DISCOURSE RELATION PARSING

NATURAL LANGUAGE PROCESSING LAB
@COLUMBIA UNIVERSITY
BY DUAA TARIQ

INTRODUCTION:

BACKGROUND: DISCOURSE RELATION PARSING IS A RESEARCH AREA IN NATURAL LANGUAGE PROCESSING FOCUSING ON ANALYZING DISCOURSE STRUCTURES BY IDENTIFYING DIFFERENT TYPES OF DISCOURSE RELATIONS (IN A SPOKEN OR WRITTEN TEXT).

PREVIOUS RESEARCH: THE STAC DATA SET (ASHER, 2016) IS AN ANNOTATED MULTI-PARTY CHAT LOG FOR AN ONLINE GAME.

PROJECT GOALS:

- CONTRIBUTE A DATASET CONSISTING OF DISCOURSE RELATIONS ANNOTATED FOR LONG DIALOGUES, SPECIFICALLY TRANSCRIPTS OF SHOW EPISODES.

- APPLY KNOWLEDGE OF DISCOURSE RELATION RECOGNITION AND LINK PREDICTION TO THE TASK

- RESEARCH AND USE TOOLS TO EFFECTIVELY COMPLETE ANNOTATIONS

- ANALYZE FREQUENCIES OF DEFINED RELATION LABELS, IDENTIFY NEW CHALLENGES AND RELATION TYPES

METHOD:

Episodes- TV show episodes from the Forever Dreaming collection (Chen & Chu, 2022). Sample size will consist of 1000 episodes.

More- Research tools, previous publications (ex: STAC corpus for an annotation trial)

Annotate- Define the discourse relations prior to the process. Annotate the spoken dialogues in the transcript using Inception, an online web application.

Revise- Revise and redefine the type of discourse relations relevant for the sample. Identify and fix any errors overseen during annotation process.

Understand- Numerically record and analyze the results. Explore further steps and ideas.

Incoming relations

i - n... ,i - 3 ,i - 2 ,i - 1 —-> i

EDUs (elementary discourse units) are any meaningful part of an utterance or statement. Discourse relations will be incoming meaning based on previous EDUs (i-1). Relations do not have to be adjacent and can be linked to any previous EDU.

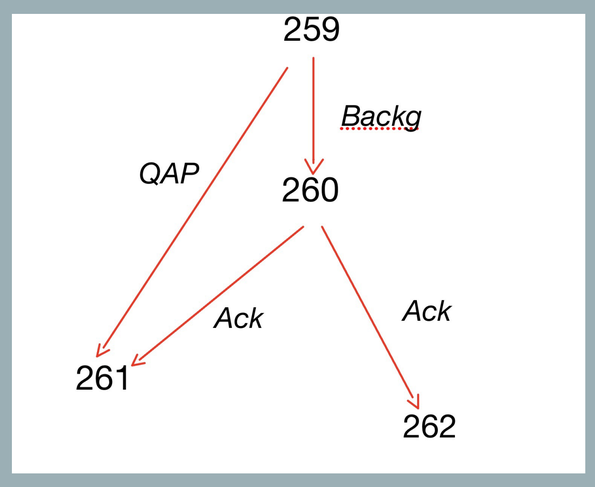
Example:

259 R: *ISN'T THIS AMAZING?*

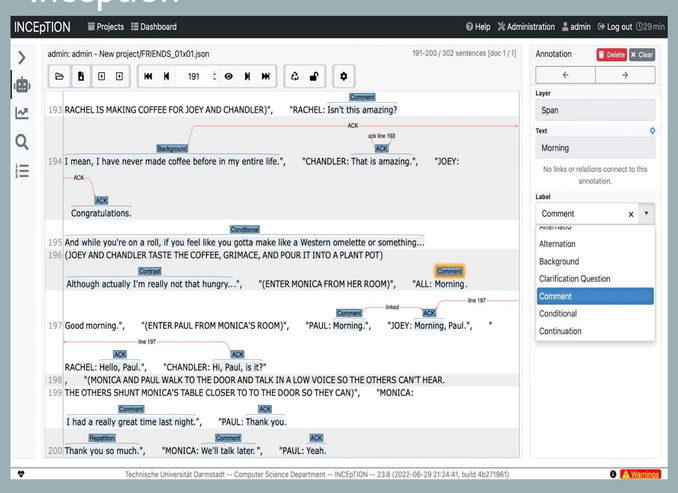
260 R: *I MEAN, I HAVE NEVER MADE COFFEE BEFORE IN MY ENTIRE LIFE*

261 C: *THAT IS AMAZING.*

262 J: *CONGRATULATIONS.*



Inception



RESULTS:

