EMNLP 2023

MultiTurnCleanup: A Benchmark for Multi-turn Spoken Conversational Transcript Cleanup





Hua Shen[1,2] Vicky Zayats,[1] Johann Rocholl,[1] Dan Walker,[1] Dirk Padfield[1]

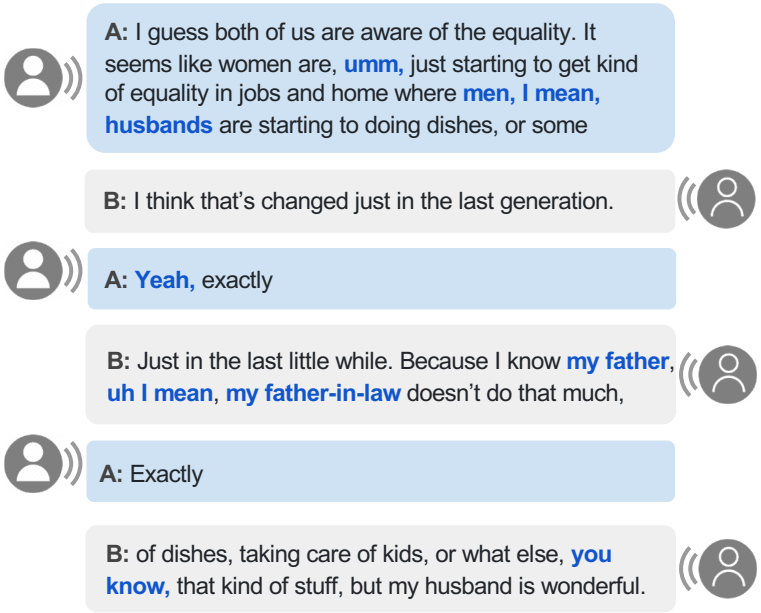
[1] Google Research [2] University of Michigan

Background and Motivation

- Compared with written language, spontaneous **spoken conversational transcripts** often contain many **irregularities**, including **filler pauses**, **repeats**, **self-corrections**, etc.

Google Research

Background and Motivation



- Compared with written language, spontaneous **spoken conversational transcripts** often contain many **irregularities**, including **filler pauses**, **repeats**, **self-corrections**, etc.
- These irregularities can **reduce the performance of downstream tasks** such as natural language understanding (NLU), or **hamper the human readability**.
- Therefore, **disfluency detection** methods are proposed to **remove disfluencies** to improve the readability of spoken conversational transcripts.

Google Research

Background and Motivation

Detect **single-turn disfluencies** successfully

B: Just in the last little while. Because I know **my father**, **uh I mean**, **my father-in-law** doesn't do that much,

Single-Turn Disfluency

BERT-based Models

B: Just in the last little while. Because I know ~~my father, uh I mean,~~ **my father-in-law** doesn't do that much,

Google Research

Background and Motivation

Detect **single-turn disfluencies** successfully

B: Just in the last little while. Because I know **my father**, **uh I mean**, **my father-in-law** doesn't do that much,

Single-Turn Disfluency

BERT-based Models

B: Just in the last little while. Because I know ~~my father, uh I mean,~~ **my father-in-law** doesn't do that much,

Neglect **cross-turn** transcript **irregularities**

B: I think that's changed **just in the last generation.**

A: ~~**Yeah, exactly**~~

B: ~~**Just in the last little while.**~~ Because I know ~~**my father, uh I mean, my father-in-law**~~ doesn't do that much,

A: ~~**Exactly**~~

? Multi-Turn Irregularities

Google Research

Research Objective

*How can we **define** and **resolve** the **task** of **cleaning up multi-turn** spoken conversational **transcripts**?*

Google Research

Challenges

A lack of …

Task Definition and Analysis

Multi-turn Cleanup Datasets

Modeling and Evaluation

Main Contributions

Formulate the **Task** and **Definitions**

Design Data Labeling **Schema and** Collected a Dataset

Built Two **Model** Pipelines and Evaluation **Benchmark**

Google Research

Task Definition and Cleanup Category

We frame the **multi-turn spoken conversational transcript cleanup task** to **remove sources of noise** from both **single-turn** disfluency and **cross-turn** irregularities **simultaneously**.

