Game-Based Video-Context Dialogue

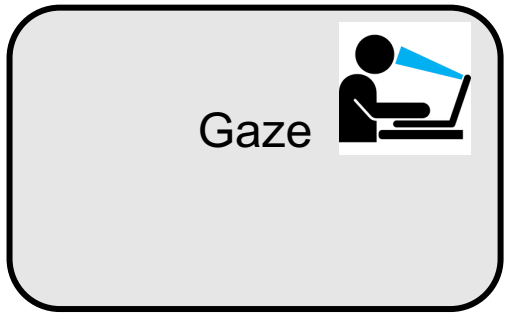
Ramakanth Pasunuru
www.rama-kanth.com

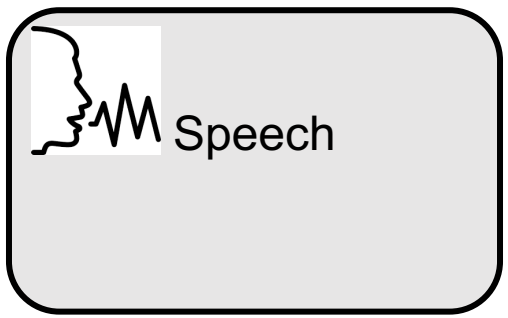
Mohit Bansal
www.cs.unc.edu/~mbansal/

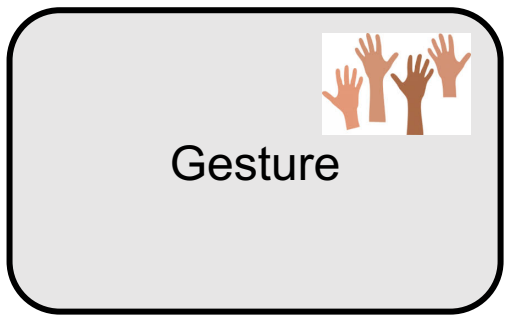
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNC NLP

