# Making a Long Story Short in Conversation Modeling

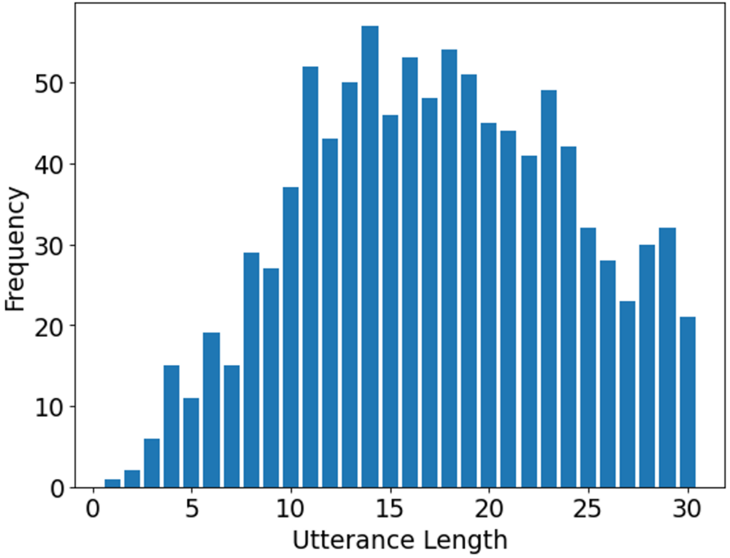
Yufei Tao[1] Tiernan Mines[2] Ameeta Agrawal[1]

**PortNLP Lab**

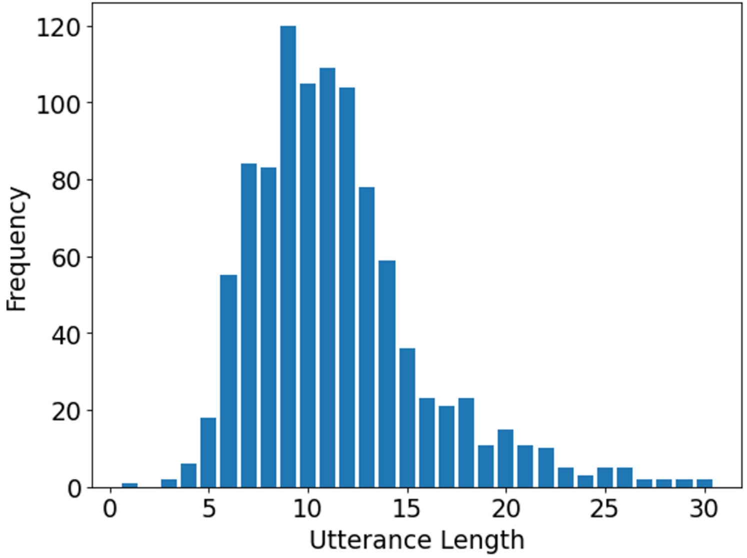
[1]Department of Computer Science, Portland State University
[2]Hello Lamp Post

