Genres: Discourse, Speech, and Tweets

Sentiment, Subjectivity & Stance
Ling 575
April 15, 2014

Roadmap

- Effects of genre on sentiment:
  - Spoken multi-party dialog
    - Guest lecturer: Valerie Freeman
  - Discourse and dialog (from text)
  - Tweets
- Examples: State-of-the-art
- Course mechanics

Sentiment in Speech

- Key contrasts:
  - Acoustic channel carries additional information
    - Speaking rate, loudness, intonation
    - Hyperarticulation
  - Conversational:
    - Utterances short, elliptical, disfluent
  - Multi-party:
    - Turn-taking, inter-speaker relations
  - Discourse factors

Discourse & Dialog

Sentiment in Discourse & Dialog

- Many sentiment-bearing docs are discourses
  - Extended spans of text or speech
    - E.g. Amazon product reviews, OpenTable, blogs, etc
- However, discourse factors often ignored
  - Structure:
    - Sequential structure
    - Topical structure
  - Dialog
    - Relations among participants
    - Relations among sides/stances

Discourse Factors

- Sentiment within a doc not simple aggregation
  - I hate the Spice Girls. ... [3 things the author hates about them] ... Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But... Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie, ... [they] act wacky as hell...the ninth floor of hell...a cheap [beep] movie...The plot is such a mess that it's terrible.

Discourse Factors

- Sentiment within a doc not simple aggregation
  - I hate the Spice Girls. ... [3 things the author hates about them] ... Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But... Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie, ... [they] act wacky as hell...the ninth floor of hell...a cheap [beep] movie...The plot is such a mess that it's terrible. **But I loved it**

Discourse Factors

- Sentiment within a doc not simple aggregation
  - I hate the Spice Girls. ... [3 things the author hates about them] ... Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But... Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie, ... [they] act wacky as hell...the ninth floor of hell...a cheap [beep] movie...The plot is such a mess that it's terrible. **But I loved it**
- What would bag-of-words say?

Discourse Factors

- Sentiment within a doc not simple aggregation
  - I hate the Spice Girls. ... [3 things the author hates about them] ... Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But... Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie, ... [they] act wacky as hell...the ninth floor of hell...a cheap [beep] movie...The plot is such a mess that it's terrible. **But I loved it**
- What would bag-of-words say? Negative
- Possible simple solution

Discourse Factors

- Sentiment within a doc not simple aggregation
  - I hate the Spice Girls. ... [3 things the author hates about them] ... Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But... Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie, ... [they] act wacky as hell...the ninth floor of hell...a cheap [beep] movie...The plot is such a mess that it's terrible. **But I loved it**
- What would bag-of-words say? Negative
- Possible simple solution: position-tagged features

Discourse Factors: Structure

- Sentiment within a doc not simple aggregation
  - I hate the Spice Girls. ... [3 things the author hates about them] ... Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But... Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie, ... [they] act wacky as hell...the ninth floor of hell...a cheap [beep] movie...The plot is such a mess that it's terrible. **But I loved it**
- What would bag-of-words say? Negative
- Possible simple solution: position-tagged features
  - Sadly no better than bag-of-words

Discourse Factors: Structure

- Summarization baseline:
  - In newswire topic summarization:

Discourse Factors: Structure

- Summarization baseline:
  - In newswire topic summarization:
    - First few sentences
      - Headline, lede
    - Often used as strong baseline in evaluations
  - In subjective reviews:

Discourse Factors: Structure

- Summarization baseline:
  - In newswire topic summarization:
    - First few sentences
      - Headline, lede
    - Often used as strong baseline in evaluations
  - In subjective reviews:
    - Last few lines
    - “Thwarted expectations”

Discourse Factors: Structure

- Summarization baseline:
  - In newswire topic summarization:
    - First few sentences
      - Headline, lede
    - Often used as strong baseline in evaluations
  - In subjective reviews:
    - Last few lines
    - "Thwarted expectations"
- Last n sentences of review much better summary
  - Than first n lines
  - Competitive with n most subjective sents overall

Discourse Factors: Cohesion

- Inspired by lexical chains in discourse analysis
  - Document cohesion influenced by topic repetition

