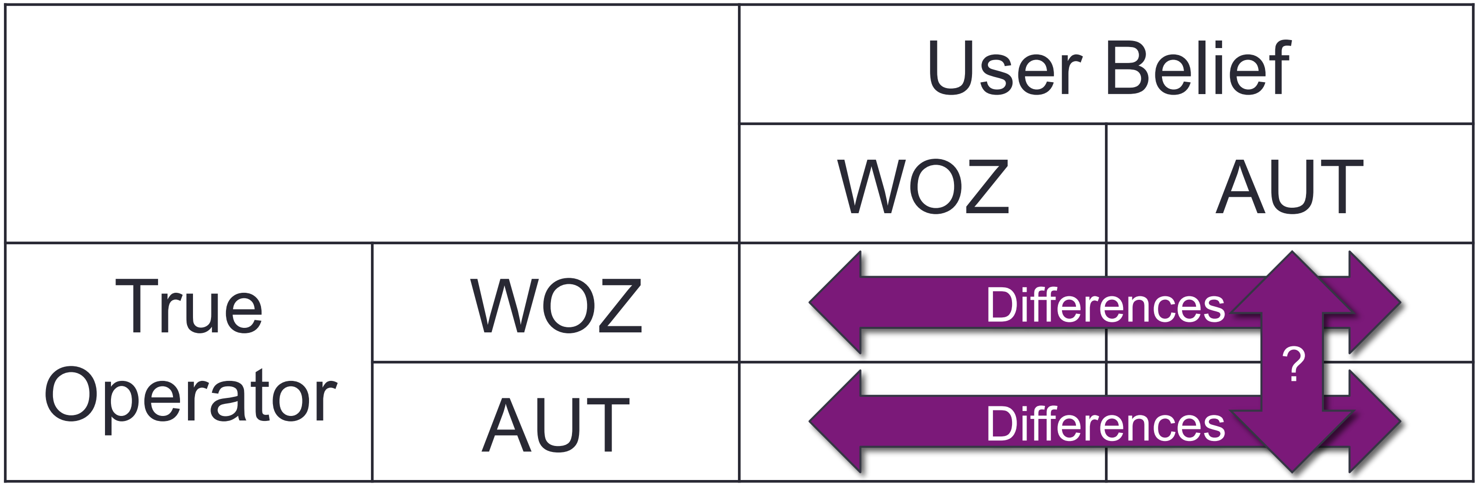
Introduction – WOZ setup

- Use hidden, human component
- WOZ experimental protocol calls for holding “all other input and output … constant so that the only unknown variable is who does the internal processing” (Paek, 2001)
- WOZ systems appear automated to user
- Gather data for fully-automated system

Introduction – WOZ performance

- Assume user behavior is similar between the WOZ and automated (AUT) setups
- In one system, training with AUT data gave rise to better performance than training WOZ data (Drummond and Litman, 2011)
- System automation differences may have caused performance gap
- Differences in user behavior may weaken automated system performance

Introduction - goal



		User Belief	
		WOZ	AUT
True Operator	WOZ		
	AUT		

- Investigate differences in WOZ and AUT user behaviors
- Hypothesized that what users say and how they say it will differ between WOZ and AUT setups

Outline

- ~~Introduction~~
- Dialogue System
- Post-hoc Experiment
- Results
- Conclusions

