Visual Discourse Parsing

Arjun Akula, Song-Chun Zhu

Center for Vision, Cognition, Learning, and Autonomy,

University of California, Los Angeles (UCLA)

June 16 2019

CVPR 2019 Workshop on Language and Vision

LONG BEACH
CALIFORNIA
June 16-20, 2019

Visual Discourse Parsing

Arjun Akula, Song-Chun Zhu

Center for Vision, Cognition, Learning, and Autonomy,

University of California, Los Angeles (UCLA)

June 16 2019

CVPR 2019 Workshop on Language and Vision

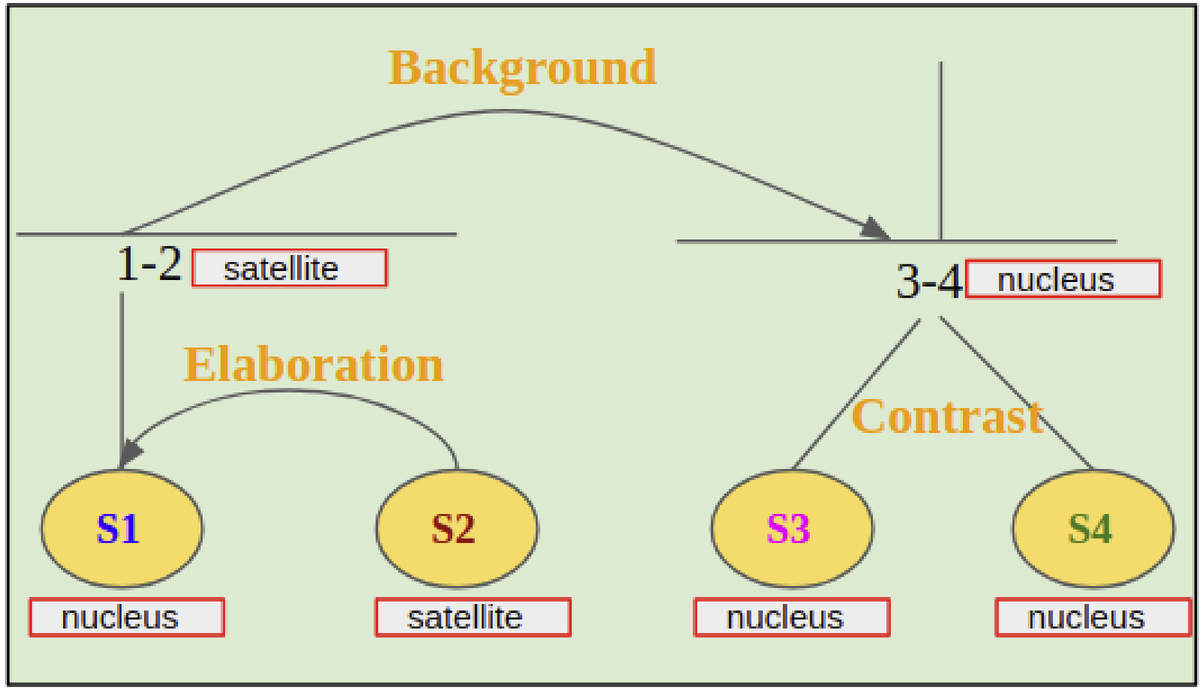
Text-Level Discourse Parsing

Text-level discourse parsing aims to unmask how two segments (or sentences) in the text are related to each other.

