Analyzing Differences in Subjective Annotations by Participants and Third-party Annotators in Multimodal Dialogue Corpus

Kazunori Komatani, Ryu Takeda (SANKEN, Osaka University), Shogo Okada (JAIST)

Background and objectives

User's subjective impression is important

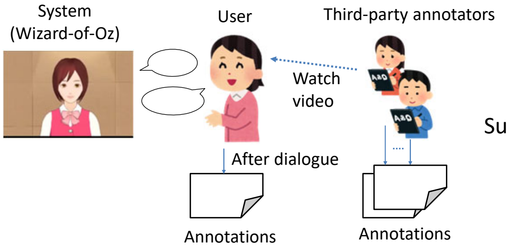
- The system should respond adaptively to it

Subjective impressions are inherently ambiguous

- not always agree among annotators (and with users themselves)

Analyses in multimodal dialogue corpus Hazumi

I. Users themselves vs. third-party annotators
II. Use of dispersion of third-party annotation results



Multimodal dialogue corpus Hazumi

Publicly available

- Movies: a written oath is required
  https://www.nii.ac.jp/dsc/idr/rdata/Hazumi/
- Annotations, feature files, etc
  https://github.com/ouktlab/Hazumi2010, etc.

Version numbers: the year and month when the data collection started

				Subjective annotations			
Version	**Recorded env.**	**No. of participants (dialogues)**	**No. of exchanges**	**Self-sentiment**	**Third-party sentiment**	**18 rapport items**	**Personality traits**
Hazumi1712	In-person	29	2,422		○		
Hazumi1902		30	2,514	○	○	○	
Hazumi1911		30	2,859	○	○	○	○
Hazumi2010	Online	33	2,798		○	○	○
Hazumi2012		63	5,334		○	○	○
Hazumi2105		29	2,235		○	○	○
Total		214	18,162				

Subjective Annotations

- Sentiment: how much the user enjoyed the dialogue in the exchange (7-point scales)
- 18 rapport items: 18 questionnaire items about the dialogue (8-point scales)
- Personality traits: the user's Big Five traits via TIPI-J inventory (10 items on 7-point scales)

Each was annotated by users themselves (self) and five third-party annotators

Analyses (I): Relationship between self- and third-party annotation results

① Personality traits

Correlation: users themselves vs. averages by third-party annotators

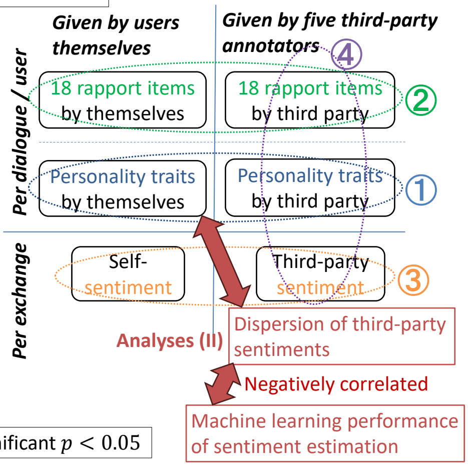
	E	*A*	*C*	*N*	*O*
Hazumi1911	**0.53**	0.08	**0.43**	0.25	0.29
Hazumi2010	**0.58**	**-0.44**	0.17	0.10	0.34
Hazumi2012	**0.39**	0.19	0.11	0.14	0.19
Hazumi2105	**0.57**	**0.37**	0.06	0.21	0.17
Total	**0.49**	0.06	**0.16**	0.15	**0.21**

$p = 8.4 \times 10^{-11}$ $p = 0.041$ $p = 0.0077$

- Extraversion (*E*): statistically significant correlations
  - Consistent with a result in psychology [Bokenau+ 2009]
  - Dialogue itself is an extroverted act
- *A* and *N* were not sufficiently expressed by users

E: Extraversion, *A*: Agreeableness, *C*: Conscientiousness, *N*: Neuroticism, *O*: Openness

Overview


Given by users themselves
Given by five third-party annotators
Per dialogue / user
Per exchange
18 rapport items by themselves
18 rapport items by third party
②
Personality traits by themselves
Personality traits by third party
①
④
Self-sentiment
Third-party sentiment
③
Analyses (II)

