WOCHAT@IWSDS2018 2018/5/15

Improving Taxonomy of Errors in Chat-oriented Dialogue Systems

Ryuichiro Higashinaka (NTT)
Masahiro Araki (Kyoto Institute of Technology)
Hiroshi Tsukahara (Denso IT Laboratory, Inc.)
Masahiro Mizukami (NTT)

Motivation

- Chat-oriented dialogue systems are attracting much attention
- But, it is difficult to evaluate chat-oriented dialogue systems and to grasp what's wrong with the systems
- We need a good taxonomy of errors for improvement
- This work is about the revision of the two previously proposed taxonomies of errors

Outline

1. Previous taxonomies of errors
   - Top-down taxonomy
   - Bottom-up taxonomy

2. Procedure to revise taxonomies

3. Revised taxonomies

Outline

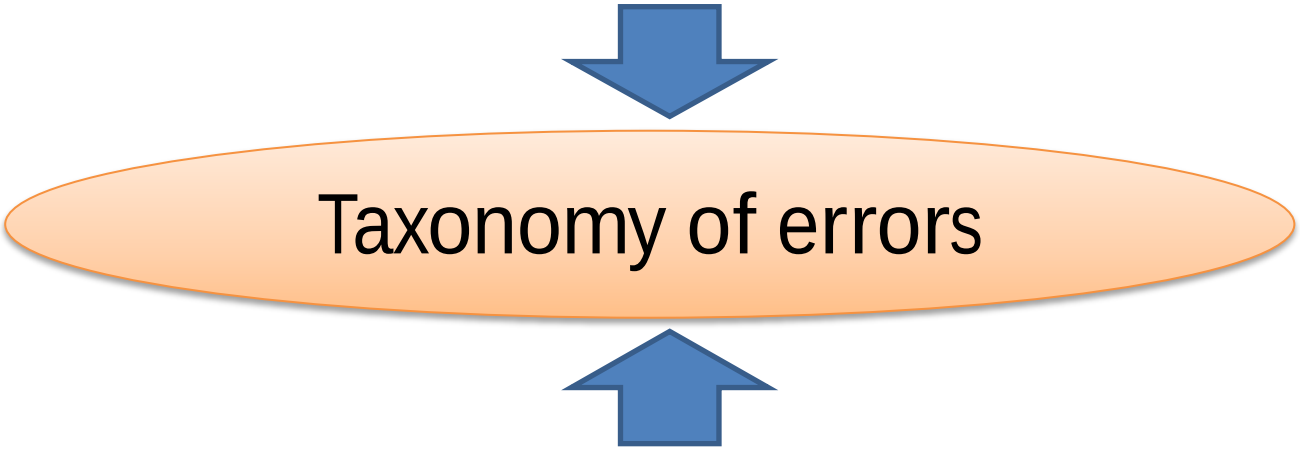
1. Previous taxonomies of errors
   - Top-down taxonomy
   - Bottom-up taxonomy

2. Procedure to revise taxonomies

3. Revised taxonomies

Approaches to creating taxonomies

Top-down approach
Researchers classify the errors on the basis of dialogue theories (e.g., Grice's maxims)



Bottom-up approach
Descriptions of errors by annotators are clustered by an automatic clustering method

Top-Down (TD) taxonomy

Specific error types under each main category

Scope of the context to which the error relates

Main category	Subcategory
Utterance	Syntactic error
	Semantic error
	Uninterpretable
Response	Excess/Lack of Information
	Non-understanding
	No relevance
	Unclear intention
	Misunderstanding
Context	Excess/Lack of proposition
	Contradiction
	Non-relevant topic
	Unclear relation
	Topic switch error
Environment	Lack of common ground
	Lack of common sense
	Lack of sociality

Error categories based on how Grice's maxims are violated

(Higashinaka+, SIGDIAL2015)

Button-up (BU) taxonomy

User	Do you know what movie will be aired on Friday night?	
System	Yes, yes.	The system did not mention a movie name. / It answered only "yes" but without saying a movie name. / This is not an answer to the question. / The conversation is not valid.
User	I haven't heard of that movie.	
System	That's good.	This is not a response to a question. / Strange answer. / It is not clear what is good. / The conversation is not valid.

Comments

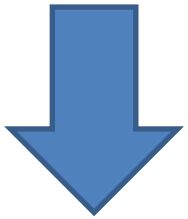
Comments describing the errors were clustered by automatic clustering

1	Not understandable
2	Ignore user question
3	General quality
4	Unclear intention
5	Ignore user utterance
6	Word usage error
7	Analysis failure
8	Inappropriate answer
9	Grammatical error
10	Violation of common sense
11	Contradiction
12	Repetition
13	Social error
14	Diversion
15	Topic-change error
16	Mismatch in conversation
17	Expression error

(Higashinaka+, EMNLP2015)

Evaluation of previous taxonomies

One way to evaluate the validity of a taxonomy is by using inter-annotator agreement (Cohen's kappa)



	Agreement	**Cohen's κ**
TD taxonomy	0.35	0.26
BU taxonomy	0.35	0.27

Kappa is low!