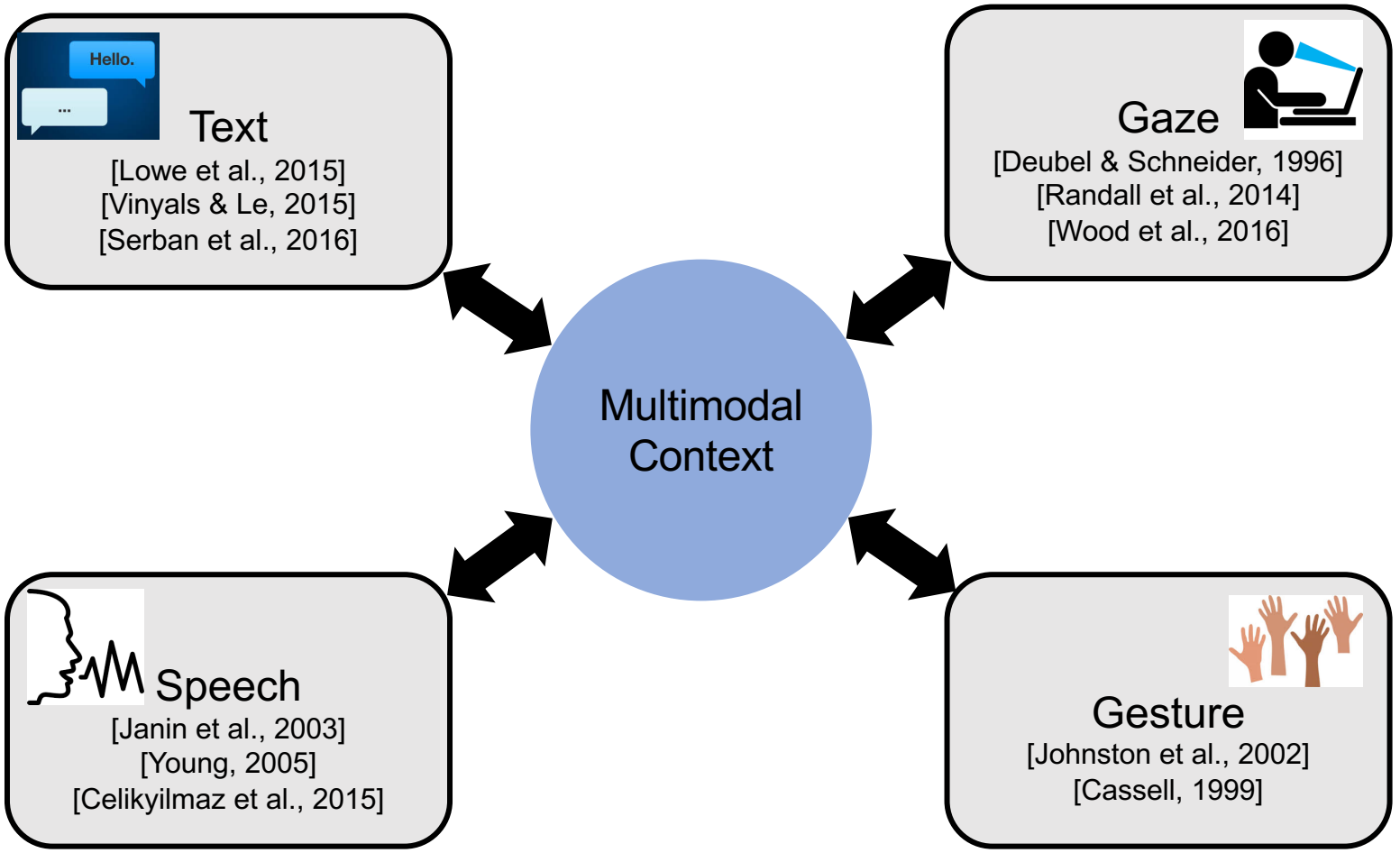
Dialogue Context



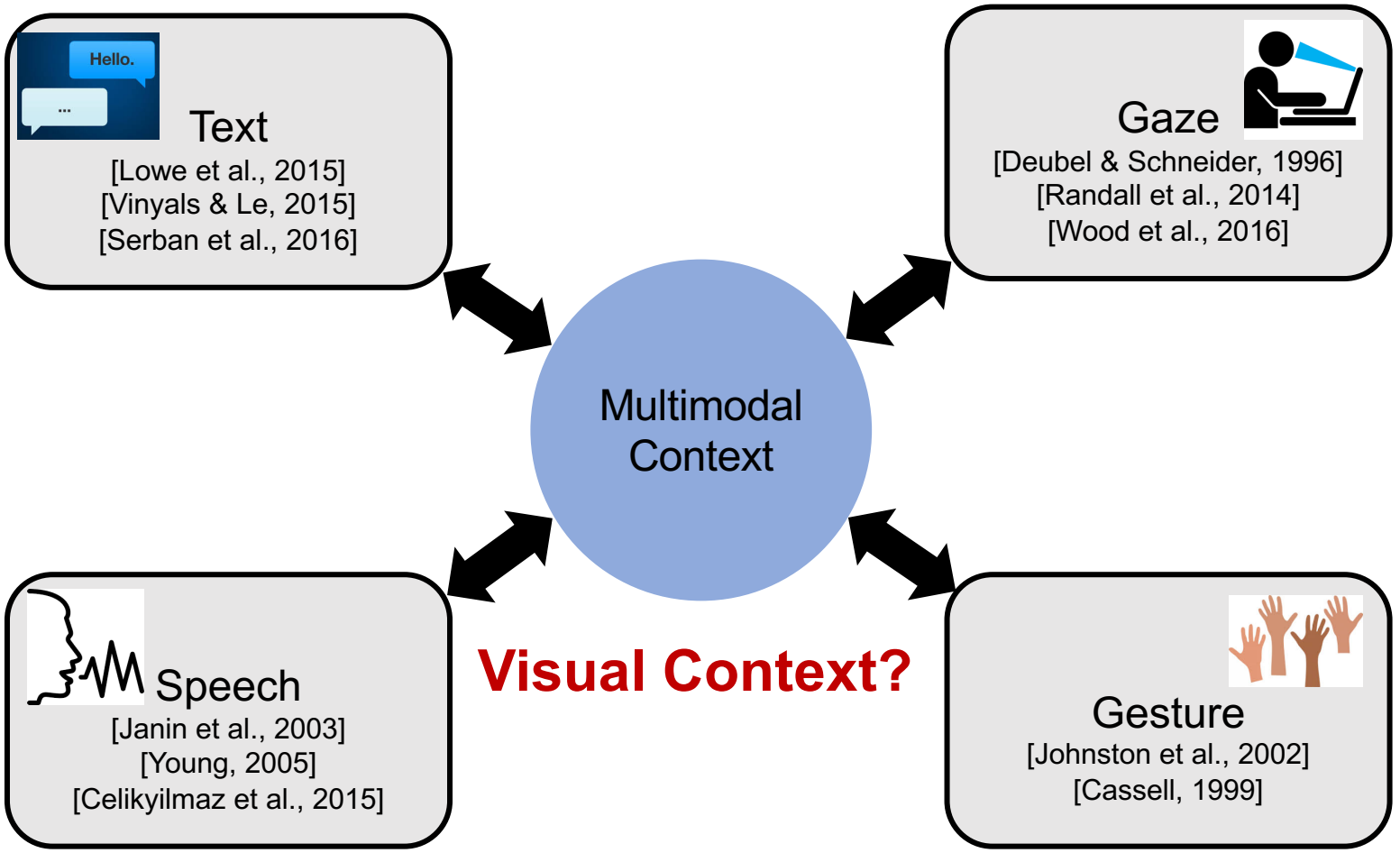




Dialogue Context



Dialogue Context



Visual Context

Image-based Context

[Das et al., 2017]

