Review Mining

Soo-Min Lim and Eduard Hovy. (2006). Automatic Identification of Pro and Con Reasons in Online Reviews. COLING-ACL-2006.

and

Oscar Tackstrom and Ryan McDonald (2011). Discovering Fine-Grained Sentiment with Latent Variable Structured Prediction Models. ECIR-2011.

Automatic Identification of Pro and Con Reasons in Online Reviews
Overview

- Goal:
    - Extract sentences that <u>explain</u> the sentiment of reviews (pros/cons)
- Difficulties:
    - No/little labeled data
    - Pros/cons may be objective sentences
        - e.g., “the battery life lasts 3 hours”
    - Domain-specificity

Automatic Identification of Pro and Con Reasons in Online Reviews
Overview

- Focus on reasons for opinions
  - reason may be objective statement
- 2 steps:
  - generate training data by aligning pros and cons with opinion-bearing sentences
  - train MaxEnt classifier to automatically identify pros and cons
- Training data: epinions.com, <review text, pros, cons> triplets
- MaxEnt classification in 2 parts:
  - identification phase
  - classification phase
    - features: lexical, positional, opinion-bearing words
- Testing data: complaints.com

Automatic Identification of Pro and Con Reasons in Online Reviews

Intuitions

- MaxEnt: “best model is the one that is consistent with the set of constraints imposed by the evidence but otherwise is as uniform as possible”
- Lexical features: “there are certain words that are frequently used in pro and con sentences which are likely to represent reasons why an author writes a review”
- Positional features: “important sentences that contain topics in a text have certain positional patterns”
- Opinion-bearing word features: capture pro and con sentences which opinion-bearing expressions (objective sentences should be captured by lex and pos features)

Automatic Identification of Pro and Con Reasons in Online Reviews
Discussion

- Novel part of paper is alignment step, but there is no explicit evaluation of this step
- Pro/con dictionary baseline for identification?
- Why where identification and classification separate steps?
    - Could do identification of cons, identification of pros
- Training set balanced differently than test set
    - epinions.com -- more positive reviews
    - complaints.com -- mostly negative
- “The average accuracy 68.0% is comparable with the pair-wise human agreement 82.1%” (baseline 59.9%) -- ???
- Best accuracy and recall on restaurant complaints, best precision on mp3 complaints
- Captured both opinion-bearing and objective pro/con statements

Discovering fine-grained sentiment with latent variable structured prediction models

Overview

- Fine-grained sentiment analysis, from coarse-grained supervision
- This is important because
  - Applications like opinion summarization and search we need analysis on fine-grained levels
  - Available data usually has document level labels
- Goal: Has better performance on sentence than lexicon based and document centric ML approaches

Discovering fine-grained sentiment with latent variable structured prediction models

Overview

- Hidden Conditional Random Fields (HCRF) model analyzes sentence-level sentiment
- Training set: 143,580 positive, negative and neutral reviews from five different domains: books, dvds, electronics, music, and videogames
- Test set: 294 positive, negative and neutral reviews

Discovering fine-grained sentiment with latent variable structured prediction models

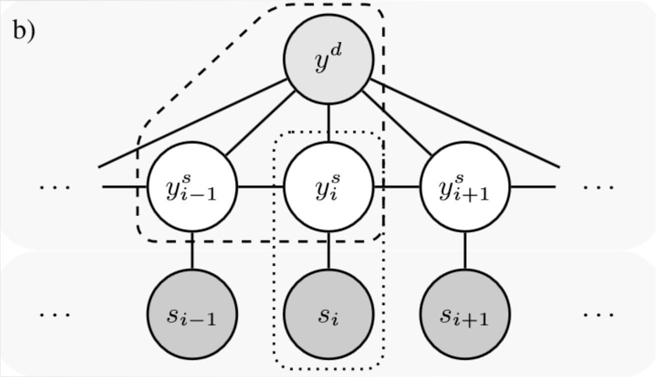
Intuitions

- Documents may have a dominant class without having uniform sentiment. Will likely have majority one sentiment, some neutral, and minority other sentiment.
- Sequential relationship between sentence sentiment
- Document sentiment is influenced by all sentences and vice versa

Discovering fine-grained sentiment with latent variable structured prediction models

Overview

- # Hidden CRF model



- y^d observable variable for document sentiment
- y^s_i (i=1..n) latent variables for sentence sentiment

    - Training: HCRF is trained on document level labels
    - Decoding: Sentence level labels are obtained from latent variables

Discovering fine-grained sentiment with latent variable structured prediction models

Discussion

- Sentence analysis without sentence level supervision
- Diverse set of review subjects
- Performance increase on larger data sets
- Comparison to baseline system trained on sentence-level sentiment data
- Little about choice of features
- Little about training process

Comparing Papers

- Both are similar tasks: sentence-level sentiment from document-level labels
- (Lim, Hovy) exploits structure of epinions.com
  - Better surface-level results, but more questionable methodology, evaluation
  - Straightforward
  - Task seems harder
- (Tackstrom, McDonald) uses machine learning model with latent variables
  - Doesn’t need special structure of text
  - Requires more data

Discovering fine-grained sentiment with latent variable structured prediction models

Optimization

- We model probability of vector: $\mathbf{y^d}$=(y^d, $\mathbf{y^s}$) conditioned on input sentences:

  $p_\theta(y^d, \mathbf{y}^s|\mathbf{s})=\exp\{<\varphi(y^d, \mathbf{y}^s, \mathbf{s}), \theta> - A_\theta(\mathbf{s})\}$
- From independence assumptions

  $\varphi(y^d, \mathbf{y}^s, \mathbf{s}) = \oplus^n{}_{i=1}\varphi(y^d, y^s{}_i, y^s{}_{i-1}, \mathbf{s})$

  $\varphi(y^d, y^s{}_i, y^s{}_{i-1}, \mathbf{s}) =\varphi(y^d, y^s{}_i, y^s{}_{i-1}) \oplus \varphi(y^s{}_i, \mathbf{s})$
- Conditional probability of observable variable

  $p_\theta(y^d|\mathbf{s})=\Sigma_{ys}p_\theta(y^d, \mathbf{y}^s|\mathbf{s})$ - marginalizing over hidden variables