Discourse Factors: Cohesion

- Inspired by lexical chains in discourse analysis
  - Document cohesion influenced by topic repetition
- Idea:
  - Neighboring sentences (often) have similar
    - Subjectivity status
    - Sentiment polarity

Discourse Factors: Cohesion

- Inspired by lexical chains in discourse analysis
  - Document cohesion influenced by topic repetition
- Idea:
  - Neighboring sentences (often) have similar
    - Subjectivity status
    - Sentiment polarity
- Approach:
  - Use baseline sentence level classifier
  - Improve with information from neighboring sentences
    - ‘sentiment flow’, min-cut (subj), other graph-based models

Discourse Factors: Dialog Participants

- Relations among dialog participants informative
- Online debates (Agrawal et al)
  - Patterns in ‘responded to’ and ‘quoted’ relations

Discourse Factors: Dialog Participants

- Relations among dialog participants informative
- Online debates (Agrawal et al)
  - Patterns in ‘responded to’ and ‘quoted’ relations
    - 74% of responses → opposing stance
      - Only 7% reinforcing
    - Quotes also generally drawn from opposing side

Discourse Factors: Dialog Participants

- Relations among dialog participants informative
- Online debates (Agrawal et al)
  - Patterns in ‘responded to’ and ‘quoted’ relations
    - 74% of responses → opposing stance
      - Only 7% reinforcing
    - Quotes also generally drawn from opposing side
- Application:
  - How can we group individuals by stance?

Discourse Factors: Dialog Participants

- Relations among dialog participants informative
- Online debates (Agrawal et al)
  - Patterns in 'responded to' and 'quoted' relations
    - 74% of responses → opposing stance
      - Only 7% reinforcing
    - Quotes also generally drawn from opposing side
- Application:
  - How can we group individuals by stance?
    - Cluster those who quote/respond to same individuals

Discourse Factors: Dialog Participants

- Beyond quoting in Congressional floor debates
  - Build on classifier for pro/con

Discourse Factors: Dialog Participants

- Beyond quoting in Congressional floor debates
  - Build on classifier for pro/con
  - Build another classifier to tag references to others as
    - Agreement/disagreement
  - Employ agreement/disagreement network as constraint

Discourse Factors: Dialog Participants

- Beyond quoting in Congressional floor debates
  - Build on classifier for pro/con
  - Build another classifier to tag references to others as
    - Agreement/disagreement
  - Employ agreement/disagreement network as constraint
  - Yields an improvement in pro/con classification alone

Sentiment in Twitter

- Reverse of discourse/dialog setting
  - Extremely short content: 140 characters
    - Related: SMS
- Distinguishing characteristics:

Sentiment in Twitter

- Reverse of discourse/dialog setting
  - Extremely short content: 140 characters
    - Related: SMS
- Distinguishing characteristics:
  - Length
  - Emoticons, Hashtags, userids
  - Retweets
  - Punctuation
  - Spelling/jargon
  - Structure

SEMEVAL 2013 Task

- Twitter sentiment task:
  - Usual shared task goals
    - Standard, available annotated corpus; fixed tasks, resource
    - Amazon Mechanical Turk labeling

SEMEVAL 2013 Task

- Twitter sentiment task:
  - Usual shared task goals
    - Standard, available annotated corpus; fixed tasks, resource
    - Amazon Mechanical Turk labeling
- Two subtasks:
  - Term-level: identify sentiment of specific term in context
  - Message-level: identify overall sentiment of message

SEMEVAL 2013 Task

- Twitter sentiment task:
  - Usual shared task goals
    - Standard, available annotated corpus; fixed tasks, resource
    - Amazon Mechanical Turk labeling
- Two subtasks:
  - Term-level: identify sentiment of specific term in context
  - Message-level: identify overall sentiment of message
- ~13K tweets: train/dev/test splits
  - ~2K SMS for comparison: test only

Overall

- Overall:
  - Term-level easier than message-level,
  - Tweets easier than SMS (adaptation)
- Total # teams: 44

