Dataset Creation for Ranking Constructive News Comment

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

Introduction

Background

- Many studies on ranking comments on an online news service (Wei+ 2016, …)
- Using users' positive feedback for a comment as the quality measure
  - Drawback1: Not be evaluated comments' quality
  - Drawback2: Be biased by where comments appear in a thread

Approach

- Propose new quality measure: **constructiveness score (C-score)**
  - Directly evaluate the quality of comments

Contributions

- Create a dataset for ranking constructive comments
  - Including 100K+ Japanese comments with constructiveness scores
  - Our datasets will be available
- Show empirical evidence that C-scores are not related to positive feedback
- Investigate how to create a dataset with awareness of *comment variation* or *article variation*

Dataset Creation

Definition for "Constructiveness"

- Definition of dictionary:
  "having or intended to have a useful or beneficial purpose."
- Definition in this work:
  Satisfying the precondition and at least one of main conditions(Kolhatkar+,2017)
  - Precondition: maintaining decency and relevance to an article
  - Main condition: typical cases of being constructive

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

Crowdsourcing Task

- Goal: labeling each comment with a graded numeric score (C-Score)
  - Difficulty: the quality of comment are ambiguous
    → Hard to answer a numerical selection task or a comparison task
- 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 dataset

- Data structure: Triplet of (article title, comment, C-Score)
- **Training dataset**: use randomly selected comments in the article
  - *Shallow*: 5 comments per article (awareness of *article variation)*
  - *Deep*: 100 comments per article (awareness of *comment variation)*
- **Test dataset**: use all comments in the article
  - Simulate a real situation
- Agreement scores (Krippendorff's alpa)
  - *Shallow*: 0.5282 *Deep*: 0.5495

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

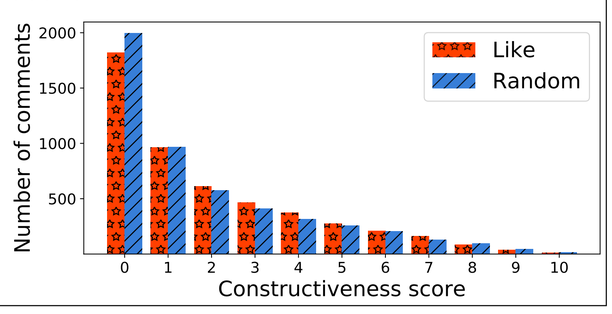
Comment	Score
Ex.1) We should build a society where people do not drink and smoke since both can lead to bad health or accidents.	9
Ex.2) If giving freedom, punishment should also be strictly given.	6
Ex.3) They are fools because they smoke, or they smoke because they are fools.	0

Table 3: Examples of comments and scores for article "Lifting the ban on drinking and smoking at 18."

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
  - Descending order of user feedback score
  - Random

Result
- Both distributions form almost the same shape
- The correlation coefficient between the user feedback and C-scores is *nearly zero* (-0.0036)



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*
  - Ranks via a RankSVM model (Lee+, 2014)
- *SVR*
  - Ranks via a support vector regression model (Vapnik+, 1997)

Preprocessing and Features

- Preprocessing
  - Split Japanese text into words
  - Replace numbers with a special token and standardized letter types
- Features for *RankSVM* and *SVR*
  - The bag-of-words of the comment
  - The cosine similarity between the comment and the title
  - The bag-of-words co-occurring in the comment and the title

Evaluation

- NDCG@k: Normalized Discounted Cumulative Gain
  - Typically calculated for the top-k ranking $\mathrm{NDCG@}k = Z_k \sum_{i=1}^{k} \frac{r_i}{\log_2(i+1)}$
- Precision@k
  - The ratio of correctly included comments in the inferred top-k comments with respect to the true top-k comments

Results

- *RankSVM* with *Deep* consistently performed the best for NDCG

	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

Table 4: Results (%) of NDCG@k and precision@k for task of ranking constructive comments.

Discussion

Relationships to Existing Communities

- Ranking comments on online news services or discussion forums
  - Main: Based on user feedback
  - New: Based on constructiveness (see **Comparison with User Feedback**)
- Analyzing constructiveness
  - Main: Classify *constructive* or *not* for a discussion thread or a comment
  - New: Rank comments in descending order of constrictiveness
- Other approaches to analyze the quality of comments
  - Sentiment analysis, hate speech detection
  - Not suitable in this task (i.e.) the simple comment ``Great!''

Future Work

- Active learning to select efficiently labeling promising comments
- Evaluate on the real world service
- Rank with considering diversity of comments

Appendix: Result of Neural Models

Models

- *RankNet*
  - Word2vec + LSTM encoder + RankNet
- *LSTMReg*
  - Word2vec + LSTM encoder + Linear Regression

Results

- Got consistent results with the results of SVM-based models

	Dataset	NDCG@1	NDCG@5	NDCG@10	Prec@1	Prec@5	Prec@10
RankNet	Shallow	73.42	73.91	75.11	**13.67**	27.40	37.81
RankNet	Deep	**75.19**	**77.17**	**78.62**	13.17	**31.72**	**41.68**
LSTMReg	Shallow	71.71	73.96	75.74	12.68	28.48	38.99
LSTMReg	Deep	69.40	72.51	74.21	10.55	26.75	36.28

