Dataset Creation for Ranking Constructive News Comments

Soichiro Fujita,† Hayato Kobayashi,‡ Manabu Okumura†

† Tokyo Institute of Technology

‡ Yahoo Japan Corporation / RIKEN AIP

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 view

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 that in articles***

Dataset Creation

Definition for "Constructiveness"

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

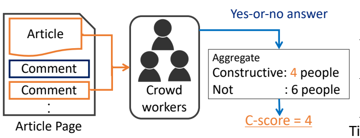
Precondition	• Related to article and not slander
Main condition	• Intent to cause discussions
	• Objective and supported by fact
	• New idea, solution, or insight
	• User's rare experience

Maintaining decency and relevance to an article

Typical cases of being constructive

Crowdsourcing Task

- Goal: Labeling each comment with a graded numeric score (C-Score)
  - Difficulty: Constructiveness includes some ambiguity
    → Hard to answer a numerical selection question 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



Comment	Score
Ex.1) We should build a society where people do not drink and smoke since both can lead to bad health or accidents.	9
Ex.2) If giving freedom, punishment should also be strictly given.	6
Ex.3) They are fools because they smoke, or they smoke because they are fools.	0

Title: *"Lifting the ban on drinking and smoking at 18."*

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
  - Simulate a real situation
- Krippendorff's alpha (relative comp.)
  - *Shallow*: 0.53, *Deep*: 0.55

	#A	#C	#C/#A	Score
Shallow	8,000	40,000	5	0 ~ 10
Deep	400	40,000	100	0 ~ 10
Test	200	42,436	212	0 ~ 40

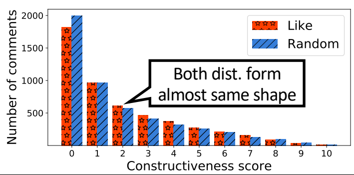
Comparison with User Feedback

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* (-0.0036)
- Constructiveness is completely different from user feedback



Ranking Constructive News Comments

Compared Methods

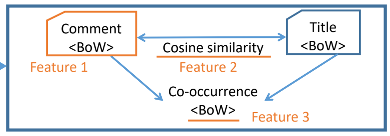
- Like, Random
  - Ranks with the user feedback score / Ranks randomly
- Length
  - Ranks in descending order on the basis of the comment length
- RankSVM (Lee+, 2014)
  - Ranks via a linear rankSVM model
  - Trained to predict relative constructiveness between two comments
- SVR (Vapnik+, 1997)
  - Ranks via a support vector regression model with a linear kernel
  - Trained to directly predict the C-score

Preprocessing and Features

Preprocessing

私は25歳です。
(I am 25 years old.)
↓
私, は, <num>, 歳, です, 。
(I, am, <num>, years, old, .)

Feature extraction



Evaluation

- NDCG@k: Normalized Discounted Cumulative Gain $Z_k \sum_{i=1}^{k} \frac{r_i}{\log_2(i+1)}$ (Normalization constant: Z_k; True C-score of the i-th ranked comment: r_i)
  - Widely used for evaluating ranking models in information retrieval tasks
  - NDCG becomes higher as the inferred ranking becomes closer to the correct ranking, especially for top ranked comments
- Precision@k
  - Ratio of correctly included comments in the inferred top-k comments with respect to the true top-k comments

Results

- Neither of *Like* and *Random* performed well
- *Length* performed better than *Like* and *Random*
- *RankSVM*: Performed better with *Deep* than with *Shallow*
  - Reason: The number of pairwise examples increases in $O(n^2)$
- *SVR*: Performed better with *Shallow* than with *Deep*
  - Reason: Features based on articles can be useful for directly inferring the C-scores without comparing comments

Overall

- NDCG: *RankSVM* with *Deep* consistently performed the best
  - Differences between NDCGs of *RankSVM* with *Deep* and *SVR* with *Shallow* were statistically significant in a paired t-test ($p < 0.05$)
- Prec: *RankSVM* with *Deep* was beaten by *SVR* with *Shallow*
  - *RankSVM* failed to find the best solutions (the most constructive comment) but obtained better solutions (fairly constructive ones)
- Note: Neural ranking model got consistent results with these finding

	Dataset	NDCG@1	NDCG@5	NDCG@10	Prec@1	Prec@5	Prec@10
Like	-	29.93	31.84	34.99	2.00	6.20	8.70
Random	-	25.85	27.90	29.06	1.10	4.60	6.50
Length	-	60.28	64.93	67.72	6.00	20.80	30.04
RankSVM	Shallow	72.24	74.63	76.79	14.50	29.40	41.24
RankSVM	Deep	**74.15**	**76.44**	**78.25**	13.00	31.60	**42.20**
SVR	Shallow	73.87	75.48	76.97	**16.50**	**32.70**	41.00
SVR	Deep	69.68	71.99	74.26	11.00	27.20	36.35

Related/Future Work

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