[Mostafazadeh et al., 2017]

Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

Is it a cow?	*Yes*
Is it the big cow in the middle?	*No*
Is the cow on the left?	*No*
On the right ?	*Yes*
First cow near us?	*Yes*

[De Vries et al., 2017]

[System]: How can i help you today ?
[User]: *Find non-fiction books by Chomsky.*
[System]: (Fetches the following books from database)

Hegemony or Survival 2003 | Manufacturing Consent:The Pol 1988 | Language and Thought 1993 | Power Systems: Conversations w 2013 | Aspects of the Theory of Syntax 1965

[User]: **"*show details for the oldest production*"** or
"*details for the syntax book*" or
"*open the last one*" or
"*i want to see the one on linguistics*" or
"*bring me Jurafsky's text book*"

[Celikyilmaz et al., 2014]

Visual Context

Image-based Context

[Das et al., 2017]

[Mostafazadeh et al., 2017]

Is it a person?	*No*
Is it an item being worn or held?	*Yes*
Is it a snowboard?	*Yes*
Is it the red one?	*No*
Is it the one being held by the person in blue?	*Yes*

Is it a cow?	*Yes*
Is it the big cow in the middle?	*No*
Is the cow on the left?	*No*
On the right ?	*Yes*
First cow near us?	*Yes*

[De Vries et al., 2017]

[System]: How can i help you today ?
[User]: *Find non-fiction books by Chomsky.*
[System]: (Fetches the following books from database)

Hegemony or Survival 2003 | Manufacturing Consent:The Pol 1988 | Language and Thought 1993 | Power Systems: Conversations w 2013 | Aspects of the Theory of Syntax 1965

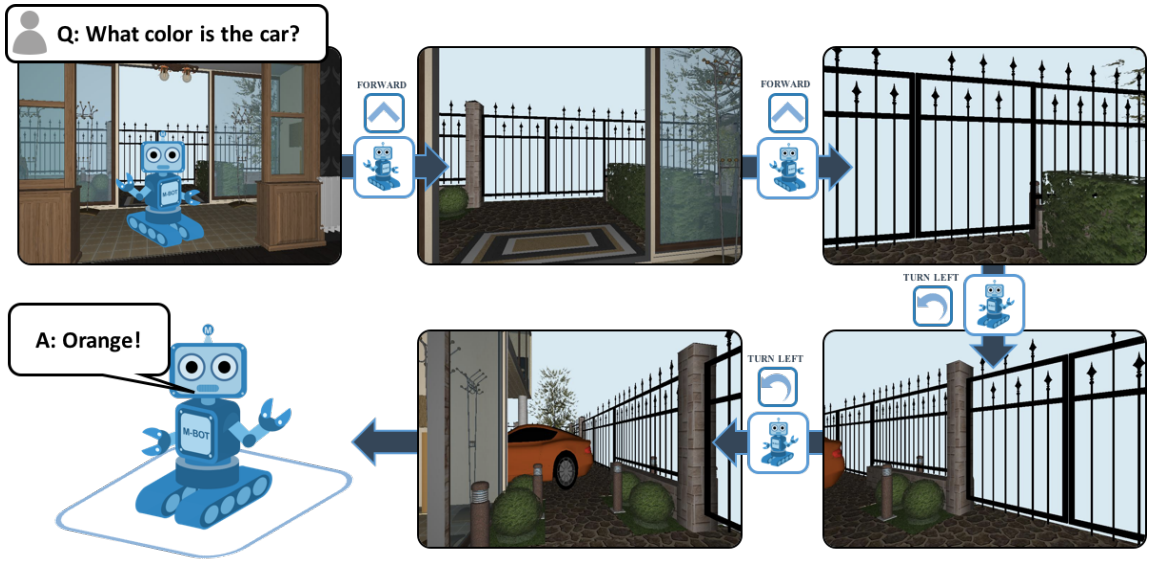
[User]: **"*show details for the oldest production*"** or
"*details for the syntax book*" or
"*open the last one*" or
"*i want to see the one on linguistics*" or
"*bring me Jurafsky's text book*"

[Celikyilmaz et al., 2014]

Dynamic Visual Context?