Outline

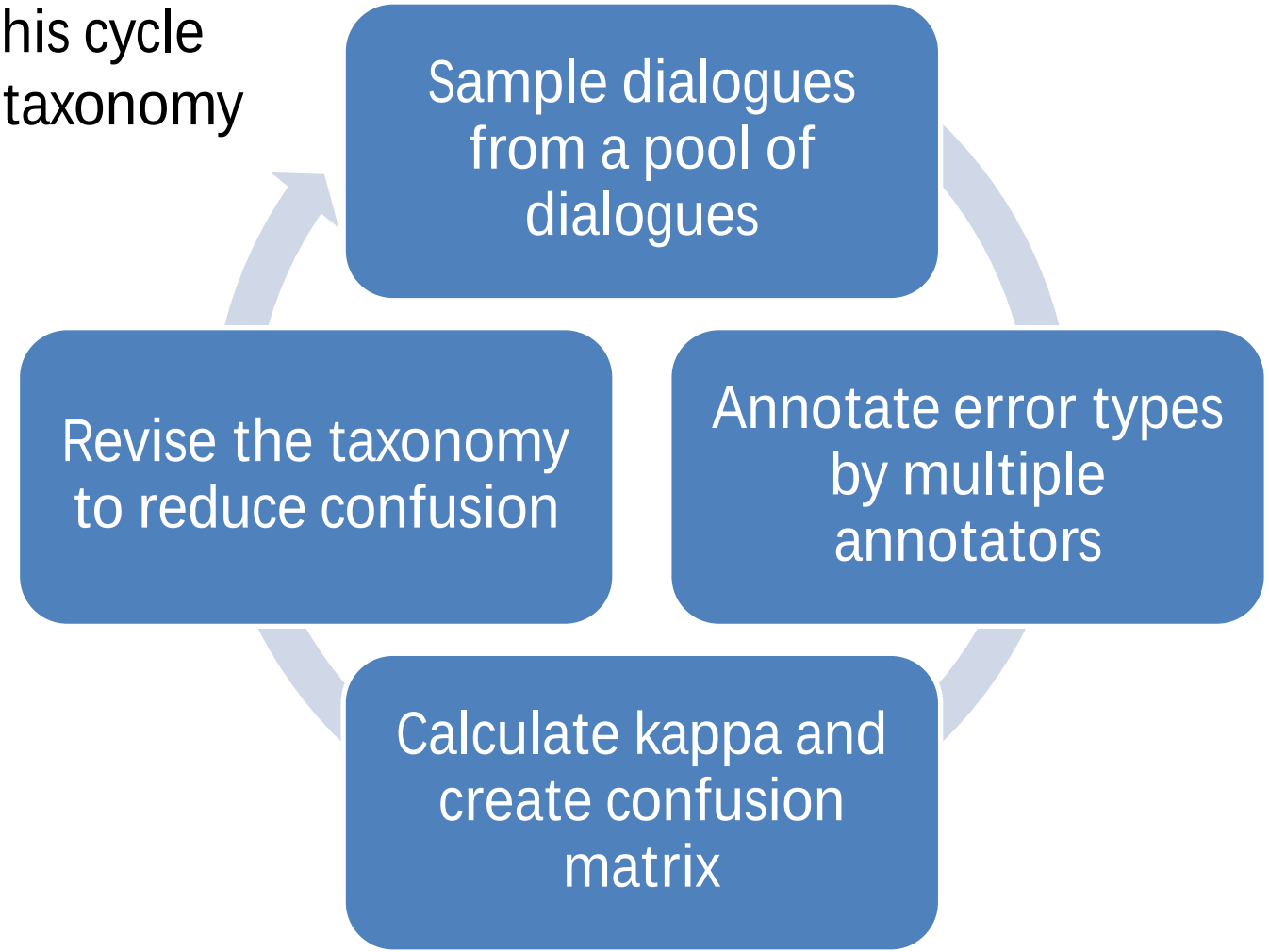
1. Previous taxonomies of errors
   - Top-down taxonomy
   - Bottom-up taxonomy

2. Procedure to revise taxonomies

3. Revised taxonomies

Procedure to revise taxonomies

We run this cycle for each taxonomy



Pool of dialogues

- Dialogue breakdown detection challenge (DBDC) dataset
  - 400 dialogues between a user and a chat-oriented dialogue system
  - Contains the data of 3 different chat-oriented dialogue systems
  - We split the data to allow multiple iterations of the improvement cycle

