Comparative Web Search Questions

Alexander Bondarenko  Pavel Braslavski  Michael Völske  Rami Aly
Maik Fröbe  Alexander Panchenko  Chris Biemann  Benno Stein  Matthias Hagen

Martin-Luther-Universität Halle-Wittenberg

www.webis.de
Comparative Web Search Questions

Which is a better pet, a cat or a dog?

- Comparative information need addresses a choice problem
- Informed choice needs arguments for and against
Comparative Web Search Questions

- People make choices every day
- Post on Stack Exchange, but also submit to search engines
- Search engines do not show a pro/con result presentation

Which programming language should I learn first?

photo credit pixabay.com
Comparative Web Search Questions

Research goals:

- Identify comparative questions
- Study the underlying information needs
- Step towards showing a pro/con result presentation
Comparative Web Search Questions

1.5 billion question queries from 2012

11 million questions from 2012

- 50,000 Yandex and 12,500 Otvety questions labeled as comparative or not
- Comparative questions labeled with more fine-grained subclasses
Comparative Web Search Questions

Comparative Yandex question queries

- Opinionated/argumentative: 66%
  Which is a better pet, a cat or a dog?

- Reason/factoid: 33%
  Which river is longer, Mississippi or Nile?

- Request preference: 70%
  What is the best pet for me?
Comparative Web Search Questions

Identifying Comparative Questions

- Change a search engine’s result presentation: intricate decision
- Identifying comparative questions: precision-oriented task
- Ensemble classifier: rules + traditional + neural
- Goal: Maximize recall at precision of 1.0

Which is a better pet, a cat or a dog? comparative
What kind of mammals are cats and dogs? non-comparative
Comparative Web Search Questions

Identifying Comparative Questions

(R1) \([\text{better}] \land \neg [\text{how}]\)

Which is a better pet, a cat or a dog?   comparative

(R8) \([\text{difference(s)}]?|\text{distinguish}] \land [\text{and}|\text{from}|\text{or}|\text{vs}]\)

How to teach a dog to distinguish A and B?  non-comparative
Comparative Web Search Questions

Identifying Comparative Questions

- Tune classifiers for a maximum precision
- Supplement handcrafted rules: extend recall
Comparative Web Search Questions

Identifying Comparative Questions

- Rules + BERT + Logistic regression
- Rules + CNN + Logistic regression
- Gradually decrease classifiers decision threshold
Comparative Web Search Questions

Yandex Log Analysis

- Test classifiers
- Ensemble-B+L recall 0.60 and Ensemble-C+L recall 0.59
- Classify 1.5 billion questions in the log: BERT days, CNN hours
- Yandex log: 2.8% are comparative questions (one per second)
## Comparative Web Search Questions

### Yandex Log Analysis

<table>
<thead>
<tr>
<th>Category</th>
<th>Quest. mln.</th>
<th>Comp. %</th>
<th>Most frequently asked question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consum. electronics</td>
<td>105.4</td>
<td>6.3</td>
<td>Which tablet is best to buy?</td>
</tr>
<tr>
<td>Cars &amp; transport.</td>
<td>143.7</td>
<td>5.2</td>
<td>Anti-radar, which one is best?</td>
</tr>
<tr>
<td>Home &amp; garden</td>
<td>166.7</td>
<td>4.0</td>
<td>Which vacuum cleaner is best to buy?</td>
</tr>
<tr>
<td>Education</td>
<td>101.8</td>
<td>3.9</td>
<td>Which pilot first surpassed the supersonic speed?</td>
</tr>
<tr>
<td>Beauty &amp; style</td>
<td>93.7</td>
<td>3.3</td>
<td>When is it best to cut hair?</td>
</tr>
</tbody>
</table>

... ... ...

