VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

NII_UIT at TRECVID 2022: Movie Summarization Task

Nam Nguyen[1, 2], Tien Hung Nguyen[1, 2], Cong Nguyen Thanh[1, 2], Hao Vo[1, 2], Khiem Le[1, 2], Tien Do[1, 2] , Tien-Dung Mai[1, 2] , An Pham Nguyen Truong[1, 2] , **Duy-Dinh Le**[1, 2] , Shin'ichi Satoh[3]

[1] *University of Information Technology (UIT), Ho Chi Minh City, Vietnam*
[2] *Vietnam National University, Ho Chi Minh City (VNU-HCM), Vietnam*
[3] *National Institute of Informatics (NII), Tokyo, Japan*

TRECVID Workshop 2022

Overview

- MSUM Task Introduction
- Challenges
- Our Approach
- Experimental Results
- Conclusion

MSUM Task Introduction

Movie Summarization Task

Summary System

Key-facts events of character Jeremy

- Charlie bullies Jeremy
- Charlie and Jeremy fight at the playground
- ...
- Jeremy passes away

Input
A movie and face queries

Output
Video summary and textual summary which include key-facts events

Collect videos and generate descriptions about critical and important information of character storyline.

Movie Summarization Task

- Goal: reduce the size concentrate the amount of high value information in the video track
  - VSUM: summarize the major life events of specific characters over a number of weeks of programming on the BBC Eastenders TV series
  - MSUM: summarize the storylines and roles of specific characters during a full movie
- MSUM Goal:
  - Efficiently capture important facts about certain persons during their role in the movie storyline.
  - Assess how well video summarization and textual summarization compare in this domain.

Movie Summarization Task

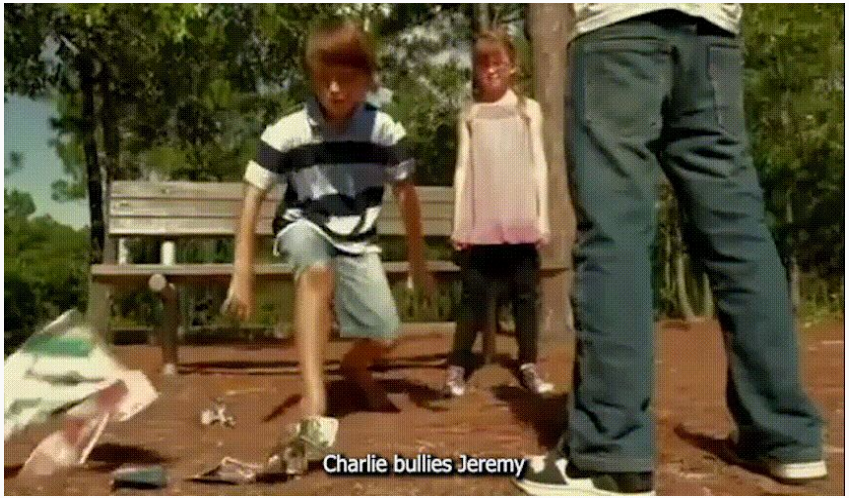
- Video Summary
  - Input:
    - a movie, a character, and image/video examples of that character
  - Output:
    - a video summary highlighting major key-fact events about the character (similar to TV20 & TV21 VSUM)
  - Video summaries will be limited by a maximum summary length
- Text Summary
  - Input:
    - a movie, a character, and image/video examples of that character
  - Output:
    - a textual summary to include key-fact events about the character role in the movie
  - Textual summaries will be limited by a maximum number of sentences and a maximum number of words

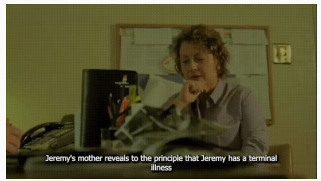
What is a Key-fact Event

- Any events that are important and critical in the character storyline.
- They should cover his/her role from the start to the end of the movie.
- Example : From the example movie “Super - Hero” – Character: Jeremy
  - *Charlie bullies Jeremy*
  - *Charlie and Jeremy fight at the playground*
  - *Jeremy's mother reveals to the principle that Jeremy has a terminal illness*
  - *Jeremy gets admitted to the hospital*
  - *Jeremy passes away*
- Key events should appear in the order in which they become apparent in the movie, and should ideally capture that characters storyline.

What is a Key-fact Event

What is a Key-fact Event





What is a Life Event (TV20 & TV21 VSUM)

Important content

- ***Major life events*** of a given character:
  - the birth of a child (rather than a short illness),
  - a divorce (rather than an argument with a loved one)
- How they ***intertwine*** with the major life events of ***other specified characters***

Input

Summary System

What happens when Janine attempts to play recording of Stacey?

What is causing Ryan to be sick in bed?

Who stabbed Janine?

Who gives Janine the recording of Stacey?

How does Janine attempt to kill Ryan while in the hospital?