Overall

- Overall:
  - Term-level easier than message-level,
  - Tweets easier than SMS (adaptation)
- Total # teams: 44
- Best system: NRC-Canada
  - Top in all but one condition
  - Message-level: 69 F-score
  - Term-level: ~89 F-score

Message-level Sentiment

- Classifier: SVM, linear kernel
- Lots of features:
  - Ngrams: word: 1,2,3,4+skip; char: 3,4,5 (binary)
    - Incorporates 'NEG' tagging

Message-level Sentiment

- Classifier: SVM, linear kernel
- Lots of features:
  - Ngrams: word: 1,2,3,4+skip; char: 3,4,5 (binary)
    - Incorporates 'NEG' tagging
  - Counts:
    - # all caps, # each POS tag, # hashtags, # contiguous punc
    - # elongated words, # negated contexts

Message-level Sentiment

- Classifier: SVM, linear kernel
- Lots of features:
  - Ngrams: word: 1,2,3,4+skip; char: 3,4,5 (binary)
    - Incorporates 'NEG' tagging
  - Counts:
    - # all caps, # each POS tag, # hashtags, # contiguous punc
    - # elongated words, # negated contexts
  - Position features:
    - Where the last token is: !/? or pos/neg emoticon

Message-level Sentiment

- Classifier: SVM, linear kernel
- Lots of features:
  - Ngrams: word: 1,2,3,4+skip; char: 3,4,5 (binary)
    - Incorporates 'NEG' tagging
  - Counts:
    - # all caps, # each POS tag, # hashtags, # contiguous punc
    - # elongated words, # negated contexts
  - Position features:
    - Where the last token is: !/? or pos/neg emoticon
  - Presence of pos/neg emoticons or Brown cluster wds

Message-level Sentiment

- Main novelty:
  - Lexical features
    - Manually constructed lexicons:
      - NRC emotion lexicon, MPQA, Bing Liu's lexicon
    - Two automatically constructed lexicons

Message-level Sentiment

- Main novelty:
  - Lexical features
    - Manually constructed lexicons:
      - NRC emotion lexicon, MPQA, Bing Liu's lexicon
    - Two automatically constructed lexicons
  - All tokens scored:
    - word unigrams, bigram, skip bigrams, POS tags, hashtags, all caps words
    - Score(w) = PMI(w, positive) – PMI(w, negative)

Message-level Sentiment

- Main novelty:
  - Lexical features
    - Manually constructed lexicons:
      - NRC emotion lexicon, MPQA, Bing Liu's lexicon
    - Two automatically constructed lexicons
  - All tokens scored:
    - word unigrams, bigram, skip bigrams, POS tags, hashtags, all caps words
    - Score(w) = PMI(w, positive) – PMI(w, negative)
    - Features include: # wds w/positive score, total score, max score, last positive score

Lexicon Induction

- Bootstrapping approach:
  - Key insight:
    - Hashtagged emotion words good cues to tweet as whole
      - E.g. joy, sadness, etc
    - Use as noisy tags for large corpus

Lexicon Induction

- Bootstrapping approach:
  - Key insight:
    - Hashtagged emotion words good cues to tweet as whole
      - E.g. joy, sadness, etc
    - Use as noisy tags for large corpus
- Strategy: Poll twitter API
  - Use collection of positive and negative seed hashtags
  - 775K tagged tweets
    - Positive if has one of the positive hashtags
    - Negative if has one of the negative hashtags
  - Use to train word-polarity association scores

Lexicons

- Applied to Twitter corpus
  - 54K unigrams, 316K bigrams, 308K pairs

- Also applied to sentiment140 corpus
  - Similar strategy, but cued on emoticons
  - 62K unigrams, 677K bigrams, 480K pairs

