Subjective Language

- **Subjective** sentences express *private states*, i.e. internal mental or emotional states
  - speculations, beliefs, emotions, evaluations, goals, opinions, judgments, ...

  (1) Jill said, "I *hate* Bill."
  (2) John *thought* he won the race.
  (3) Jane *hoped* for good weather.

Opinion Extraction and Summarization

- Extract subjective information from text
  - Basic, low-level relations
- Organize into **opinion frames**
  - One frame for each opinion expression
- Summarize within and across documents
  - Format: graphs, timelines

  Goal: provide insights that would not otherwise be easily accessible

Plan for the Talk

- Analysis of subjective language
  - Identifying and summarizing *opinions*
    - Examples
- Opinion topic identification
  - Definitions
  - Corpus
  - Approach
  - Results
Fine-grained Opinion Extraction

“The Australian Press launched a bitter attack on Italy”

- Five components
  - Opinion trigger
  - Polarity
    - positive
    - negative
    - neutral
  - Strength/intensity
    - low..extreme
  - Opinion holder (source)
  - Target (topic)

Plan for the Talk

- Analysis of subjective language
  - Identifying and summarizing opinions
    - Examples

Opinion topic identification

- Definitions
- Corpus
- Approach
- Results

Examples

(1) \[\text{[OH John]}\] likes Marseille for its weather and cultural diversity.

(2) \[\text{[OH Al]}\] thinks that the government should tax more in order to curb CO2 emissions.

Definitions

- **Topic** - the real-world object, event or abstract entity that is the subject of the opinion as intended by the opinion holder

- **Topic span** - the closest, minimal span of text that mentions the topic

- **Target span** - the span of text that covers the syntactic surface form comprising the contents of the opinion
### Definitions

1. (OH John) *likes* Marseille for its weather and cultural diversity.

2. (OH Al) *thinks* that the government should tax more in order to curb CO2 emissions.

3. Although he doesn’t like government-imposed taxes, he thinks that a fuel tax is the only effective solution.

### Related Work

- **Product reviews**
  - Limit “topics” to mentions of product names, components, and their attributes
  - Lexicon look-up
  - Focused on methods for lexicon acquisition

- **MPQA corpus** (Wiebe, Wilson & Cardie, 2004)
  - Topic annotation deemed too difficult
  - Target span annotation is underway

- **Kim & Hovy (2006)**
  - Target span extraction using semantic frames
  - Limited evaluation

### Related Work

- **Text segmentation** (e.g. Choi (2000), Malioutov and Barzilay (2006))
  - bears some resemblance to opinion topic identification
  - goal: partition a text into a linear sequence of topically coherent segments
  - typically assume that fragments of text (e.g. sentences) with similar lexical distribution are about the same topic
  - opinion topics
    - not necessarily spatially coherent
    - there may be two opinions in the same sentence on different topics
Issues in Opinion Topic Identification

- Multiple potential topics mentioned within a single target span

(2) \( \text{[Oh Al] thinks that [TARGET SPAN [TOPIC SPAN? the government] should [TOPIC SPAN? tax gas] more in order to [TOPIC SPAN? curb [TOPIC SPAN? CO2 emissions]]}. \)

- Requires context

**Topic of an opinion** is the entity that comprises the main information goal of the opinion based on the discourse context.

A Coreference Approach

- Hypothesize that the notion of topic coreference will facilitate identification of opinion topics

Two opinions are **topic-coreferent** if they share the same opinion topic.

- Easier than specifying the topic of each opinion in isolation

Issues in Opinion Topic Identification

- Opinion topics are not always explicitly mentioned

(4) \( \text{[Oh John] believes the violation of Palestinian human rights is one of the main factors. **Topic: ISRAELI-PALESTINIAN CONFLICT**} \)

(5) \( \text{[Oh I] disagree entirely!} \)

Plan for the Talk

- Analysis of subjective language
  - Identifying and summarizing opinions
  - Examples
- Opinion topic identification
  - Definitions
  - Corpus
  - Approach
  - Results
Creating an Opinion Topic Corpus