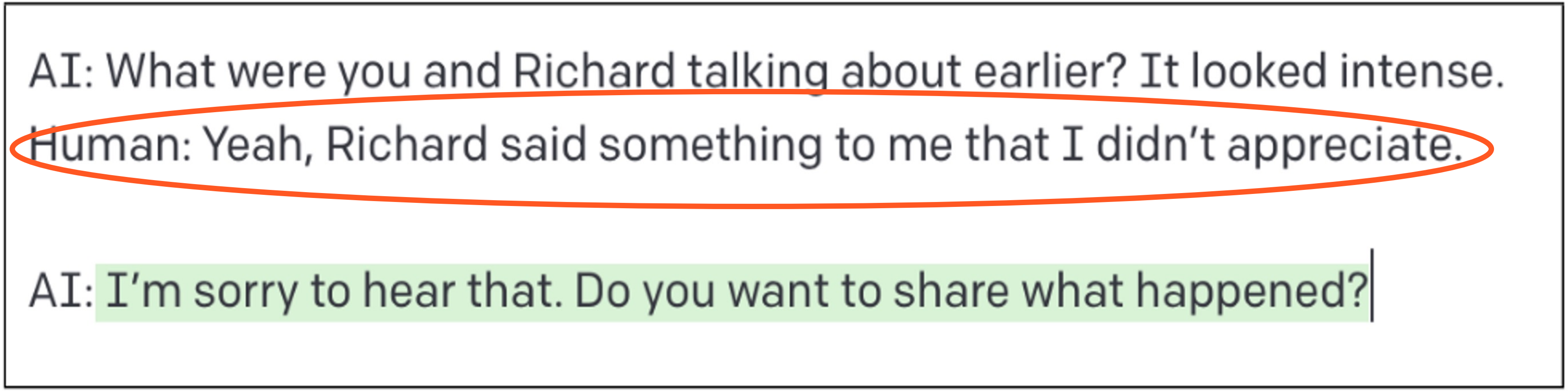
# *Utterance Lengths in Conversations*

Topical-Chat



PROSOCIALDIALOG



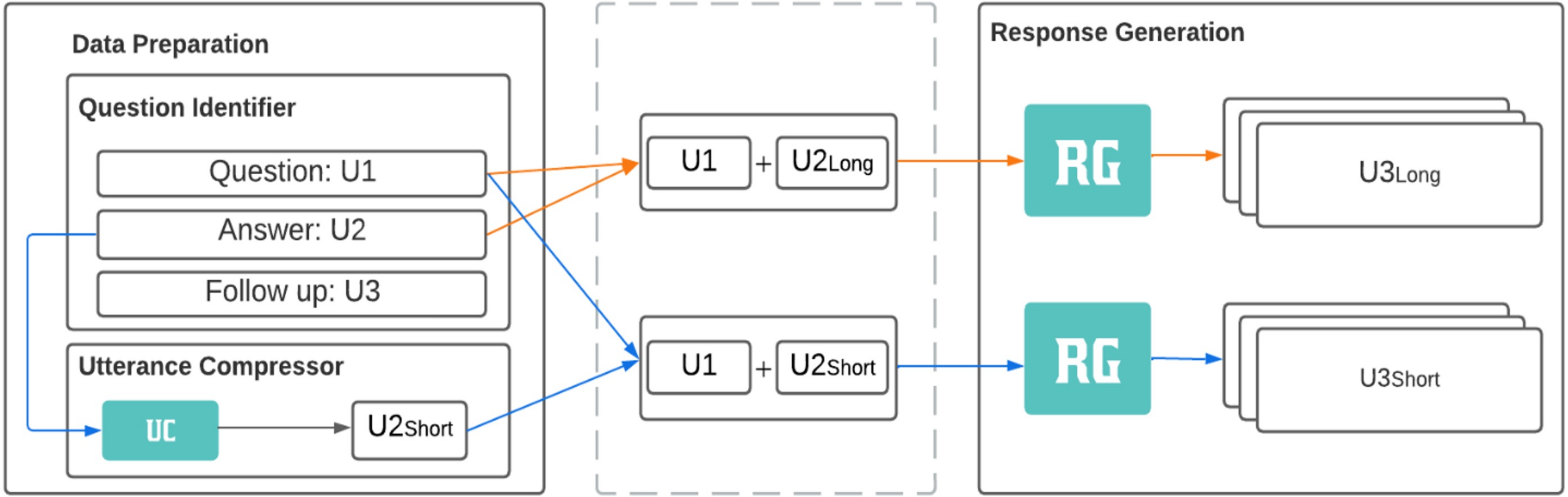


AI: What were you and Richard talking about earlier? It looked intense.

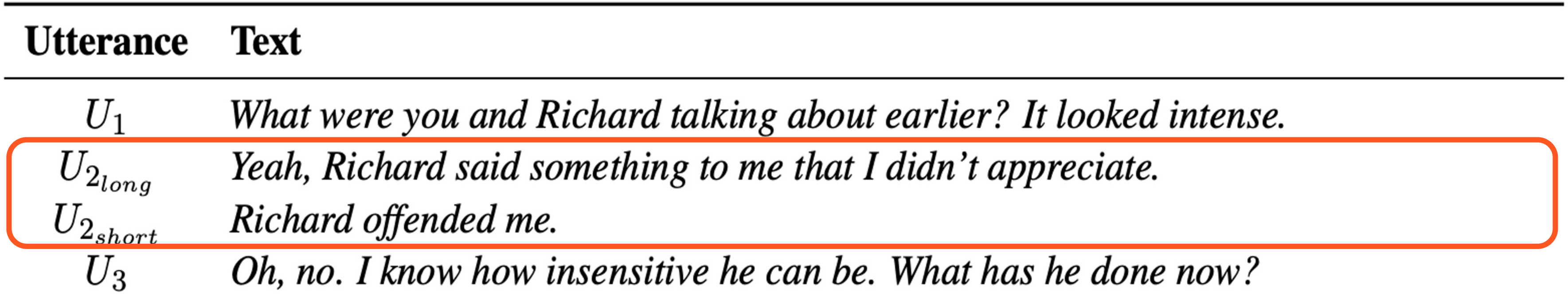
Human: Yeah, Richard said something to me that I didn't appreciate.

