

Seminar Natural Language Processing (NLP) — Part 7 (example talk)

A Dialogue Corpus for Learning to Construct Explanations

Henning Wachsmuth, Milad Alshomary

<https://ai.uni-hannover.de>



Leibniz
Universität
Hannover

Motivation



Explainer

Milad, what is an explanation?

It's giving details to make clear why some fact holds, how to perform some action, or similar.

Oh, I actually meant in the context of explainable AI.

I see. Well, an explanation in XAI may, for example, reason why some classification decision was made.

Classification... uhhh... what's that again?



Explainee

▪ Explaining

- Pervasive communicative process, aimed at explainee's understanding
- Depends on the explainee's prior knowledge
- Explainer needs to react and adjust to explainee's responses
- So far, nearly all XAI research in NLP sees explanations as monological

Presented paper (Wachsmuth et al., 2022)

▪ A dialogue corpus for learning to explain

- Developed together with experts from humanities
- Controlled human explaining dialogues of varying complexity
- Multiple manual annotations at dialogue turn level



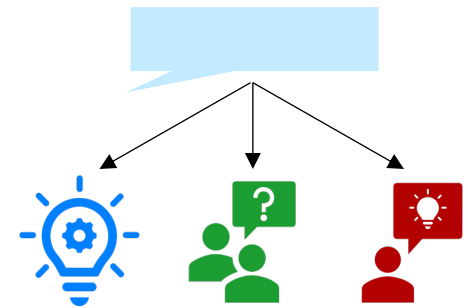
▪ Insights into dialogical explanations

- Differences across explainees' proficiency levels
- Interaction of topics, dialogue acts, and explanation moves
- Language of explainers and explainees



▪ Experiments on turn classification

- BERT-based baselines for the annotated dimensions
- Impact of sequence and task dependencies
- Results lay the ground for more human-centered XAI



Source data

- **Video series “5 Levels”** <https://www.wired.com/video/series/5-levels>
 - Expert explains science-related topic to five different explainees
 - Controlled dialogical explaining setting



Child



Teenager



Undergrad



Grad



Colleague

- **Acquired source data**
 - **13 topics.** Including blockchain, machine learning, sleep, ...
 - **65 dialogues.** With subtitles available
 - **1550 turns.** 23.8 turns per dialogue, manually segmented
 - **51344 words.** 33.1 words per turn

Dialogues for four more topics unannotated in corpus

Annotations

Annotation process

- Annotation scheme developed with experts from humanities
- Five Upwork freelancers manually labeled each turn for three dimensions

Topic



Main topic

Subtopic

Related topic

Other/No topic

Dialogue act



Check question

What/How question

Other question

Confirming answer

Disconfirming answer

Other answer

Agreeing statement

Disagreeing statement

Informing statement

Other

Explanation move



Test understanding

Test prior knowledge

Provide explanation

Request explanation

Signal understanding

Signal non-understanding

Provide feedback

Provide assessment

Provide extra information

Other

Explaining Dialogue Corpus <https://github.com/webis-de/COLING-22>



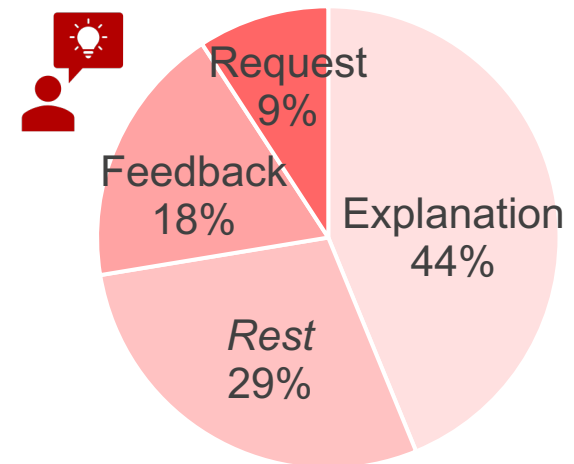
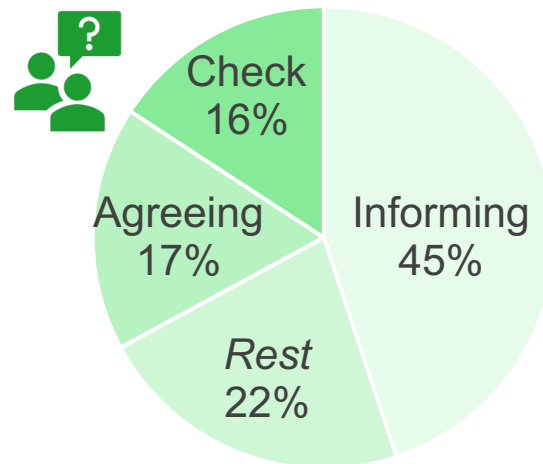
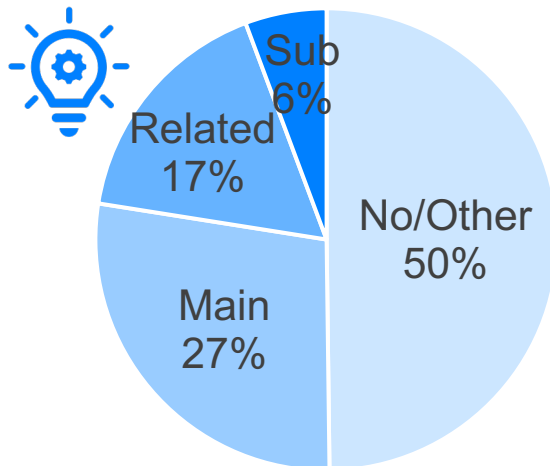
Annotation reliability