Dialogue System - ITSPOKE

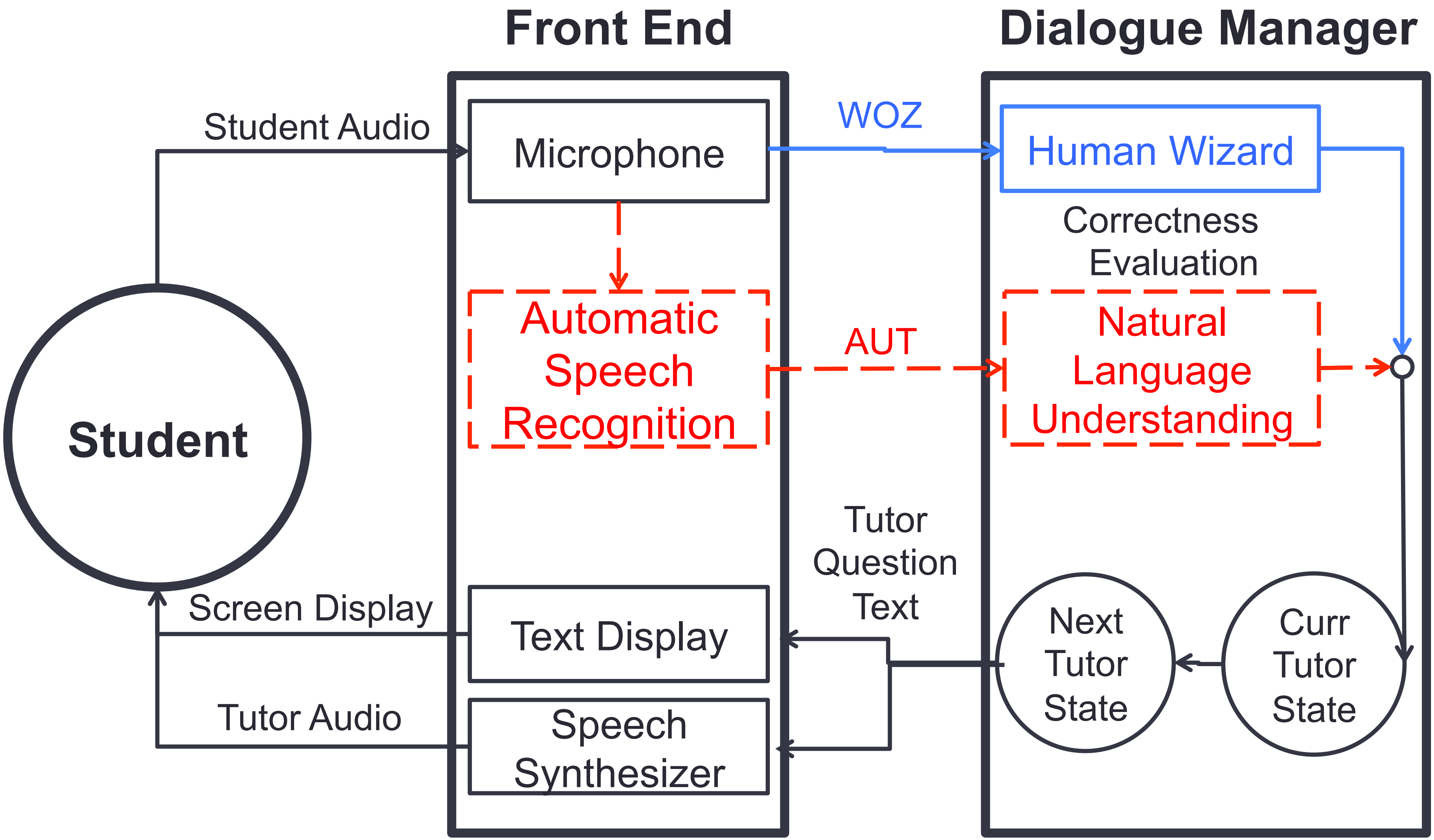
- Our data comes from the Intelligent Tutoring Spoken Dialogue System (ITSPOKE)
- We draw from two prior experiments (one WOZ, one AUT) (Forbes-Riley and Litman(a), 2011; Forbes- Riley and Litman(b), 2011)
- Baseline, non-adaptive conditions of those experiments
- Users tutored in basic Newtonian physics
- Dialogues illustrated one or more basic physics concepts

Dialogue System – sample dialogue

- Tutor text is shown on a screen and read aloud via text-to-speech, and the user responds verbally to the tutor’s queries

Tutor	So what are the forces acting on the packet after it’s dropped from the plane?
Student	um gravity then well air resistance is negligible just gravity
Tutor	Fine. So what’s the direction of the force of gravity on the packet?
Student	vertically down

Dialogue System - workflow



Dialogue System – two user groups

- Setups varied by component for understanding and evaluating responses
  - One human, one automated
- Each student participated in only one setup
  - Students were not informed whether the system was fully automated
- Distinct student group responses constitute data

Outline

- ~~Introduction~~
- ~~Dialogue System~~
- Post-hoc Experiment
- Results
- Conclusions

