A Case Study of In-House Competition for Ranking Constructive Comments in a News Service

Hayato Kobayashi\textsuperscript{1}, Hiroaki Taguchi\textsuperscript{1}, Yoshimune Tabuchi\textsuperscript{1}, Chahine Koleeejan\textsuperscript{1}, Ken Kobayashi\textsuperscript{1}, Soichiro Fujita\textsuperscript{2}, Kazuma Murao\textsuperscript{3}, Takeshi Masuyama\textsuperscript{1}, Taichi Yatsuka\textsuperscript{1}, Manabu Okumura\textsuperscript{2}, Satoshi Sekine\textsuperscript{4}

\textsuperscript{1}Yahoo Japan Corporation, \textsuperscript{2}Tokyo Institute of Technology, \textsuperscript{3}VISITS Technologies Inc., \textsuperscript{4}RIKEN
Background

- Ranking user comments is important for online news services because comment visibility directly affects the user experience.
- There have been many studies on comment ranking by user feedback.
  - (Hsu+ 2009, Das Sarma + 2010; Brand&V. D. Merwe 2014; Wei+ 2016)
- However, user feedback does not always represent comment quality.

![Comments on Yahoo! JAPAN News](image.png)

Figure 1: Comments on Yahoo! JAPAN News for article “Lifting the ban on drinking/smoking at 18.”

(e.g., by position bias)
Fujita et al. (2019) introduced the concept of constructiveness in argument analysis for ranking comments without biased user feedback.

- Constructiveness has no correlation with user feedback (Like/Dislikes).

Table 1: Conditions for constructive comments.

<table>
<thead>
<tr>
<th>Pre</th>
<th>Related to article and not libelous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main</td>
<td>Intended to stimulate discussions</td>
</tr>
<tr>
<td></td>
<td>Objective and supported by fact</td>
</tr>
<tr>
<td></td>
<td>New idea, solution, or insight</td>
</tr>
<tr>
<td></td>
<td>User’s unique experience</td>
</tr>
</tbody>
</table>

- Maintain decency and relevance
- Represent typical cases of being constructive
This Work

Approach
- Take Fujita et al.’s study one step further towards practical application.
  - Key aspect: Performance improvement by in-house competition.

Contributions
- Report the details of the in-house competition in Yahoo! JAPAN News.
  - 2.73% improvement in performance (NDCG) against the baseline.
- Consider several ensembles of the submitted various models.
  - 0.62% improvement in NDCG against the best single model.
In-House Competition

Task

- Ranking comments based on their constructiveness scores (C-scores).
  - C-score = a graded numeric score representing the level of constructiveness.

Dataset

- 59,120 comments (9,845 articles with about 6 comments).
  - Including 995 long comments (with 126-400 characters).

Evaluation

- NDCG: \( \frac{1}{K} \sum_{k=1}^{K} \text{NDCG}@k \)
  \[ \text{NDCG}@k = Z_k \sum_{i=1}^{k} \frac{2^r_i - 1}{\log_2(i+1)} \]
- NDCG-L: NDCG only for the long comments (sub measure).
  - To avoid sloppy methods that determine long comments to be constructive.

Table 1: Conditions for constructive comments.
Submission Trend

- Number of submissions increased at the beginning of work (where time is more available) and on the day of the deadline.
- 8 individuals submitted:
  - 14 models during the competition period (before the deadline).
  - +4 models after the deadline.
- Total 18 models for research.

Figure 2: Cumulative number of submissions over the competition period.
- Many models performed better than Baseline.
- Highest performance increase was 2.73% by Model-17 for NDCG.
- Use of the leaderboard had a positive effect for participants submitting high-performance models for both measures in the latter half of the competition.

Figure 3: Increase (%) in NDCG (top) and NDCG-L (bottom) for each model compared to Baseline.

**Baseline**: A linear rankSVM model with features based on term-frequency vectors.
High-performance Models

- **Model-4**: Highest NDCG (before the deadline).
  - A gradient boosting model with features based on pretrained word embeddings.

- **Model-11**: Highest sum of NDCG and NDCG-L.
  - A linear rankSVM model with features based on C-score prediction (= stacking) and the distance between an article and its comment.

- **Model-14**: Highest NDCG-L.
  - A gradient boosting model with features based on maximal substrings and words.

- **Model-17**: Highest NDCG (after the deadline).
  - A variant of the RankNet model (BiLSTM+GCNN) with features based on subwords.
Ensemble of Submitted Models (Trial after Competition)

- Prepared 4 simple and 2 recent ensemble methods.

- **NormAve**: Use the average of the predicted scores of all models after normalizing the scores (Burges+ 2011).

- **WeightEval**: Use the weighted average of the top-k promising predictions (Fujita+ 2020), which is a hybrid of (continuous) majority voting and averaging.

