ISA-19, Nancy, France
June 20, 2023

Colorado State University

How Good is Automatic Segmentation as a Multimodal Discourse Annotation Aid?

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, **Nikhil Krishnaswamy**, and Nathaniel Blanchard

Introduction

- Modern AI systems depend on annotated training data
- Most systems rely on “oracle” (gold-standard, human) annotations
- However, real-world deployments increasingly use some automation in preprocessing

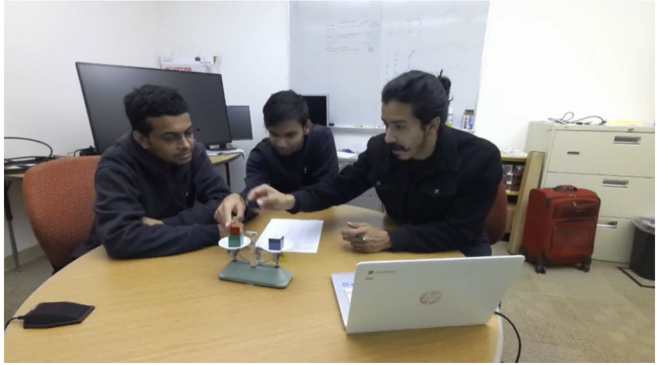
Research Questions

- How does such automated preprocessing affect downstream annotation?
- Can annotators rely on automated preprocessing?
- How should developers of annotation specs account for the use of automated preprocessing (e.g., speech recognition)?

Study Domain

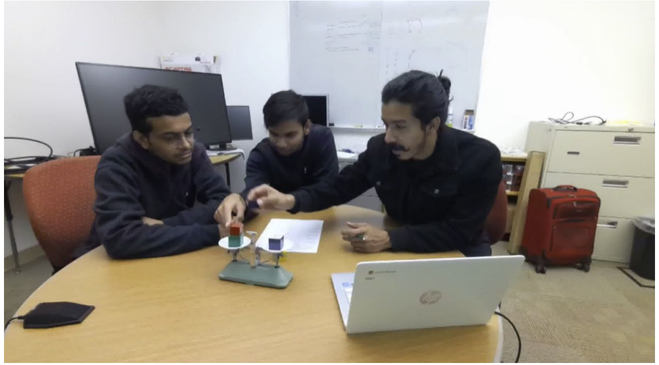
- We focus on a group collaborative problem solving (CPS) task
- Multiple modalities are indicated (speech, gesture, action, etc.)
- Annotation of CPS is performed at utterance level
- Specs assume that utterances have been segmented and transcribed by humans

Use Case: Weights Task



- 10 groups of 3 volunteers, 170 mins. of video
- Determine the weights of several colored cubes
- Discuss and record the weights discovered, infer the pattern
- Requires manipulating objects with group collaboration

Use Case: Weights Task



- Collaborative Problem Solving
- Schema breaks down *facets*, *sub-facets*, and *indicators*
- Facets: *Constructing Shared Knowledge*, *Negotiation/Coordination*, and *Maintaining Team Function*

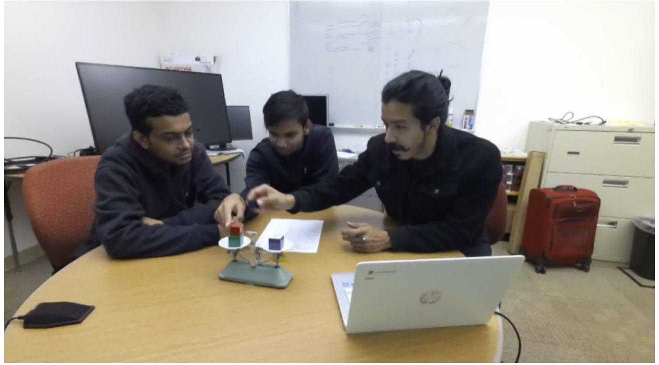
Use Case: Weights Task

Table 1
Proposed generalized competency model of facets, sub-facets, and indicators.