Post-hoc Experiment

- Determine whether differences exist between WOZ and AUT responses
- Compared features of user turns to each question individually
- The table below shows the number of users and dialogue turns they took for each setup over 111 questions asked in both setups

System	#Users	#Turns
WOZ	21	1542
AUT	25	2034

Post-hoc Experiment - features

- Prosodic features: length of the pause before speech began, speech duration, pitch, and energy (RMS)
- Pitch and energy: maximum, minimum, mean, and standard deviation
- 10 total prosodic features
- Normalized each prosodic feature using same algorithm as live system

Post-hoc Experiment - features

- Lexical features: *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker et al., 2001)
  - *Tentative(T):* "maybe", "perhaps", and "guess"
  - *Prepositions(P):* "to", "with", and "above"
  - Utterance "Maybe above" would receive feature vector:
    - <0, …, 0, T=50, 0, …, 0, P=50, 0, …, 0>
- Used human transcriptions for all utterances
- 69 total LIWC lexical category features

Post-hoc Experiment

- Looked for response feature differences for each question in two ways:
- 1) A statistical comparison of features
- 2) Response classification via machine learning

Outline

- ~~Introduction~~
- ~~Dialogue System~~
- ~~Post-hoc Experiment~~
- Results
  - Statistical Comparison of Features
  - Response Classification Experiments
- Conclusions

Statistical Comparison of Features

- For each question, all features between WOZ and AUT responses were compared
- Welch’s unpaired, two-tailed t-tests

Statistical Comparison of Features

- Possible that differences were inherent in WIZ/AUT student groups
- Created control groups with evenly mixed, randomly selected WIZ/AUT students
- We report only questions for which at least one feature differed between WOZ and AUT but not between these two control groups

Statistical Comparison of Features

- The number of questions for which at least one feature differed statistically significantly ($p < 0.05$) between WOZ and AUT responses

Feature Set	#Questions	%Corpus by Turns
Prosodic	42	46.22%
Lexical	33	35.46%
Either	61	66.86%

Statistical Comparison of Features

- 10/10 prosodic, 29/69 lexical features differed significantly ($p < 0.05$) for at least one question
- Features differing for at least 10% of the corpus:

Feature	%Corpus	#Questions	#WOZ>AUT
Duration	22.15%	19	1
RMS Min	16.86%	15	14
Dictionary Words	15.13%	13	11
pronoun	12.56%	10	10
social	11.35%	9	8
funct	10.99%	9	9
Six Letter Words	10.91%	9	0

Statistical Comparison of Features

- Users used more words with the wizarded system

Feature	%Corpus	#Questions	#WOZ>AUT
Dictionary Words	15.13%	13	11
pronoun	12.56%	10	10
social	11.35%	9	8
funct	10.99%	9	9

- There exist features which differ for a substantial number of questions

Statistical Comparison of Features

- A question for which the *Dictionary Words* feature was greater for WOZ responses:

Tutor	So how do these two forces’ directions compare?		
Most common responses		**Longest responses**	

WOZ(9)
AUT(2)
	they are opposite	WOZ	the relationship between the two forces’ directions are towards each other since the sun is pulling the gravitational force of the earth

WOZ(3)
AUT(8)
	opposite	AUT	they are opposite directions

Outline

- ~~Introduction~~
- ~~Dialogue System~~
- ~~Post-hoc Experiment~~
- Results
  - ~~Statistical Comparison of Features~~
  - Response Classification Experiments
- Conclusions

