The OPAL System at NTCIR 8 MOAT

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P.

NTCIR 8 MOAT, 18 June, 2009 Tokyo
Outline

- Motivation
- Contribution
- Related Work
- English Monolingual Subtask
- English-Chinese Cross-Lingual Subtask
- Evaluation
- Discussion
Motivation

Challenges

Growth in the access to technology +
Development of the Social Web

Exponential increase of the subjective data on the Web:
factoid & opinionated info
Big impact on lives/business

NEED FOR AUTOMATIC PROCESSING METHODS FOR
SUBJECTIVE TEXT: Sentiment Analysis (SA)

NTICIR 8 MOAT, 18 June, 2009 Tokyo
Motivation

• Factod Question Answering (QA): long tradition
• Opinionated QA: remains a challenge
• Need for OQA

• Need for specialised tools and methods
Contribution

- Deep study of the performance of new sentiment-topic detection methods

- Introduction of specialised tools
  - temporal expressions and
  - anaphora resolution and analysis of their effect

- New retrieval techniques (using Wikipedia with LSA), improving the performance.
Related Work

• Cardie et al., 2003: separated opinions from facts and summarized them as answer to opinion questions.
• Kim and Hovy, 2005: identified opinion holders, which are frequently asked in opinion questions.
• TAC 2008 Opinion QA track: a collection of factoid and opinion queries -“rigid list” (factoid) and “squishy list” (opinion) - to which the traditional systems had to be adapted.
• The Alyssa system (Shen et al. 2007): a SVM classifier trained on the MPQA corpus, English NTCIR data and rules based on the subjectivity lexicon
• Varma et al., 2008: query analysis to detect the polarity of the question using defined rules.
• The QUANTA system (Li et al, 2008): OQA by detecting the opinion holder, the object and the polarity of the opinion. It uses a semantic labeler based on PropBank and manually defined patterns. Sentiment classification: extract and classify the opinion words. Answer retrieval: score the retrieved snippets depending on the presence of topic and opinion words and only choose as answer the top ranking results
English Monolingual Subtask

• 20 topics
• For each topic:
  – a question & short query were given
  – Expected Polarity & Period of time
• Each topic:
  – set of documents split into sentences
  – opinion units
    • for polarity,
    • opinion target
    • source task
• We submitted 3 runs of the OpAL system for: opinionated, relevance and polarity judgement tasks

NTICIR 8 MOAT, 18 June, 2009 Tokyo
English Monolingual subtask

*Judging sentence opinionatedness*

- The system has to assign YES –opinionated-/NO-factoid- to each sentence
- We employed
  - system run 1
  - system runs 2 & 3
- Rule based but with different resources
- Opinionated sentence: at least 2 opinion words or one opinion word preceeded by a modifier

NTICIR 8 MOAT, 18 June, 2009 Tokyo
English Monolingual subtask

Judging sentence opinionatedness

• 1\textsuperscript{st} approach: opinion word from:
  – General Inquirer,
  – Micro WordNet Opinion and
  – Opinion Finder

• 2\textsuperscript{nd} approach: opinion word from:
  – General Inquirer
  – Micro WordNet Opinion
English Monolingual subtask

Determining sentence relevance

- The system had to output for each sentence an assessment of relevance
- 3 strategies:
  - 1st JIRS (Java Information Retrieval System) to find relevant snippets JIRS retrieves passages based on the question structures (not keywords)
English Monolingual subtask

Determining sentence relevance

- 2nd Faceted Search Wikipedia and LSA, to find words related to the topic –find concepts-
- We match the query words to a category in Wikipedia and two consecutive words, 3, 4 ...the concepts found are the topic components.
- For each topic component we employ LSA to the first 20 documents (retrieved by Yahoo) to determine the most related words (LSA we employ Infomap)
- We expand the query using words similar to the topic
English Monolingual subtask

*Determining sentence relevance*

- 3rd we judge a part for the topic relevance, the temporal appropriateness
- We employ TERSEO
- We filter the sentences obtained
Polarity and topic-polarity classification for judging sentence answerness

• Required the system to assign a value of POS, NEG or NEU to each of the sentences in the documents provided.
• Polarity of the sentences: we passed each sentence through an opinion mining system employing SVM ML over the NTCIR 7 MOAT corpus, the MPQA corpus and EmotiBlog.
• Each sentence is preprocessed using Minipar
• System training:
  – the part of speech (POS)
  – opinionatedness/intensity - if the word is annotated as opinion word, its polarity, i.e. 1 and -1
  – syntactic relatedness with other opinion word – if it is directly dependent of an opinion word or modifier (0 or 1), plus the polarity/intensity and emotion of this word (0 for all the components otherwise).

NTICIR 8 MOAT, 18 June, 2009 Tokyo
Polarity and topic-polarity classification for judging sentence answer-ness

- The difference between the submitted runs consisted in the lexicons used to determine whether a word was opinionated or not.
- 1\textsuperscript{st} run: General Inquirer, MicroWordNet and the Opinion Finder opinion resources.
- 2\textsuperscript{nd}: aside from these three sources, the “emotion trigger” resource (Balahur and Montoyo, 2008).

