Automatic Evaluation of Discussion Quality using Topic Relevance and Participants' Performance

Yoko Nishihara (Ritsumeikan University, Japan)
Wataru Sunayama (The University of Shiga Prefecture, Japan);
Shiho Imashiro (Recruit Management Solutions Co., Ltd., Japan)

Background

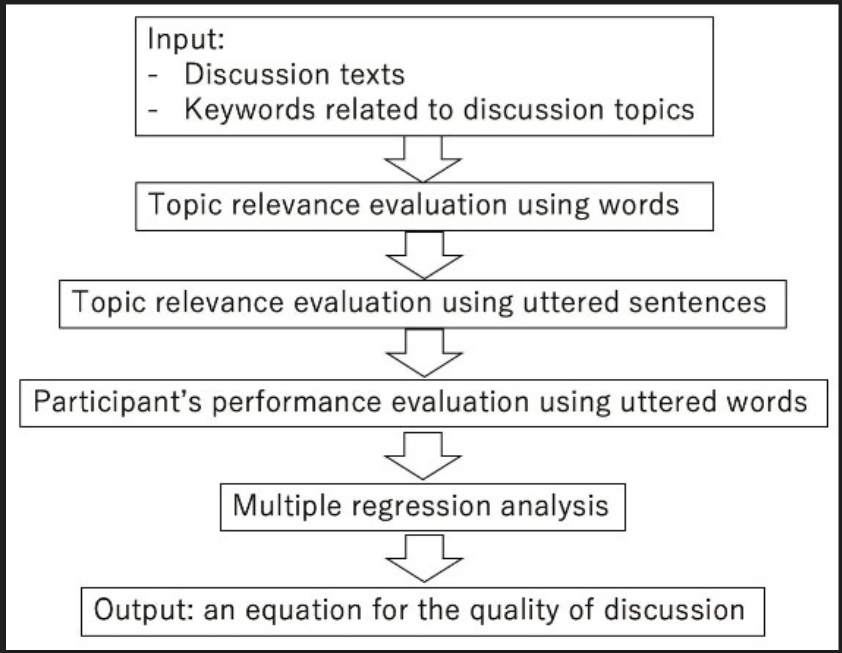
- People often have discussions with others in working cooperatively.
  - Sharing their opinions
  - Deciding their goals in the work
- It would be nice to have high-quality discussion in a limited time duration.
- It is difficult for the discussion members to objectively assess the discussion's quality.
- It is also challenging to find what is the cause of low-quality discussion.

Objective of the research

- We propose an automatic evaluation method of the discussion quality.
- We have two assumptions:
  - (1) If a discussion has much information related to discussion topics, the quality of discussion must become high.
  - (2) If more participants give their opinion on discussion topics, the quality of discussion must become high.

Proposed method

- Input: Discussion texts and keywords
- Output: An equation for the quality of discussion
- The quality of discussion is evaluated from three points:



Flowchart of the proposed method.

Input: Discussion texts and keywords related to discussion topics

- A set of discussion text and keywords related to discussion topics are inputted.
- In the discussion text, one line has two types of information:
  - (1) participant's name
  - (2) his/her utterance text
- The keywords are given by a method user.
  - A list of words with their frequency is given.
  - The user selects words as keywords from the list.

Discussion text

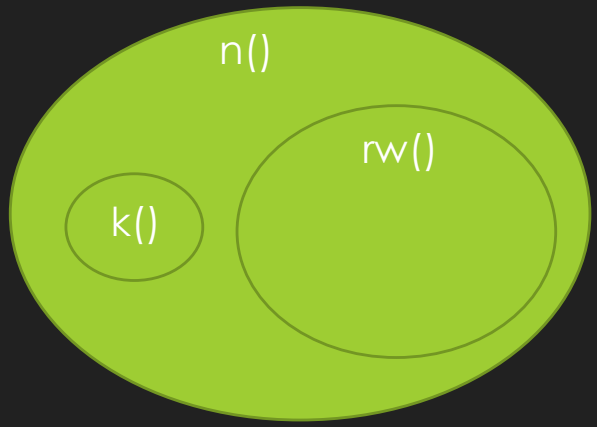
Participant 1: utterance 1
Participant 2: utterance 2
Participant 3: utterance 3
Participant 2: utterance 4
Participant 1: utterance 5
…