Text

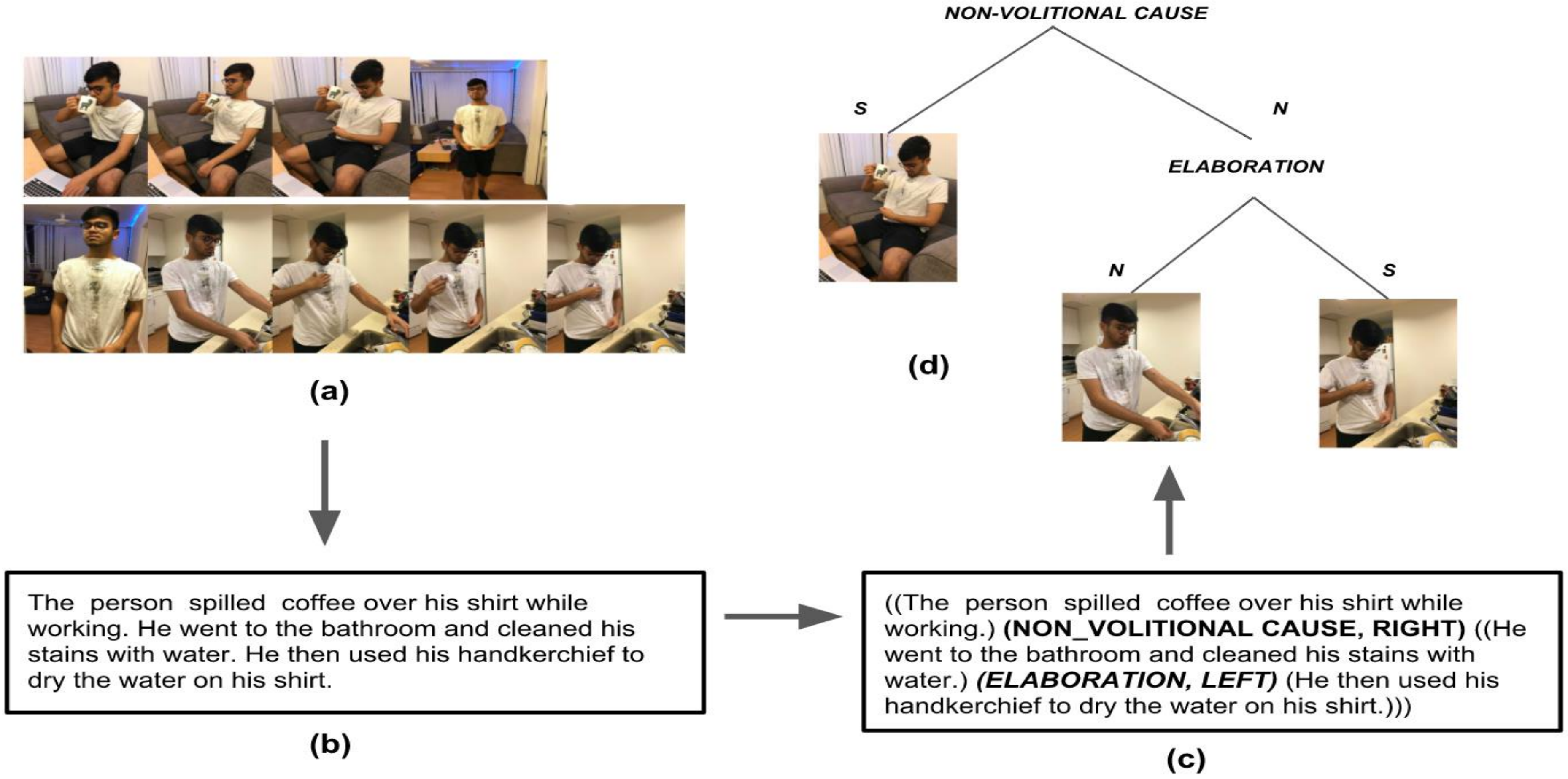
Lactose is a milk sugar (S1). The enzyme Lactase breaks it down (S2). For want of Lactase most adults cannot digest milk (S3). In populations that drink milk the adults have more lactase perhaps through natural selection (S4).

Discourse Parse (RST)



Visual Discourse Parsing (VDP)

We introduce a new AI task - Visual Discourse Parsing, where the AI agent needs to understand discourse relations among the scenes in a video.



The person spilled coffee over his shirt while working. He went to the bathroom and cleaned his stains with water. He then used his handkerchief to dry the water on his shirt.

((The person spilled coffee over his shirt while working.) **(NON_VOLITIONAL CAUSE, RIGHT)** ((He went to the bathroom and cleaned his stains with water.) ***(ELABORATION, LEFT)*** (He then used his handkerchief to dry the water on his shirt.)))