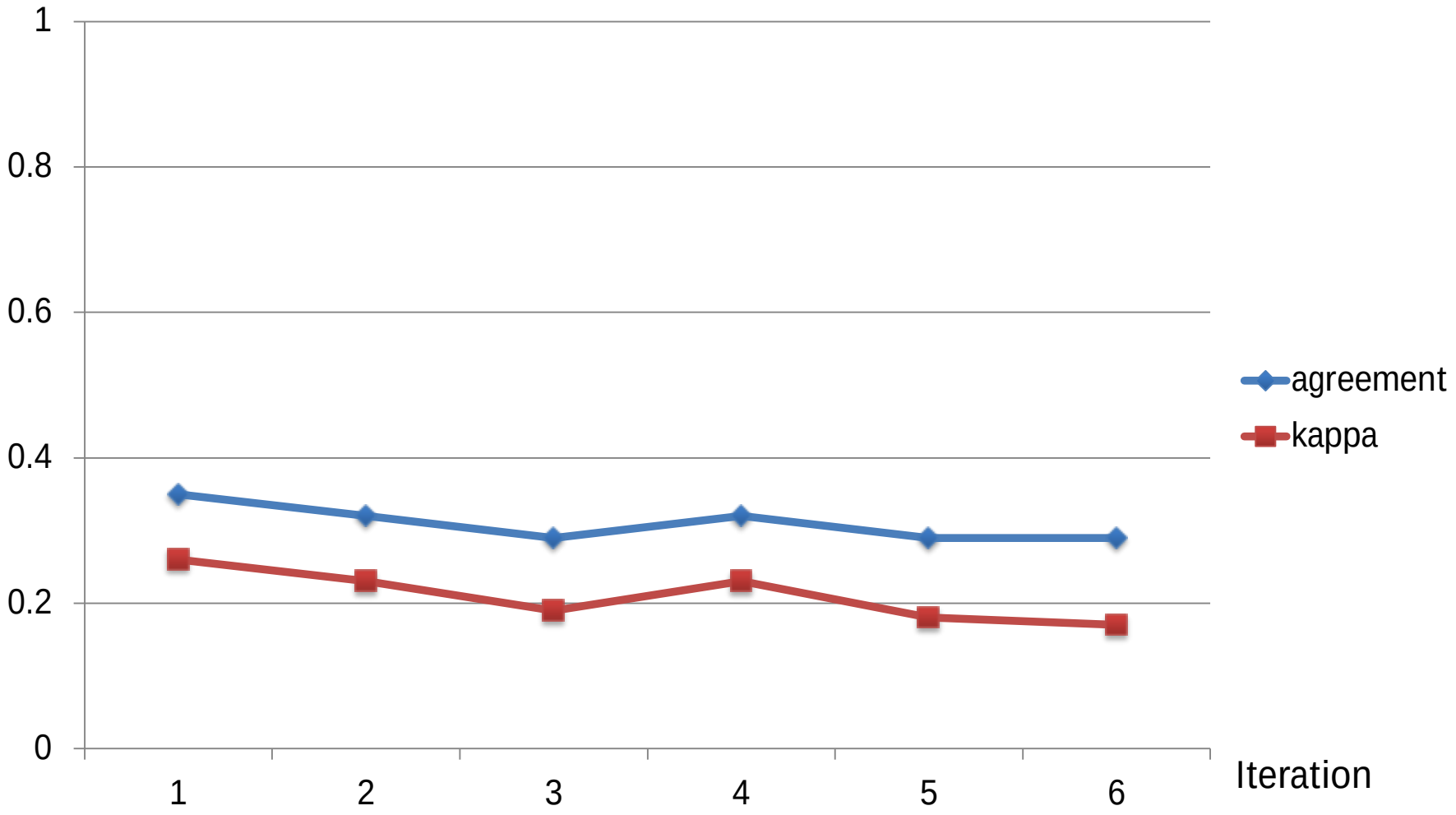
Dataset available at: https://dbd-challenge.github.io/dbdc3/data/

(Higashinaka+, LREC2016, DSTC6)

