An Overview of Opinionated Tasks and Corpus Preparation

Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan

What is an opinion?

- Opinion is a subjective information
- Opinion usually contains an opinion holder, an attitude, and a target, but not obligatory
- A sentential clause or a meaningful unit (in Chinese) is the smallest unit of an opinion.
Why opinion processing is important?

- There is explosive information on the Internet, and it’s hard to extract opinions by humans.
- Opinions of the public is an important index of companies and the government.
- Opinions change over time, so to keep track of opinions automatically is an important issue.

Opinionated Applications

- Sentiment word mining
- Opinionated sentence extraction
- Opinionated document extraction
- Opinion summarization
- Opinion tracking
- Opinionated question answering
- Multi-lingual/Cross-lingual opinionated issues
Fact-based vs. Opinion-based

Examples:
• Circular vs. Happy
• He is an engineer. vs. He thinks that his boss is a kind person.
• Why the sky is blue? vs. Do people support the government?

Previous Work (1)

• English:
  – Sentiment words
    (Wiebe et al., Kim and Hovy, Takamura et al.)
  – Opinion sentence extraction
    (Riloff and Wiebe, Kim and Hovy)
  – Opinion document extraction
    (Wiebe et al., Pang et al.)
  – Opinion summarization: reviews and products
    (Hu and Liu, Dave et al.)
Previous Work (2)

- Japanese
  - Opinion extraction (Kobayasi et al.: reviews, at word/sentence level)
  - Opinion summarization (Morinaga et al.: product reputations, Seki, Eguchi, and Kando)
- Chinese
  - Opinion extraction (Ku, Wu, Li and Chen)
  - Opinion summarization (Ku, Li, Wu and Chen)
  - News and Blog Corpora (Ku, Liang and Chen)
- Korean?

Corpus Preparation (1)

- Quantity
  - How much materials should we collect?
    - Words/Sentences/Documents
- Source
  - What source should we pick? Mining opinions from general documents or the obvious opinionated documents? (ex. Discussion group)
  - News, Reviews, Blogs, …
Corpus Preparation (2)

- Different granularity
  - Word level
  - Sentence level
  - Clause level
  - Document level
  - Multi-documents (summarization)
- Different sources
- Different languages

Previous Work
(Corpus Preparation 1)

- Example: NRRC Summer Workshop on Multiple-Perspective QA
  - People involved: 1 researcher, 3 graduate students, 6 professors
  - Collect 270,000 documents, over 11-month periods, retrieve documents relevant to 8 topics, more than 200 documents of each topic
Previous Work
(Corpus Preparation 2)

- Source: news documents
  (World News Connection - WNC)

<table>
<thead>
<tr>
<th>8 Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>The presidential election in Zimbabwe</td>
</tr>
<tr>
<td>Passage of the Kyoto Protocol</td>
</tr>
<tr>
<td>The U.S. annual human rights report</td>
</tr>
<tr>
<td>The political upheaval in Venezuela</td>
</tr>
<tr>
<td>Relation between Taiwan and China</td>
</tr>
<tr>
<td>Israeli settlements in the West Bank</td>
</tr>
<tr>
<td>The U.S. holding of al Quaida and</td>
</tr>
<tr>
<td>Taliban detainees at Guantanamo Bay</td>
</tr>
<tr>
<td>Reactions to the U.S. characterization</td>
</tr>
<tr>
<td>of certain countries as an “axis of evil.”</td>
</tr>
</tbody>
</table>

- In another work on word level: 2,615 words

NTCIR-5, 2005

Previous Work
(Corpus Preparation 3)

- Example: Using NTCIR Corpus (Chinese)
  - Reusable
  - NTCIR2, news documents
  - Retrieve documents relevant to 6 topics
  - On average, 34 documents for each topic
  - At Word level: 838 words
  - Experiments using NTCIR3 are ongoing