Visual Context

Dynamic-Visual Context



[Das et al., 2018]

Our Twitch-FIFA Dataset

2:36:07 L7Gasm : unsub

2:36:10 Flame_96 : then maybe ea would wake up and make a good game

2:36:19 melvin109 : !record

2:36:19 Moobot : 11 - 4

2:36:20 L7Gasm : JK i love u @InceptionXx

2:36:21 slake7 : Your mic is picking up a lot of static background noise, have you got mic boost turned on?

2:36:21 Anselm2 : yeah me too

2:36:22 Matt344 : @Flame_96 Imagine everyone PTB, most games on Champions

Our Twitch-FIFA Dataset

2:36:07 L7Gasm : unsub

2:36:10 Flame_96 : then maybe ea would wake up and make a good game

2:36:19 melvin109 : !record

2:36:19 Moobot : 11 - 4

2:36:20 L7Gasm : JK i love u @InceptionXx

2:36:21 slake7 : Your mic is picking up a lot of static background noise, have you got mic boost turned on?

2:36:21 Anselm2 : yeah me too

2:36:22 Matt344 : @Flame_96 Imagine everyone PTB, most games on Champions

Video + Chat based Context

Multiple speakers

Task

S1: what an offside trap OMEGALUL

S2: Lol that finish bro

S3: suprised you didn't do the extra pass

S4: @S10 a drunk bet? 

S5: @S11 thanks mate

S6: could have passed one more

S7: Pass that

S1: record now!

S8: !record

S9: done a nother pass there

The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right)

Task

S1: what an offside trap OMEGALUL

S2: Lol that finish bro

S3: suprised you didn't do the extra pass

S4: @S10 a drunk bet? 

S5: @S11 thanks mate

S6: could have passed one more

S7: Pass that

S1: record now!

S8: !record

S9: done a nother pass there

The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right)

Applications of Video-Grounded Dialogue

- Personal Assistants
- Intelligent tutors
- Human-robot Collaboration

Twitch-FIFA Dataset Collection

- To extract triples (instances) of video context, chat context, and response, we divide the videos based on the fixed time frames
  - 20-sec context windows to extract video clips and users utterances
  - Chat utterances in the next 10-sec window are potential responses

- We select the response that has at least some good coherence and relevance with the chat context's topic

Filtering Process

- Discourage frequent responses
- we choose the first (earliest) response that has high similarity with some other utterance in this response window (using 0.5 BLEU threshold, based on manual inspection)

	Relevance to Video+Chat
filtered response wins	34%
1st response wins	3%
Non-distinguishable	63% (56 both-good, 7 both-bad)

Human evaluation of our dataset, comparing our filtered responses versus the first response in the window (for relevance w.r.t. video and chat contexts)

Twitch-FIFA Dataset Statistics

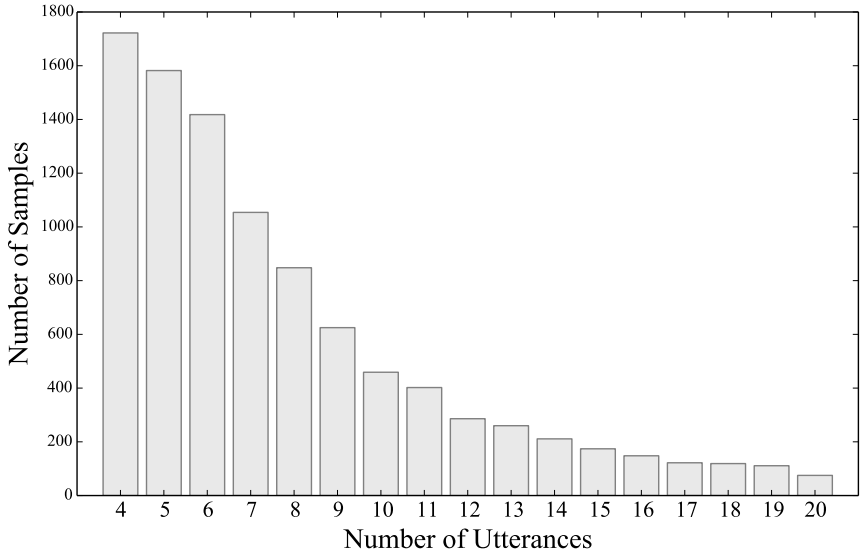
UNC NLP

Statistics	Train	Val	Test
#Videos	33	8	8
Total Hours	58.4	11.9	15.4
Final Filtered #Instances	10,510	2,153	2,780
Avg. Chat Context Length	69.0	63.5	71.2
Avg. Response Length	6.5	6.5	6.1

