Accurate Product Name Recognition from User Generated Content

Team: ISSSID
Sen Wu, Zhanpeng Fang, Jie Tang
Department of Computer Science
Tsinghua University
ISSSID Team

• ISSSID
  – “In Science, Self-Satisfaction Is Death”
Task

SubTask1: Identify the product mentions

SubTask2: Align mentions to product entries
Challenges

• Heterogeneous data
  – Product name, category & price

• Informal text
  – from Forum

• Product recognition subtasks
  – Recognition & Alignment
Result

- 0.30379 (public score) / 0.22041 (private score)
- 1st place winner

<table>
<thead>
<tr>
<th>#</th>
<th>Δ1w</th>
<th>Team Name</th>
<th>* in the money</th>
<th>Score</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>ISSSID</td>
<td>*</td>
<td>0.22041</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>↓1</td>
<td>Olexandr Topchylo</td>
<td>*</td>
<td>0.19883</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>↓2</td>
<td>8000</td>
<td>*</td>
<td>0.18780</td>
<td>49</td>
</tr>
<tr>
<td>4</td>
<td>↓3</td>
<td>Balazs Godeny</td>
<td></td>
<td>0.18778</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>↓4</td>
<td>Labeler</td>
<td></td>
<td>0.16444</td>
<td>3</td>
</tr>
</tbody>
</table>
Figure 2. Approach Framework.
1st Basic Model: Standard Matching

- Directly use the annotated information in the training data
- Extract terms that are annotated as products in the train data
- Find corresponding terms/symbols in the test data
2nd Basic Model: Rule Templates

• Product mentions
  – Identify product mentions by rules

• Rules for recognition
  – Special words

• Rules for filter
  – Semantic patterns
  – General list
Rule Templates: Special Words

• Products’ names: combination of specific characters
  – Denon 3808CI receiver
  – Marantz VP11S2

• Special words
  – One-gram non-standard words
  – Appear no more than 20 times in the catalogs
Rule Templates: Special Words

• Find special words in text

• Identify the whole product mention
  – Construct a name tokens set using products’ names which contains the special word
  – Expand the special word on both sides if the neighbor tokens are in the set

Four million special words found?
Rule Templates: Special Words

- Too many incorrect mentions
  - ‘Goto page 1, 2, 3Next Page 1 of 3’
  - ‘Replied by mohmony’
  - ‘SIGN UP 25 MJanosh 490 Thu May 17’
  - ‘all speakers to small and raise the crossovers up to 80hz’

How to filter the product mentions?
Rule Templates: Semantic Patterns

• Most products follow a pron, prep or quantifier
  – ‘my mac’, ‘the Xbox’, ‘one GTR’…

• Preposition ‘for’ in product mentions
  – ‘Seidio Innocase 360 for BlackBerry Curve 8900’

• Words following ‘by’
  – Usually represent a person rather than a product name
  – ‘Posted by jbooker82’
Rule Templates: General List

• Several categories of words are not helped for special words
  – Stop words (e.g., his, her)
  – Capitalized nouns (e.g., January, Monday)
  – Common abbreviations (e.g., mins, kg)

• Filter special words in the above categories
  – ‘speakers to small and raise the crossovers up to 80hz’
  – ‘Then the resulting M2TS is 23fps’
Rule Templates: Other Filter Rules

• Length limitation: 2~15 characters per token

• Filter product mentions beside particular words
  – Views, replies, posts & pages

• Mixture word contains both number and letters
3\textsuperscript{rd} Basic Model: Conditional Random Field

- "Mallet"
  - A machine learning for language toolkit

- Sequence tagging model

- Three Categories
  - ‘B’: beginning of a product mention
  - ‘I’: inside a product mention
  - ‘O’: outside a product mention
# CRF: Features

<table>
<thead>
<tr>
<th>Features</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOKEN</td>
<td>Current token</td>
</tr>
<tr>
<td>FC</td>
<td>Upper-case/Lower-case of first character</td>
</tr>
<tr>
<td>CHARCNT</td>
<td>#characters</td>
</tr>
<tr>
<td>UCCNT</td>
<td>#upper-case characters</td>
</tr>
<tr>
<td>NUMCNT</td>
<td>#numeric characters</td>
</tr>
<tr>
<td>LCCNT</td>
<td>#lower-case characters</td>
</tr>
<tr>
<td>DSHCNT</td>
<td>#dash-characters</td>
</tr>
<tr>
<td>SLSHCNT</td>
<td>#slash-characters</td>
</tr>
<tr>
<td>PERIODCNT</td>
<td>#periods characters</td>
</tr>
<tr>
<td>GRWRDCNT</td>
<td>#matching grammatical words</td>
</tr>
<tr>
<td>BRNDWRDCNT</td>
<td>#matching English common words</td>
</tr>
<tr>
<td>ENWRDCNT</td>
<td>#matching brand words</td>
</tr>
<tr>
<td>P_TOKEN*</td>
<td>Previous token</td>
</tr>
<tr>
<td>P_PREP*</td>
<td>If the previous token is a preposition</td>
</tr>
<tr>
<td>PF*</td>
<td>Pattern feature</td>
</tr>
</tbody>
</table>

CRF1: the same as baseline 2; CRF2: include additional features(*)
Blending Process

- Filter the mentions that CRF recognizes by rule templates method

- Filter conflicted mentions by following priority: SM > CRF > RT

- Blend all the mentions together
Product Alignment

• Select the product items whose name contains the product mention

• Utilize product category data
  – Every product mention only belong to one category, CE or AU
  – Conformity principle
## Experiment

<table>
<thead>
<tr>
<th>No.</th>
<th>Models</th>
<th>Public Leaderboard</th>
<th>Private Leaderboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Standard Matching</td>
<td>0.14557</td>
<td>0.09005</td>
</tr>
<tr>
<td>2</td>
<td>Rule Templates</td>
<td>0.15844</td>
<td>0.11365</td>
</tr>
<tr>
<td>3</td>
<td>CRF1</td>
<td>0.16328</td>
<td>0.15775</td>
</tr>
<tr>
<td>4</td>
<td>CRF2</td>
<td>0.12168</td>
<td>0.14390</td>
</tr>
<tr>
<td>5</td>
<td>3 + 4</td>
<td>0.17375</td>
<td>0.17465</td>
</tr>
<tr>
<td>6</td>
<td>1 + 2</td>
<td>0.26525</td>
<td>0.17909</td>
</tr>
<tr>
<td>7</td>
<td>3 + 6</td>
<td>0.30656</td>
<td>0.20526</td>
</tr>
<tr>
<td>8</td>
<td>4 + 7</td>
<td>0.30379</td>
<td>0.22041(+0.02)</td>
</tr>
</tbody>
</table>

- Performance of each model is limited
- Combination can significantly improves the performance
Summary

• “Tricks” on how to win the contest
  – Rules + Statistics (+5%)
  – Blending (+6.5%)
  – Pruning (+1-5%)
Thank you!
Questions?