Build on the MPQA corpus (www.cs.pitt.edu/mpqa)

1. Annotator maintains a list of the opinion expressions **REMAINING** to be annotated and a list of the current topic-coreferent groupings (i.e. **CLUSTERS**)
2. For each opinion expression, \( o \), in **REMAINING**, if \( o \) is topic-coreferent with the opinions of one of the \( e \) (existing) **CLUSTERS**, add \( o \) to \( e \);
   else start a new cluster, \( n \); insert \( o \) in \( n \); add \( n \) to **CLUSTERS**
3. Label each cluster
4. Mark the **TOPIC SPAN** of each opinion.

Interannotator Agreement

- **Annotator 1**
  - 150 of the 535 MPQA documents
- **Annotator 2**
  - 20 of these 150
- IAG measures from noun phrase coreference resolution

<table>
<thead>
<tr>
<th></th>
<th>B^3</th>
<th>( \alpha )</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>all opinions</td>
<td>0.64</td>
<td>0.55</td>
<td>0.69</td>
</tr>
<tr>
<td>sentiment-bearing opinions</td>
<td>0.72</td>
<td>0.73</td>
<td>0.80</td>
</tr>
<tr>
<td>strong opinions</td>
<td>0.74</td>
<td>0.77</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Plan for the Talk

- Analysis of subjective language
  - Identifying and summarizing **opinions**
    - Examples
- Opinion topic identification
  - **Definitions**
  - **Corpus**
  - **Approach**
  - **Results**

A Coreference Approach to Opinion Topic Identification

- Find the clusters of coreferent opinions
  - the critical step for topic identification
- Label the clusters with the name of the topic
  - we currently ignore this step
  - address in future work through frequency analysis of the terms in each of the clusters
The Topic Coreference Algorithm

- Adapt a standard machine learning-based approach to NP coreference resolution (Soon et al., 2001; Ng & Cardie, 2002)
  1. identify topic spans
  2. perform pairwise classification of the associated opinions as to whether or not they are topic-coferent
  3. cluster the opinions according to the results of (ii)

Identifying Topic Spans

- Sentence
  - Use the whole sentence containing the opinion
- Automatic
  - Rule-based method that depends on the syntactic constituent type of the opinion expression
  - Requires a parser and grammatical role labeling algorithm
- Manual
  - Use the topic span marked by the human annotator
- Modified Manual
  - Returns the manually identified topic span only when it is within the sentence of the opinion expression; otherwise returns the opinion sentence

Identifying Topic Spans

- MPQA_{topic} corpus
  - 4976 opinions in 150 documents
- Topic spans
  - 86% are in the opinion sentence
  - 73% are within the span extracted by the AUTOMATIC method
  - 3% (173) are outside the paragraph that contains the opinion

Pairwise Topic Coreference Classification

- Given a pair of opinions (and their associated polarity, source, and topic span), determine whether or not the opinions are topic-coreferent.

- Given a training document, construct a training example for every pair of opinions
  - each pair is represented as a feature vector
- Label the example as positive if the associated opinions belong to the same topic cluster, and negative otherwise.
Features

- Positional
  - Same sentence
  - Same paragraph
  - Consecutive sentences
  - Consecutive paragraphs
  - Distance apart in sentences
  - Distance apart in paragraphs

- Lexico-semantic (topic-span-based)
  - tf.idf cosine similarity of term vectors
  - Word overlap (binary)
  - Coreferent NPs (simple rule-based system)
  - Named Entity overlap (alias)

- Opinion-based
  - Sources match
  - Polarities match
  - Same source but conflicting polarities

Clustering

- Pairwise classification provides an estimation of how likely two opinions are to be topic-coreferent.
- To form the topic clusters, we use **single-link clustering**
  - Assign opinion pairs with a topic-coreference score above a threshold, $t$, to the same topic cluster
  - Perform transitive closure of the clusters

The Topic Coreference Algorithm

1. identify topic spans
2. perform pairwise classification of the associated opinions as to whether or not they are topic-related
3. cluster the opinions according to the results of (ii)
Plan for the Talk