Response Classification Experiments

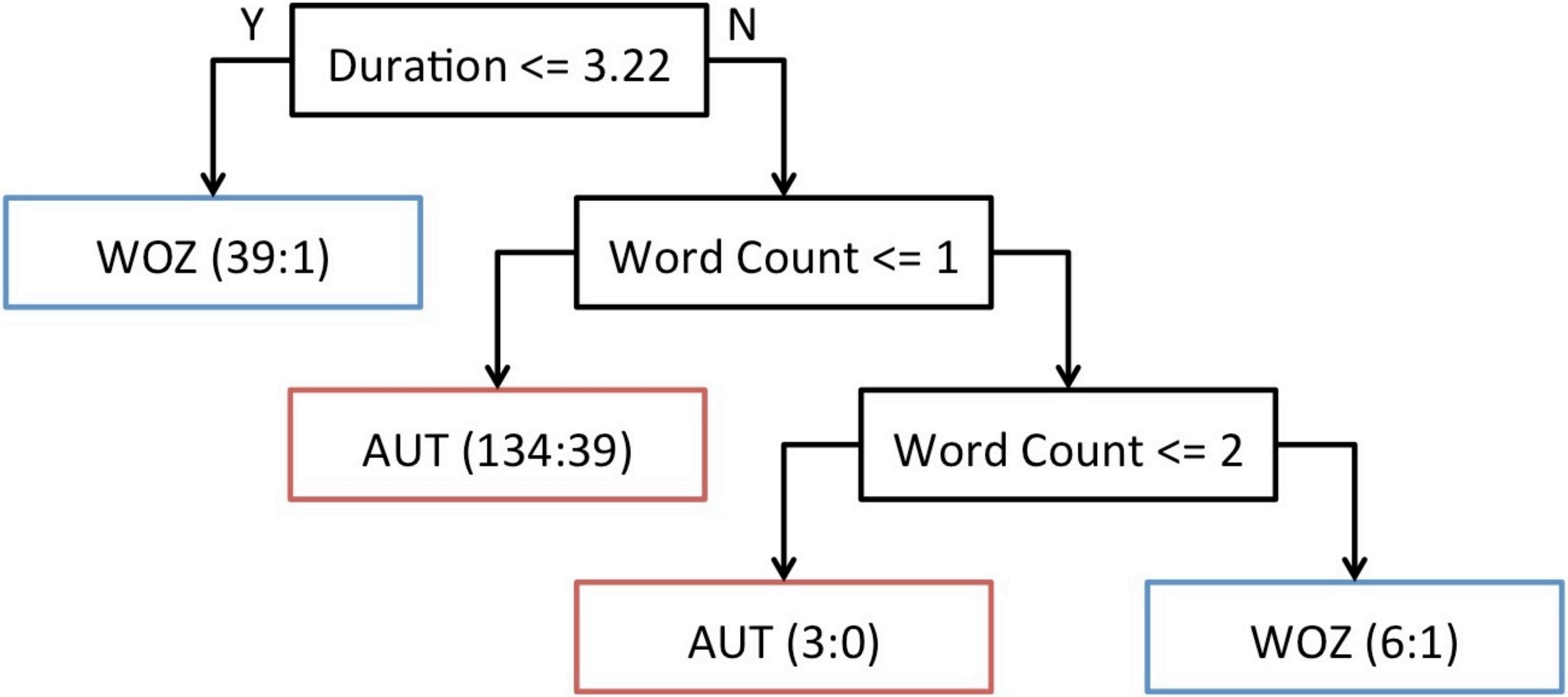
- Use classification models to distinguish WOZ/AUT setup
- J-48 model was trained and tested for each question
- Accuracy compared against a majority-class baseline

Response Classification Experiments

- 97 questions considered in total
- 21/97 outperformed the majority-class baseline
- 32.79% of the corpus by turns

Response Classification Experiments

- “Would you like to do another problem?”