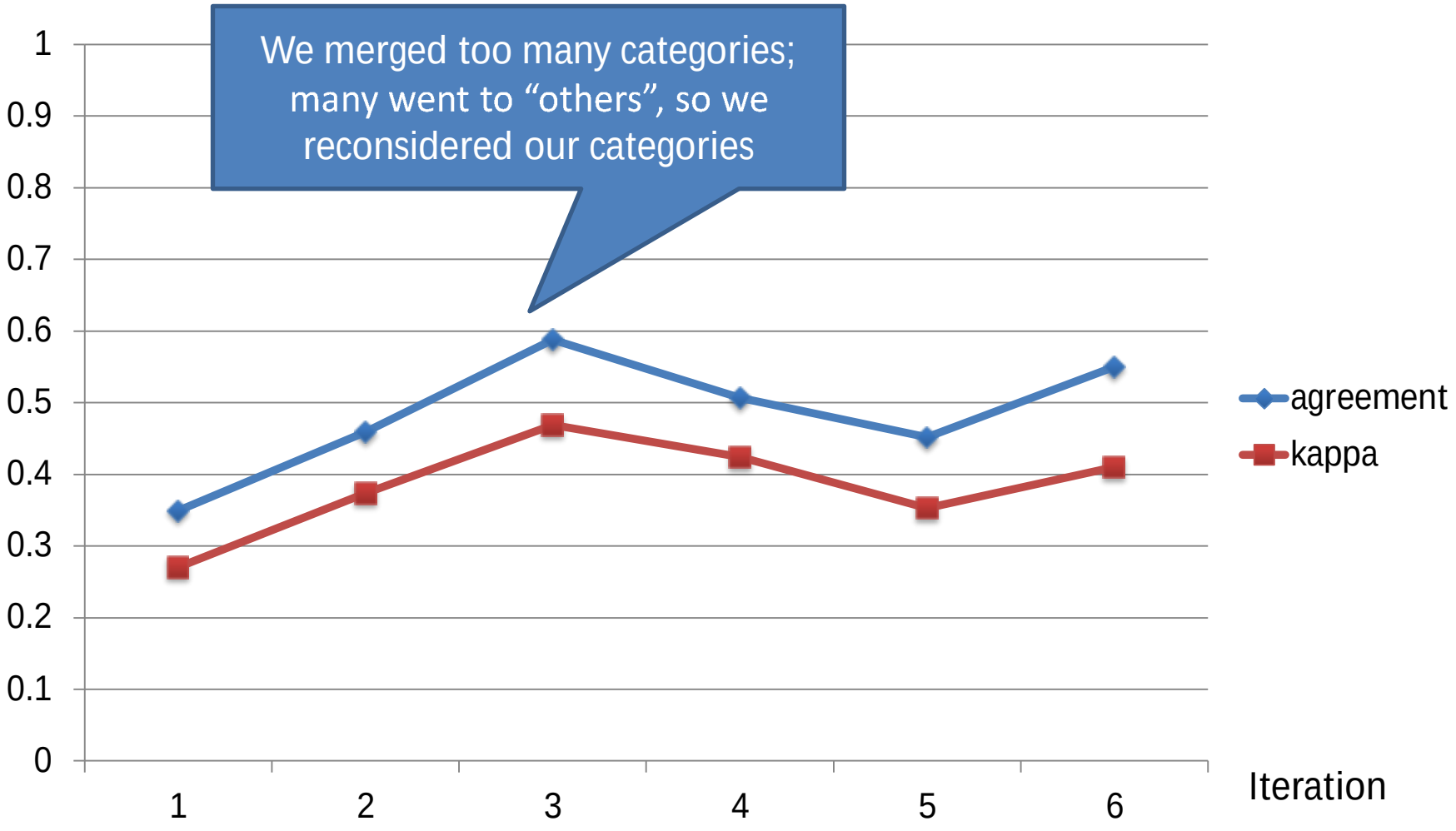
Revisions and iterations

- Revisions
  - Merge confusing categories
  - Introduction of a decision flow
    - To give orders to possible overlapping categories
    - Orders decided by the ease of decision
- Iterations
  - Five iterations by two expert annotators
  - Final iteration by ten crowd workers
    - To ensure that the errors can be reliably annotated by anyone

Transitions of kappa (TD Taxonomy)



Transitions of kappa (BU Taxonomy)



Improvement in kappa

Previous taxonomies

	Agreement	Cohen's κ
TD taxonomy	0.35	0.26
BU taxonomy	0.35	0.27

Revised taxonomies

	Agreement	Cohen's κ
TD taxonomy (revised)	0.32	0.24
BU taxonomy (revised)	0.54	0.44