AI: I'm sorry to hear that. Do you want to share what happened?

# Model Overview

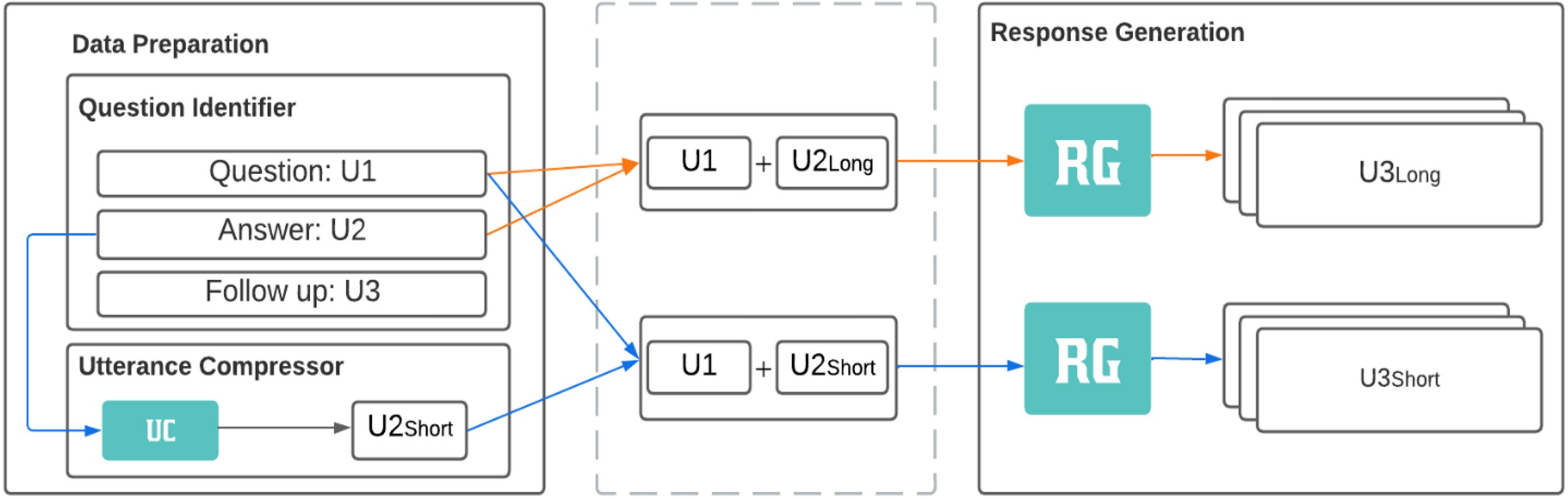


# Data Preparation

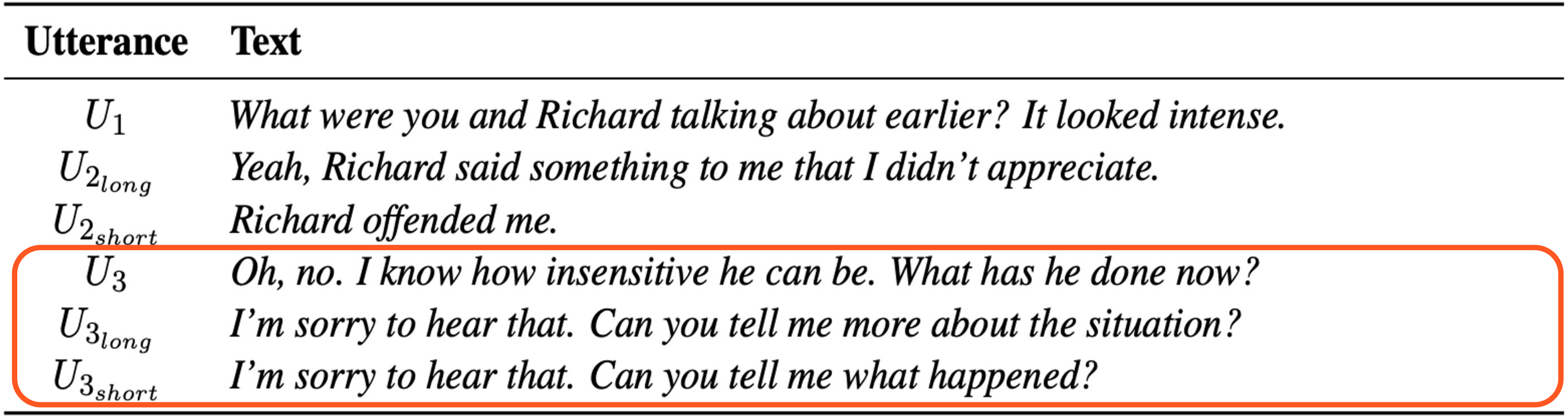


| Utterance | Text |
|---|---|
| $U_1$ | *What were you and Richard talking about earlier? It looked intense.* |
| $U_{2_{long}}$ | *Yeah, Richard said something to me that I didn't appreciate.* |
| $U_{2_{short}}$ | *Richard offended me.* |
| $U_3$ | *Oh, no. I know how insensitive he can be. What has he done now?* |

# Model Overview



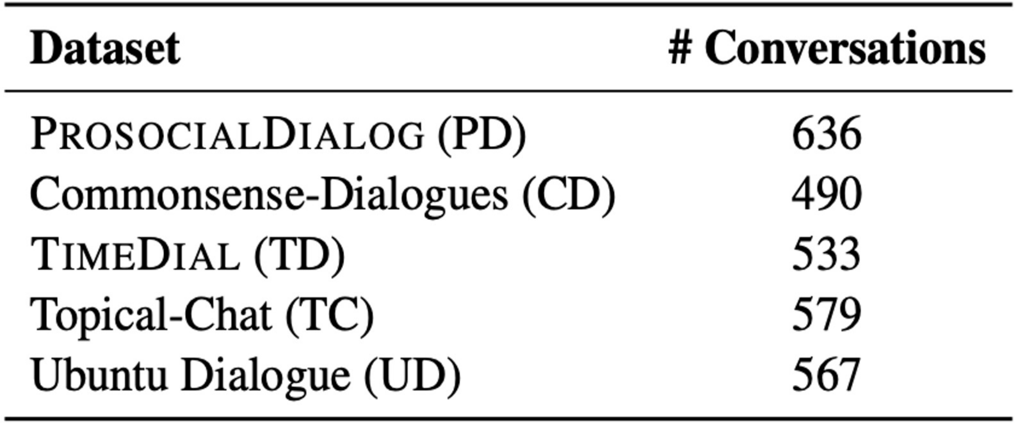
# Response Generation



| Utterance | Text |
|---|---|
| $U_1$ | *What were you and Richard talking about earlier? It looked intense.* |
| $U_{2_{long}}$ | *Yeah, Richard said something to me that I didn't appreciate.* |
| $U_{2_{short}}$ | *Richard offended me.* |
| $U_3$ | *Oh, no. I know how insensitive he can be. What has he done now?* |
| $U_{3_{long}}$ | *I'm sorry to hear that. Can you tell me more about the situation?* |
| $U_{3_{short}}$ | *I'm sorry to hear that. Can you tell me what happened?* |

# Datasets

- Five multi-turn conversation datasets
  - PROSOCIALDIALOG
  - Commonsense-Dialogues
  - TIMEDIAL
  - Topical-Chat
  - Ubuntu Dialogue



| Dataset | # Conversations |
|---|---|
| PROSOCIALDIALOG (PD) | 636 |
| Commonsense-Dialogues (CD) | 490 |
| TIMEDIAL (TD) | 533 |
| Topical-Chat (TC) | 579 |
| Ubuntu Dialogue (UD) | 567 |