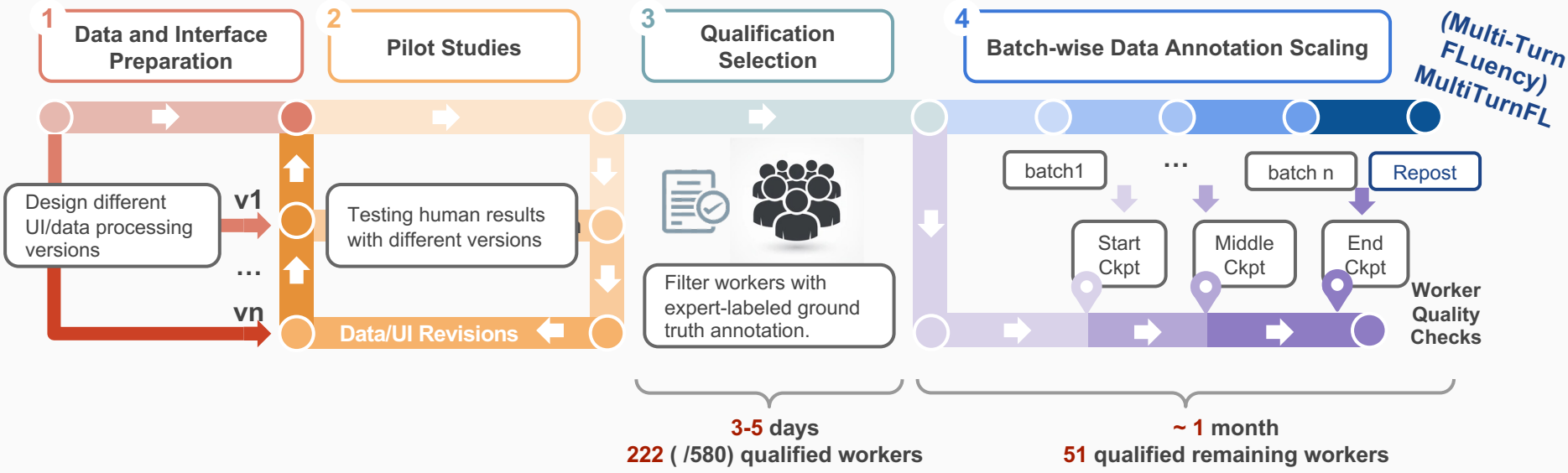
Five multi-turn cleanup **categories**:

Category	Definition	Conversation Instance
Acknowledgment and Confirmation	Speakers show they are listening to and agree with the other speakers.	**A:** I guess both of us are very much aware of the equality. it seems like women are, just starting to get kind of equality in jobs and the home where husbands are starting to doing dishes, ~~or some~~ **B:** I think that's changed just in the last generation. **A:** ~~Exactly.~~ **B:** ~~Just in the last little while.~~ Because I know my father-in-law doesn't do that much, **A:** ~~Exactly.~~ **B:** of dishes, taking care of kids, ~~or what else,~~ ~~you know,~~ that kind of stuff but my husband is wonderful. **A:** that's the way my husband is too. it doesn't bother him to do the dishes, ~~it doesn't bother him to do~~ the laundry verses, men from way back, ~~there is that,~~ if you did that you were henpecked
Repetition and Paraphrase	Speakers may repeat or paraphrase their words during the conversation.	
Think aloud	Speakers talk to themselves during thinking instead of talking to others.	
Incomplete Sentences	Speakers may also say incomplete sentences due to interruption, changing topics, etc.	
Others	All the remaining categories.	