List of words with their frequency

Word 1: frequency
Word 2: frequency
Word 3: frequency
Word 4: frequency
…

Evaluation point 1: Topic relevance evaluation using words

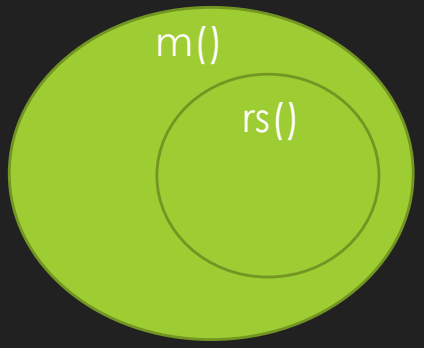
- If a discussion is highly related to discussion topics, the amount of words related to it becomes high.
- The topic relevance of the discussion is evaluated by the rate of words related to the topics by TRW(di).
- $TRW(d_i) = \frac{k(d_i)+rw(d_i)}{n(d_I)}$
  - di: discussion text
  - k(): number of keywords
  - rw(): number of words related to keywords
  - n(): number of words in a discussion text
- More the sum of k() and rw() is, more TRW() is.



Relation of three word sets

Evaluation point 2: Topic relevance evaluation using uttered sentences

- If a discussion is highly related to discussion topics, the amount of uttered sentences related to discussion topics becomes high.
- The topic relevance of the discussion is evaluated by the rate of uttered sentences related to the topics by TRS().
- $TRS(d_i) = \frac{rs(d_i)}{m(d_i)}$
  - rs(): the number of uttered sentences related to a topic
  - m(): the number of uttered sentences in a discussion text



Relation of two sentences sets

Evaluation point 3: Participant's performance evaluation using uttered words

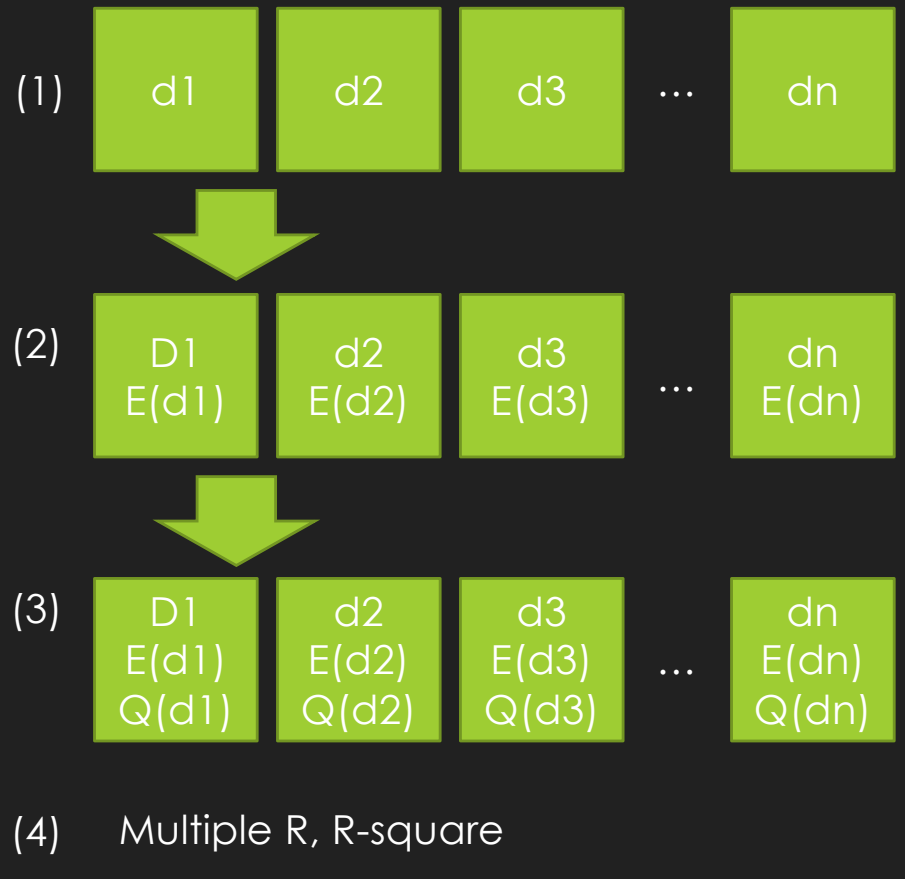
- If the quality of discussion is high, each participant gives topic-related utterances.
- The participant's performance is evaluated by the number of words related to the topics in utterances by PP(di).
- $PP(d_i) = minimum(rs2(p_i))$
  - pi: a participant
  - rs2(): the number of uttered sentences related to a topic
  - minimum: a function to obtain the lowest value among the participants
- PP() evaluates the lowest performance participant in a group.