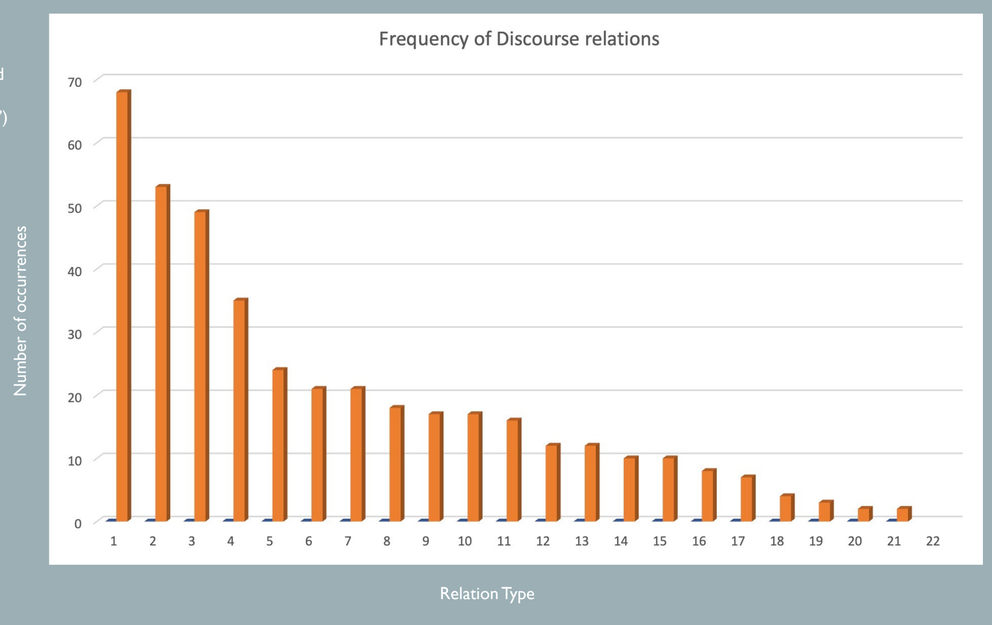
FREQUENCIES OF RELATIONS IN EXAMPLE EPISODE

Relation Type	#
Comment	68
Continuation	53
Acknowledgment	49
Question-Answer Pair	35
Elaboration	24
Clarification Question	21
Background	21

Contrast	18
Explanation	17
Q-Elab	17
Repetition	16
Present Narration	12
Parallel	12
Alternation	10
Rhetorical Question	10

Narrate-Elab	7
Conditional	4
Recount Narration	3
Correction	2
Interruption	2
Total	409

Relations are labelled from 1 ("comment") to 21 ("interruption")



ANALYSIS:

OBSERVATIONS:

IN LONG DIALOGUES, "COMMENT" , "CONTINUATION" AND "ACKNOWLEDGEMENT" ARE THE MOST FREQUENT RELATIONS DUE TO CONSISTENT CONVERSATION BETWEEN CHARACTERS.

"CORRECTION" AND "INTERRUPTION" ARE RARE OCCURRENCES IN THE EXAMPLE EPISODE, LIKELY DUE TO CHANCE.

EACH SCENE BEGINS WITH "COMMENT" SINCE THEY ARE USUALLY UNRELATED TO PREVIOUS EDUS.

"REPETITION" AND "RHETORICAL QUESTION" ARE NEW RELATIONS DUE TO THE ADDED USE OF RHETORIC IN TV EPISODES (ENTERTAINMENT).

"REPETITION" AND "ACKNOWLEDGEMENT" WERE PICKED UP BY INCEPTION AND BECAME PREDICTABLE.

CHALLENGES:

STAC HAS PLAYERS FOCUSED ON ONE OBJECTIVE, WHICH WAS THE GAME. TV EPISODES CONSISTS OF A VARIETY OF SCENARIOS.

TV SHOWS ARE MULTI-MODAL, SO IT IS MORE DIFFICULT TO DETECT SARCASM (OR TONE OF THE SPEAKER) FROM JUST THE WRITTEN TEXT

THE STAC CORPUS CONSISTED OF CHAT ONLY, WHEREAS THERE IS A LOT MORE ACTION AND SPONTANEITY (RESULTING IN INTERRUPTION, ETC.) ON TV SHOWS.

INSTANCES WHERE CHARACTERS SPEAK ON THE PHONE, AND YOU CAN ONLY ASSUME WHAT THE OTHER PERSON IS SAYING. ONE-SIDED UTTERANCES ARE STILL A PART OF THE DIALOGUE.

Future Work:

Friends is a limited genre, so a variety of shows and episodes are needed.

Larger sample size of at least 1000 episodes.

Relations may still need to be redefined, multiple relations for single utterances should be explored.

ACKNOWLEDGEMENTS/ REFERENCES

Mentor: Yilun (Bobby) Hua

PI: Professor Kathleen McKeown

Resources:

Tools- Inception: Technische Universität Darmstadt -- Computer Science Department -- INCEpTION -- 23.8 (2022-06-29 21:24:41, build 4b271961

Forever Dreaming Transcripts: Chen, M., Chu, Z., Wiseman, S., & Gimpel, K. (2022, June 6). *SummScreen: A dataset for abstractive screenplay summarization*. arXiv.org. Retrieved July 28, 2022, from https://arxiv.org/abs/2104.07091

STAC data set: Asher, N., Hunter, J., Morey, M., Farah, B., & Afantenos, S. (n.d.). *Discourse structure and dialogue acts in multiparty dialogue: The stac corpus*. ACL Anthology. Retrieved July 28, 2022, from https://aclanthology.org/L16-1432/

Li, J., Liu, M., Qin, B., & Liu, T. (2022, January 20). *A survey of discourse parsing - frontiers of computer science*. SpringerLink. Retrieved July 28, 2022, from https://link.springer.com/article/10.1007/s11704-021-0500-z

Shi, Z., & Huang, M. (n.d.). *A deep sequential model for discourse parsing on multi-party dialogues*. Proceedings of the AAAI Conference on Artificial Intelligence. Retrieved July 28, 2022, from https://ojs.aaai.org/index.php/AAAI/article/view/4680