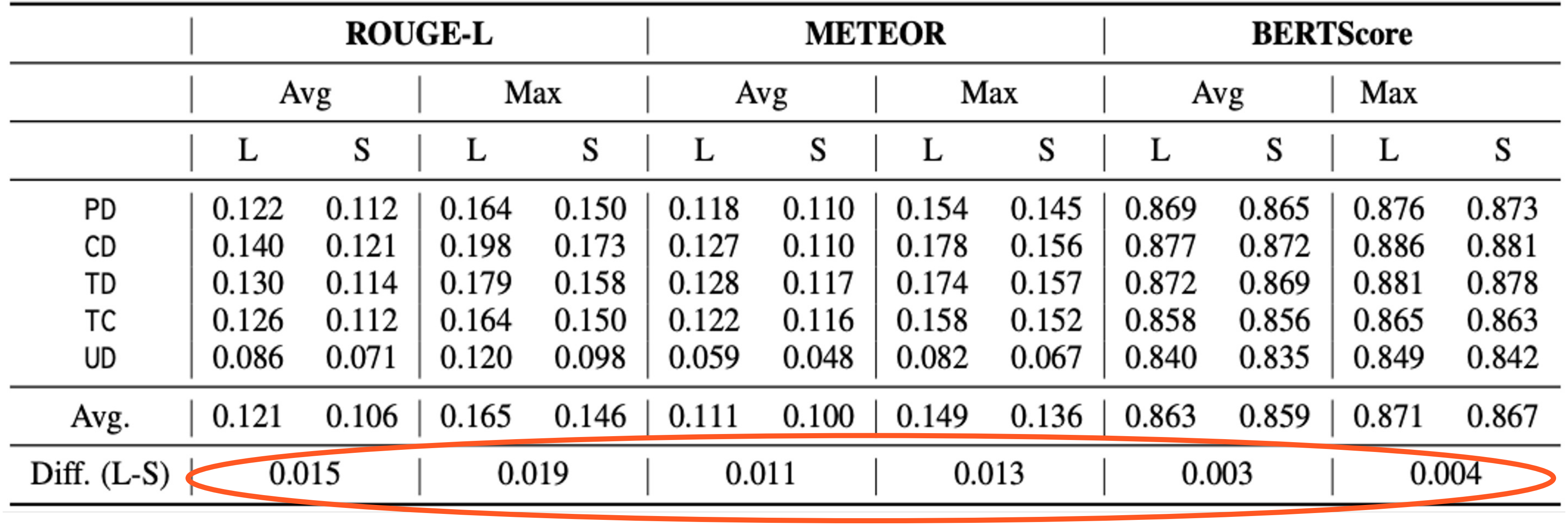
# Evaluation Metrics

- Automatic Evaluation
    - ROGUE-L
    - METEOR
    - BERTScore

# Evaluation Metrics

- Human Evaluation
  - We randomly selected 8 samples from each dataset for a total of 40 evaluation samples.
  - Four annotators were asked whether U3long or U3short is more similar to U3, if both of them were equally similar (**both**), or if neither of them was similar to U3 (**neither**).
  - Compute inter-annotator agreement using Fleiss' Kappa

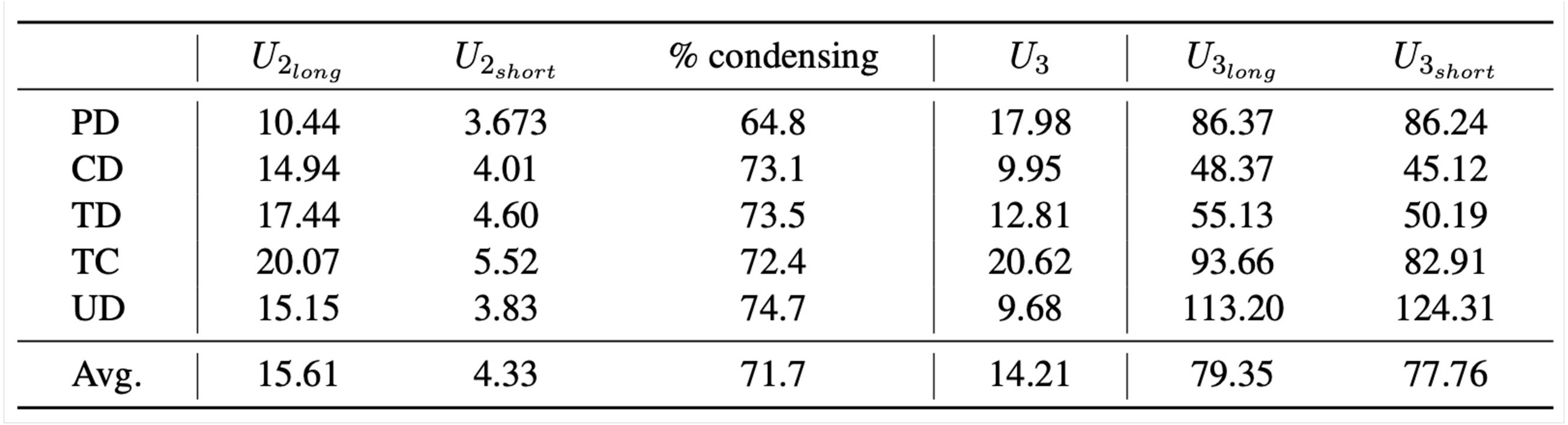
# Results (automatic)