- Analysis of subjective language
  - Identifying and summarizing opinions
    - Examples
- Opinion topic identification
  - Definitions
  - Corpus
  - Approach
  - Results

Experiments

- Three classifiers
  - averaged perceptron (Freund and Schapire, 1998)
  - SVM (Joachims’ SVMlight, 1998)
  - a rule-learner, RIPPER (Cohen, 1995)
- 10-fold cross validation
  - 8 for training, 1 for validation (to determine clustering threshold), 1 for testing
- Three evaluation measures
  - $B^3$ (Bagga and Baldwin, 1998)
  - Krippendorf’s $\alpha$ (Passoneau, 2004)
  - CEAF (Luo, 2005)

Baselines

- all-in-one
  - assigns all opinions to the same cluster
- 1 opinion per cluster
  - assigns each opinion to its own cluster
- same paragraph
  - opinions in the same paragraph are assigned to the same cluster
- Choi 2000
  - state-of-the-art text segmentation system
  - parameter controls the average number of sentences per segment
  - we report results using the best parameter setting on the test data

Results

<table>
<thead>
<tr>
<th></th>
<th>$B^3$</th>
<th>$\alpha$</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-in-one</td>
<td>.37</td>
<td>-.10</td>
<td>.30</td>
</tr>
<tr>
<td>1 opinion per cluster</td>
<td>.29</td>
<td>.22</td>
<td>.27</td>
</tr>
<tr>
<td>same paragraph</td>
<td>.55</td>
<td>.31</td>
<td>.50</td>
</tr>
<tr>
<td>Choi 2000</td>
<td>.54</td>
<td>.37</td>
<td>.54</td>
</tr>
</tbody>
</table>
Interannotator Agreement (revisited)

<table>
<thead>
<tr>
<th></th>
<th>B$^3$</th>
<th>$\alpha$</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-in-one</td>
<td>.37</td>
<td>-.10</td>
<td>.30</td>
</tr>
<tr>
<td>1 opinion per cluster</td>
<td>.29</td>
<td>.22</td>
<td>.27</td>
</tr>
<tr>
<td>same paragraph</td>
<td>.55</td>
<td>.31</td>
<td>.50</td>
</tr>
<tr>
<td>Choi 2000</td>
<td>.54</td>
<td>.37</td>
<td>.54</td>
</tr>
<tr>
<td>all opinions</td>
<td>.64</td>
<td>.55</td>
<td>.69</td>
</tr>
<tr>
<td>sentiment-bearing opinions</td>
<td>.72</td>
<td>.73</td>
<td>.80</td>
</tr>
<tr>
<td>strong opinions</td>
<td>.74</td>
<td>.77</td>
<td>.82</td>
</tr>
</tbody>
</table>

Results

<table>
<thead>
<tr>
<th></th>
<th>B$^3$</th>
<th>$\alpha$</th>
<th>CEAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-in-one</td>
<td>.37</td>
<td>-.10</td>
<td>.30</td>
</tr>
<tr>
<td>1 opinion per cluster</td>
<td>.29</td>
<td>.22</td>
<td>.27</td>
</tr>
<tr>
<td>same paragraph</td>
<td>.55</td>
<td>.31</td>
<td>.50</td>
</tr>
<tr>
<td>Choi 2000</td>
<td>.54</td>
<td>.37</td>
<td>.54</td>
</tr>
<tr>
<td>sentence spans</td>
<td>.57</td>
<td>.40</td>
<td>.54</td>
</tr>
<tr>
<td>automatic (heuristics)</td>
<td>.57</td>
<td>.41</td>
<td>.54</td>
</tr>
<tr>
<td>modified manual topic spans</td>
<td>.64</td>
<td>.51</td>
<td>.61</td>
</tr>
<tr>
<td>manual topic spans</td>
<td>.71</td>
<td>.66</td>
<td>.62</td>
</tr>
</tbody>
</table>

Conclusions

- New definition of opinion topics in the context of fine-grained subjectivity analysis.
- Approach to opinion topic identification that relies on the identification of topic-coreferent opinions
- Initial results look quite promising