Output

Annotation and Assessment

- Human annotators will:
  - Watch each movie
  - For selected characters, extract key-fact events about them
- Video Summary evaluation:
  - Assessors will watch submitted summaries (subject to max duration)
  - Systems are rewarded for including the key-fact events
  - Scoring is based on the percentage of correct key-facts included in the summaries
  - Subjective evaluation will also be conducted (contextuality, redundancy, etc.)
- Textual Summary evaluation:
  - Systems will submit a summary of up to X sentences and Y words
  - Assessors will read the submitted textual summary and mark correctly retrieved key-facts
  - Objective evaluation of retrieved key-facts (regardless of any filler sentences)
  - Subjective evaluation will also be conducted (readability, contextuality, redundancy, etc.)

Dataset

- Dataset
  - 10 full movies from a licensed movie dataset from Kinolorberedu.
- Topics (Characters to Summarize):
  - Each topic will consist of a movie, the character to summarise the key-fact events for, and a set of image/video examples of that character.
  - For video summaries, a max summary time (in seconds) will be specified for each character.
  - While for text summaries, the max sentences limit will be specified for each character as well. A sentence for text summary can be either a key-fact (the focus of the task), or a filler sentence. The max sentences a run can submit for a given character will include all key-facts and filler sentences.

Dataset

type	video_name	time_of_movie *(hh:mm:ss)*	# scene	time_of_scene *(s)*		
				min	max	avg
train	Calloused_Hands	1:37:16	**65**	20	247	96
	Liberty_Kid	1:31:42	56	**12**	299	**94**
	Llike_me	1:23:56	28	47	300	**167**
	losing_ground	1:25:38	40	29	246	120
	Memphis	**1:18:39**	47	17	294	97
test	Archipelago	**1:50:04**	57	21	389	113
	Bonneville	1:32:39	41	19	269	124
	ChainedforLife	1:29:28	38	15	370	136
	heart_machine	1:23:37	**28**	22	**451**	158
	Little_Rock	1:22:48	39	24	289	121
Mean		1:29:35	43.9	22.6	315.4	122.6
Total		14:55:47	439	226	3154	1226

Dataset - Training Set

- A key-fact for every 3-4 mins

video_name.character	time_of_movie *(hh:mm:ss)*	# scene	time_of_scene *(s)*			# key-facts
			min	max	avg	
Calloused_Hands.Byrd	**1:37:16**	**65**	20	247	96	**35**
Calloused_Hands.Debbie						24
Liberty_Kid.DERRICK	1:31:42	56	**12**	299	**94**	27
Llike_me.Burt_Walden	1:23:56	**28**	47	**300**	**167**	**4**
Like_me.Kiya						12
losing_ground.Sarah_Rogers	1:25:38	40	29	246	120	15
Memphis.willis	**1:18:39**	47	17	294	97	13
Mean	1:27:26	47.2	25	277.2	114.8	26

Submission - Video Result

```
<!--    Example movie summarization video results for a msum run     -->
<MovieSummarizationVideoResults>
 <MovieSummarizationVideoRunResult pid="SiriusCyberCo" priority="2" desc="This automatic run uses algorithm 1">
  <MovieSummarizationVideoTopicResult target="Jeremy" movie="Super Hero" length="25">
    <item seqNum="1" startTime="0" endTime="4"/>
    <item seqNum="2" startTime="4" endTime="9"/>
    <item seqNum="3" startTime="9" endTime="15"/>
    <item seqNum="4" startTime="15" endTime="21"/>
    <item seqNum="5" startTime="21" endTime="25"/>
  </MovieSummarizationVideoTopicResult>
  <!--    ...    -->
  <MovieSummarizationVideoTopicResult target="Jeremy's Mother" movie="Super Hero" length="19">
    <item seqNum="1" startTime="0" endTime="5"/>
    <item seqNum="2" startTime="5" endTime="9"/>
    <item seqNum="3" startTime="9" endTime="13"/>
    <item seqNum="4" startTime="13" endTime="19"/>
  </MovieSummarizationVideoTopicResult>
 </MovieSummarizationVideoRunResult>
</MovieSummarizationVideoResults>
```

Submission - Text Result

```
<!--    Example movie summarization text results for a msum run     -->
<MovieSummarizationTextResults>
 <MovieSummarizationTextRunResult pid="SiriusCyberCo" priority="2" desc="This automatic run uses algorithm 1">
  <MovieSummarizationTextTopicResult target="Jeremy" movie="Super Hero" numLines="5">
   <item seqNum="1" type="K" text="Charlie bullies Jeremy"/>
   <item seqNum="2" type="K" text="Charlie and Jeremy fight at the playground"/>
   <item seqNum="3" type="K" text="Jeremy's mother reveals to the principle that Jeremy has a terminal illness"/>
   <item seqNum="4" type="F" text="Jeremy gets admitted to the hospital"/>
   <item seqNum="5" type="K" text="Jeremy passes away"/>
  </MovieSummarizationTextTopicResult>
  <!--    ...    -->
  <MovieSummarizationTextTopicResult target="Jeremy's Mother" movie="Super Hero" numLines="4">
   <item seqNum="1" type="F" text="Jeremy's Mother reads a bedtime story to Jeremy."/>
   <item seqNum="2" type="K" text="Jeremy's mother reveals to the principle that Jeremy has a terminal illness."/>
   <item seqNum="3" type="F" text="Jeremy's mother awaits anxiously in the hospital."/>
   <item seqNum="4" type="F" text="Jeremy's mother cries after Jeremy passes away"/>
  </MovieSummarizationTextTopicResult>
 </MovieSummarizationTextRunResult>
</MovieSummarizationTextResults>
```