Reasonable kappa for the BU taxonomy

Outline

1. Previous taxonomies of errors
   - Top-down taxonomy
   - Bottom-up taxonomy

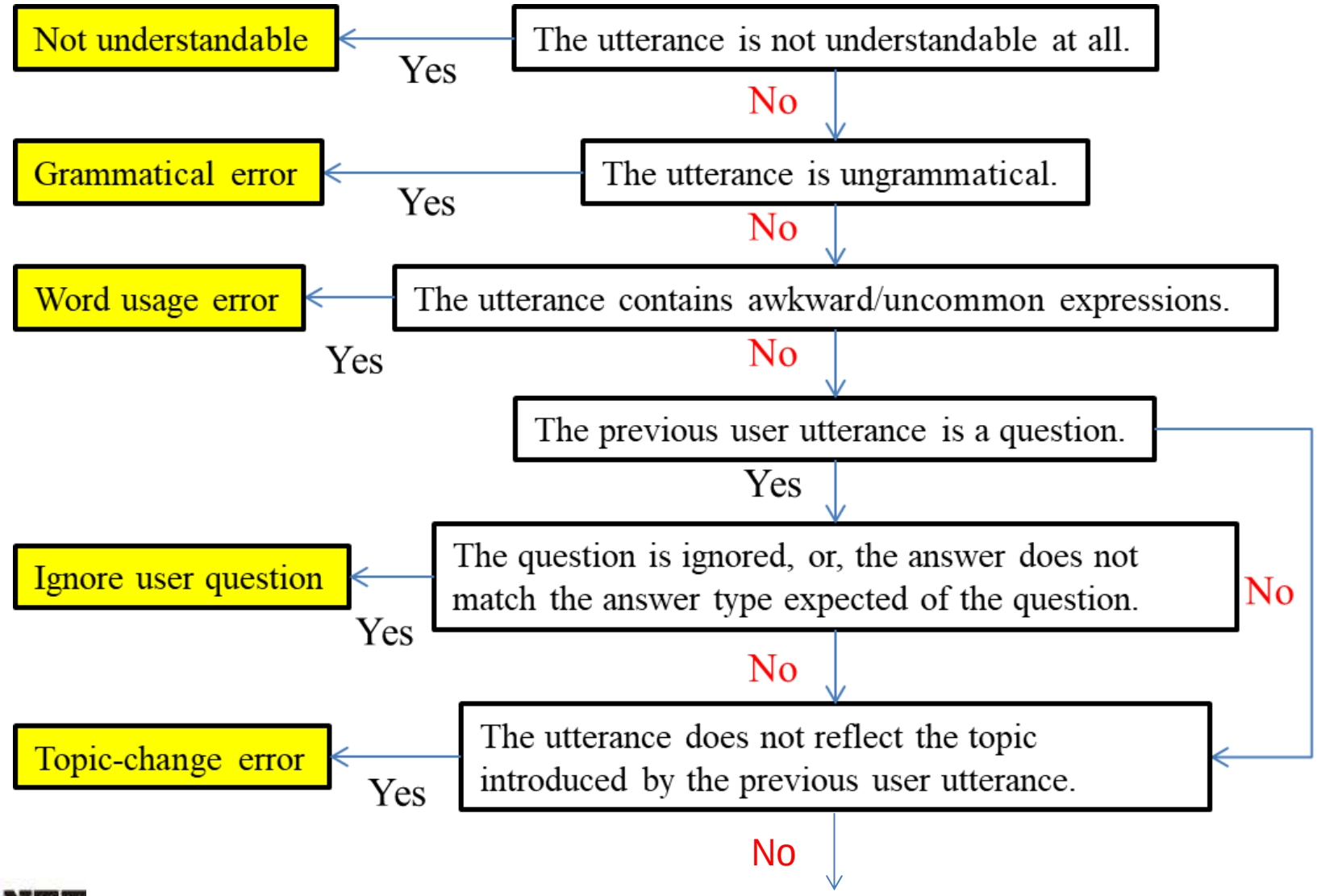
2. Procedure to revise taxonomies

3. Revised taxonomies

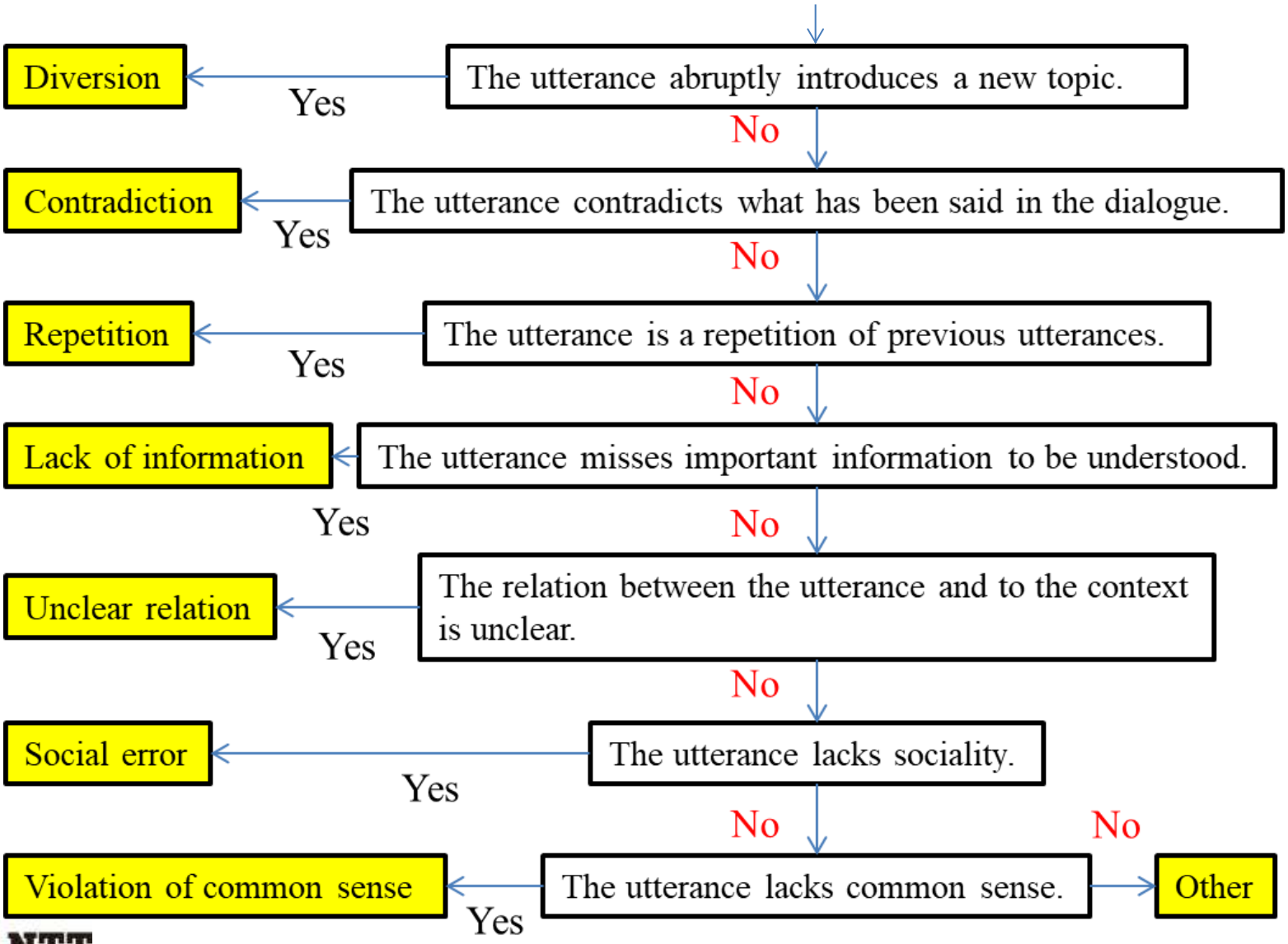
Revised BU taxonomy

	Category	Explanation
1	Not understandable	Utterance is un-interpretable.
2	Grammatical error	Utterance is ungrammatical.
3	Word usage error	Words that do not fit context are used in utterance.
4	Ignore user question	System ignores or fails to answer question.
5	Topic-change error	Utterance does not reflect topic introduced by user.