| | ROUGE-L | | | | METEOR | | | | BERTScore | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Avg | | Max | | Avg | | Max | | Avg | | Max | |
| | L | S | L | S | L | S | L | S | L | S | L | S |
| PD | 0.122 | 0.112 | 0.164 | 0.150 | 0.118 | 0.110 | 0.154 | 0.145 | 0.869 | 0.865 | 0.876 | 0.873 |
| CD | 0.140 | 0.121 | 0.198 | 0.173 | 0.127 | 0.110 | 0.178 | 0.156 | 0.877 | 0.872 | 0.886 | 0.881 |
| TD | 0.130 | 0.114 | 0.179 | 0.158 | 0.128 | 0.117 | 0.174 | 0.157 | 0.872 | 0.869 | 0.881 | 0.878 |
| TC | 0.126 | 0.112 | 0.164 | 0.150 | 0.122 | 0.116 | 0.158 | 0.152 | 0.858 | 0.856 | 0.865 | 0.863 |
| UD | 0.086 | 0.071 | 0.120 | 0.098 | 0.059 | 0.048 | 0.082 | 0.067 | 0.840 | 0.835 | 0.849 | 0.842 |
| Avg. | 0.121 | 0.106 | 0.165 | 0.146 | 0.111 | 0.100 | 0.149 | 0.136 | 0.863 | 0.859 | 0.871 | 0.867 |
| Diff. (L-S) | 0.015 | | 0.019 | | 0.011 | | 0.013 | | 0.003 | | 0.004 | |

# Results (automatic)

Length differences of U2 and U3 across five datasets



| | $U2_{long}$ | $U2_{short}$ | % condensing | $U_3$ | $U3_{long}$ | $U3_{short}$ |
|---|---|---|---|---|---|---|
| PD | 10.44 | 3.673 | 64.8 | 17.98 | 86.37 | 86.24 |
| CD | 14.94 | 4.01 | 73.1 | 9.95 | 48.37 | 45.12 |
| TD | 17.44 | 4.60 | 73.5 | 12.81 | 55.13 | 50.19 |
| TC | 20.07 | 5.52 | 72.4 | 20.62 | 93.66 | 82.91 |
| UD | 15.15 | 3.83 | 74.7 | 9.68 | 113.20 | 124.31 |
| Avg. | 15.61 | 4.33 | 71.7 | 14.21 | 79.35 | 77.76 |

# Results (human assessment)

- 54% of the annotations were marked as ‘both’ or ‘neither’ of $U_{3long}$ and $U_{3short}$ are/is similar to the original $U_3$, suggesting that the qualitative analysis of the generated responses were similar.
- Fleiss’ Kappa score was 0.5848: moderate level of agreement.

# Conclusion & Future Work

- With shorter utterances, GPT-3 can still produce coherent and contextually appropriate responses, indicating potential for model efficiency without quality loss.
- Further exploration of token reduction and linguistic nuances in compressed inputs.
- Expanding to contexts beyond question responses and using advanced models like GPT-4 and others.
- Alternative evaluation methods like G-Eval and MEEP to capture a broader range of conversational dynamics (engagingness, etc.).

# *Limitations*

- Quality assessment based on single reference utterance comparison.
- Generated U3 responses tended to be much longer than the riginal U3's in the dataset.
- The analysis focuses on utterances preceded by a question, would be interesting to extend this to other types of conversational contexts.
- Compressing U2 using GPT-3 may not be the most efficient approach; a heuristic method could be more ideal considering efficiency.
- Utilization of GPT-3; future work could explore GPT-4 or other models for improved insights.

# *Thanks*

Yufei Tao
yutao@pdx.edu