Challenges

Key-fact Event Representation

- *'Daryl broke up with his girlfriend over breakfast'* is more likely to be a major key fact than *'Daryl had eggs and toast for breakfast'*
- A key-fact event regarding a character does not necessarily require that character to be visible in the scene.
    - Example: Jeremy's mother revealed to the principle that Jeremy had a terminal illness. This would clearly count as key-fact regarding *Jeremy even though he was not present in the scene*.

Key-facts of Jeremy character

- Charlie **bullies** Jeremy
- Charlie and Jeremy **fight at the playground**
- Jeremy's mother **reveals** to the principle that **Jeremy has a terminal illness**
- Jeremy **gets admitted to the hospital**
- Jeremy **passes away**

Audio Information

- Hard to clearly generate text from character's dialogue because the audio in a movie may contain music and sound effects.
- Transcript may not contain the information about character.
- It is difficult to determine whether a generated text is a key-fact or not

A shot video of Memphis movie

Keyfacts event (GT)

Willis is on a TV show called Memphis Sounds and declares that he imagined his success into existence because he is a wizard.

Transcript generated by audio

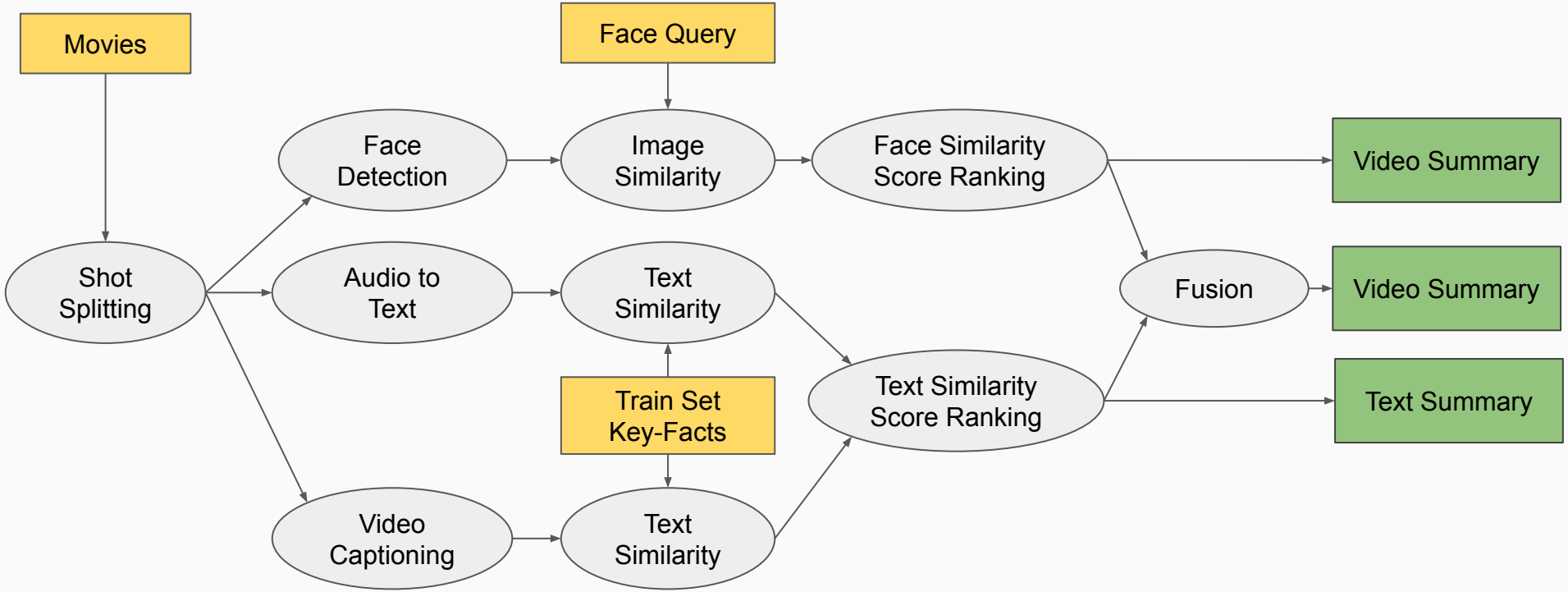
"did you ever think that you'd have an abo going to be a motion picture yes i did you did very often ah and in very many ways i created this i imagined it into existence the thing about sorcery which is exactly what that was cause i consider myself to be a wizard is that you create a reality that you invision you use magic and then the magic comes true but the magic doesn't fulfil everything you thought it would fulfil"