↓ old version w/o appendix

Dataset Creation for Ranking Constructive News Comment

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

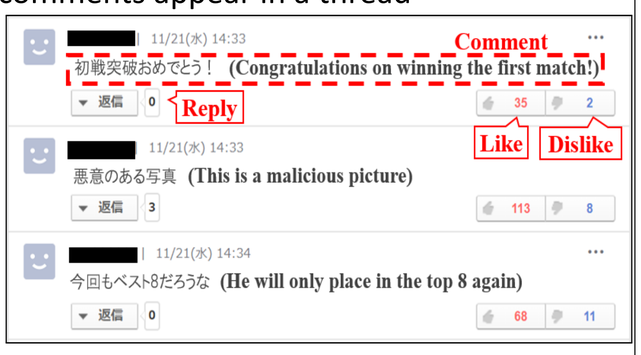
Introduction

Background

- Many studies on ranking comments on an online news service (Wei+ 2016, …)
- Using users' positive feedback for a comment as the quality measure
  - Drawback1: Not be evaluated comments' quality
  - Drawback2: Be biased by where comments appear in a thread

Approach

- Propose new quality measure: **constructiveness score (C-score)**
  - Directly evaluate the quality of comments



Contributions

- Create a dataset for ranking constructive comments
  - Including 100K+ Japanese comments with constructiveness scores
  - Our datasets will be available
- Show empirical evidence that C-scores are not related to positive feedback
- Investigate how to create a dataset with awareness of *comment variation* or *article variation*

Dataset Creation

Definition for ``Constructiveness''

- Definition of dictionary:
  "having or intended to have a useful or beneficial purpose."
- Definition in this work:
  Satisfying the precondition and at least one of main conditions
  - Precondition: maintaining decency and relevance to an article
  - Main condition: typical cases of being constructive

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

Crowdsourcing Task

- Goal: labeling each comment with a graded numeric score (C-Score)
  - Difficulty: the quality of comment are ambiguous
    → Hard to answer a numerical selection task or a comparison task
- 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
- Filtering rules for labeling:
  1. news articles with more than 100 comments
  2. Comments with 10 to 125 Japanese characters long

Training and Test dataset

- Data structure: Triplet of (article title, comment, C-Score)
- **Training dataset**: use randomly selected comments in the article
  - Deep: 100 comments per article (awareness of *comment variation)*
  - Shallow: 5 comments per article (awareness of *article variation)*
- **Test dataset**: use all comments in the article
  - Simulate a real situation
- Agreement scores (Krippendorff's alpa)
  - Shallow: 0.5282 Deep: 0.5495

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

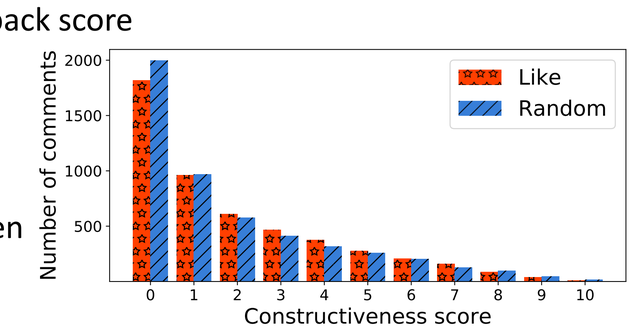
Comment	Score
Ex.1) We should build a society where people do not drink and smoke since both can lead to bad health or accidents.	9
Ex.2) If giving freedom, punishment should also be strictly given.	6
Ex.3) They are fools because they smoke, or they smoke because they are fools.	0

Table 3: Examples of comments and scores for article "Lifting the ban on drinking and smoking at 18."

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
  - Descending order of user feedback score
  - Random

Result
- Both distributions form almost the same shape
- The correlation coefficient between the user feedback and C-scores is *nearly zero* (-0.0036)



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
  - Ranks via a RankSVM model (Lee+, 2014)
  - Trained to predict relative constructiveness between two comments
- SVR
  - Ranks via a support vector regression model (Vapnik+, 1997)
  - Trained to directly predict the C-score

Preprocessing and Features

Preprocessing
- Split Japanese text into words
  - MeCab (Kudo+, 2004) / Neologd (sato+, 2017)
- Replace numbers with a special token and standardized letter types

Features for *RankSVM* and *SVR*
- The bag-of-words of the comment
- The cosine similarity between the comment and the title
- The bag-of-words co-occurring in the comment and the title

Evaluation

- NDCG@k: Normalized Discounted Cumulative Gain
  - Typically calculated for the top-k ranking $\mathrm{NDCG@}k = Z_k \sum_{i=1}^{k} \frac{r_i}{\log_2(i+1)}$
  - NDCG becomes higher as the inferred ranking becomes closet to the correct ranking, especially for top ranked comments
- Precision@k
  - The ratio off 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*
- **Overall**
  - *RankSVM* with *Deep* consistently performed the best for NDCG
  - The differences between NDCGs of *RankSVM* with *Deep* and *SVR* with *Shallow* were statistically significant in a paired t-test ($p < 0.05$)
- **RankSVM**
  - Performed better with *Deep* than with *shallow*
- **SVR**
  - Performed better with *Shallow* than with *Deep*

	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

Table 4: Results (%) of NDCG@k and precision@k for task of ranking constructive comments.

Discussion

Relationships to Existing Communities

- Ranking comments on online news services or discussion forums
  - Main: Based on user feedback
  - New: Based on constructiveness (see **Comparison with User Feedback**)
- Analyzing constructiveness
  - Main: Classify *constructive* or *not* for a discussion thread or a comment
  - New: Rank comments in descending order of constrictiveness
- Other approaches to analyze the quality of comments
  - Sentiment analysis, hate speech detection
  - Not suitable in this task (i.e.) the simple comment ``Great!''

Future Work

- Active learning to select efficiently labeling promising comments
- Evaluate on the real world service
- Rank with considering diversity of comments