NTICIR 8 MOAT, 18 June, 2009 Tokyo
English-Chinese cross-lingual subtask

- Systems have to output, for each of the twenty topics and their corresponding questions (in a language), the list of sentences containing answers (in another language).
- The topics and questions are given in English;
- the output of the system contains the sentences in set of documents in Traditional Chinese which contain an answer to the given topics.

- 3 runs of OpAL

NTICIR 8 MOAT, 18 June, 2009 Tokyo
English-Chinese cross-lingual subtask

- To tokenize the Chinese texts using LingPipe
- We applied a technique known as “triangulation” to obtain opinion and subjectivity resources for Chinese to obtain resources for different languages, starting from correct parallel resources in 2 initial languages.
English-Chinese cross-lingual subtask

• This technique requires the existence of two correct parallel resources in two different languages to obtain correct resources for a third language.
• We have previously translated and cleaned the General Inquirer, MicroWordNet and Opinion Finder lexicons for Spanish.
• The “emotion triggers” resource is available for English and Spanish--translated into Chinese (Google translator).
• We performed the intersection of the obtained translations –the corresponding words translated in the same manner (from English as well as from Spanish).
• We removed words that we translated differently from English and Spanish.
English-Chinese cross-lingual subtask

• We mapped each of these resources to four classes, depending on the score they are assigned in the original resource – of “high positive”, “positive”, “high negative” and “negative” and we give each word a corresponding value (4, 1, -4 and 1), respectively.
• We translated the topic words determined in English using LSA.
• For each of the sentence, we compute a score, given by the sum of the values of the opinion words that are matched in it.
• Conditions to be considered an answer:
  – to contain at least one topic word
  – the polarity determined corresponds to the required polarity, as given in the topic description.
• 1st: General Inquirer and MicroWordNet
• 2nd : we added the “emotion trigger resource”
• 3rd : only the Opinion Finder lexicon.

NTICIR 8 MOAT, 18 June, 2009 Tokyo
# Evaluation & Discussion

Results of system runs for opinionated

<table>
<thead>
<tr>
<th>System RunID</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPAL1</td>
<td>17.99</td>
<td>45.16</td>
<td>25.73</td>
</tr>
<tr>
<td>OPAL2</td>
<td>19.44</td>
<td>44</td>
<td>26.97</td>
</tr>
<tr>
<td>OPAL3</td>
<td>19.44</td>
<td>44</td>
<td>26.97</td>
</tr>
</tbody>
</table>

Results of system runs for relevance

<table>
<thead>
<tr>
<th>System RunID</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPAL1</td>
<td>82.05</td>
<td>47.83</td>
<td>60.43</td>
</tr>
<tr>
<td>OPAL2</td>
<td>82.61</td>
<td>5.16</td>
<td>9.71</td>
</tr>
<tr>
<td>OPAL3</td>
<td>76.32</td>
<td>3.94</td>
<td>7.49</td>
</tr>
</tbody>
</table>

NTICIR 8 MOAT, 18 June, 2009 Tokyo
Results of system runs for polarity

<table>
<thead>
<tr>
<th>System RunID</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPAL1</td>
<td>38.13</td>
<td>12.82</td>
<td>19.19</td>
</tr>
<tr>
<td>OPAL2</td>
<td>50.93</td>
<td>12.26</td>
<td>19.76</td>
</tr>
</tbody>
</table>

Results of system runs for the cross-lingual task – agreed measures, Traditional Chinese

<table>
<thead>
<tr>
<th>System RunID</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPAL1</td>
<td>3.54</td>
<td>56.23</td>
<td>6.34</td>
</tr>
<tr>
<td>OPAL2</td>
<td>3.35</td>
<td>42.75</td>
<td>5.78</td>
</tr>
<tr>
<td>OPAL3</td>
<td>3.42</td>
<td>72.13</td>
<td>6.32</td>
</tr>
</tbody>
</table>
## Results of system runs for the cross-lingual task – non-agreed measures, Traditional Chinese

<table>
<thead>
<tr>
<th>System RunID</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPAL1</td>
<td>14.62</td>
<td>60.47</td>
<td>21.36</td>
</tr>
<tr>
<td>OPAL2</td>
<td>14.64</td>
<td>49.73</td>
<td>19.57</td>
</tr>
<tr>
<td>OPAL3</td>
<td>15.02</td>
<td>77.68</td>
<td>23.55</td>
</tr>
</tbody>
</table>
Evaluation & Discussion

• Extensive filtering for topic and the temporal restrictions increases the system precision, but dramatic drop in the recall.
• Simpler methods in the cross-lingual task yielded better results, the OpAL cross-lingual run 3 obtaining the highest F score for the non-agreed measures and ranking second according to the agreed measures.
• The LSA-based method to determine topic-related words is not enough to perform this task.
• The terms obtained by employing this method are correct and useful
  – they should be expanded using language models, to better account for the language variability.
• Systems performing finer tasks, (temporal expression, anaphora resolution), are not mature enough and led to lower performances of the system and dramatic loss in recall.

NTICIR 8 MOAT, 18 June, 2009 Tokyo
Thank you for your attention