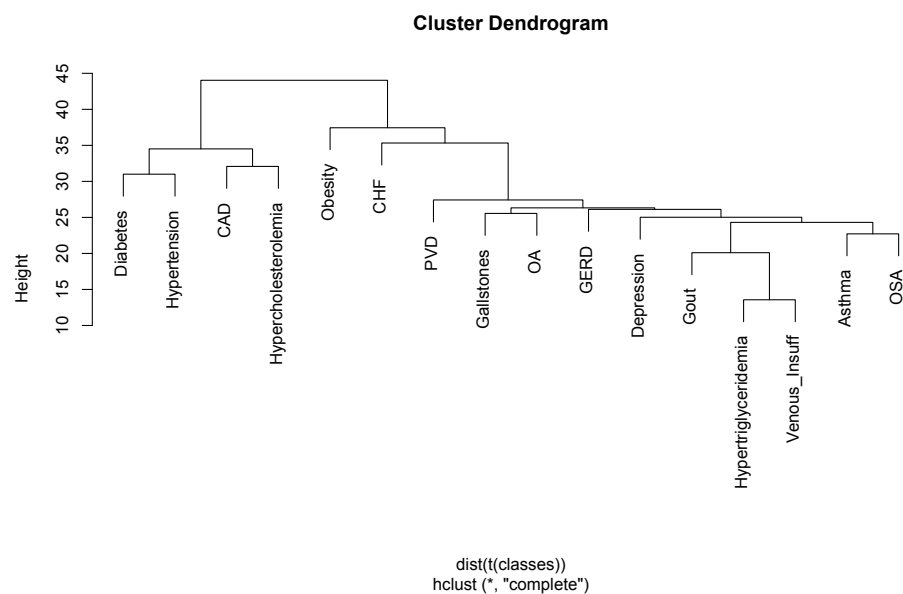


Figure : Simple Clustering

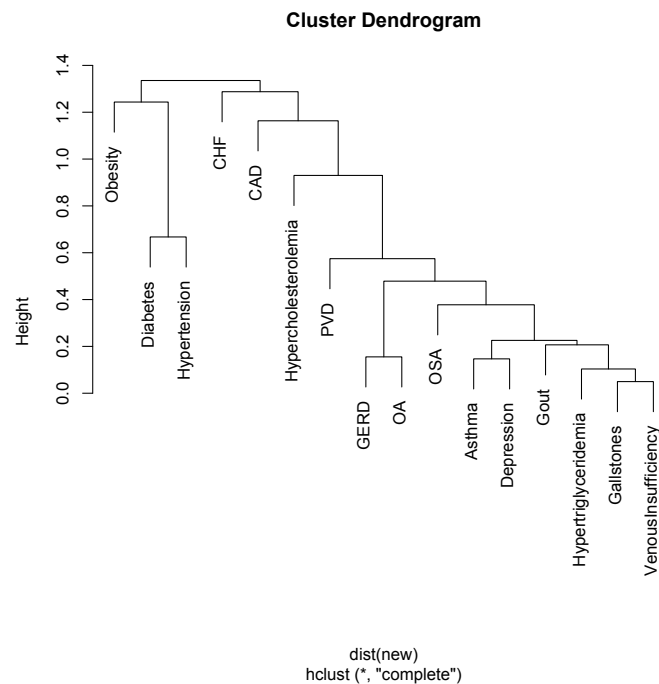


Figure : Five Factor PCA Model

# Final Write-Up

## Further items to research:

- Can combining both textual and intuitive features provide a better basis for diagnosis?
- Can other features be added to improve subjectivity accuracy?
- Can a decision tree be developed to arrive in the most likely disease cluster versus ending up with multiple diagnoses?



LING 575

Aspect Based Feature Selection  
For Review Sentiment Analysis

T.J. Trimble & Yi-Shu Wei

# Intro

- Create a baseline classifier for sentence level sentiment analysis using MALLET
- Find aspect based features using MALLET LDA
- Do feature selection over unigram features using LDA models

# Task

- Restaurant review data (Ganu et al., 2009)
  - Sentence level sentiment annotations with aspects
- Classify sentences as positive, negative, or neutral
- Throw out “conflicted” sentences

# Data

- 3400 sentences in total, ~5 per review
  - Each sentence labeled with positive, negative, neutral, or conflicted
  - Each sentence labeled with aspect, such as “food”, “service”, “price”, etc.
- Reviews randomly split into 60/20/20 train, dev, test sets

# Outline

- Baseline classifier: MALLET
  - Experiment with features
    - Unigrams, bigrams, removing stop words
    - “not” and “but” shallow analysis
  - Experiment with feature selection
    - Sentiment lexicon
  - Experiment with algorithms
    - MaxEnt, Decision Tree, Naïve Bayes

# Baseline Classifier

- Unigram + bigram features, binarized
  - Bigrams didn't help, so didn't try trigrams
- MALLET's --remove-stop-words feature
- Kept “!,.?”

# Baseline Classifier

- Shallow structural analysis
  - Replace all words after “but” and “not” with “but\_word” and “not\_word”
  - No improvement
    - maybe can be improved?

# Baseline Classifier

- Sentiment Lexicon (Hu and Liu 2004)
  - Tried two things:
    - In negative sentences, only keep negative words, and vice versa
    - In negative sentences, keep neutral words and negative words, and vice versa
      - In neutral sentences, only keep neutral words



# Baseline Classifier

- Sentiment Lexicon
  - Keeping more words helped

# Baseline Classifier

- Algorithms (from MALLET)
  - Decision Tree
  - MaxEnt
  - NaiveBayes
- No tuning, but MaxEnt worked best out of the box!