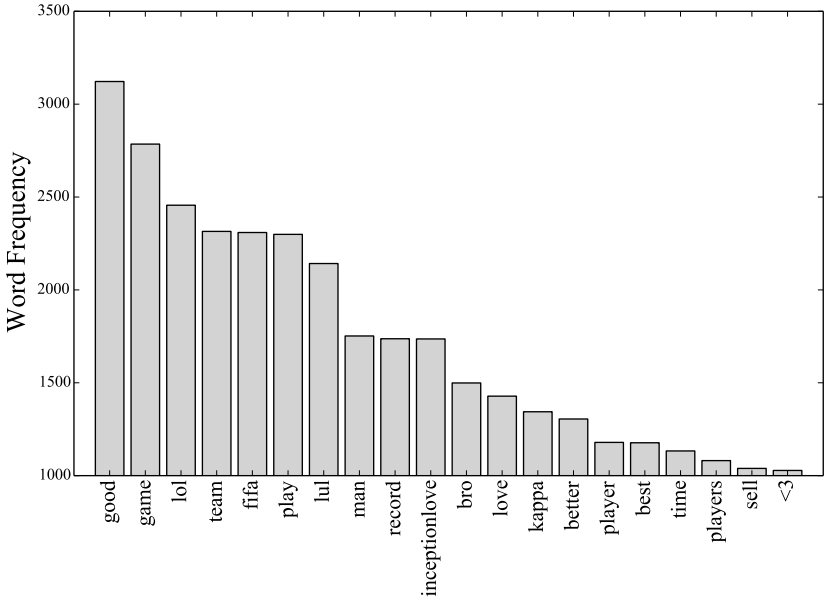
Twitch-FIFA dataset's chat statistics (lengths are defined in terms of number of words)

- Anonymized user identities

Dataset Statistics



Distribution of #utterances in chat context (w.r.t. the #training examples for each case)



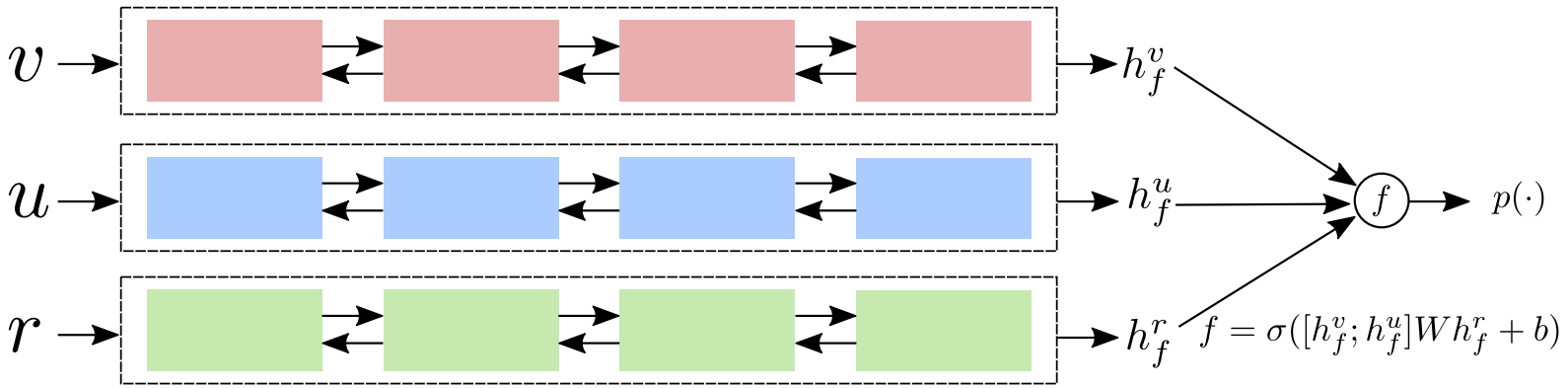
Frequent words in our Twitch-FIFA dataset