- Final dimension labels obtained with MACE (Hovy et al., 2013)

Measure	Topic	Dialogue Act	Explanation Move
Fleiss' κ	0.35	0.49	0.43
MACE competence	0.30–0.76	0.58–0.82	0.45–0.85

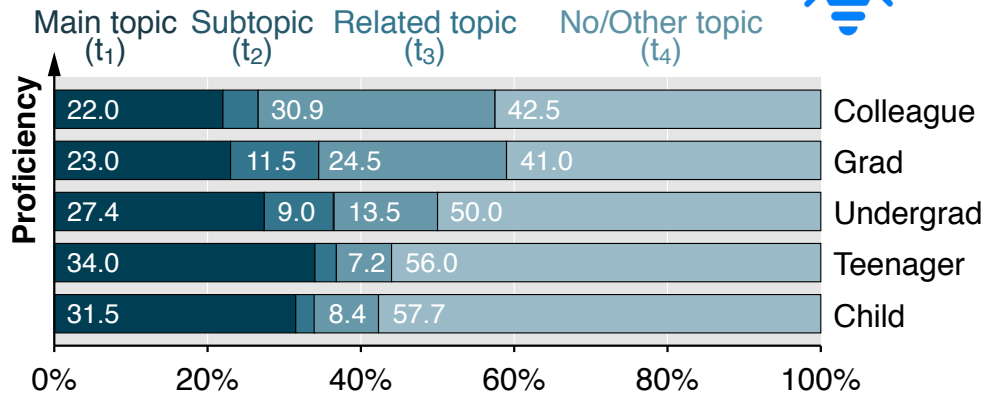
Corpus statistics

- Full distribution in the paper, including explainer/explainee differences



Analysis (selection)

Topics across proficiency levels



Participant-specific language

I want to know if you agree, sleep is the coolest thing you've ever heard of.



Explainer
to teenager

Interaction of acts and moves



Approach	Explainer	Explainee	Total
Informing Explanation	45.9%	31.3%	38.8%
Agreeing Feedback	3.9%	14.2%	9.0%
Agreeing Understanding	3.5%	9.1%	6.3%
Check Prior knowledge	10.5%	–	5.4%
Check Request	2.7%	6.8%	4.7%

So all kind of older logic and stuff like that. So, I mean, it's sort of based on, like, you're presented the little MUX chip.



Explainee
(colleague)

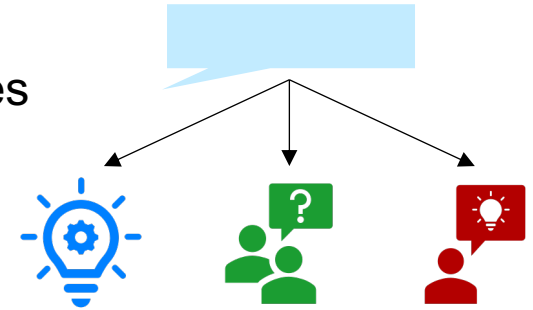
Experiments (selection)

▪ Experimental setup

- Classify each turn dimension with different approaches
- Evaluate in 13-fold cross-topic validation

▪ Classification approaches

- **BERT-basic**. Main topic and turn text as input
- **BERT-sequence**. Additionally all previous turn texts and labels as input
- **BERT-multitask**. Main topic and turn text as input, all three labels as output



▪ Macro F_1 results

Approach	Topic	Dialogue Act	Explanation Move
BERT-basic	0.51	0.44	0.41
BERT-sequence	0.52	0.47	0.43
BERT-multitask	0.41	0.38	0.34
Majority baseline	0.17	0.06	0.06

Takeaways

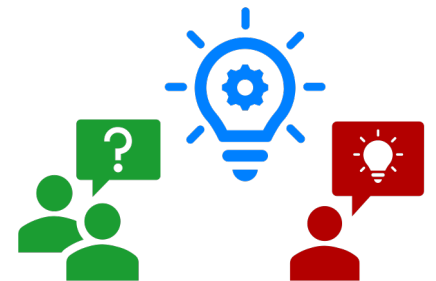
▪ Explaining

- Pervasive communicative process, aimed at understanding
- Strongly dependent on the explainer and explainee involved
- Dialogical explanations barely studied so far in NLP



▪ Presented paper

- A dialogue corpus for learning how humans explain
- Insights into explanations for different proficiency levels
- Baselines for classifying topics, acts, and moves of turns



▪ Discussion

- Groundwork for learning to construct explanations
- Needed for XAI that interacts with different users
- Future work should increase scale and heterogeneity



<https://trr318.upb.de/en>

References

- **Hovy et al. (2013).** Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In Proceedings of NAACL-HLT 2013, pages 1120–1130.
- **Wachsmuth and Alshomary (2022).** Henning Wachsmuth and Milad Alshomary. “Mama always had a way of explaining things so I could understand”: A dialogue corpus for learning to construct explanations. In Proceedings of the 29th International Conference on Computational Linguistics, pages 344–354, 2022.