Finding and Extracting Opinions in On-line Text

Plan for the Talk
- Subjectivity and sentiment in language
- Opinion extraction
  - definition and examples
- Algorithms and evaluation
- Demo

Subjective Language
- **Subjective** sentences express private states, i.e. internal mental or emotional states
  - speculations, beliefs, emotions, evaluations, goals, opinions, judgments, ...
    - Jill said, "I hate Bill."
    - John *thought* about whom to vote for.
    - Claire *hoped* her lecture would go well.

Subjectivity vs. Sentiment
- **Sentiment** expressions are a type of subjective expression
  - expressions of positive and negative emotions, judgments, evaluations, ...
    - Jill said, "I hate Bill."
    - John *thought* about whom to vote for.
    - Claire *hoped* her lecture would go well.

In this talk, opinion = any subjective language
Why Study Opinions?

- Web Queries of a Subjective Nature
  - How have business views towards global climate change varied over the past decade?
  - What is the reaction in Asia to the Bush policy towards the Kyoto Protocol?
  - How have consumers and businesses responded to Gore’s “An Inconvenient Truth”?
  - Who were the first people to propose bailout options for banks in the current economic crisis?
  - What does Sarah Palin think about <X>?

Research Trend

- Factual and Event-based Text
  - Pang & Lee [ACL 2002, ACL 2004, ...]

- Subjective Text
  - sentiment analysis tome [Pang & Lee, 2008]

Plan for the Talk

- Subjectivity and sentiment in language
- Opinion extraction
  - definition and examples
- Algorithms and evaluation
- Demo
Fine-grained Opinion Extraction

...The Australian press has launched a bitter attack on Italy after seeing their beloved Socceroos eliminated on a controversial late penalty. Italian coach Lippi has been blasted for his comments after the game.

In the opposite camp, Lippi is preparing his side for the upcoming game with Ukraine. He hailed 10-man Italy's determination to beat Australia and said their winning penalty was rightly given.

...
Example – Opinion Summary

What makes this hard?

- MPQA corpus
  - 2812 opinion expressions (medium or higher intensity)
  - 4282 content word tokens
  - 49% are unique
- For words in these expressions that appear > 1 time
  - 38% appear in both subjective and objective contexts
    - achieved (2 subjective, 4 objective);
    - against (15 subjective, 40 objective);
    - considering (3 subjective, 7 objective);
    - difficult (7 subjective, 8 objective);
    - fact (14 subjective, 7 objective);

Plan for the Talk

- Subjectivity and sentiment in language
- Opinion extraction
  - definition and examples
- Algorithms and evaluation
- Demo
Opinion Extraction and Aggregation

- Generate opinion frames
  - Source/opinion holder
  - Opinion expression
    - Polarity
  - Topic/target
- Group related opinions together
  - By Source
  - By Topic

Apply machine learning / statistical learning algorithms...

Research Trend: ML for NLP

<table>
<thead>
<tr>
<th>Year</th>
<th>ML/Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992 ACL</td>
<td>24% (8/34)</td>
</tr>
<tr>
<td>1994 ACL</td>
<td>35% (14/40)</td>
</tr>
<tr>
<td>1996 ACL</td>
<td>39% (16/41)</td>
</tr>
<tr>
<td>1999 ACL</td>
<td>60% (41/69)</td>
</tr>
<tr>
<td>2001 NAACL</td>
<td>87% (27/31)</td>
</tr>
<tr>
<td>2005 ACL</td>
<td>96% (74/77)</td>
</tr>
</tbody>
</table>

Supervised Inductive Learning

Examples
(features + class)

ML Algorithm

(novel) examples
(features)

Statistical model
(program)

Class
Supervised Inductive Learning

examples of (Source) NPs in context [features + class]

ML Algorithm

(novel) NPs in context [features]
statistical model (program)
Source? Not Source?

Source Annotations

Australian press> has launched a bitter attack on Italy after seeing <their> beloved Socceroos eliminated on a controversial late penalty. Italian coach Lippi has also been blasted for his comments after the game.

In the opposite camp Lippi is preparing his side for the upcoming game with Ukraine. <He> hailed 10-man Italy's determination to beat Australia and said the penalty was rightly given.

Identifying Sources of Opinions ...
as an Information Extraction task

- Sequence tagging

The Washington Post reported Obama's view on the oil crisis.