Dispersion of third-party sentiments

Negatively correlated

Machine learning performance of sentiment estimation

Bold: the correlation is statistically significant $p < 0.05$

② Rapport 18 items

Correlation between self- and third-party annotations

'*' denotes inverted items

5*	Unsatisfying	**0.38**
9	Engrossing	**0.35**
2*	Boring	**0.32**
17	Worthwhile	**0.29**
8*	Awkward	**0.27**
16*	Dull	**0.25**
10*	Unfocused	**0.23**
6*	Uncomfortably paced	**0.23**
1	Well-coordinated	**0.22**
12*	Intense	**0.21**
11	Involving	**0.21**
14	Active	**0.20**
4	Harmonious	**0.20**
7*	Cold	**0.19**
18*	Slow	**0.17**
13	Friendly	0.13
15	Positive	0.09
3	Cooperative	0.07
Average of 18 items		**0.34**

$p < 0.05$

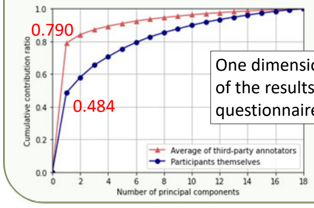
Content of the dialogue

Manner of the dialogue

Feeling and atmosphere of the dialogue

Correlated significantly ($p = 1.23 \times 10^{-5}$)

PCA results for the 18 items



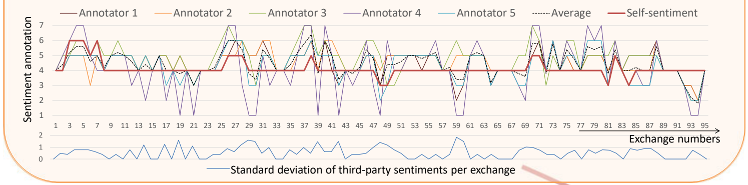
One dimension can explain 79% of the results of third-party questionnaire items

③ Sentiment

Correlation between self-sentiments and averages of third-party sentimentsp

	Macro average	(max, min)
Hazumi1902	0.45	(0.69, 0.11)
Hazumi1911	0.41	(0.79, 0.01)

- Self-sentiments are not always expressed and perceived by the annotators
- Large individual differences among users



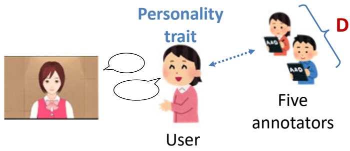
④ Relation among results by third-party annotators

	Personality traits					
	E	*A*	*C*	*N*	*O*	Average sentiments
Average 18 rapport items	**0.53**	**0.68**	**0.21**	**-0.22**	**0.52**	**0.55**
Average sentiments	**0.21**	**0.30**	0.12	0.00	**0.36**	

- Correlation among average sentiments, average 18 rapport items, and *E*, *A*, and *O*
- The third-party annotation results had some correlations, but the self-annotation results seem more complex because not all factors are expressed during dialogues.

Analyses (II): Use of dispersion of third-party sentiments

For what users can the estimation results be reliable?



Dispersion of sentiments

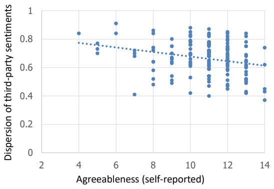
- A user's dispersion: time average of standard deviation of third-party sentiments per exchange

- Annotation results agree → higher machine learning performance
  - Emotion recognition for children speech [D. Seppi+ 2008]
  - Our preliminary results: regression errors and the dispersions correlated

Correlation between the dispersions and personality traits

	E	*A*	*C*	*N*	*O*
Hazumi1911	0.24	**-0.44**	0.16	0.12	0.27
Hazumi2010	**0.38**	**-0.38**	-0.15	0.11	0.04
Hazumi2012	-0.13	-0.20	0.00	0.08	-0.05
Hazumi2105	-0.20	-0.13	0.29	-0.04	0.03
Total	-0.05	**-0.26**	-0.05	0.03	-0.04

$p = 9.1 \times 10^{-4}$



The dispersions were negatively correlated with Agreeableness (self-reported)

- More agreeable users had smaller dispersion of third-party sentiments
- Such users may express their sentiments in a way that others can perceive

→ Sentiment estimation results for agreeable users would be reliable