Specifically, given a video, the task is to identify a scene's relation with the context.

Applications of VDP

- Video Summarization.

- Video Paragraph Captioning: aids in generating coherent paragraph descriptions of videos.

- Visual Sentiment Analysis.

- Visual Dialog and Visual Story-telling.

VDP Data Collection

- Developed a VDP dataset containing 310 videos.
- Shot at various places in UCLA such as playing sports (Table Tennis, Frisbee, Tennis, Rugby), bus stop, dining hall, elevator, class-room, library, garden and study room.
- On average, the length of each video is about 19 seconds.
- We first manually generated descriptions of each video and then annotated the discourse structure of these descriptions - with the help of 6 graduate students.
- Each video is annotated by at least 2 students.

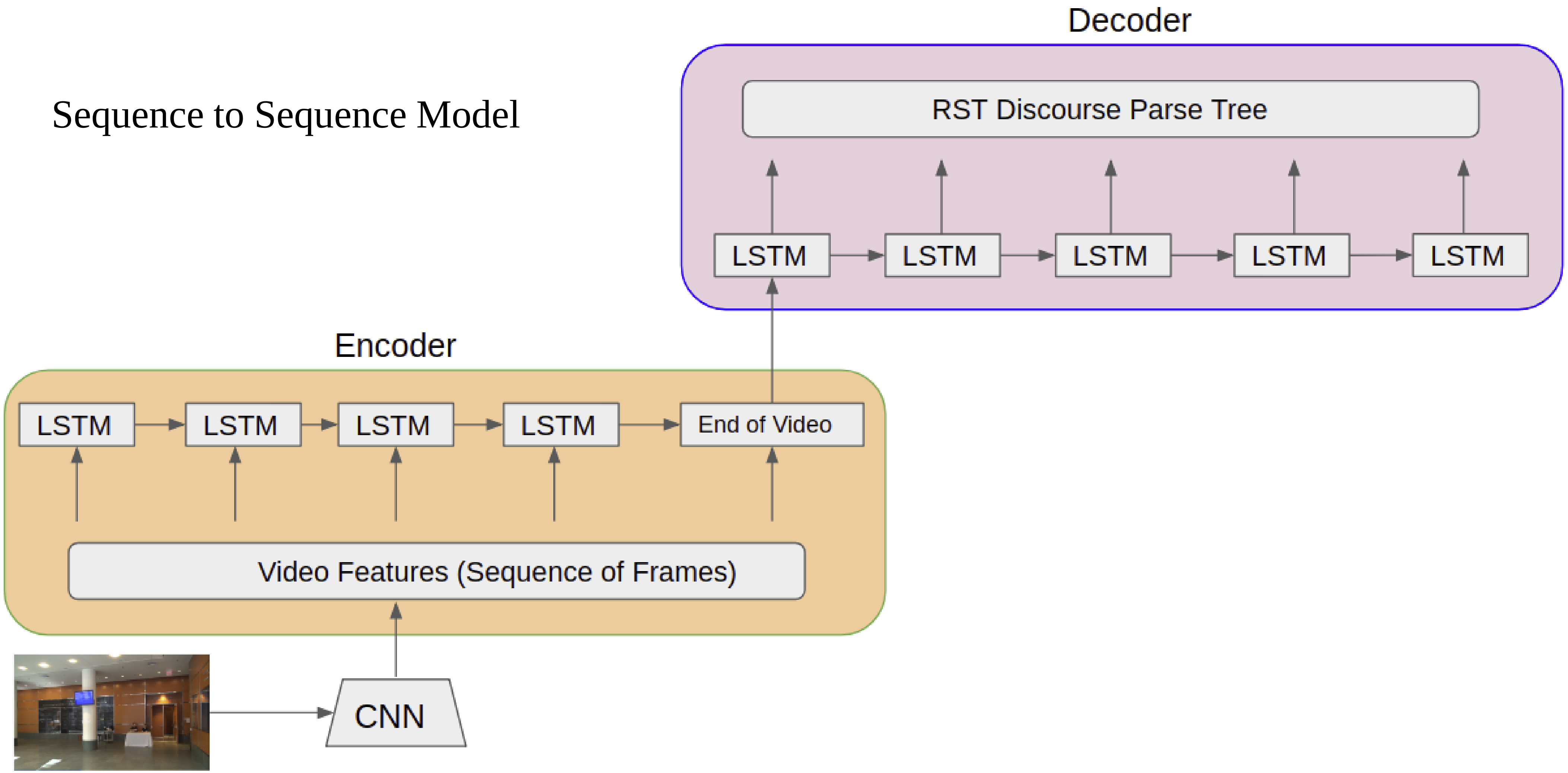
Example

Video Description: The person was watching a video on Youtube. Another person came in to complain about the noise. The person then put on his headphone.

RST annotation: <RIGHT_CONFIG> S1 <background> <RIGHT_DIR> S2 <cause> <RIGHT_DIR> S3