Our Approach

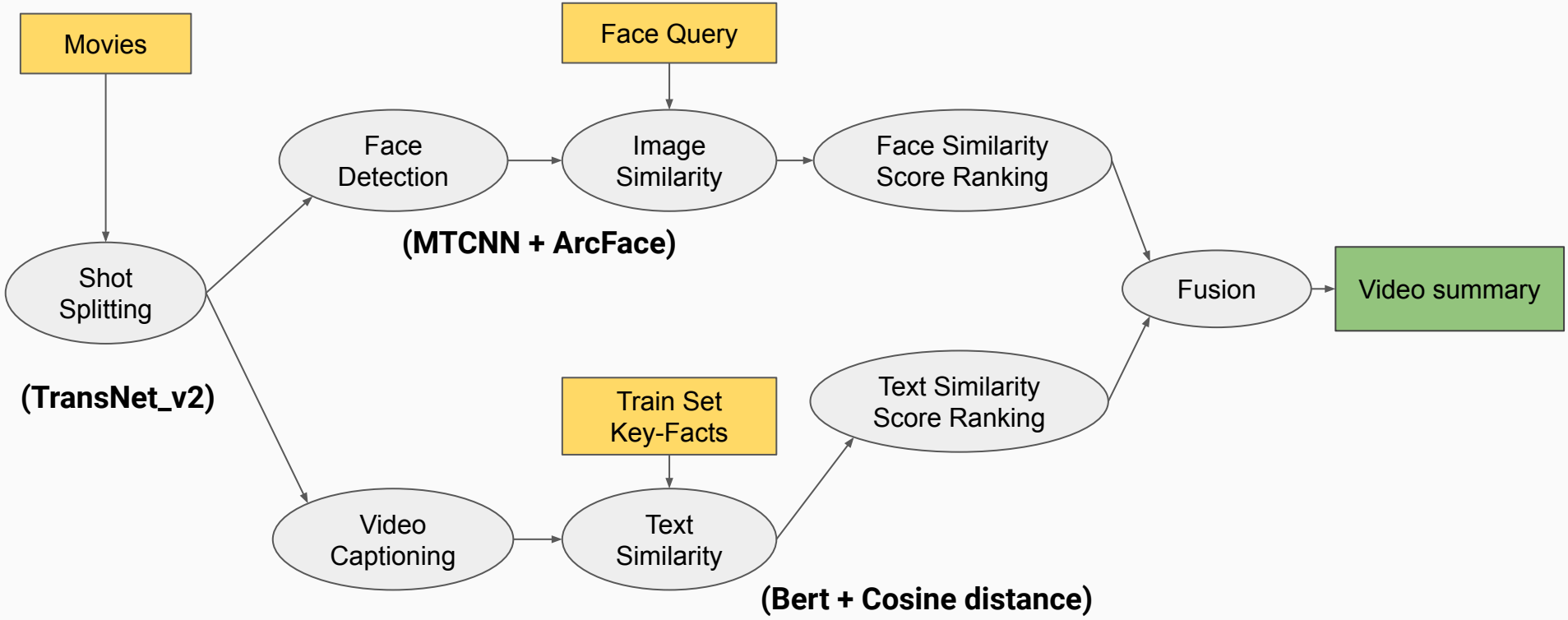
Problems - Proposed Methods

- Shot segment:
  - Problem: How to detect and filter long and unnecessary shot?
  - Solution:
    - Shot detection: TransNetV2.
- Face search:
  - Problem: How to detect and compare similarity between faces?
  - Solution:
    - Face Detection: MTCNN
    - Face Similarity: ArcFace
- Text search:
  - Problem: How to generate text from visual and audio contents
  - Solution:
    - Audio to Text
    - Video Captioning
- Fusion → finding appropriate weights

Overall Pipeline



Fusion of Face Recognition and Text



Experimental Results

Video Summarization

Submitted Runs with Fusion Weights

$$score = \begin{bmatrix} face_{score} \\ audio_{score} \\ caption_{score} \\ 1 - (time_{shot}/time_{limit}) \end{bmatrix} \times \begin{bmatrix} w_{face} & w_{audio} & w_{caption} & w_{time} \end{bmatrix}$$

- **RUN1**: Only Face Recognition.
- **RUN2**: Face Recognition + Video Captioning.
- **RUN3**: Face Recognition + Audio2Text.
- **RUN4**: Face Recognition + Audio2Text + Video Captioning