Conditional Random Fields

- Discriminative training
- Incorporation of arbitrary non-independent features (past + future)
  - semantic class, suffixes, constituent type, etc.

- Perform better than related classification and generative models (e.g. HMMs)
  - Part-of-speech tagging [Lafferty et al., 2001]
  - Noun phrase chunking [Sha and Pereira, 2003]
  - Human protein name tagging [Bunescu et al. 2004]
Features for Source Extraction

- **Syntactically**...
  - *mostly* noun phrases
- **Semantically**...
  - entities that can bear opinions
- **Functionally**...
  - linked to opinion expressions

Features for Source Extraction

- Words [-4,+4]
- Capitalization
- Part-of-speech tags [-2,+2]
- Opinion phrase lexicon
  - Derived from training data
  - Wiebe et al.’s [2002] 500+ word lexicon
- Shallow semantic class information
  - Sundance partial parser and named entity tagger
  - WordNet hyponym
- Constituent type
- Grammatical role
  - Collins’ parser
- Task-specific combinations
  - E.g., Parent contains opinion word

Evaluation

- MPQA data set (www.cs.pitt.edu/mpqa)
  - ~550 documents
  - Manually annotated w.r.t. fine-grained opinion information
  - Provides gold standard
- Automatically derive training/test examples
- 10-fold cross-validation
- Evaluation measures
  - Precision
  - Recall
  - F-measure

Results: Opinion Holders

- >82% precision (accuracy)
- ~60% recall (coverage)
- 69.4 F-measure

So there’s a lot of room for improvement...
Errors

- False positives
  - Perhaps this is why Fidel Castro has not spoken out against what might go on in Guantanamo.
- False negatives
  - And for this reason, too, they have a moral duty to speak out, as Swedish Foreign Minister Anna Lindh, among others, did yesterday.
  - In particular, Iran and Iraq are at loggerheads with each other to this day.

Extracting and Linking to Opinions

- To be useful, we need to link sources to their opinions
  - <source> expresses <opinion>

Research Trend: Structured Learning

- Beyond simple classification tasks
- Dependent/output variable has an internal structure
- Multiple dependent/output variables with dependencies or constraints among them

E.g. syntactic parse tree, source-expresses-opinion relation

Opinion Frame Extraction via CRFs and ILP

- Joint extraction of entities and relations

  k-best 80F

  all source-opinion pairs

  [Roth & Yih, 2004]

  [Choi et al., EMNLP 2006]
Constraints

- Binary integer variables $O_i$, $S_j$, $L_{i,j}$
  - Weights for $O_i$, $S_j$, $L_{i,j}$ are based on probabilities from individual classifiers

Constraints

\[ \forall i, \ O_i = \sum_j L_{i,j} \]  : link coherency (only one link from each opinion)

\[ \forall j, \ S_j + A_j = \sum_i L_{i,j} \]  : link coherency (up to two links from each source)

\[ \forall j, \ A_j - S_j \leq 0 \]  : link coherency (preferably one link from each source)

\[ \forall i, j, i < j, \ X_i + X_j = 1, \ X \in \{S, O\} \]  : entity coherency (for all pairs of entities with overlapping spans)

Objective function

\[ f = \sum_i (w_i, O_i) + \sum_i (w_i, O_i) + \sum_j (w_j, S_j) + \sum_j (w_j, S_j) + \sum_i (w_i, L_{i,i}) + \sum_j (w_i, L_{i,j}) \]

Opinion Frame Extraction via CRFs and ILP

- Joint extraction of entities and relations

Sources and Coreference

**Australian press** has launched a bitter attack on **Italy** after seeing their beloved **Socceroos** eliminated on a controversial late penalty. Italian coach Lippi has also been blasted for his comments after the game.

In the opposite camp, Lippi is preparing his side for the upcoming game with Ukraine: He hailed 10-man Italy's determination to beat Australia and said the penalty was rightly given.

Somewhat different techniques...different talk 😊.

Appinions Demo

- **www.jodange.com**
- **Opinion frames**
  - Opinion expressions
  - Opinion holder
  - Polarity
  - Topic
- **Aggregation**
  - w.r.t. opinion holder or topic
  - over time
Summary

- Techniques for extracting fine-grained opinion information from on-line text
  - Machine learning methods
  - Fairly shallow linguistic knowledge
- Incorporation into an application for browsing and summarizing opinions on the web