Multiple regression analysis with three evaluation values

- Multiple regression analysis is conducted to generate an equation for the discussion quality.
- The three values (TRW(), TRS(), and PP()) are normalized to use for the analysis.
- The equation is given by Q(di):
- $Q(d_i) = \alpha \times C_1\big(TRW(d_i)\big) + \beta \times C_2\big(TRS(d_i)\big) + \gamma \times C_3\big(PP(d_i)\big) + \delta$
  - C_1, C_2, are C_3 are normalized functions.
  - α, β, γ, and δ are weights.

Evaluation experiment

- Procedures:
  - (1) Discussion data is prepared and formatted.
  - (2) A human annotator gives a correct value for the quality to each discussion as E(dn).
  - (3) Proposed method generates a equation for the quality and gives a value as Q(dn).
  - (4) Multiple R and R-square given by MRA are checked.



Discussion data (1/2)

- The discussion data was obtained from a workshop held by a business company.
  - The workshop aimed to build a new business plan for a given issue.
- There were 20 groups.
  - The number of group member ranged from 3 to 5.
- The duration was about 20 minutes for a discussion.
- Each discussion were converted to as transcripts by human annotators.

Discussion data (2/2)

- Each discussion was evaluated for research purpose by a person in charge of each discussion from the following three points:
  - E1: Overall evaluation, 4-scale points (4: excellent, 1: not good)
  - E2: Activity evaluation for group members, 3-scale points
  - E3: Evaluation for constructive discussion, 3-scale points
- The sum of E1, E2, and E3 was given as the correct discussion quality.
- There were two annotators. One annotator observed and evaluated the discussion of each of the 10 groups.

Discussion data
and scores given by annotators

Group	Participants	E1	E2	E3	Annotated score
A	4	3	3	2	8
B	4	4	3	3	10
C	4	1	2	2	5
D	4	2	2	2	6
E	4	4	2	2	8
F	4	1	2	1	4
G	4	4	2	2	8
H	4	4	3	2	9
I	5	1	1	1	3
J	4	4	3	3	10
K	4	1	1	2	4
L	4	4	3	3	10
M	3	3	3	3	9
N	4	3	3	3	9
O	4	2	2	1	5
P	4	2	2	2	6
Q	3	3	3	3	9
R	4	2	2	2	6
S	4	4	2	3	9
T	4	3	3	2	8

Correct discussion quality ↑

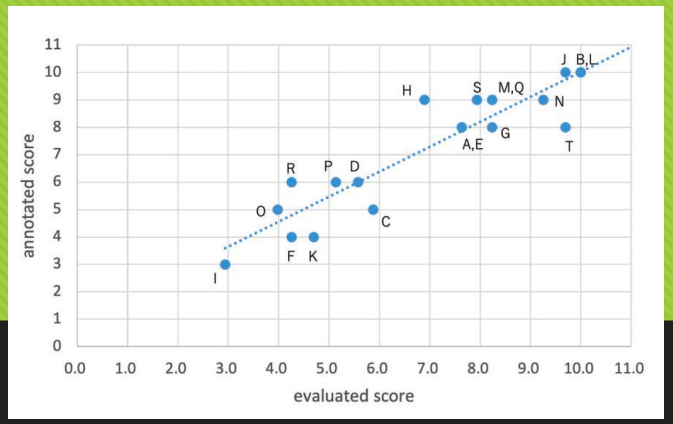
Selected keywords as discussion topics

- An experimenter read the discussion texts and selected keywords as the discussion topics.
- The number of keywords ranged from 2 to 6.

