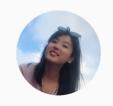


MultiTurnCleanup: A Benchmark for Multi-turn Spoken Conversational Transcript Cleanup











Hua Shen, Vicky Zayats, Johann Rocholl, Dan Walker, Dirk Padfield





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





A: I guess both of us are aware of the equality. It seems like women are, umm, just starting to get kind of equality in jobs and home where men, I mean, husbands are starting to doing dishes, or some

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,





B: of dishes, taking care of kids, or what else, **you know**, that kind of stuff, but my husband is wonderful.



- 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.

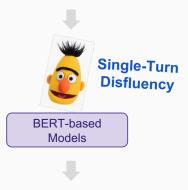




Detect single-turn disfluencies successfully



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





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



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

B: Just in the last little while. Because I know my father up.1 mean -my father-in-law doesn't do that much,

Models

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









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

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**

Task Definition and Multi-Turn Cleanup Category



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:

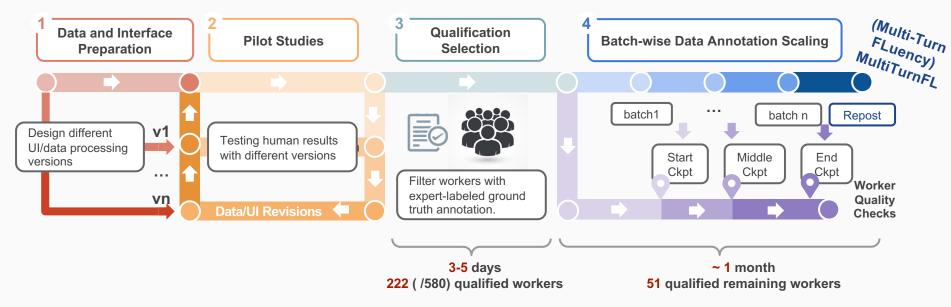
Category	Definition	Conversation Instance
Acknowledgment and Confirmation	Speakers show they are listening to and agree with the other speakers.	A: I guess both of us are very much aware of the equality. it seems like women are, just starting to get kind of equality in jobs and the home where husbands are starting to doing dishes, or some
Repetition and Paraphrase	Speakers may repeat or paraphrase their words during the conversation.	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
Think aloud	Speakers talk to themselves during thinking instead of talking to others.	doesn't do that much, A: Exactly. B: of dishes, taking care of kids, or what else, you know, that kind
Incomplete Sentences	Speakers may also say incomplete sentences due to interruption, changing topics, etc.	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
Others	All the remaining categories.	way back, there is that, if you did that you were henpecked

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**.





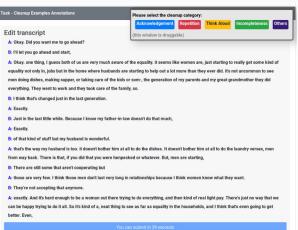
User Interface Demo



conversation.

A Demo of the Human Annotation Interface

Conversation Transcript Cleanup In this HIT, you will read one instruction followed by three subtasks. For each subtask, you are asked to "clean up" a conversation transcript snippet. The speakers A and B might stop or change thoughts during the conversation, or repeat their words, or correct something they've already said, etc. Your task is to CROSS OUT all the unnecessary phrases without affecting the overall text meaning. As a result, the cleaned conversations would be more coherent and concise, similar to the movie scripts or podcasts. Meanwhile, you will also annotate which category of cleanup you think the phrase is - by clicking the category button first and then cross out the unnecessary phrases. We show four typical cleanup categories and the 'other' type including all the rest cases in the below examples. We will evaluate your annotated transcripts to decide the qualification based on: 1) If your annotations are accurate and agree with others' annotations; 2) If your categories are reasonable. If you have any questions and feedbacks, please reach out to disfluencylabelingmturk@gmail.com Thank you! Review Transcript Cleanup Examples (Click to Hide Examples) Conversation Example B: I really do. It's going to take, the police, I don't think can do it alone, A: That's I think you're B: it's going to take all of us getting together, I'm trying to, my kid is asking for a kiss and a hug so he can go to bed. B: It's going to take all of us, getting together and just saying, you know, we're not going to take it any more. A: we won't allow this in our neighborhood B: that's that's A: and it said they're going to other neighborhoods from there, oh what's the neighborhood name, from the LakeView neighborhood, A: and then guite often they'll get shot, because they're horning in on sombody else's territory. Cleanup Types and Definition 1. Acknowledgement and Confirmation Speakers use acknowledgement and confirmation words/sentences to show they are listening to and agree with the other speakers, such as "That's I think you're!, "You're kidding!, "My-god!", etc. 2. Repetition and Paraphrase Speakers may repeat or paraphrase their words/sentences during the conversation. For example, the speaker B repeated to say 41's going to take all of us getting together. Also, there is a repetition of from there and from the LakeView neighborhood. 3. Think aloud Think aloud includes words / sentences that the speakers say to themself during their thinking but not for talking to the other speaker for communication. For instance, speakerA is thinking while saying "oh what's the neighborhood name," Speakers may also say incomplete sentences due to interruption or changing topics, etc. For example, you should remove the incomplete sentence "I'm trying to," said by speaker B. 5. Others All the rest categories of cleanups are included in the Others type. For example, "you know! is the filler words which does not add meaning to the



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

. . . .



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

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
Repetition and Paraphrase	Speakers may repeat or paraphrase their words during the conversation.	30k (21%)	 B: I think that's changed just in the last generation. A: Exactly. B: Just in the last little while. Because
Think aloud	Speakers talk to themselves during thinking instead of talking to others.	15.7k (11%)	I know my father-in-law doesn't do that much, A: Exactly. B: of dishes, taking care of kids, or what else, were known that kind of stuff but my bushend is
Incomplete Sentences	Speakers may also say incomplete sentences due to interruption, changing topics, etc.	47.2k (33%)	wonderful. A: that's the way my husband is too. it doesn't bother him to do the dishes, it doesn't bother
Others	The remaining discontinuity categories.	25.8k (18%)	him to do the laundry verses, men from way back , there is that, if you did that you were henpecked.

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			
Fleiss' Kappa	. =0.0	Train	Dev	Test	All
	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

Stage1

Single-turn Detector

V

Cross-turn Detector

Stage2

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

A Combined Model Pipeline

Multi-turn Combined Detector

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



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	Baseline	15.5	8.77	65.8
CLEANUP	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!





Hua shen <u>huashen@umich.edu</u> <u>huashen218</u>

Vicky Zayats <u>vzayats@google.com</u>

Johann Rocholl jcrocholl@google.com

Dan Walker <u>danwalkeriv@google.com</u>

Dirk Padfield <u>padfield@google.com</u>