Models

- Discriminative Models

- Generative Models

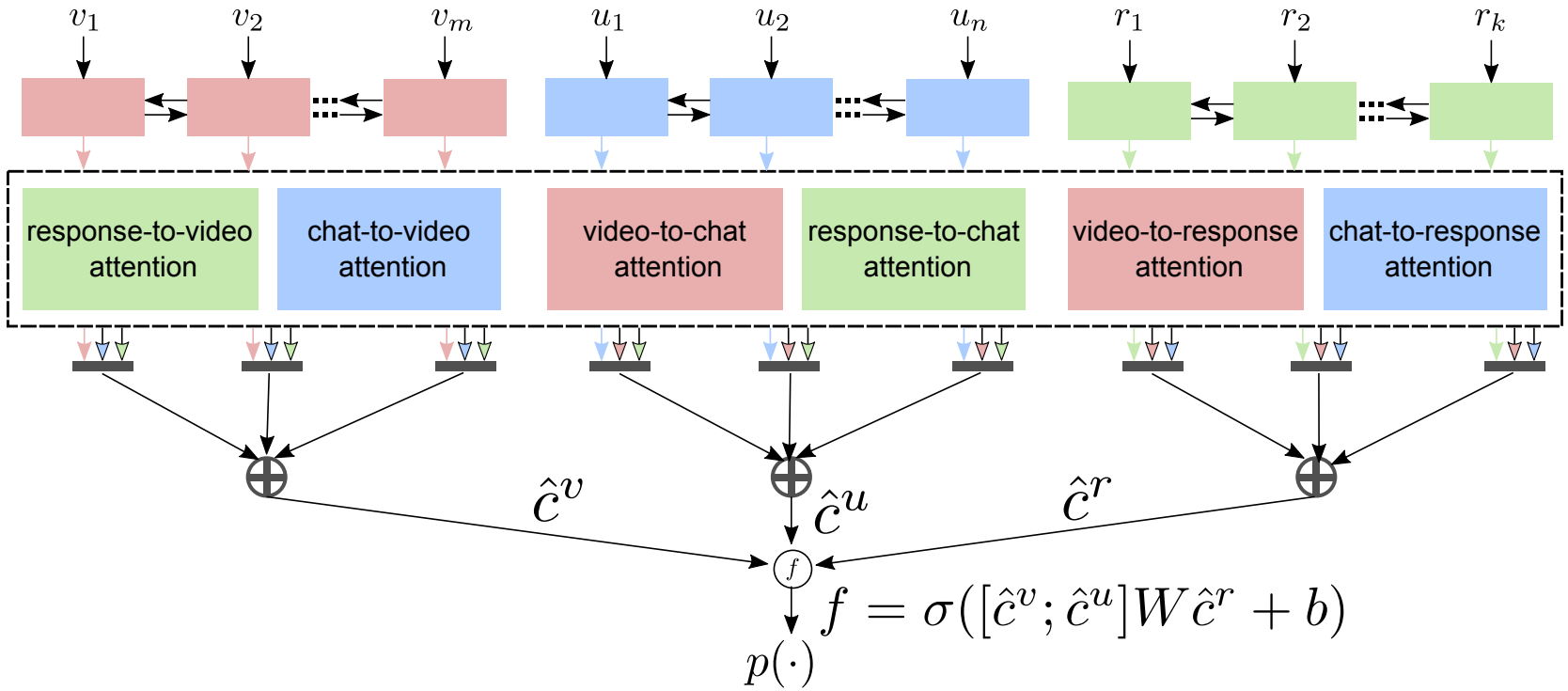
Discriminative Model



Our **Triple Encoder** discriminative model with bidirectional LSTM-RNN encoders for video, chat context, and response

[Lowe et al., 2015]

