Enhance Japanese Opinionated Sentence Identification Using Linguistic Features: Experiences of the IISR Group at NTCIR-8 MOAT Task
Guo-Hau Lai, Jyun-Wei Huang, Chia-Pei Gao, Richard Tzong-Han Tsai*
Yuan Ze University, Taiwan, R.O.C.

We have described a feature-based system that is designed to recognize opinionated sentences or not. Our system utilizes various features: headlines in newspapers, Japanese sentence patterns, dependency pairs, numeral features and some related to newspapers opinionated words. The experiments show that our feature-based system is feasible and effective.

System Overview
Our system contains two parts:
1) Generate the features.
2) Recognize opinionated sentences.
Figure 1 shows our system architecture.

Recognition Method

Dataset for training
<table>
<thead>
<tr>
<th></th>
<th>Opinion</th>
<th>Non-opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR6 OAT</td>
<td>3,026</td>
<td>6,682</td>
</tr>
<tr>
<td>NTCIR7 MOAT</td>
<td>1,562</td>
<td>4,323</td>
</tr>
<tr>
<td>NTCIR7&amp;8 Sample</td>
<td>499</td>
<td>1,102</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5,087</td>
<td>12,107</td>
</tr>
</tbody>
</table>

Opinion Score
We extracted the features in the feature sets that was useful for recognizing opinionated sentences. These feature sets consist of a list of patterns or words and we calculate the score for each item by following equation:

\[
\text{score}(\text{item}_i) = \begin{cases} 
\text{Op}(\text{item}_i) - \text{On}(\text{item}_i) & \text{if } \text{item}_i \text{ appeared in opinionated sentences} \\
\text{Op}(\text{item}_i) + \text{On}(\text{item}_i) & \text{if } \text{item}_i \text{ appeared in non-opinionated sentences}
\end{cases}
\]

We constructed four tables that record opinion score for sentence patterns, keywords, dependency pairs, and numeral features.

Opinionated Sentence Recognition
1° Our system recognize opinionated sentence according to extracted headlines.
2° For non-headline sentences, the sentence will be analyzed by CaboCha for dependency analysis.
3° Our system looks up the features in the dictionaries and sums up the score of each feature appeared in a sentence.
4° If the score calculated by 3° is greater than 0, our systems determines the sentence as opinion.

Dataset Analysis
We analyzed NTCIR7 and NTCIR8 sample data (contains 1,601 sentences in total and 499 opinionated sentences) to find useful features for the opinion judgment subtask. We founded these features as follow.

Headline
We used some special structures and opinion words for opinionated sentence recognition in headlines. 1) A person says something. 2) Opinion words.

Japanese Sentence Patterns
These following sentence patterns are examples that were frequently used in opinionated sentences.

- について  この兵士の死とウラン弾の因果関係について、イタリア国防省は否定した。
- そうだ  この兵士の死とウラン弾の因果関係について。

Keywords
For counting the frequency of each word, we used MeCab to perform morphological analysis. Then we compile a opinionated word list and expanded manually by using Japanese WordNet.

Dependency Pairs
We used the CaboCha as our Japanese dependency parser, then extracted all dependency pairs and pairs with distances greater than one were not used.

A dependency pair consists of the source element and the sink element. The source element is the named entity (person, organization, etc.) or noun type semantic primitive. However, we only focused on the sink element and extract verbal noun (sahen type semantic primitive. However, we only focused on the sink element and extract verbal noun (sahen type semantic primitive.

Statistics and Date
The non-opinionated sentences tend to have numbers and these numbers may be statistics, period of time, or a specific time.

- Sentence containing a statistics data
  パリ島駐在の日本人向け旅行估计によると、パリを訪れる年間約30万人の日本人観光客のうち6割以上が女性客。
- Sentence containing a period of time
  日本でも在日米軍が95年12月から96年1月に沖ノ鳥島の演習で劣化ウラン弾1520発を誤射したことが判明。
- Sentence containing a specific time
  昨年10月12日、インドネシア・バリ島のディスコで爆破テロが発生、日本人夫婦を含む外国人観光客2000人以上が死亡、300人以上が負傷した。

Evaluation Result

Formal Run
We finally submitted three runs as the formal run:
Run1 Using all features and training without the sample data
Run2 Delete the headline features and training without the sample data
Run3 Using all features and all training dataset

The evaluation results are presented in following table:

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run1</td>
<td>67.30</td>
<td>49.86</td>
<td>57.28</td>
</tr>
<tr>
<td>Run2</td>
<td>67.74</td>
<td>47.65</td>
<td>55.95</td>
</tr>
<tr>
<td>Run3</td>
<td>67.86</td>
<td>51.53</td>
<td>58.58</td>
</tr>
</tbody>
</table>

Effects of feature sets for opinion judgment subtask
We also conducted the experiment to confirm the effects of each feature sets for the opinion judgment subtask. Training data is the same of above and the formal run data is used for the test.

According to the results of formal run, we have confirmed that the headline features are effective. We made comparison between the result of using all feature sets and the result of deleting one feature set (Sentence pattern, keyword, dependency pair, numeral). The following table shows deleted feature sets, each result and difference of F-measure. Larger difference means more effective feature sets.

<table>
<thead>
<tr>
<th>Deleted feature</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>68.52</td>
<td>51.28</td>
<td>58.66</td>
<td></td>
</tr>
<tr>
<td>Sentence pattern</td>
<td>67.81</td>
<td>37.04</td>
<td>47.91</td>
<td>10.75</td>
</tr>
<tr>
<td>Keyword</td>
<td>73.28</td>
<td>38.61</td>
<td>50.57</td>
<td>8.09</td>
</tr>
<tr>
<td>Dependency pair</td>
<td>68.37</td>
<td>47.91</td>
<td>56.34</td>
<td>2.32</td>
</tr>
<tr>
<td>Numerical</td>
<td>66.94</td>
<td>53.86</td>
<td>59.53</td>
<td>-0.87</td>
</tr>
</tbody>
</table>

The results show numeral feature improve precision but didn’t improve recall and F-measure. In some opinionated sentences, these numeral features appear more times than other features and our system output these sentences as non-opinionated sentences.