(The other methods are omitted due to time constraint.)
Results of Ensemble Models

- WeightEval performed the best for the main measure NDCG.
  - 0.62% improvement against the best single model.

- NormAve is the most promising for practical use (no parameter tuning).

### Table 2: NDCG variants (%) and precision (%) for (a part of) the submitted models and their ensembles.

<table>
<thead>
<tr>
<th>Model</th>
<th>NDCG</th>
<th>NDCG-L</th>
<th>NDCG@3</th>
<th>Prec@3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>81.63</td>
<td>86.74</td>
<td>81.09</td>
<td>73.30</td>
</tr>
<tr>
<td>Model-4</td>
<td>83.60</td>
<td>82.15</td>
<td>82.79</td>
<td>73.98</td>
</tr>
<tr>
<td>Model-11</td>
<td>83.35</td>
<td>88.34</td>
<td>82.93</td>
<td>73.20</td>
</tr>
<tr>
<td>Model-14</td>
<td>82.53</td>
<td><strong>88.77</strong></td>
<td>81.83</td>
<td>72.86</td>
</tr>
<tr>
<td>Model-17</td>
<td>83.86</td>
<td>88.24</td>
<td>83.27</td>
<td>72.01</td>
</tr>
<tr>
<td>ScoreAve</td>
<td>83.85</td>
<td>86.66</td>
<td>83.20</td>
<td>73.40</td>
</tr>
<tr>
<td>NormAve</td>
<td>84.33</td>
<td>88.41</td>
<td><strong>84.01</strong></td>
<td><strong>74.11</strong></td>
</tr>
<tr>
<td>RankAve</td>
<td>83.46</td>
<td>88.25</td>
<td>82.92</td>
<td>73.30</td>
</tr>
<tr>
<td>TopkAve</td>
<td>84.35</td>
<td>88.35</td>
<td>83.31</td>
<td>73.54</td>
</tr>
<tr>
<td>PostEval</td>
<td>84.32</td>
<td>88.64</td>
<td>83.88</td>
<td>73.91</td>
</tr>
<tr>
<td>WeightEval</td>
<td><strong>84.38</strong></td>
<td><strong>88.30</strong></td>
<td><strong>84.18</strong></td>
<td>74.04</td>
</tr>
</tbody>
</table>
Towards Practical Use

- Qualitative evaluation from the perspective of service.
  - 3 service experts ranked the comment lists created by candidate models.
  - Criterion: Which list should be provided as a service?

- Two cases:
  - Baseline vs. naive methods.
  - Baseline vs. submitted models.
  - Service preferred not to use ensemble models because it would be unreasonable to maintain different models.
Baseline vs. Naive Methods

- **Feedback**: Descending/ascending order of number of Likes/Dislikes.
- **Latest**: Descending order of comment date.
- **Length**: Descending order of comment length.

- Baseline (C-score) clearly performed better than the other methods.
- Constructiveness is useful even in human evaluation, while the previous study (Fujita+ 2019) used NDCG only.

<table>
<thead>
<tr>
<th></th>
<th>Average Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback</td>
<td>2.61</td>
</tr>
<tr>
<td>Latest</td>
<td>3.42</td>
</tr>
<tr>
<td>Length</td>
<td>2.20</td>
</tr>
<tr>
<td>Baseline (C-score)</td>
<td>1.77</td>
</tr>
</tbody>
</table>

Table 3: Qualitative evaluation results of Baseline and naive methods (lower ranks are better).
Baseline vs. Submitted Models

- Prepared the four high-performance single models.
  - Model-4 (GBM with word embeddings), Model-11 (rankSVM with stacking), Model-14 (GBM with maximal substrings), Model-17 (RankNet with subwords).

- Best single model (Model-17) also had the best average rank.

- Competition format is effective even in a service-level judgment.

<table>
<thead>
<tr>
<th>Model</th>
<th>Average Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>3.86</td>
</tr>
<tr>
<td>Model-4</td>
<td>3.64</td>
</tr>
<tr>
<td>Model-11</td>
<td>3.63</td>
</tr>
<tr>
<td>Model-14</td>
<td>3.41</td>
</tr>
<tr>
<td>Model-17</td>
<td><strong>3.11</strong></td>
</tr>
</tbody>
</table>

Table 4: Qualitative evaluation results of submitted models and Baseline (lower ranks are better).
Conclusion

Summary
- Reported the details of the in-house competition in Yahoo! JAPAN News.
  - 2.73% improvement in performance (NDCG) against the baseline.

Discussion
- Service decision suggests that while an ensemble of different models is promising in an academic sense, it still has challenges in an industrial sense.
  - Model unification/distillation for improving maintainability and latency?
Thank you!