6	Diversion	Utterance abruptly introduces different topic.
7	Contradiction	Content of utterance contradicts what has already been said.
8	Repetition	Utterance is just repeated without new information.
9	Lack of information	Utterance misses important information.
10	Unclear relation	Relation between utterance and context is unclear.
11	Social error	Utterance lacks politeness.
12	Violation of common sense	Utterance lacks common sense.
13	Others	Miscellaneous error

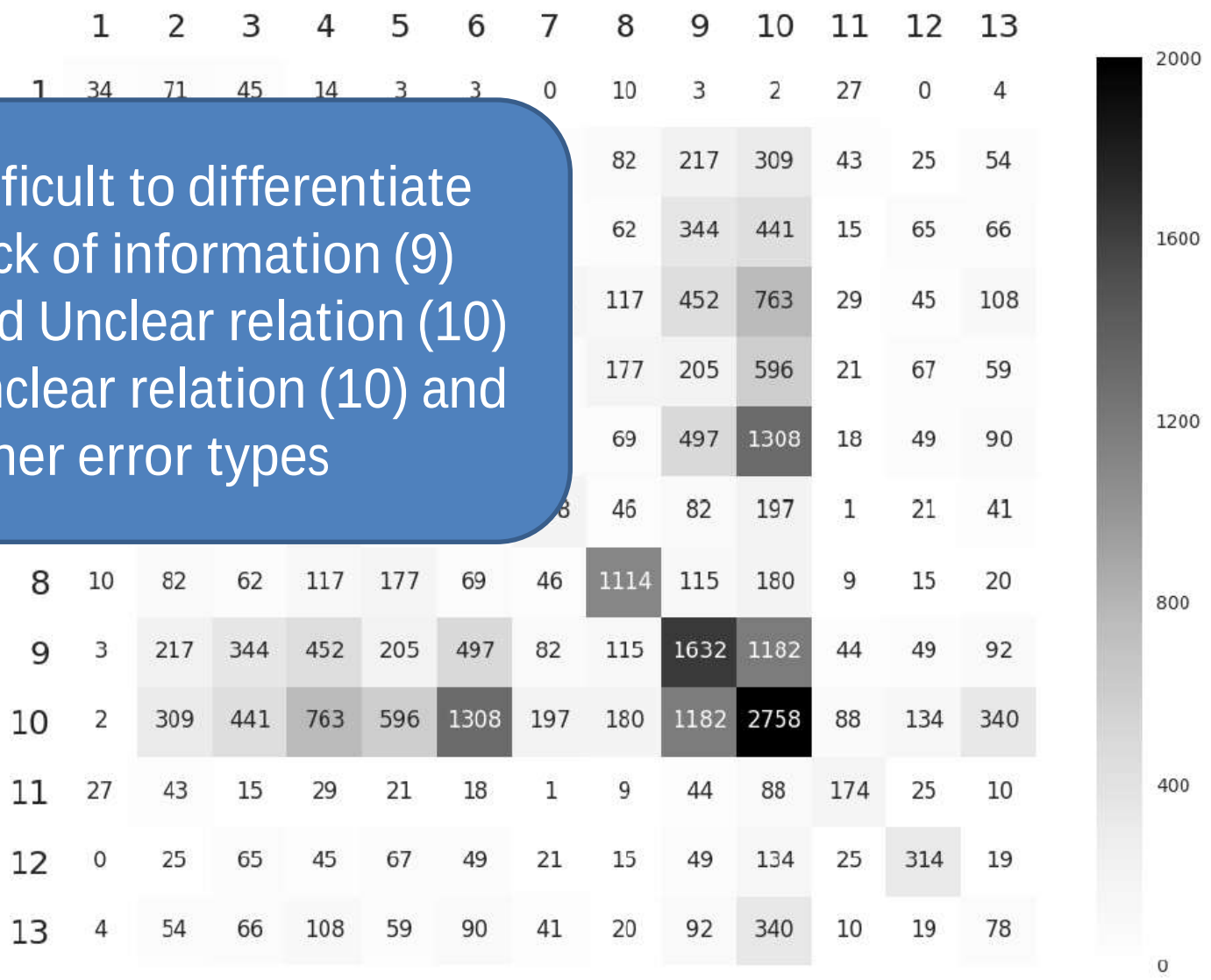
Decision flow for the BU taxonomy



Decision flow for the BU taxonomy (cont'd)



Confusion matrix (BU taxonomy)