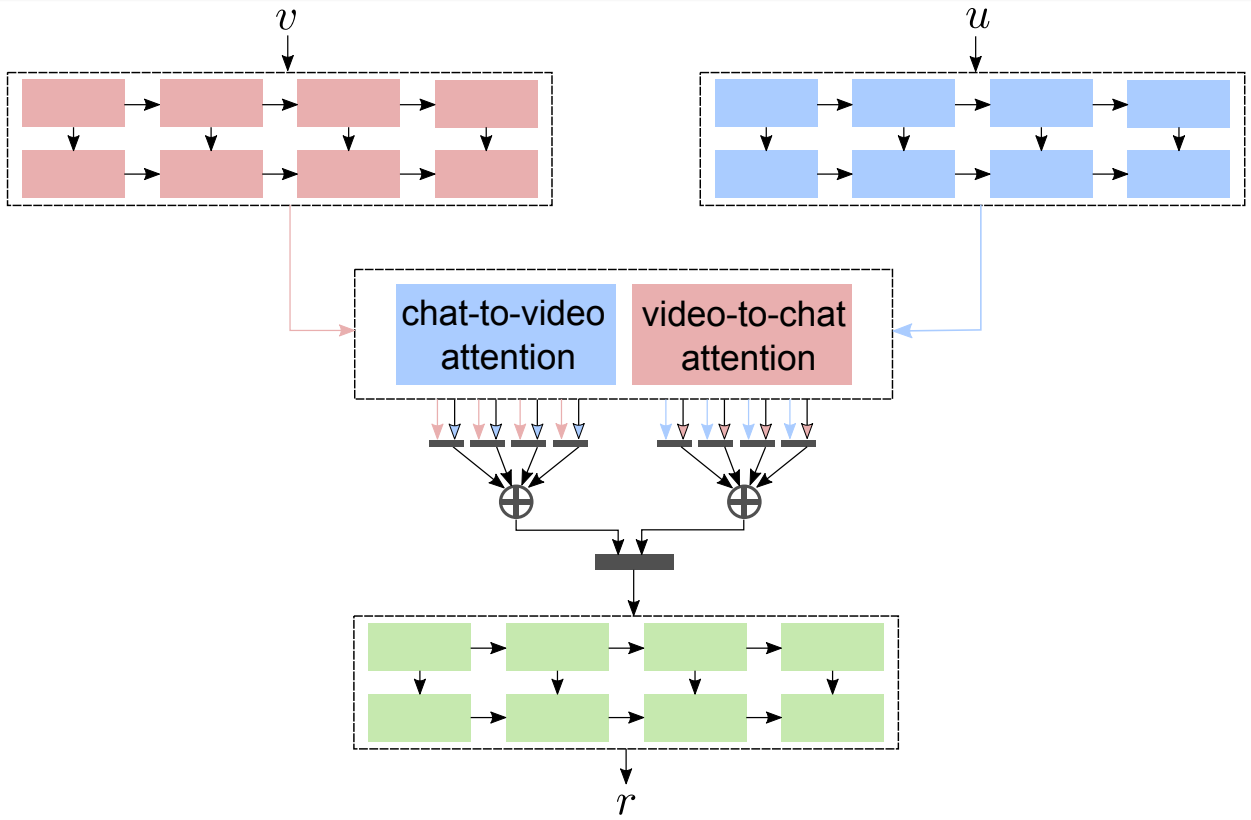
Discriminative Model



Our **Tri-Directional Attention Flow (TriDAF)** model with all pairwise modality attention modules, as well as self attention on video context, chat context, and response as inputs

[Seo et al., 2017; Lin et al., 2017]

Generative Model



Our **BiDAF-Generative** model with bidirectional attention flow between video context and chat context during response generation

[Seo et al., 2017; Luong et al., 2015]

Evaluation

- Retrieval-based recall@k scores
  - Discriminative models: re-rank responses (9 negative, 1 positive)
  - Generative models: re-rank based on log probability score of the generated response
- Phrase-matching metrics (Generative models)
  - METEOR
  - ROUGE
- Human Evaluation

Negative Samples

- Negative samples do not come from the video corresponding to positive response

- Training:
  - 3 random negative triples with only one modality being negative (for both discriminative and generative models)

- Testing/Validation:
  - 9 random negative responses (for recall@k eval)

Results

Models	r@1	r@2	r@5
BASELINES			
Most-Frequent-Response	10.0	16.0	20.9
Naive Bayes	9.6	20.9	51.5
Logistic Regression	10.8	21.8	52.5
Nearest Neighbor	11.4	22.6	53.2
Chat-Response-Cosine	11.4	22.0	53.2

Performance of our baselines, discriminative models, and generative models for recall@k metrics on our Twitch-FIFA test set. C and V represent chat and video context, respectively.

Results

Models	r@1	r@2	r@5
BASELINES			
Most-Frequent-Response	10.0	16.0	20.9
Naive Bayes	9.6	20.9	51.5
Logistic Regression	10.8	21.8	52.5
Nearest Neighbor	11.4	22.6	53.2
Chat-Response-Cosine	11.4	22.0	53.2
DISCRIMINATIVE MODEL			
Dual Encoder (C)	17.1	30.3	61.9
Dual Encoder (V)	16.3	30.5	61.1
Triple Encoder (C+V)	18.1	33.6	68.5
TriDAF+Self Attn (C+V)	20.7	35.3	69.4

Performance of our baselines, discriminative models, and generative models for recall@k metrics on our Twitch-FIFA test set. C and V represent chat and video context, respectively.