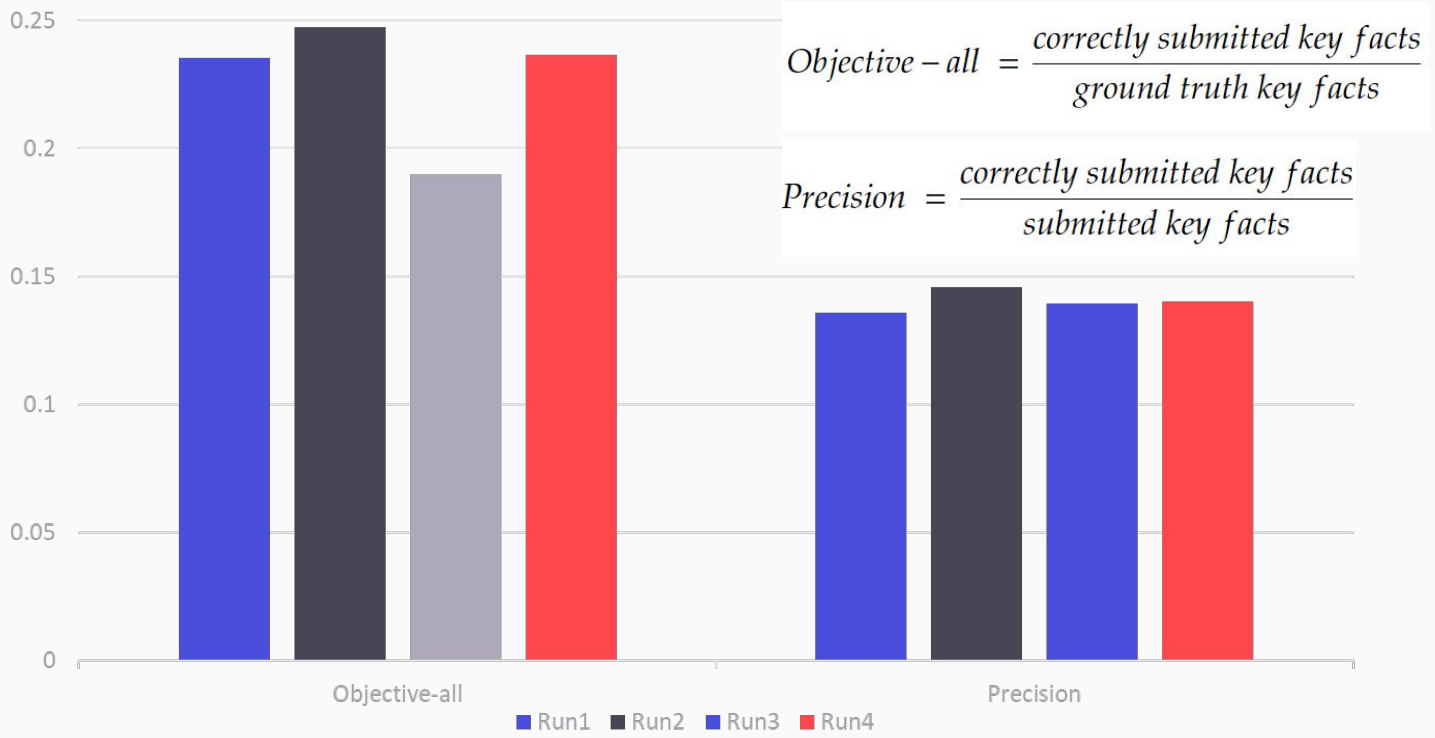
$$W_{run01} = [0.7 \quad 0.0 \quad 0.0 \quad 0.3]$$