NTCIR-5, 2005
### Previous Work
**Corpus Preparation 4**

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Topic Title</th>
<th>Doc #</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZH021</td>
<td>Civil ID Card</td>
<td>37</td>
</tr>
<tr>
<td>ZH024</td>
<td>The Abolishment of Joint College Entrance Examination</td>
<td>55</td>
</tr>
<tr>
<td>ZH026</td>
<td>The Chinese-English Phonetic Transcription System</td>
<td>37</td>
</tr>
<tr>
<td>ZH027</td>
<td>Anti-Meinung Dam Construction</td>
<td>15</td>
</tr>
<tr>
<td>ZH028</td>
<td>Hewing Down of Chinese Junipers in Chilan</td>
<td>23</td>
</tr>
<tr>
<td>ZH036</td>
<td>Surrogate Mother</td>
<td>37</td>
</tr>
</tbody>
</table>

### Previous Work
**Corpus Preparation 5**

- Example: Using reviews from Web (Japanese)
  - Specific domains: cars and games
  - 15,000 reviews (230,000 sentences) for cars, 9,700 reviews (90,000 sentences) for games
  - Using topic words (ex. Companies of cars and games)
  - Semi-automatic methods for collecting opinion terms (with patterns)
Corpus Annotation

• Annotation types (1)
  – Support/Non-support
  – Sentiment/Non-sentiment
  – Positive/Neutral/Negative
  – Strong/Medium/Weak

• Annotation types (2)
  – Opinion holder/Attitude/Target
  – Nested opinions

Previous Work
(Corpus Annotation 1)

• Example: NRRC Summer Workshop on Multiple-Perspective QA (English)
  – Total 114 documents annotated
  – 57 with deep annotations, 57 with shallow annotations
  – 7 annotators
Previous Work
(Corpus Annotation 2)

• Tags
  – Opinion: on=implicit/formally declared/…
  – Fact: onlyfactive=yes/no
  – Subjectivity: strength=high/medium/low
  – Attitude: neg-attitude/pos-attitude
  – Writer: opinion holder information

Previous Work
(Corpus Annotation 3)

• Example: Using NTCIR Corpus (Chinese)
  – Total 204 documents are annotated
  – 3 annotators
  – Using XML-style tags
  – Define types, but no strength (considering the agreement issue)
Previous Work
(_corpus Annotation 4)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Level</th>
<th>Attribute</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;DOC_ATTITUDE&gt;&gt;DOC_ATTITUDE&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Document Attitude: Define the opinion polarity of the whole document</td>
</tr>
<tr>
<td>&lt;SEN_ATTITUDE&gt;&gt;SEN_ATTITUDE&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Sentence Attitude: Define the opinion polarity of one sentence</td>
</tr>
<tr>
<td>&lt;OPINION_SEG&gt;&gt;OPINION_SEG&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Opinion Segment: Define the scope of one opinion</td>
</tr>
<tr>
<td>&lt;OPINION_SRC&gt;&gt;OPINION_SRC&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Opinion Source: Define the opinion holder of a specific opinion</td>
</tr>
<tr>
<td>&lt;SENTIMENT_KW&gt;&gt;SENTIMENT_KW&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Sentiment Keyword: Define the opinion polarity of a single word</td>
</tr>
<tr>
<td>&lt;OPINION_OPR&gt;&gt;OPINION_OPR&gt;</td>
<td></td>
<td></td>
<td></td>
<td>Opinion Operator: Define the keyword of expressing an opinion</td>
</tr>
</tbody>
</table>

Corpus Evaluation (1)

- How to choose materials?
  - Filter out candidates whose annotations are too diverse among annotators? (Agreements?)
  - How many annotators are needed for one candidate? (More annotators, lower agreements)
  - How to build the gold standard?
    - Voting
    - Use instances with consistent annotations
Corpus Evaluation (2)

• How to evaluate a corpus for a subjective task?
  – Agreement (Is it enough?)
  – Kappa value (To what agreement level?)
    • Almost perfect agreement
    • Substantial agreement
    • Moderate agreement
    • Fair agreement
    • Slight agreement
    • Less than change agreement

Previous Work
Corpus Evaluation

• Different languages/annotations may have different agreements.
  – Kappa: 0.32-0.65 (onlyfactivity, English)
  – Kappa: 0.40-0.68 (word level, Chinese)
• Different annotators with different background may have different agreements.
What are needed for this work?

- What kind of documents? News? Others?
- All relevant documents?
- Provide only the type of documents, or fully annotated documents for training?
- Provide some sentiment words as clues?
- To what granularity? Word, clause, sentence, document, or multi-document?
- In which language? Mono-lingual, multi-lingual or cross-lingual?

Any comments are welcome