Message Classification Feature Analysis

Experiment	Tweets	SMS
all features	69.02	68.46
all - lexicons	60.42 (**-8.60**)	59.73 (**-8.73**)
all - manual lex.	67.45 (-1.57)	65.64 (-2.82)
all - auto. lex.	63.78 (-5.24)	67.12 (-1.34)
all - Senti140 lex.	65.25 (-3.77)	67.33 (-1.13)
all - Hashtag lex.	65.22 (-3.80)	70.28 (1.82)
all - ngrams	61.77 (-7.25)	67.27 (-1.19)
all - word ngrams	64.64 (-4.38)	66.56 (-1.9)
all - char. ngrams	67.10 (-1.92)	68.94 (0.48)
all - negation	67.20 (-1.82)	66.22 (-2.24)
all - POS	68.38 (-0.64)	67.07 (-1.39)
all - clusters	69.01 (-0.01)	68.10 (-0.36)
all - encodings (elongated, emoticons, punctuations, all-caps, hashtags)	69.16 (0.14)	68.28 (-0.18)

Term Classification Feature Analysis

Experiment	Tweets	SMS
all features	89.10	88.34
all - ngrams	83.86 (**-5.24**)	80.49 (**-7.85**)
all - word ngrams	88.38 (-0.72)	87.37 (-0.97)
all - char. ngrams	89.01 (-0.09)	87.31 (-1.03)
all - lexicons	85.15 (-3.95)	83.70 (-4.64)
all - manual lex.	87.69 (-1.41)	86.84 (-1.5)
all - auto lex.	88.24 (-0.86)	86.65 (-1.69)
all - negation	88.38 (-0.72)	86.77 (-1.57)
all - stopwords	89.17 (0.07)	88.30 (-0.04)
all - encodings (elongated words, emoticons, punctns., uppercase)	89.16 (0.06)	88.39 (0.05)
all - target	72.97 (**-16.13**)	68.96 (**-19.38**)
all - context	85.02 (-4.08)	85.93 (-2.41)

Discussion

- Message level analysis
- Lexical features very important
  - Automatically induced lexicons significant contribution
    - Roughly 5 points of F-score
      - For tweets only: doesn't carry over to SMS

Discussion

- Message level analysis
- Lexical features very important
  - Automatically induced lexicons significant contribution
    - Roughly 5 points of F-score
      - For tweets only: doesn't carry over to SMS
- N-grams next most important
  - Words somewhat higher (in tweets)

Discussion

- Message level analysis
- Lexical features very important
  - Automatically induced lexicons significant contribution
    - Roughly 5 points of F-score
      - For tweets only: doesn't carry over to SMS
- N-grams next most important
  - Words somewhat higher (in tweets)
- Encoding features:
  - Minimal impact:

Discussion

- Message level analysis
- Lexical features very important
  - Automatically induced lexicons significant contribution
    - Roughly 5 points of F-score
      - For tweets only: doesn’t carry over to SMS
- N-grams next most important
  - Words somewhat higher (in tweets)
- Encoding features:
  - Minimal impact:
    - Redundant with lexical and ngram features

Discussion II

- Term-level analysis
- Similar features, but focused on target term
  - Along with term length, term split features

Discussion II

- Term-level analysis
- Similar features, but focused on target term
  - Along with term length, term split features
- N-gram features biggest impact
  - 5-7 points F-score
    - Data fits well: 85% of target terms seen in training

Discussion II

- Term-level analysis
- Similar features, but focused on target term
  - Along with term length, term split features
- N-gram features biggest impact
  - 5-7 points F-score
    - Data fits well: 85% of target terms seen in training
- Lexicon features next
  - Impact of manual lexicon less clear cut

Summary

- Sentiment classification is not just text classification
  - Differs in response to many factors
    - Tokenization, stemming, POS tagging, negation…
- Baseline ML polarity classification
  - Built on (adapted) bag-of-words models
  - Draws on machine learning approaches
- Can be enhanced through improved linguistic, context features
- Similarities & differences across genres

Course Mechanics

- Individual
  - Critical reading assignments
    - Weekly – one paper
- Groups of 2-3
  - Lead topic presentation/discussion: once
    - Select from list of topics, readings
    - Analyze, discuss in class
  - Term project
    - Explore specific topic in depth
      - Can implementation or analysis + write-up
      - Linguistics elective: talk to me

Datasets

- Diverse data sets:
  - Web sites: Lillian Lee's and Bing Liu's
- Movie review corpora
- Amazon product review corpus
- Online and Congressional floor debate corpora
- Multi-lingual corpora: esp. NTCIR
- MPQA subjectivity annotation news corpus