A Novel Data Labeling Schema

Efficient Schema for High-quality Data Collection

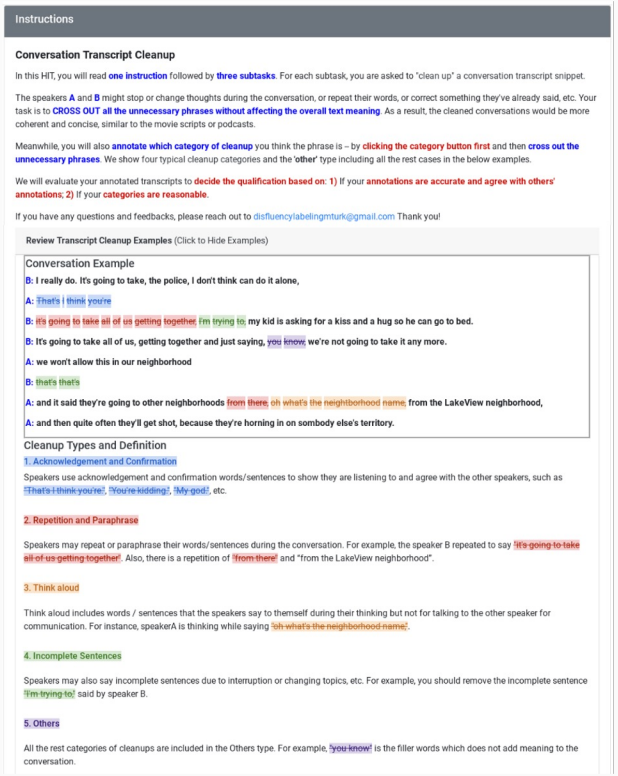
We further propose a **four-step** data labeling **schema** to label both multi-turn **cleanups** and **categories**.

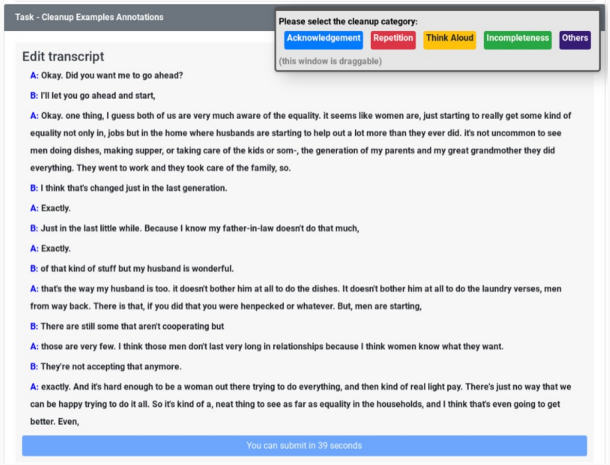


Google Research

User Interface Demo

A Demo of the Human Annotation Interface





- Detailed and foldable Instructions.
- Clear Examples.
- Annotation with Different Categories.
- Minimum Time Constraint to Submit.

....

Google Research

Collected MultiTurnCleanup Data Source and Statistics

MultiTurnCleanup Dataset Statistics

We build the dataset based on widely-adopted **Switchboard** dataset – a human-human telephone conversation transcript corpus.

- After **preprocessing** with the **existing single-turn disfluency labels** on Switchboard;
- We further collect **143k multi-turn cleanup labels** in total.

Data Groups	Filenames	#Conv	#Turns	#Tokens	#Cleanup
Train set	sw2* + sw3*	932	74k	1M	132k
Dev set	sw4[5-9]*	86	3.7k	60k	6.1k
Test set	sw40* and sw41*	64	2.9k	43k	5k
Sum	—	**1082**	**81k**	**1.1M**	**143k**

Table. Statistics of our collected MultiTurnCleanup dataset

Google Research

Label Count and Percentage for Each Category

MultiTurnCleanup Dataset Category and Distribution

We compute the **count** and **percentage** of each multi-turn category. The categories are reasonably distributed.