$$W_{run02} = [0.5 \quad 0.0 \quad 0.3 \quad 0.2]$$

$$W_{run03} = [0.5 \quad 0.3 \quad 0.0 \quad 0.2]$$

$$W_{run04} = [0.5 \quad 0.15 \quad 0.15 \quad 0.2]$$

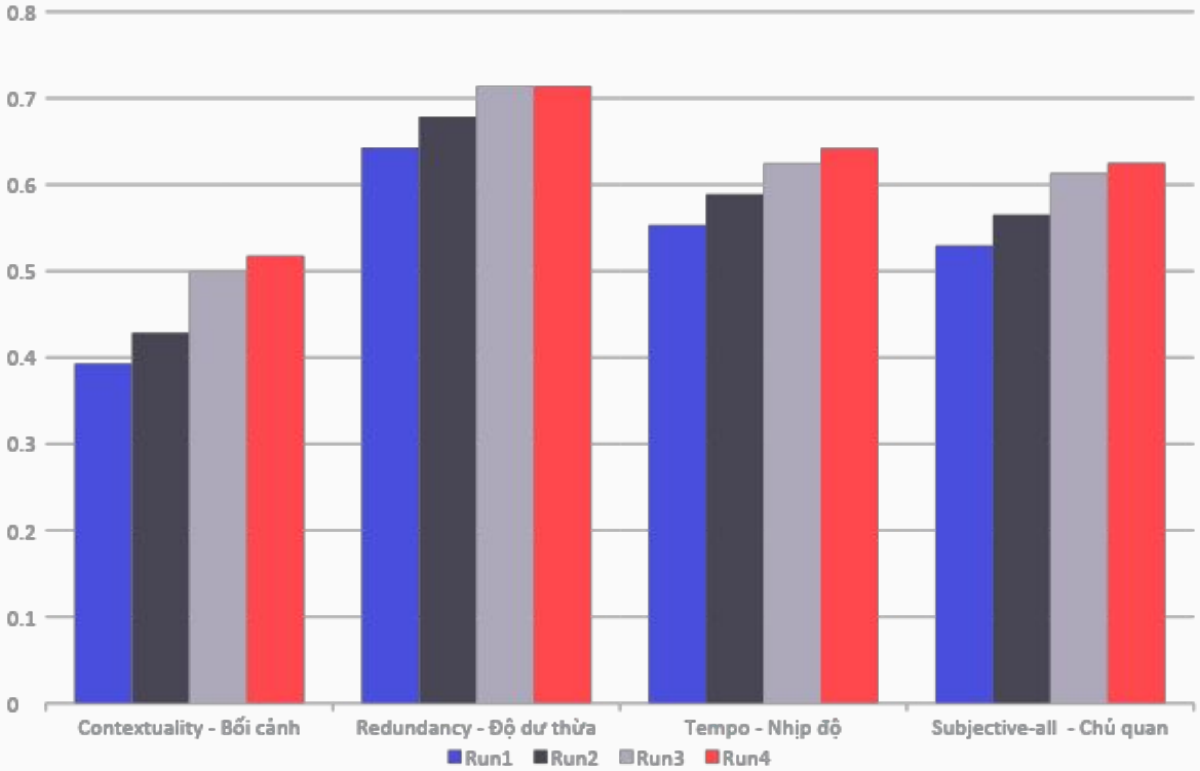
Objective Result



$$Objective - all = \frac{correctly\ submitted\ key\ facts}{ground\ truth\ key\ facts}$$

$$Precision = \frac{correctly\ submitted\ key\ facts}{submitted\ key\ facts}$$

Subjective Result



Run 1

	Submitted Keyfacts	Submitted Keyfacts đúng	Ground Truth Keyfacts	Objective-All	Precision
Archipelago-Cynthia	7	2	10	0.200	0.286
Archipelago-Edward	7	0	Unknown	0	0
ChainedforLife-Mabel	19	4	13	0.308	0.211
HeartMachine-cody	25	5	16	0.312	0.2
HeartMachine-virginia	25	2	11	0.182	0.080
bonneville-Arvilla	44	3	19	0.158	0.068
littlerock-Atsuko	56	9	19	**0.474**	0.161
littlerock-cory	49	4	16	0.25	0.082

Run 2

	Submitted Keyfacts	Submitted Keyfacts đúng	Ground Truth Keyfacts	Objective-All	Precision
Archipelago-Cynthia	7	2	10	0.2	0.286
Archipelago-Edward	7	0	Unknown	0	0
ChainedforLife-Mabel	19	4	13	0.308	0.211
HeartMachine-cody	25	5	16	0.312	0.2
HeartMachine-virginia	25	1	11	0.091	0.04
bonneville-Arvilla	41	2	19	0.105	0.049
littlerock-Atsuko	45	10	19	0.526	0.222
littlerock-cory	44	7	16	0.438	0.159

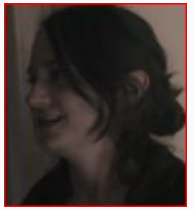
Run 3

	Submitted Keyfacts	Submitted Keyfacts đúng	Ground Truth Keyfacts	Objective-All	Precision
Archipelago-Cynthia	7	2	10	0.2	0.286
Archipelago-Edward	7	0	Unknown	0	0
ChainedforLife-Mabel	19	4	13	0.308	0.211
HeartMachine-cody	24	4	16	0.25	0.167
HeartMachine-virginia	24	2	11	0.182	0.083
bonneville-Arvilla	38	1	19	0.053	0.026
littlerock-Atsuko	47	10	19	0.526	0.213
littlerock-cory	39	5	16	0.312	0.128

Run 4

	Submitted Keyfacts	Submitted Keyfacts đúng	Ground Truth Keyfacts	Objective-All	Precision
Archipelago-Cynthia	7	2	10	0.2	0.286
Archipelago-Edward	7	0	Unknown	0	0
ChainedforLife-Mabel	19	4	13	0.308	0.211
HeartMachine-cody	25	5	16	0.312	0.2
HeartMachine-virginia	24	2	11	0.182	0.083
bonneville-Arvilla	42	1	19	0.053	0.024
littlerock-Atsuko	49	10	19	0.526	0.204
littlerock-cory	44	5	16	0.312	0.114

Archipelago-Cynthia

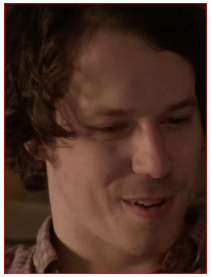


Clip 1:
Face reg: 0.55
Audio to Text: 0.163
Video Caption: 0.332
Time: 26.36s

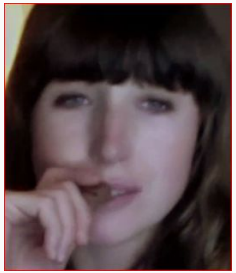
Clip 2:
Face reg: 0.52
Audio to Text: 0.192
Video Caption: 0.261
Time: 16.88s