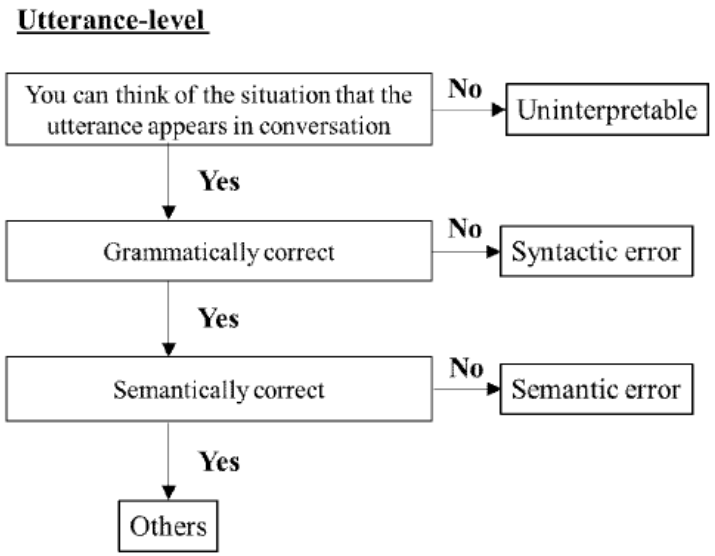
Difficult to differentiate

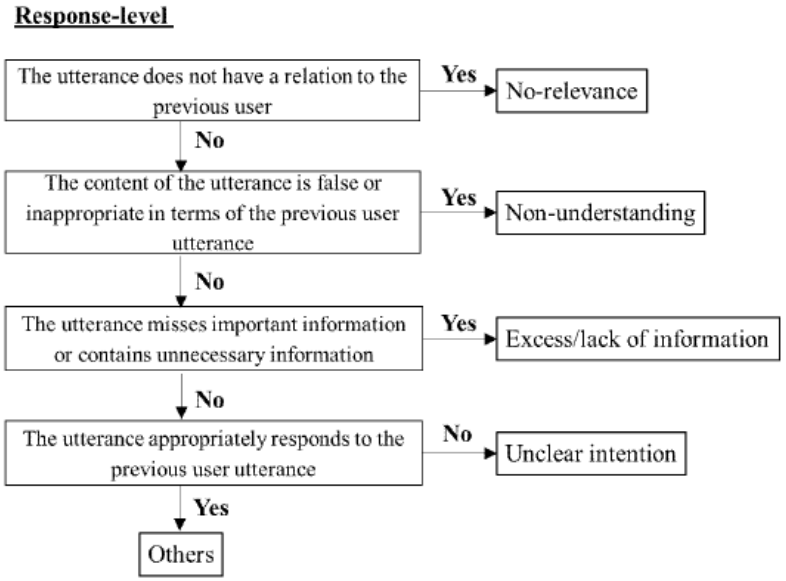
- Lack of information (9) and Unclear relation (10)
- Unclear relation (10) and other error types

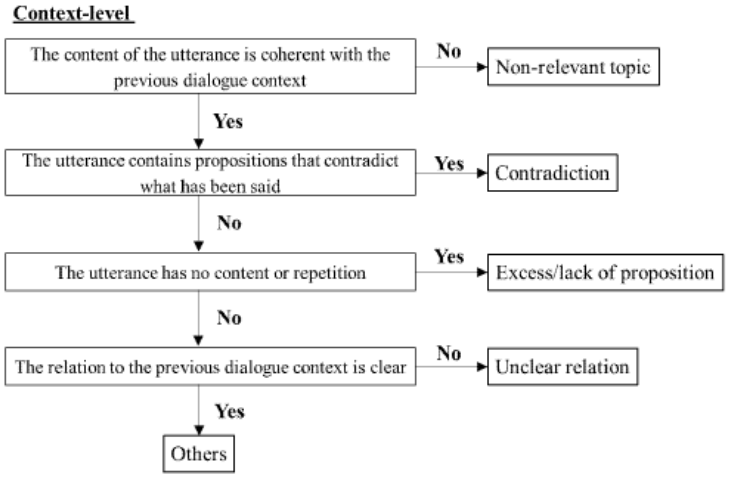
Revised TD taxonomy

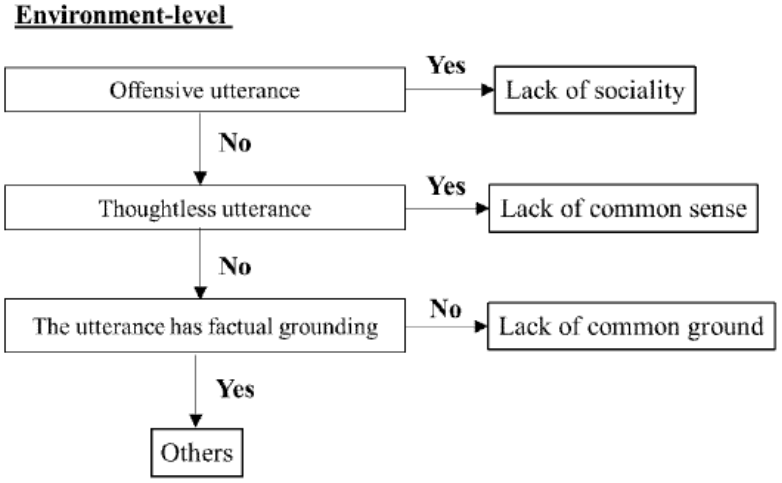
Main category	Subcategory	Explanation
Utterance	Syntactic error	Grammatically invalid utterance
	Semantic error	Semantically invalid utterance
	Un-interpretable	Not understandable
	Others	Other utterance-level error
Response	Excess/lack of information	Utterance misses important information or contains unnecessary information
	Non-understanding	Content of utterance is false or inappropriate regarding previous user utterance
	No-relevance	Utterance does not have any relation to previous user utterance
	Unclear intention	Relation to previous user utterance is not clear
	Others	Other response-level error
Context	Excess/lack of proposition	Utterance is just empty words or repetition
	Contradiction	Utterance contains propositions that contradict what has been said
	Non-relevant topic	Topic of utterance is irrelevant to current context
	Unclear relation	Relation to previous dialogue context is not clear
	Others	Other context-level error
Environment	Lack of common ground	Utterance has no factual grounding
	Lack of common sense	Thoughtless utterance
	Lack of sociality	Offensive utterance
	Others	Other environment-level error