Selected keywords as discussion topics

Group	Keywords as discussion topics
A	bonsai (a potted dwarf tree) and 2 other words
B	craftwork and 2 other words
C	child and 1 other word
D	student and 1 other word
E	online and 3 other words
F	environmental load and 1 other word
G	waste and 5 other words
H	Indonesia and 4 other words
I	internship and 1 other word
J	education and 1 other word
K	health and 1 other word
L	China and 5 other words
M	tourist and 3 other words
N	capable person and 4 other words
O	volunteer and 3 other words
P	industry and 3 other words
Q	online and 4 other words
R	ability and 4 other words
S	company and 3 other words
T	business and 3 other words

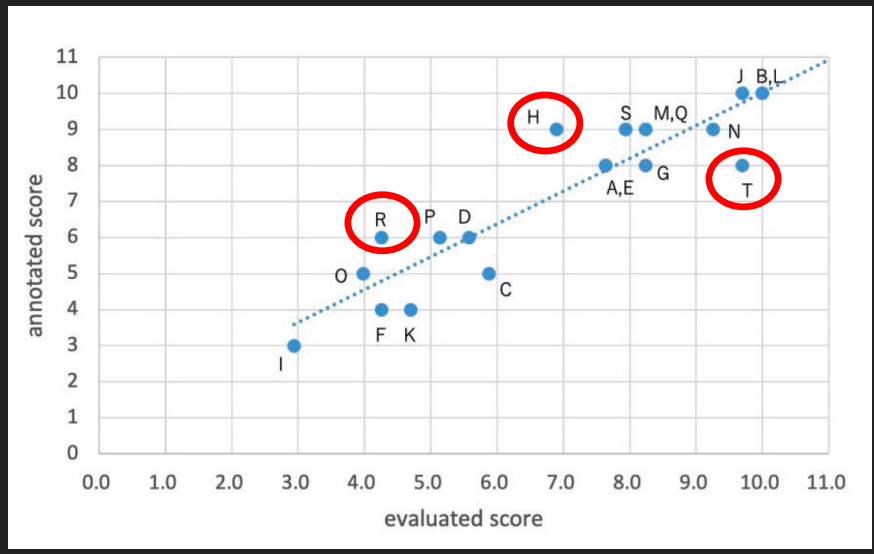
Experimental result



- Obtained Q() was:
- $Q(d_i) = \alpha \times C_1\big(TRW(d_i)\big) + \beta \times C_2\big(TRS(d_i)\big) + \gamma \times C_3\big(PP(d_i)\big) + \delta$
  - α: 1.60
  - β: 0.40
  - γ: 0.27
  - δ: 0.91
- Multiple R: 0.92, R-square: 0.85, standard error: 0.93
- We obtained high-accurate equation for the quality of discussion.

Discussion: evaluation of generated equation Q()

- Figure shows a correlation between annotated scores and evaluated scores.
- Most of the discussion texts fit a line by Q().
- Texts of H, R, and T were far from the line.
  - Because the value PP() of H, R, and T were different from others.
  - PP() evaluates participants' performance by their utterances related to discussion topics.
  - The length of an utterance was depend on speakers.
  - We will consider the length of utterance in the evaluation.



Discussion: number of keywords for topic

- We assume if a discussion has much information related to topics, the discussion quality must become high.
- However, the number of keywords were not considered in the evaluation experiment.
- Not only the granularity but also the relation of topics may be related to the quality.
- We will survey the effect of the relation of topic keywords for the discussion quality.

Selected keywords as discussion topics

Group	Keywords as discussion topics
A	bonsai (a potted dwarf tree) and 2 other words
B	craftwork and 2 other words
C	child and 1 other word
D	student and 1 other word
E	online and 3 other words
F	environmental load and 1 other word
G	waste and 5 other words
H	Indonesia and 4 other words
I	internship and 1 other word
J	education and 1 other word
K	health and 1 other word
L	China and 5 other words
M	tourist and 3 other words
N	capable person and 4 other words
O	volunteer and 3 other words
P	industry and 3 other words
Q	online and 4 other words
R	ability and 4 other words
S	company and 3 other words
T	business and 3 other words

Conclusion

- This paper proposed an automatic evaluation method for the discussion quality.
- We conducted an evaluation experiments using the real discussion data.
- It was found that our method could evaluate the quality of discussion with high-accuracy.
- We will improve our method to use in evaluation team's activity.