ChainedforLife-Mabel

HeartMachine-Cody



HeartMachine-Virginia



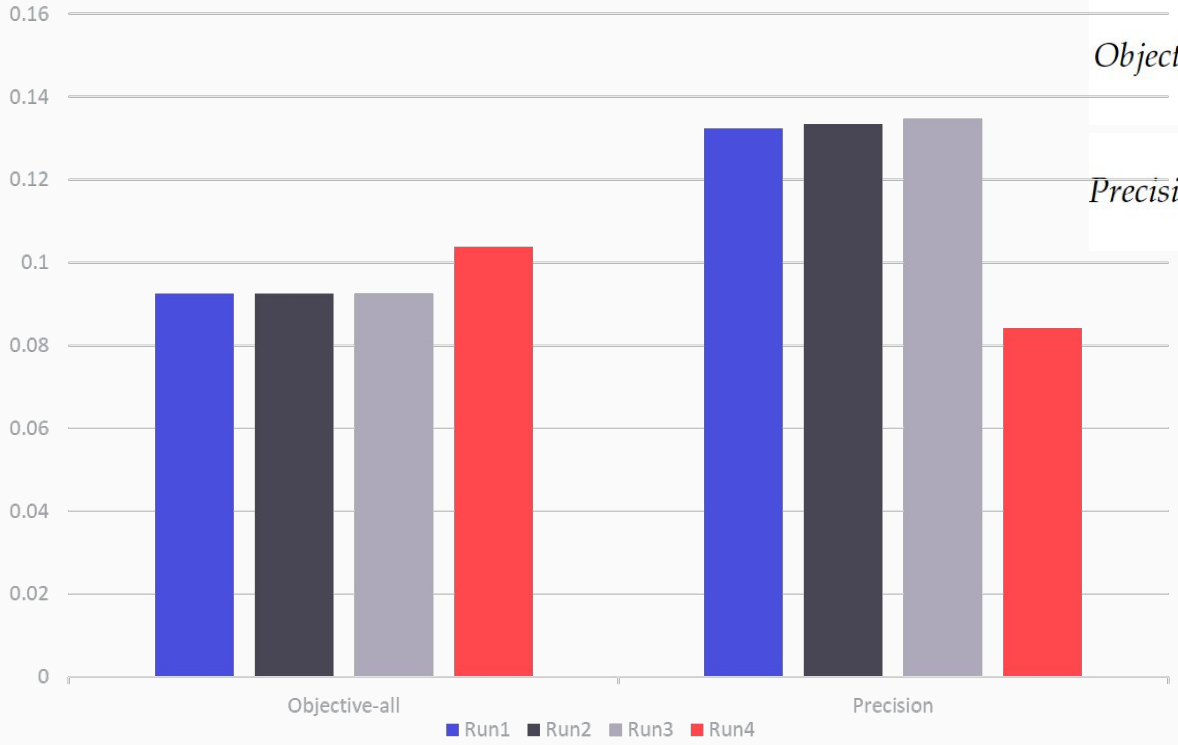
Bonneville-Arvilla

Littlerock-Atsuko

Littlerock-Cory

Text Summarization

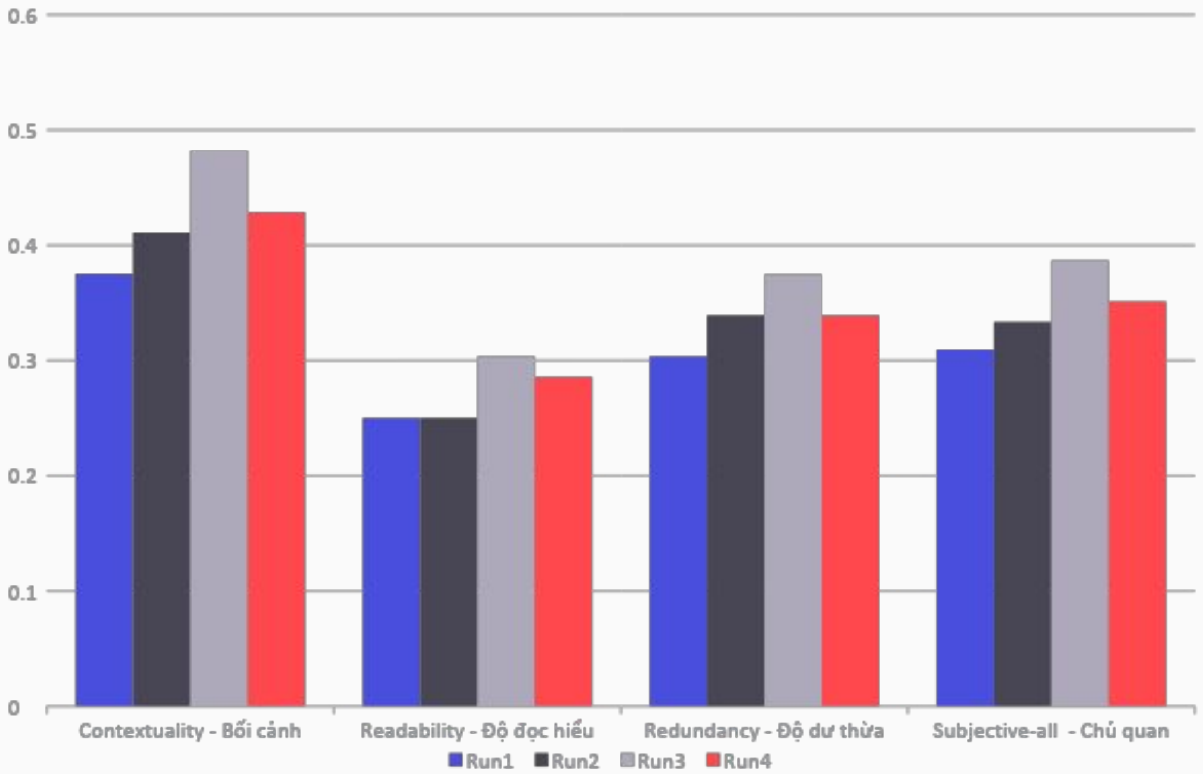
Objective Result



$$Objective-all = \frac{correctly\ submitted\ key\ facts}{ground\ truth\ key\ facts}$$

$$Precision = \frac{correctly\ submitted\ key\ facts}{submitted\ key\ facts}$$

Subjective Result



Run 1

	Submitted Keyfacts	Correctly Submitted Keyfacts	Objective-All	Precision
Archipelago-Cynthia	38	0	0	0
Archipelago-Edward	28	0	0	0
ChainedforLife-Mabel	26	2	0.154	0.286
HeartMachine-cody	32	3	0.188	0.273
HeartMachine-virginia	22	3	0.273	0.375
bonneville-Arvilla	38	0	0	0
littlerock-Atsuko	38	0	0	0
littlerock-cory	32	2	0.125	0.125

Run 2

	Submitted Keyfacts	Correctly Submitted Keyfacts	Objective-All	Precision
Archipelago-Cynthia	38	0	0	0
Archipelago-Edward	28	0	0	0
ChainedforLife-Mabel	26	2	0.154	0.286
HeartMachine-cody	32	3	0.188	0.273
HeartMachine-virginia	22	3	0.273	0.375
bonneville-Arvilla	38	0	0	0
littlerock-Atsuko	38	0	0	0
littlerock-cory	32	2	0.125	0.133

Run 3

	Submitted Keyfacts	Correctly Submitted Keyfacts	Objective-All	Precision
Archipelago-Cynthia	38	0	0	0
Archipelago-Edward	28	0	0	0
ChainedforLife-Mabel	26	2	0.154	0.286
HeartMachine-cody	32	3	0.188	0.273
HeartMachine-virginia	22	3	0.273	0.375
bonneville-Arvilla	38	0	0	0
littlerock-Atsuko	38	0	0	0
littlerock-cory	32	2	0.125	0.143

Run 4

	Submitted Keyfacts	Correctly Submitted Keyfacts	Objective-All	Precision
Archipelago-Cynthia	38	0	0	0
Archipelago-Edward	28	0	0	0
ChainedforLife-Mabel	26	2	0.154	0.143
HeartMachine-cody	32	3	0.188	0.150
HeartMachine-virginia	22	4	0.364	0.286
bonneville-Arvilla	38	0	0	0
littlerock-Atsuko	38	0	0	0
littlerock-cory	32	2	0.125	0.095

ChainedforLife-Mabel