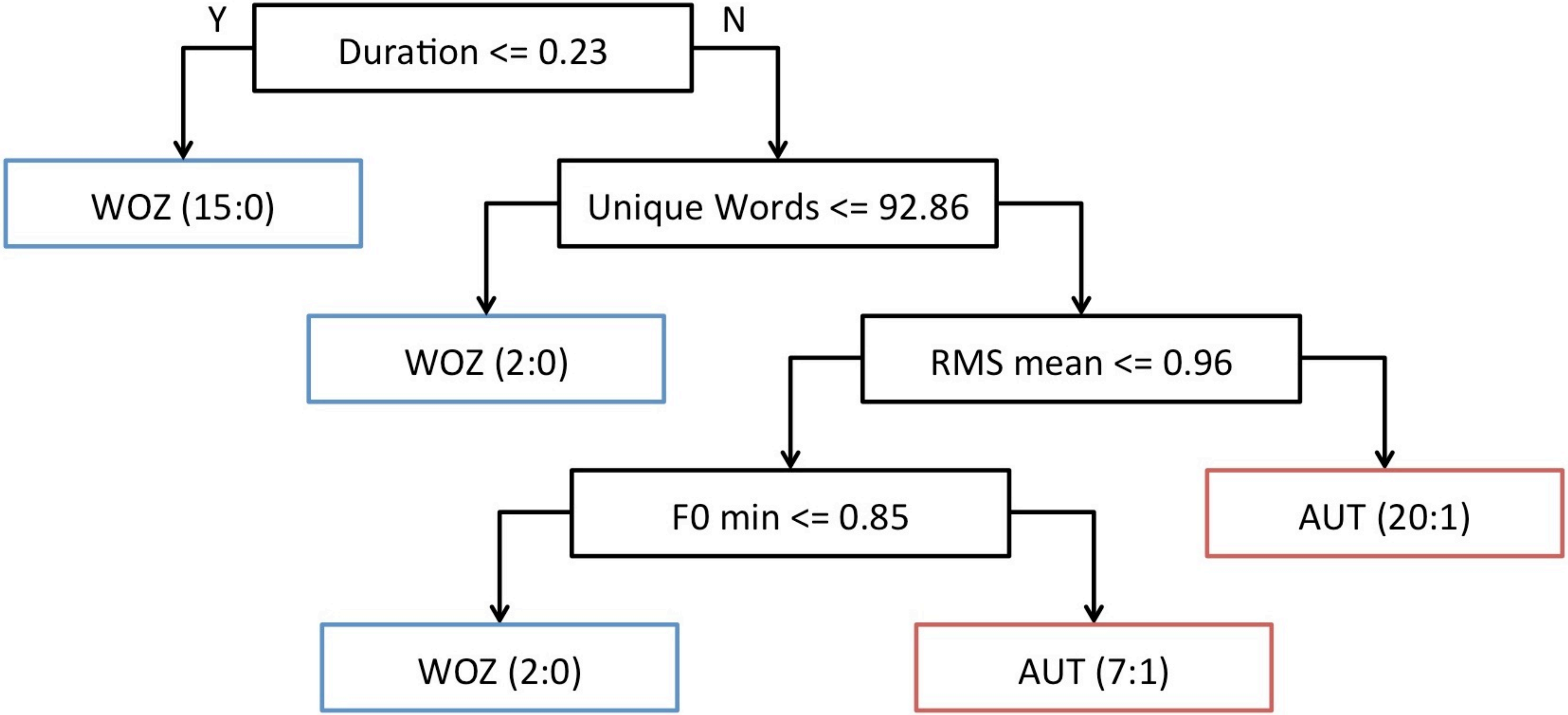


Response Classification Experiments

- This result is consistent with literature (Schechtman and Horowitz, 2003; Rosé and Torrey, 2005) that suggests that users interacting with automated systems will be more curt

Response Classification Experiments

“Now let’s find the forces exerted on the car in the vertical direction during the collision. First, what vertical force is always exerted on an object near the surface of the earth?”



Outline

- ~~Introduction~~
- ~~Dialogue System~~
- ~~Post-hoc Experiment~~
- ~~Results~~
  - ~~Statistical Comparison of Features~~
  - ~~Response Classification Experiments~~
- Conclusions

Discussion

- There exist significant differences between user responses to a wizarded and an automatic dialogue system’s questions
- Contribution of the wizard was limited to speech recognition and correctness evaluation

Discussion

- Results suggest that user speech changes as a result of user confidence in the system’s accuracy
- Relationship between user confidence and user speech may be analogous to observed differences in past experiments
- These results suggest ways in which raw wizarded data may fall short of ideal for training an automated system

Future Work - exploration

- Measure how the observed differences change over the course of the dialogue
- Use different methods of normalization for user speech values

Future Work - solutions

- Intentional wizard error could be introduced to frustrate the user; analogous to intentional errors produced in user simulation (Lee and Eskenazi, 2012)
- Generalizable statistical classification domain adaptation (Daumé and Marcu, 2006) and adaptation demonstrated to work well in NLP-specific domains (Jiang and Zhai, 2007)

DIFFERENCES IN USER RESPONSES TO A WIZARD-OF-OZ VERSUS AUTOMATED SYSTEM

Jesse Thomason and Diane Litman

University of Pittsburgh