PKUTM Experiments in NTCIR-8 MOAT Task

Author: Chenfeng Wang*, Tengfei Ma*, Liqiang Guo, Xiaojun Wan and Jianwu Yang

Affiliation: Insitute of Computer Science & Technology of Peking University

Speaker: Tengfei Ma
• Aspects of Opinion Analysis
  – Is it opinionated?
  – Is the opinion positive or negative?
  – What is the opinion?
  – Who gives the opinion and who does the opinion point to?
  – How to summarize all the opinions?
  – ...
Background of Opinion Analysis

- Aspects of Opinion Analysis
  - Is it opinionated?
  - Is the opinion positive or negative?
  - What is the opinion?
  - Who gives the opinion and who does the opinion point to?
  - How to summarize all the opinions?
  - ...
Background of Opinion Analysis

- The trend of opinion analysis
  - Coarse-grain to fine-grain
    - Holder/target extraction
  - General to domain-specific and domain-transfer
    - Opinion analysis in news, product reviews, movie reviews
    - Cross-Lingual, transfer learning
  - Publisher-predominate to interactive
Background of Opinion Analysis

• The trend of opinion analysis
  – Coarse-grain to fine-grain
    • Holder/target extraction
  – General to domain-specific and domain-transfer
    • Opinion analysis in news, product reviews, movie reviews
    • Cross-Lingual, transfer learning
  – Publisher-predominate to interactive
Our tasks in NTCIR-8 Moat

- Opinionated subtask.
- Opinion holder extraction.
- Opinion target extraction.
I. DETECTION OF SUBJECTIVE SENTENCES
Detection of subjective sentences

• Equivalent to a classification problem

Our method:
  – Some combined datasets
  – Some special opinion features
  – A general classifier and an improvement
Detection of subjective sentences

• Data Preprocessing
  – Choosing the training Datasets
    • NTCIR6/NTCIR7 corpora and NTCIR8’s samples
    • Containing both simplified and traditional Chinese
  – Translate traditional Chinese to Simplified Chinese
  – POS, NER
  – Building Lexicons
Source:

- expanded Hownet by using the Synonymy Thesaurus + MPQA(English->Chinese) + NTU + our in-house labeled corpora

Types:

- **Opinion Operators** e.g. 声称 (claim)
- **Opinion Indicators** e.g. 但是 (but)
- **Degree Adverbs** e.g. 非常 (very), 缺乏 (lack of)
- **Opinion Words** (28421 opinion words)
- **Strong Opinion Words** (6471 words)
### Feature Selection

<table>
<thead>
<tr>
<th>Punctuations Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of quotation marks like “”, 「」, ’ and ”</td>
</tr>
<tr>
<td>Presence of colon followed by quotation marks</td>
</tr>
<tr>
<td>Percentage of punctuations in sentences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Words and Entities Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>The percentage of numeral words</td>
</tr>
<tr>
<td>The presence of pronoun</td>
</tr>
<tr>
<td>The presence of a named entity</td>
</tr>
<tr>
<td>The presence of a word which indicates a sequence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lexical Subjective Clues</th>
</tr>
</thead>
<tbody>
<tr>
<td>The presence of opinion operator</td>
</tr>
<tr>
<td>The presence of opinion indicator</td>
</tr>
<tr>
<td>The logarithm of percentage of opinion words</td>
</tr>
<tr>
<td>The logarithm of percentage of strong opinion words</td>
</tr>
<tr>
<td>The presence of degree verb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Collocation Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>The presence of collocations between named entities and opinion operators</td>
</tr>
<tr>
<td>The presence of collocations between pronouns or nouns and opinion operators</td>
</tr>
<tr>
<td>The presence of collocations between opinion operators and opinion words</td>
</tr>
<tr>
<td>The presence of collocations between pronouns and opinion words</td>
</tr>
<tr>
<td>The presence of collocations between nouns or pronouns and opinion words</td>
</tr>
<tr>
<td>The presence of collocations between degree adverbs and opinion operators</td>
</tr>
<tr>
<td>The presence of collocations between degree adverbs and opinion words</td>
</tr>
<tr>
<td>The presence of collocations between nouns or named entities and opinion words</td>
</tr>
</tbody>
</table>
Detection of subjective sentences