# Baseline Results: trainers

U: unigrams; B: bigrams; Bin: binarized; S: stop-words removed;  
Lex: sentiment lexicon used; Struct: shallow structure analysis

	Decision Tree		Naive Bayes		MaxEnt	
	Train	Dev	Train	Dev	Train	Dev
U	0.598	0.544	0.852	0.643	0.979	<b>0.672</b>
U + Bin	0.598	0.544	0.847	0.637	0.979	0.651
B	0.602	0.548	0.988	0.578	0.994	0.594
B + Bin	0.602	0.548	0.988	0.576	0.994	0.595
U + B	0.601	0.544	0.960	0.639	0.997	<b>0.672</b>
U + B + Bin	0.601	0.544	0.960	0.651	0.997	0.661

# Baseline Results: lexicon

U: unigrams; B: bigrams; Bin: binarized; S: stop-words removed;  
Lex: sentiment lexicon used; Struct: shallow structure analysis

	MaxEnt	
	Train	Dev
U	0.979	0.672
U + Struct	0.983	0.669
U + B + Struct	0.997	0.670
<b>U + Lex</b>	0.984	<b>0.730</b>
U + B + Lex	0.998	0.697
U + Struct + Lex	0.987	0.714
U + Bin + Struct + Lex	0.988	0.717
U + B + Bin + Struct + Lex	0.998	0.691

# LDA

- Detect topics in data (unsupervised)
- Each document is a mixture of topics
- Find keywords for each topic

# LDA

- Use mallet train-topics
  - number of topics = 10; we did not tune hyperparameters
  - can choose word- or phrase-based...
  - “MAP” LDA: split training data into 10 parts based on the most probable topic
  - each part has 145-243 sentences

- 0 ve place dinner times nyc area side bit time years lunch live restaurant bad romantic small expect st couple
- 1 service food great prices excellent staff friendly attentive quality decor good atmosphere delicious restaurant wonderful money fun thing makes
- 2 food place restaurant restaurants love indian thai authentic city favorite cuisine japanese chinese pay places street give neighborhood italian
- 3 chicken rice fish hot spicy dish sauce thai ordered shrimp special beef rolls fresh dishes soup fried tuna curry
- 4 delicious menu food dessert portions steak pasta made huge appetizer appetizers ve del fresh large salad order dishes cheap
- 5 back restaurant night place friends time recommend highly dinner family check friend recommended reviews saturday day coming sushi boyfriend
- 6 pizza good sum dim taste bagels sushi nyc cold menu places sandwich lobster slice cheese rest overpriced make price
- 7 table back wait times ca restaurant people asked waiter experience minutes order seated time left group bar make wrong
- 8 place great good food service pretty spot average atmosphere excellent late lunch cool night perfect found sit date ambiance
- 9 great wine good worth amazing meal food eat list visit menu deal selection priced price drinks glass decent house

# Feature selection

- Assign a score to each token based on its mutual information with “positive” or “negative” labels
- Do a general feature selection, and then do for each of the 10 parts



# Feature selection

great not delicious best n't was worst excellent  
do good just rude slow overpriced attentive no  
order nothing twice table our bland wonderful  
to is recommend we perfect seem horrible past  
although told us try after authentic work that  
took were average tables oily together annoying  
disappointment awful watery unless bill section  
problem tasted loud friendly she asked bring  
minutes but amazing

- 0 best i worst twice not cozy level took out the way was go as never once lived years return 's spot first service employees sitting tried noise putting then them also without overpriced his tables come expensive these editorial bucks better an were about been over
- 1 n't at do great slow friendly attentive and nothing problem money or very something least because am bo should takes visit broke green over lava nyc g times twenty say clean probably commend middle now skeptical seating oh o slightly official distract cake came when
- 2 no with best be italian one ' not pay much authentic msg lousy using overrated all foods who folks my to would new live call quite why than actually ordering favorite compared simple do out city have indian chinese makes idea many am among definitely great like asian
- 3 was were oily best bland try dry is since get classic into recommend ordered love small disgusting watery got nothing fun could here delicious a section just and not flavor from tasted even dish there its an would sea seafood where couple i chinese seasoning few
- 4 delicious n't parisian fondue tasted joke expensive drink ok did main fresh appetizer been steak huge pasta eat selection lot now all would entree was were small great is stinks dip cosette cozy path mozzarella changes medium excuse share after loyalty artisanal

- 5 back after n't recommend went service the half friend by great finally very door past not about highly again what work experience did worst order entire could disappointment off know while first decided made reservation everyone we for girls pm so has my night
- 6 the nite best overpriced good food try menu slice th must special sell eaten wish like ever worst nothing busy n't not was you sure ok late there no so in of is nyc it for really when plain out all bagels 's taste a at with family pastrami joe both two went walk snapple
- 7 go our to back or waitress service great definitely best was never over rude loud small order table all could glasses at promptly night and water understand us awful she someone ca every appetizers begin believe line asked up reservations as conversation 'm enough
- 8 horrible great good service and not away stay average other be crowd ues sweetness heard should tiffin big late loved if yeah due same is n't very perfect i fact no cost rude took expensive tables around down excellent from did restaurants can lunch been amount
- 9 not off restaurant such rao nothing know 've dinner amazing least twice italian i great ent rees annoying new disappointed out it to wine well be wait just and worth are month 're when on weird after pay only limited thing wanted could table would without take never

# Conclusion

- Unigrams very effective
- MaxEnt works great
- Sentiment lexicon very effective
  - [Hopefully] topic specific feature selection will work even better!