Results

Models	r@1	r@2	r@5
BASELINES			
Most-Frequent-Response	10.0	16.0	20.9
Naive Bayes	9.6	20.9	51.5
Logistic Regression	10.8	21.8	52.5
Nearest Neighbor	11.4	22.6	53.2
Chat-Response-Cosine	11.4	22.0	53.2
DISCRIMINATIVE MODEL			
Dual Encoder (C)	17.1	30.3	61.9
Dual Encoder (V)	16.3	30.5	61.1
Triple Encoder (C+V)	18.1	33.6	68.5
TriDAF+Self Attn (C+V)	20.7	35.3	69.4
GENERATIVE MODEL			
Seq2seq +Attn (C)	14.8	27.3	56.6
Seq2seq +Attn (V)	14.8	27.2	56.7
Seq2seq + Attn (C+V)	15.7	28.0	57.0
Seq2seq + Attn + BiDAF (C+V)	16.5	28.5	57.7

Performance of our baselines, discriminative models, and generative models for recall@k metrics on our Twitch-FIFA test set. C and V represent chat and video context, respectively.

Results

Models	METEOR	ROUGE-L
MULTIPLE REFERENCES		
Seq2seq + Atten. (C)	2.59	8.44
Seq2seq + Atten. (V)	2.66	8.34
Seq2seq + Atten. (C+V) ⊗	3.03	8.84
⊗ + BiDAF (C+V)	3.70	9.82

Performance of our generative models on phrase matching metrics

Human Evaluation

Models	Relevance
BiDAF wins	41.0 %
Seq2seq + Atten. (C+V) wins	34.0 %
Non-distinguishable	25.0 %

Human evaluation (two annotators with 50 task instances each) comparing the baseline and BiDAF generative models

Analysis: Negative training pairs

Models	recall@1	recall@2	recall@5
1 neg.	18.21	32.19	64.05
3 neg.	22.20	35.90	68.09

Ablation (dev) of one vs. three negative examples for TriDAF self-attention discriminative model

Analysis: Loss functions

Models	recall@1	recall@2	recall@5
Classification loss	19.32	33.72	66.60
Max-margin loss	22.20	35.90	68.09

Ablation of classification vs. max-margin loss on our TriDAF discriminative model (on dev)

Models	recall@1	recall@2	recall@5
Cross-entropy (XE)	13.12	23.45	54.78
XE+Max-margin	15.61	27.39	57.02

Ablation of cross-entropy loss vs. cross-entropy+maxmargin loss for our BiDAF-based generative model (on dev)

Discriminative Output Example



bloodtrail bloodtrail bloodtrail bloodtrail bloodtrail || yoooo || kappapride || xxuxx skillzzzz , favourite player you have used this year ? || pl3ad aa9love || are you playin with ksi ? ? kappa xxuxx || bought okocha cuz of you ant . first game 2 goals 3 assists ! game changer thank you m8 || play || ! pause || resume || twerkchoke twerkchoke twerkchoke || lul

1) good pass jebaited
2) shawn mendez kreygasm kreygasm
3) can say that i am american
4) ! camera
5) can you notice me
6) do you have a main squad
7) otw nelson for 47k imma buy right now on xbox
8) do *
9) inceptionderp inceptionlove
10) bpl is over priced

Output retrieval example from TriDAF model

Generative Output Example

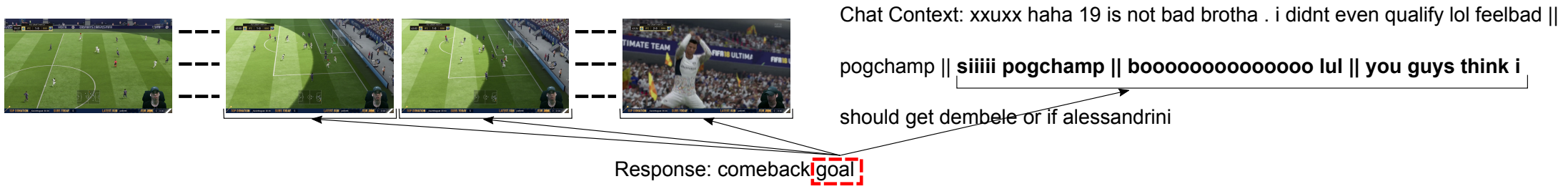
chat is aids || where has all thr challenges gone aswell ? || did mat yet messi ? || hellllllllllllllllllllllllllllllllllllllllllllo || put messi on get in behind if u can || chris is getting ronaldo and messi || no one wants jamies coctail sausage haha || free kick with messi

Ground-truth: play it to messi he makes good runs

Generated: get messi for the other team

Output generative example from BiDAF model

Attention Visualization



Attention visualization: generated word ‘goal’ in response is intuitively aligning to goal-related video frames (top-3-weight frames highlighted) and context words (top-10-weight words highlighted)

Thanks!

Data/code available at https://github.com/ramakanth-pasunuru/video-dialogue

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNC NLP

Acknowledgment: DARPA YFA17-D17AP00022, ARO-YIP W911NF-18-1-0336, Google, Bloomberg, NVidia