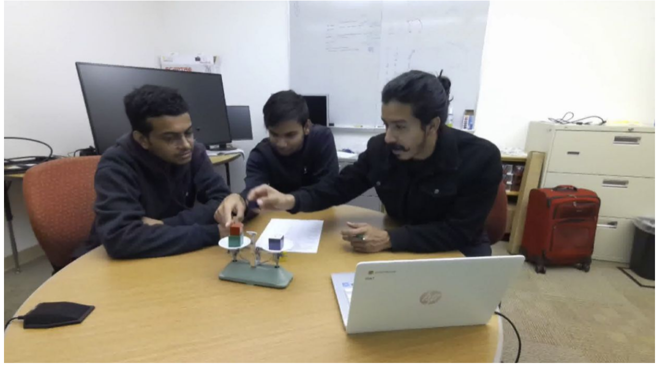
Facet	Sub-facet	Indicators
Constructing shared knowledge—expresses one's own ideas and attempts to understand others' ideas	Shares understanding of problems and solutions	Talks about specific topics/concepts and ideas on problem solving ● Proposes specific solutions ● Talks about givens and constraints of a specific task ● Builds on others' ideas to improve solutions
	Establishes common ground	Recognizes and verifies understanding of others' ideas • Confirms understanding by asking questions/ paraphrasing Repairs misunderstandings ● Interrupts or talks over others as intrusion (R)
Negotiation/Coordination—achieves an agreed solution plan ready to execute	Responds to others' questions/ideas	● Does not respond when spoken to by others (R) ● Makes fun of, criticizes, or is rude to others (R) ● Provides reasons to support/refute a potential solution ● Makes an attempt after discussion
	Monitors execution	● Talks about results ● Brings up giving up the challenge (R)
Maintaining team function—sustains the team dynamics	Fulfills individual roles on the team	● Not visibly focused on tasks and assigned roles (R) ● Initiates off-topic conversation (R) ● Joins off-topic conversation (R)
	Takes initiatives to advance collaboration processes	● Asks if others have suggestions ● Asks to take action before anyone on the team asks for help ● Compliments or encourages others

Note. “R” next to an indicator means that it is reverse coded.

Sun et al., 2020

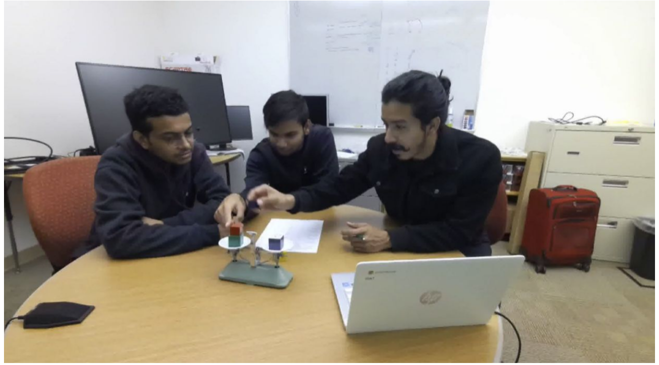


Use Case: Weights Task



- Multimodal task: not every communicative act is spoken
- In-person, situated collaboration
- Cross-talk, interruptions, incomplete sentences all pose challenges for ASR

Use Case: Weights Task



- CPS annotation involves both listening to the audio and watching the video
- It may be unclear what collaborative moves are made without full context (multiple utterances, full situational information)
- How much information is lost with automatic segmentation/transcription?

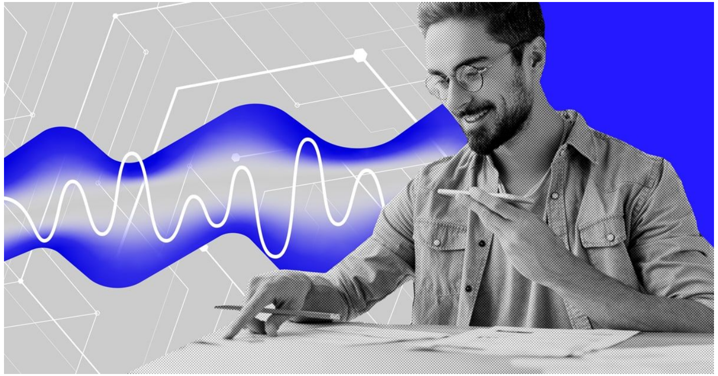
Methodology

- Manually segment and transcribe A/V data
- Annotate manually-segmented data
- Automatically segment and transcribe A/V data
- Map annotations to automatically-preprocessed data
- Evaluate the differences

