Introduction

Background

- Task: Ranking comments in each article w.r.t. a quality measure
- Motivation: Improve comment visibility for the user experience

Previous work:
- Quality measure = users’ positive feedback (e.g., ‘Like’)
- Drawback1: Biased by where the comment appears (Position Bias)
- Drawback2: Biased by the majority of users, especially for political views

Approach

- Directly evaluate the quality of comments
  - Constructiveness score (C-Score)
  - Investigate how to label comments
    - i.e., which to pay attention: Comment or article variation

Contributions

- Create a dataset for ranking constructive comments
  - Including 100K Japanese comments with constructiveness scores
  - Our datasets will be available (https://research-lab.yahoo.co.jp/en/software)
- Show empirical evidence that C-scores aren’t always related to user feedback
- Clarify the performance of pairwise ranking models tends to be more enhanced by the variation in comments than in articles

Dataset Creation

Definition for “Constructiveness”

- Definition of dictionary:
  - “Having or intending to have a useful or beneficial purpose.”
- Definition in this work:
  - Digested version of the definition in (Kolhatkar+, 2017)

Crowdsourcing Task

- Goal: Labeling each comment with a graded numeric score (C-Score)
  - Difficulty: Constructiveness includes some ambiguity
    → Hard to answer a numerical question such as or a comparison question (e.g., “How constructive is it?” / “Which is more constructive?”)
  - CS Task: Judge a comment to be constructive by a yes-or-no (binary) question
  - Label: # of crowdsourcing workers who judged the comment to be constructive

Training and Test Datasets

- Data structure: (article, comment, C-score)
  - Training dataset: Randomly selected comments in each article
    - Shallow: 40K comments with article variation (5 comments × 8K articles)
    - Deep: 40K comments with comment variation (100 comments × 400 articles)
  - Test dataset: All comments in each article
    - Krippendorff’s alpha (relative comp.)
      - Shallow: 0.53, Deep: 0.55

Comparison with User Feedback

Setting

- Investigate the relationship between constructiveness and user feedback
- Comparing C-scores of 5K comments (5 comments × 1K articles) extracted by
  - Like: Descending order of user feedback score
  - Random

Result

- The correlation coefficient between user feedback scores and C-scores was nearly zero (-0.036)
- Constructiveness is completely different from user feedback

Dataset Creation for Ranking Constructive News Comments

Soichiro Fujita,† Hayato Kobayashi,‡ Manabu Okumura†
† Tokyo Institute of Technology
‡ Yahoo Japan Corporation / RIKEN AIP

Related Work

- Ranking comments on news/discussion services
  - Previous studies (Wei+, 2016, …) only used user feedback
  - User feedback is completely different from constructiveness
- Analyzing constructiveness
  - Previous studies (Kolhatkar+, 2017, …) addressed binary classification tasks
  - Our task is a ranking task based on graded numeric scores
- Other approaches to analyze the quality of comments
  - Sentiment analysis (Fan+, 2010, …), hate speech detection (Kwok+, 2013, …)
  - Not suitable in this task (e.g., “Great!” is not constructive)

Future Work

- Labeling promising comments via active learning
- Evaluation with A/B testing on the running service
- Ranking constructive comments while keeping their diversity