- More than just comparing products to buy
- Holidays in summer
- People compare vacation destinations starting from May
- In autumn differences between eatable and poisonous mushrooms
Comparative Web Search Questions

Answering Comparative Questions

- More than 65% of comparative questions are non-factoid
- Search for answers to Yandex questions on Otvety
- 48% have an answer on Otvety
Comparative Web Search Questions

Summary

- Yandex log: 2.8% are comparative questions (one per second)
- Comparison intents: beyond just comparing products to buy
- 65% of comparative questions are clearly non-factoid
- Half of them could be answered with CQA answers
- How can we answer non-factoid comparative web search questions?
Comparative Web Search Questions

Summary

- Yandex log: 2.8% are comparative questions (one per second)
- Comparison intents: beyond just comparing products to buy
- 65% of comparative questions are clearly non-factoid
- Half of them could be answered with CQA answers
- How can we answer non-factoid comparative web search questions?

Future Work

- Extract compared items and aspects
- Summarization for answers to the non-factoid questions
Comparative Web Search Questions

Summary

- Yandex log: 2.8% are comparative questions (one per second)
- Comparison intents: beyond just comparing products to buy
- 65% of comparative questions are clearly non-factoid
- Half of them could be answered with CQA answers
- How can we answer non-factoid comparative web search questions?

Future Work

- Extract compared items and aspects
- Summarization for answers to the non-factoid questions

thank you!
## Experimental Results

Classification results on the test set. Goal: Best recall at a precision of 1.0

<table>
<thead>
<tr>
<th>Individ. model</th>
<th>Recall</th>
<th>F1</th>
<th>Ensembles</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>0.54</td>
<td>0.70</td>
<td>Ens.-B+L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN</td>
<td>0.52</td>
<td>0.68</td>
<td>Ens.-C+L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT</td>
<td>0.44</td>
<td>0.61</td>
<td>Ens.-B+C+L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rules (R1–7)</td>
<td>0.44</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparative subclass classifications on the test. Goal: Maximize F1

<table>
<thead>
<tr>
<th>Subclass</th>
<th>CNN Prec.</th>
<th>CNN Rec.</th>
<th>CNN F1</th>
<th>BERT Prec.</th>
<th>BERT Rec.</th>
<th>BERT F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion/argument</td>
<td>0.93</td>
<td>0.90</td>
<td>0.92</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Reason/factoid</td>
<td>0.85</td>
<td>0.79</td>
<td>0.82</td>
<td>0.82</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>Context/aspect</td>
<td>0.88</td>
<td>0.52</td>
<td>0.62</td>
<td>0.75</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>Method</td>
<td>0.79</td>
<td>0.80</td>
<td>0.79</td>
<td>0.75</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>Preference</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.96</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>Direct</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>Superlative</td>
<td>0.93</td>
<td>0.79</td>
<td>0.86</td>
<td>0.92</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>Micro average</td>
<td>0.92</td>
<td>0.88</td>
<td>0.90</td>
<td>0.90</td>
<td>0.93</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Absolute and relative frequencies of the comparative question subclasses, which are not mutually exclusive

<table>
<thead>
<tr>
<th>Comparative</th>
<th>Yandex</th>
<th>Otvety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opinion</td>
<td>916 (65%)</td>
<td>1,469 (94%)</td>
</tr>
<tr>
<td>Argumentative</td>
<td>676 (48%)</td>
<td>586 (37%)</td>
</tr>
<tr>
<td>Reason</td>
<td>83 (6%)</td>
<td>10 (&lt;1%)</td>
</tr>
<tr>
<td>Factoid</td>
<td>378 (27%)</td>
<td>101 (6%)</td>
</tr>
<tr>
<td>Method</td>
<td>106 (8%)</td>
<td>41 (3%)</td>
</tr>
<tr>
<td>Superlative</td>
<td>180 (13%)</td>
<td>287 (18%)</td>
</tr>
<tr>
<td>Direct</td>
<td>603 (43%)</td>
<td>893 (57%)</td>
</tr>
<tr>
<td>Aspect</td>
<td>302 (22%)</td>
<td>546 (35%)</td>
</tr>
<tr>
<td>Context</td>
<td>238 (17%)</td>
<td>405 (26%)</td>
</tr>
<tr>
<td>Preference (requested)</td>
<td>985 (70%)</td>
<td>1,281 (82%)</td>
</tr>
<tr>
<td>(stated)</td>
<td>18 (1%)</td>
<td>77 (5%)</td>
</tr>
</tbody>
</table>