Methodology: ASR

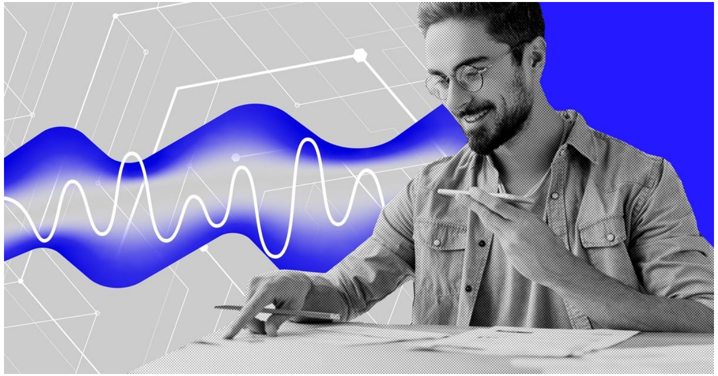
- Automatic segmentation and transcription
- Google ASR
- Whisper

Methodology: Annotation



- Videos transcribed by hand
- Marking each person's speech (.1 sec. intervals)
- Adding CPS annotation codes to each utterance (multiple codes allowed - binary task)

Methodology: Annotation



- Mapped manual annotations (oracle)

  to automatically-segmented utterances from Google and Whisper

- Multiclass binary CPS labels mapped from oracle to automatic utterances

  by temporal overlap

  - If oracle utterance overlaps with multiple ASR utterances, biggest overlap is chosen

Results: Count of Utterances

Group	1	2	3	4	5	6	7	8	9	10
Whisper	297	201	391	293	406	278	311	354	136	346
Google	139	151	254	128	146	153	380	235	90	146
Oracle	229	207	337	195	237	227	590	338	134	379

Table 1: # of utterances per group determined by each segmentation method. Totals: Whisper - 3,013 utterances; Google - 1,822 utterances; Oracle - 2,873.

Results: Count of Utterances

- Almost uniformly, Whisper segments more utterances than the oracle, and Google creates fewer
- Google performs well at not transcribing silence
- Whisper may invent an utterance to fill space (hallucination seems to be a common problem with OpenAI products?)

Results: Intrinsic ASR Metrics

	Google				Whisper			
Group	WER	Sub. rate	Del. rate	Ins. rate	WER	Sub. rate	Del. rate	Ins. rate
1	0.571	0.252	0.113	0.206	0.534	0.193	0.045	0.296
2	0.459	0.211	0.128	0.120	0.416	0.177	0.040	0.200
3	0.539	0.236	0.117	0.186	0.527	0.177	0.047	0.303
4	0.529	0.267	0.154	0.170	0.572	0.201	0.040	0.332
5	0.631	0.262	0.173	0.195	0.581	0.175	0.060	0.346
6	0.581	0.252	0.077	0.252	0.525	0.191	0.041	0.293
7	0.610	0.260	0.155	0.196	0.650	0.209	0.064	0.377
8	0.532	0.259	0.137	0.137	0.486	0.200	0.048	0.238
9	0.571	0.274	0.180	0.118	0.514	0.229	0.084	0.202
10	0.645	0.306	0.087	0.252	0.612	0.202	0.054	0.356
Average	0.573	0.259	0.132	0.183	0.542	0.195	0.052	0.294

Table 2: WER, substitution rate, deletion rate, and insertion rate by group.