Decision flow for the TD taxonomy

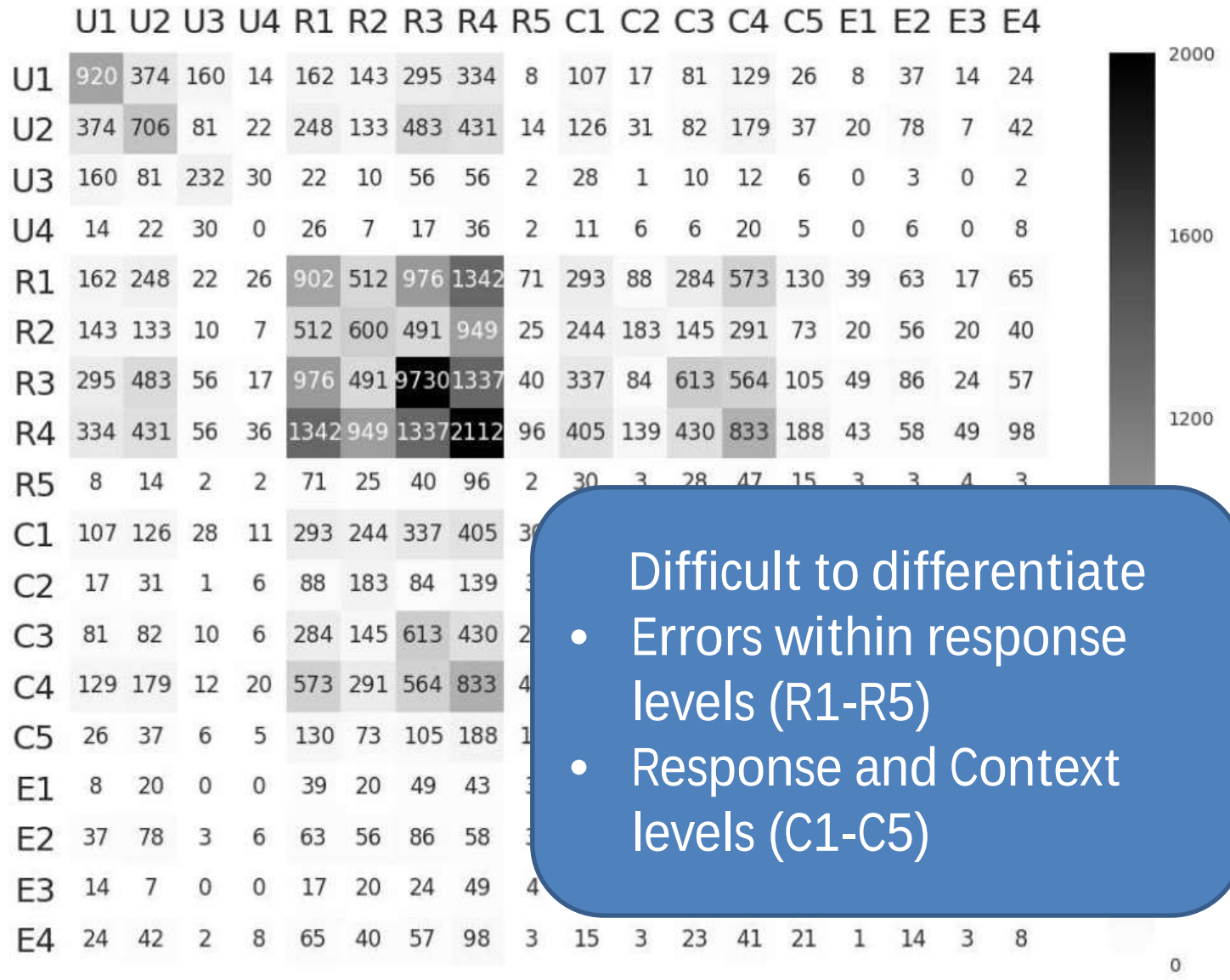








Confusion matrix (TD taxonomy)



Difficult to differentiate

- Errors within response levels (R1-R5)
- Response and Context levels (C1-C5)

Summary and future work

- We revised two previously proposed taxonomies of errors in chat-oriented dialogue systems
- Revised BU taxonomy achieved reasonable inter-annotator agreement of 0.44 in Cohen's kappa
- Future work
  - Investigate the reason behind the poor inter-annotator agreement of the TD taxonomy
  - Merge TD/BU taxonomies into one gold taxonomy