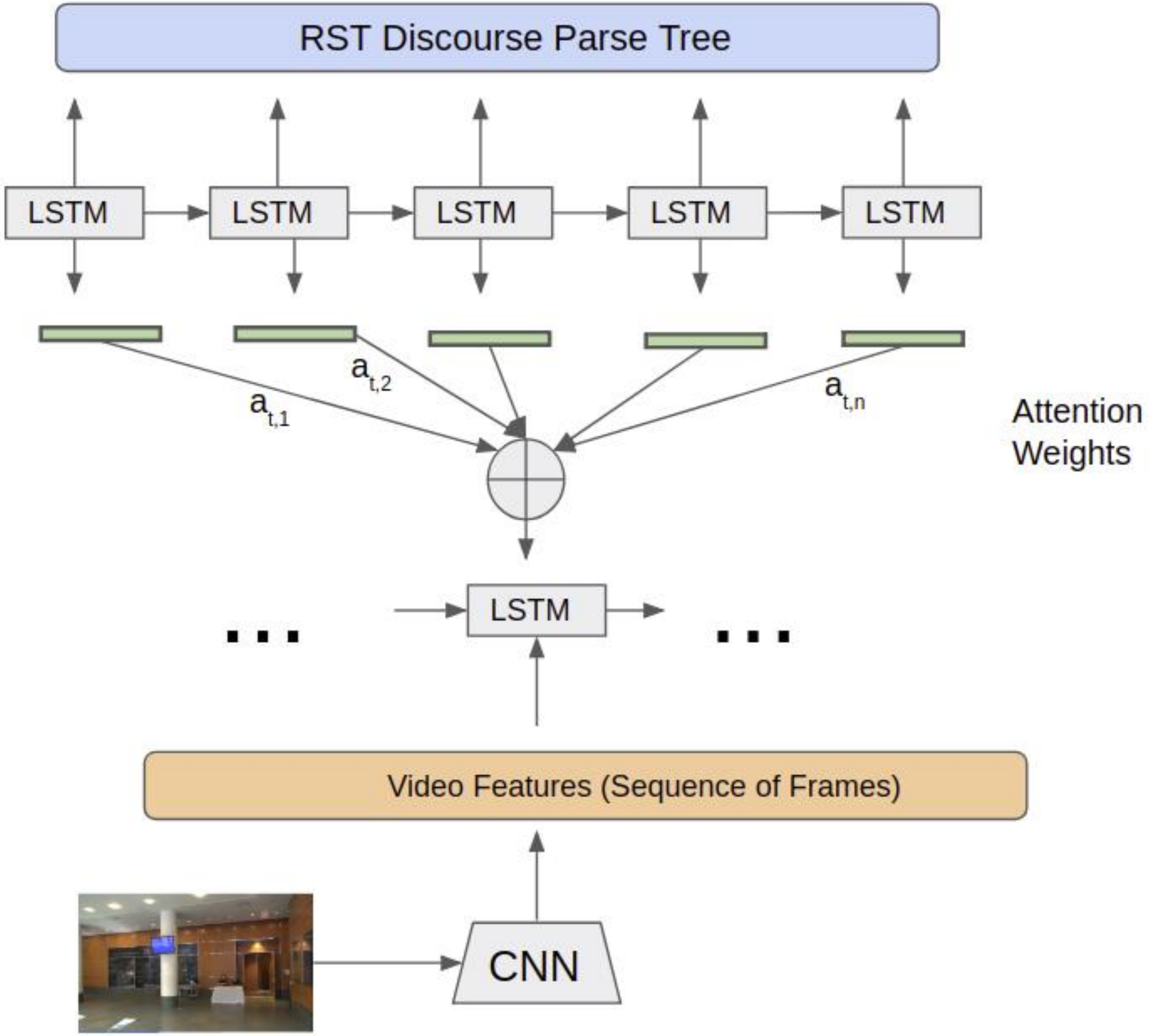
Baseline Models

Sequence to Sequence Model



Baseline Models

Sequence to Sequence Model + Soft Attention



Evaluation Metrics

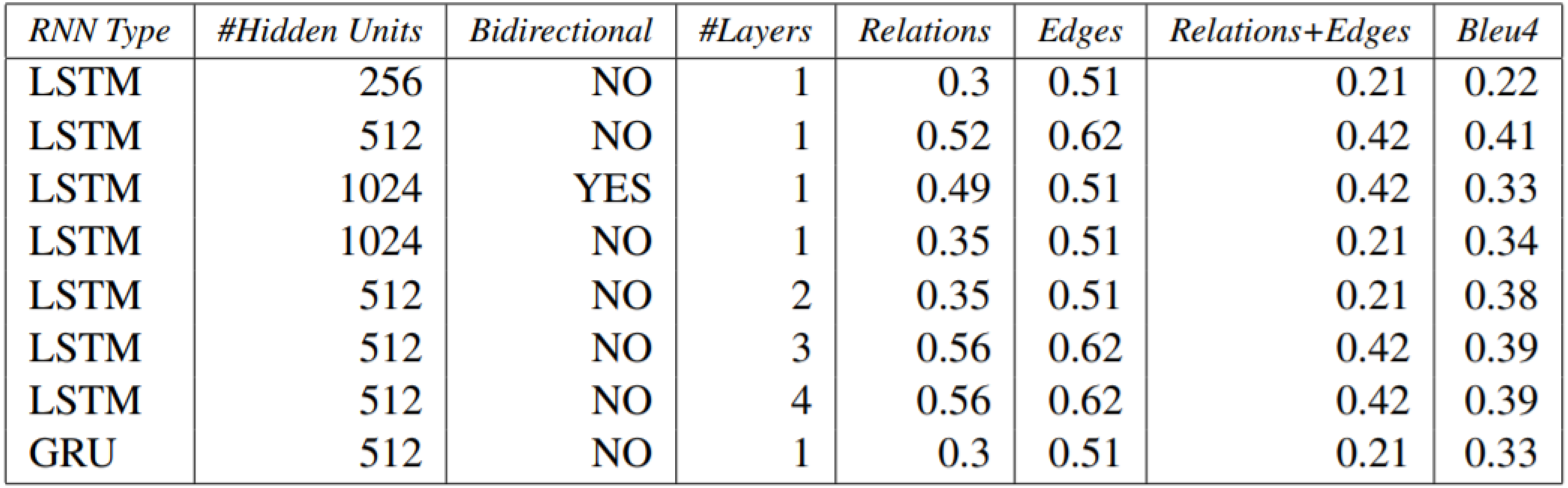
- **BLEU score:** to evaluate the translation quality of the discourse structure generated from the videos.

- **Relations Accuracy:** total number of relations correctly predicted by the model.

- **Edges Accuracy:** total number of edges (i.e. RST node nuclearity directions) correctly predicted by the model.

- **Relations+Edges Accuracy**: the predicted discourse structure will be considered correct only if all the relations and the edges are correctly predicted by the model.