Results: Intrinsic ASR Metrics

- Evaluation of ASR after automatic segmentation is a proxy for information lost during segmentation process
- Google: significantly more deletion and substitution errors
- Whisper: significantly more insertion errors
- Follows patterns established in utterance counts

Results: Difference in Annotations

- Example: Interruptions (CPS indicator #5)
- Automatic segmentation may split or lump utterances separated by interruption
- Annotations at utterance level may miss interruption entirely

Segment	Label
Weren't those both thirty or no only one of them twenty and thirty	• Confirms understanding
No this is twenty you're off the team	• Interrupts • Initiates off-topic conversation
Twenty and then	None
Weren't those both thirty or no only one of them twenty and thirty No this is twenty you're off the team Twenty and then	• Confirms understanding • Interrupts • Initiates off-topic conversation
Weren't those both thirty	• Confirms understanding
No this is twenty	• Interrupts • Initiates off-topic conversation
Twenty and thirty	• Confirms understanding
Twenty and then	None
You're off the team	• Interrupts • Initiates off-topic conversation

Figure 1: Overlap between oracle (top), Google (middle), and Whisper (bottom) segments. Right column shows the CPS indicator annotated for each utterance.

Results: Difference in Annotations

- Google segmentation causes “Interrupts” to be assigned to the first and last utterances
- Whisper segmentation results in extra “Interrupts” and “Off-topic conversation annotations”

Segment	Label
Weren't those both thirty or no only one of them twenty and thirty	• Confirms understanding
No this is twenty you're off the team	• Interrupts • Initiates off-topic conversation
Twenty and then	None
Weren't those both thirty or no only one of them twenty and thirty No this is twenty you're off the team Twenty and then	• Confirms understanding • Interrupts • Initiates off-topic conversation
Weren’t those both thirty	• Confirms understanding
No this is twenty	• Interrupts • Initiates off-topic conversation
Twenty and thirty	• Confirms understanding
Twenty and then	None
You’re off the team	• Interrupts • Initiates off-topic conversation

Figure 1: Overlap between oracle (top), Google (middle), and Whisper (bottom) segments. Right column shows the CPS indicator annotated for each utterance.

Results: Difference in Annotations

- Example: person speaks, pauses, completes sentence

  “*Think it just feels like it’s*”—0.3 seconds—”*a lot heavier…*”

- Single utterance, split in two
- Should be coded “Discussing results,” instead neither utterance is coded as anything

Discussion

- Collaborative Problem Solving Requirements
- Quality Loss
- Annotation Priority

Discussion

- CPS is a challenging task (6 months to train annotator)
  - Any method of speeding up the process is valuable
- We focused on segmentation of audio
- Annotations themselves require viewing video, listening to intonation, and including temporal context

Discussion

- If annotations performed with access to multimodal information are applied to automatically-segmented audio
- Information loss can be severe

Discussion

Table 2: Weighted average AUROC for binary classification

Modalities	Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.814	.804	.829	.788	.783	.791	.712	.689	.678
Prosodic	.832	.796	.714	.730	.710	.595	.661	.649	.598
Verbal + Prosodic	.840	.818	.794	.785	.794	.760	.720	.699	.645

Table 3: Standard deviations of weighted average AUROC across all 10 groups for binary classification

Modalities	Construction of shared knowledge			Negotiation/ Coordination			Maintaining team function		
	RF	AB	NN	RF	AB	NN	RF	AB	NN
Verbal	.044	.037	.040	.054	.052	.057	.082	.079	.079
Prosodic	.038	.051	.118	.055	.056	.094	.077	.074	.091
Verbal + Prosodic	.035	.044	.143	.054	.052	.099	.076	.088	.095

Bradford et al., 2023

Discussion

- Preparing annotations over oracle utterances and then transferring to automatically-segmented utterances (e.g., through rules or a classifier) may obscure semantic information captured at oracle level
- Backs up previous conclusions, such as need for annotators to agree on both spans and annotations

Conclusion

- As AI systems trained over annotated data proliferate, inference will necessarily be performed over automatically preprocessed data
- Future models will benefit from task-aware annotation specs that account for noise introduced by imperfect preprocessing

Conclusion

- CPS example: if multiple labels may be lumped into a single utterance, should one be allowed to “dominate”?
- Which information is most important to capture if some is assumed to be lost in preprocessing
- Characterizing potential preprocessing tools and accounting for their effects in the annotation scheme

Thank you

Supported by NSF DRL #2019805, “Institute for Student-AI Teaming”

Colorado State University