• Classifier
  –Basic classifiers
    • such as SVM, Naive Bayes, Max Entropy and Decision Tree
    • The comparison is shown in the following section
  –Improved classifier
    • Iterative classifier using former results of detecting subjective sentences
Detection of subjective sentences

• Results in NTCIR8

<table>
<thead>
<tr>
<th>Run</th>
<th>Datasets and Classifiers</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run1</td>
<td>All datasets + iterative classifier</td>
<td>0.3721</td>
<td>0.8370</td>
<td>0.5152</td>
</tr>
<tr>
<td>Run2</td>
<td>NTCIR7 + NTCIR8 simplified Chinese + basic classifier</td>
<td><strong>0.4134</strong></td>
<td>0.8335</td>
<td><strong>0.5527</strong></td>
</tr>
<tr>
<td>Run3</td>
<td>Run2 + NTCIR7 traditional dataset</td>
<td>0.3405</td>
<td><strong>0.9062</strong></td>
<td>0.4950</td>
</tr>
</tbody>
</table>

• Additional Tests (Comparison of different classifiers)

![Lenient and Strict Performance](image)
Detection of subjective sentences

• Discussion of the results
  – Training data
    • More ≠ Better
    • When and how to leverage translated datasets
  – Classifier
    • Iterative → risk
  – Problem
    • Ambiguous definition
    • Ambiguous words
Holder/target task

EXTRACTION OF OPINION HOLDERS AND TARGETS
• Common methods
  – Parsing and direct training (Bethard)
  – Maximum Entropy ranking (Kim and Hovy)
  – Labeling

• Our method
  – Chunking and heuristic rules
Extracting opinion holders/targets

- Advantage of Chunking
  - Better than parsing in Chinese
  - Easier to control and modify than shallow parsing

- Process:
  - Training data: proposition bank
  - Modifying training data
  - Training and labeling by CRF
Extracting opinion holders/targets

• Heuristic rules for opinion holder extraction
  – before an opinion operator (include a colon) or following a quotes.
  – not governed by a preposition
  – in other sentences sometimes
  – using nouns or pronouns as candidates to complement the upper missing cases
  – author
Extracting opinion holders/targets

- Heuristic rules for opinion target extraction
  - Similar to opinion holder extraction
  - Mainly existing in the opinion clause or as the object of an opinion operator
  - Coherent with neighbor sentences
### Evaluations Results for Opinion Targets

#### Holder Extraction

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only for opinionated sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run1</td>
<td>0.550</td>
<td>0.434</td>
<td>0.485</td>
</tr>
<tr>
<td>Run2</td>
<td>0.554</td>
<td>0.431</td>
<td>0.485</td>
</tr>
<tr>
<td>Run3</td>
<td>0.548</td>
<td>0.473</td>
<td>0.508</td>
</tr>
<tr>
<td>For all sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run1</td>
<td>0.204</td>
<td>0.434</td>
<td>0.277</td>
</tr>
<tr>
<td>Run2</td>
<td>0.232</td>
<td>0.431</td>
<td>0.301</td>
</tr>
<tr>
<td>Run3</td>
<td>0.186</td>
<td>0.473</td>
<td>0.267</td>
</tr>
</tbody>
</table>

#### Target Extraction

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only for opinionated sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run1</td>
<td>0.892</td>
<td>0.736</td>
<td>0.806</td>
</tr>
<tr>
<td>Run2</td>
<td>0.896</td>
<td>0.732</td>
<td>0.805</td>
</tr>
<tr>
<td>Run3</td>
<td>0.877</td>
<td>0.792</td>
<td>0.832</td>
</tr>
<tr>
<td>For all sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run1</td>
<td>0.339</td>
<td>0.736</td>
<td>0.464</td>
</tr>
<tr>
<td>Run2</td>
<td>0.385</td>
<td>0.732</td>
<td>0.504</td>
</tr>
<tr>
<td>Run3</td>
<td>0.307</td>
<td>0.792</td>
<td>0.442</td>
</tr>
</tbody>
</table>
Extracting opinion holders/targets

• Discussion
  – Limited by the parsing technique
  – Features are complex for machine learning
  – Future research (See (Ma, Coling10))
    • Adding semantic information
    • Adding syntactic rules to leverage relevant information (e.g. reviews--news)
Thank you~

Any questions?