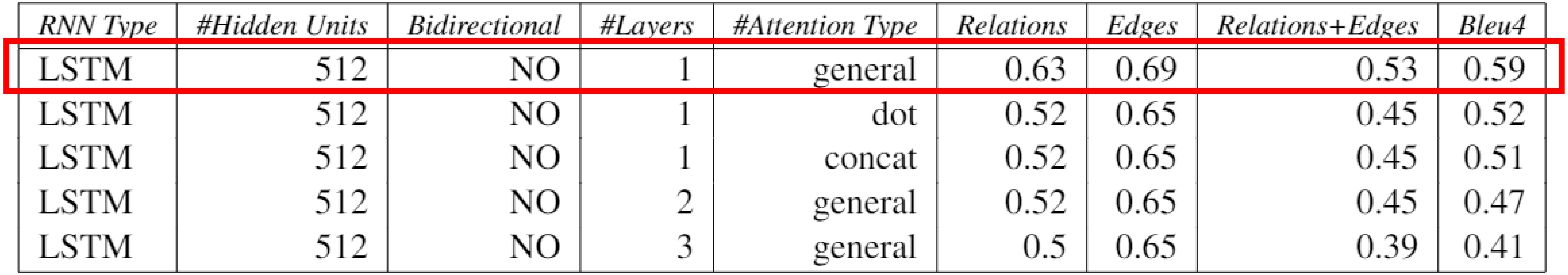
Results



RNN Type	*#Hidden Units*	*Bidirectional*	*#Layers*	*Relations*	*Edges*	*Relations+Edges*	*Bleu4*
LSTM	256	NO	1	0.3	0.51	0.21	0.22
LSTM	512	NO	1	0.52	0.62	0.42	0.41
LSTM	1024	YES	1	0.49	0.51	0.42	0.33
LSTM	1024	NO	1	0.35	0.51	0.21	0.34
LSTM	512	NO	2	0.35	0.51	0.21	0.38
LSTM	512	NO	3	0.56	0.62	0.42	0.39
LSTM	512	NO	4	0.56	0.62	0.42	0.39
GRU	512	NO	1	0.3	0.51	0.21	0.33

Evaluation using sequence-to-sequence model without Attention.

Results



RNN Type	*#Hidden Units*	*Bidirectional*	*#Layers*	*#Attention Type*	*Relations*	*Edges*	*Relations+Edges*	*Bleu4*
LSTM	512	NO	1	general	0.63	0.69	0.53	0.59
LSTM	512	NO	1	dot	0.52	0.65	0.45	0.52
LSTM	512	NO	1	concat	0.52	0.65	0.45	0.51
LSTM	512	NO	2	general	0.52	0.65	0.45	0.47
LSTM	512	NO	3	general	0.5	0.65	0.39	0.41

Evaluation using sequence-to-sequence model with Soft Attention.

Qualitative Analysis

	Target Prediction	Output Prediction
Okay	<left_config> the person was eating some rice <sequence> <bi_dir> he poured some soy sauce on the rice <sequence> <bi_dir> he continued eating the rice	<left_config> the person eating <sequence> <bi_dir> he then some sauce <sequence> <bi_dir> he continued eating
Okay	<left_config> a person was throwing a frisbee high up <background> <right_dir> he caught it himself <sequence> <bi_dir> he threw it again	<left_config> the person was throwing a frisbee <background> <right_dir> he ran it up a frisbee <sequence> <bi_dir> he threw again
Not bad	<left_config> the person was looking at the corner <background> <right_dir> another person showed up <justify> <right_dir> they greeted each other and sat together	<left_config> the person was throwing a frisbee <background> <right_dir> another person came and a seat and him to the person
worst	<right_config> the person was drinking up his juice <sequence> <bi_dir> he went to get a cup of coffee <elaboration> <left_dir> he put some sugar into it and went back to his seat	<left_config> the person was looking at the computer <background> <right_dir> he put to to to to to to to to to to to to

Ongoing Efforts and Challenges

- Annotation Aggreement

- Scaling up our annotations by leveraging Existing Datasets
  - ActivityNet which enforced coherence while collecting from humans.