Ground-Truth	Mabel in discussion with make-up artist and others about perceived imperfections in women including facial hair, small breasts, scars, tattoos, etc.	Mabel sits quietly next to Rosenthal in the auditorium while the entire cast watch the screen dailies.
Run 1	Mabel sitting on a chair with a large crowd of people watching on the	A man sits on a chair with a woman standing in front of a large crowd
Run 2		
Run 3		
Run 4		

HeartMachine-Cody

Ground-Truth	Cody flirts with Virginia over Skype.	Cody starts searching to East Village for clues of Virginia.	Over Skype, Cody teases Virginia.
Run 1	cody sitting on a chair with a camera and leads into a woman sitting	cody walking along a large yard while others watch on the side	cody sitting in a chair and speaking to the camera
Run 2			
Run 3			
Run 4			

HeartMachine-Virginia

Ground-Truth	Virginia flirts with Cody over Skype and Cody tells her he almost saw her on subway.	Over Skype, Virginia tells Cody that she had dated a barista from the East Village.	Over Skype, Virginia is evasive.	Over Skype, Virginia tells Cody she loves him.
Run 1	virginia speaking to the camera while holding a rag and leads into a man	virginia sitting in a chair and speaking to the camera		virginia sitting on a bed with a person standing in front of a camera.
Run 2				
Run 3			A man and woman are sitting on a chair	
Run 4				

HeartMachine-Cory

Ground-Truth	cory sitting on a bed with a woman standing in front of a camera. : Cory meets Atsuko and her brother Rintaro at a party in a motel room.	cory walking into a room and looking at the camera. : Cory is observing his work in the art opening and is happy.
Run 1	cory sitting on a bed with a woman standing in front of a camera	cory walking into a room and looking at the camera
Run 2		
Run 3		
Run 4		walking into a room and looking at the camera

Conclusion

Conclusion

- A simple baseline is proposed for MSUM Task
  - Face recognition can find scenes that key-fact events likely happen
- Difficult task due to high semantic information of key-fact events