Category	Definition	Count (%)	Conversation Instance
Acknowledgment and Confirmation	Speakers show that they are listening to and agree with the other speakers	24.3k (17%)	**A**: I guess both of us are very much aware of the equality. it seems like women are, just starting to get kind of equality in jobs and the home where husbands are starting to doing dishes, ~~or some~~ **B**: I think that's changed just in the last generation. **A**: ~~Exactly.~~ **B**: ~~Just in the last little while.~~ Because I know my father-in-law doesn't do that much, **A**: ~~Exactly.~~ **B**: of dishes, taking care of kids, ~~or what else, you know,~~ that kind of stuff but my husband is wonderful. **A**: that's the way my husband is too. it doesn't bother him to do the dishes, ~~it doesn't bother him to do~~ the laundry verses, men from way back , ~~there is that,~~ if you did that you were henpecked.
Repetition and Paraphrase	Speakers may repeat or paraphrase their words during the conversation.	30k (21%)	
Think aloud	Speakers talk to themselves during thinking instead of talking to others.	15.7k (11%)	
Incomplete Sentences	Speakers may also say incomplete sentences due to interruption, changing topics, etc.	47.2k (33%)	
Others	The remaining discontinuity categories.	25.8k (18%)	

Inter-Rater Reliability

Fleiss' Kappa Agreement

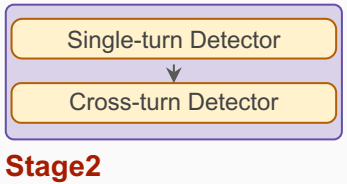
We further compute the **average Fleiss' Kappa** agreement scores of **all** the conversational **turns** in the train / dev / test datasets.

IRR	Experts	MTurk Worker Agreement			
		Train	Dev	Test	All
Fleiss' Kappa	**0.596**	0.5577	0.5927	0.5464	0.5587

- MTurk annotators achieve **Moderate Agreement comparable to** our experts' performance.

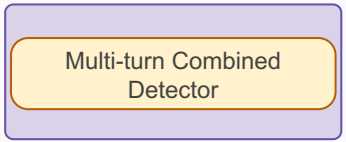
Two Model Pipelines

A Two-stage Model Pipeline



- Trained **single-turn** detector with traditional **Disfluency** dataset
- Trained **cross-turn** detector with **MultiTurnCleanup** dataset

A Combined Model Pipeline



- Created a **Union Labeling Dataset** (Disfluency + MultiTurnCleanup)
- Trained the **combined pipeline** with Union labeling set

Google Research

Evaluation and Benchmark

Model Performance

We evaluate the **F1 score** (F1), **Recall** (R), **Precision** (P) performance on these datasets and models .

- Table. Performance of the two model pipelies

	Model	F1	R	P
Multi-Turn Cleanup Task	BASELINE	58.2	42.5	92.3
	Two-Stage	68.2	64.6	72.3
	Combined	74.9	72.9	76.9

- Table. Model performance on the two subtasks.

Sub-tasks	Model	F1	R	P
DISFLUENCY	STD	89.8	88.3	91.2
MULTITURN CLEANUP	BASELINE	15.5	8.77	65.8
	MTD	56.8	55.4	58.3

- When evaluating the **two model pipelines** on the **MultiTurnCleanup task** (i.e., remove **both single**-turn and **multi**-turn **cleanups**), the **combined model** performed **best**, and **both** pipelines **outperformed** the **baseline**.
- When evaluating the **two sub-tasks**, including **single**-turn **disfluency** and **multi**-turn **discontinuity** detection, we found the **MTD(MultiTurn Detector) model** trained with our **collected** dataset could also **outperform** the **Baseline** model.

Google Research

Takeaway

We defined a **novel** “Multi-turn Conversational Transcript Cleanup” **task**, proposed a **labeling schema** for getting high-quality data via MTurk, and collected the **MultiTurnCleanup dataset**.

Further, we proposed **two model pipelines** and the **evaluation benchmark** for future research to address this novel task.

Thank You!

Paper

Dataset

- Hua shen huashen@umich.edu huashen218
- Vicky Zayats vzayats@google.com
- Johann Rocholl jcrocholl@google.com
- Dan Walker danwalkeriv@google.com
- Dirk Padfield padfield@google.com

Google Research