

# A New Approach for Evaluating Movie Summarization

George Awad

National Institute of Standards and Technology  
Gaithersburg, Maryland, USA  
george.awad@nist.gov

Keith Curtis

Technological University of the Shannon  
Athlone, Ireland  
keith.curtis@tus.ie

## ABSTRACT

An important need in many situations involving video collections (archive video search/reuse, personal video organization/search, movies, tv shows, etc.) is to summarize the video in order to reduce the size and concentrate the amount of high value information in the video track. The Movie Summarization (MSUM) track in the annual TRECVID (TREC Video Retrieval Evaluation) benchmark [9] was proposed to evaluate automatic systems working on the movie summarization domain. This track used a licensed movie dataset from Kinolorberedu<sup>1</sup>, in which the goal is to summarize the storylines and roles of specific characters during a full movie. The goal of this track is twofold: Efficiently capture important facts about certain persons during their role in the movie storyline, and assess how well video summarization and textual summarization compare in this domain.

## CCS CONCEPTS

• **Information systems** → *Multimedia content creation*; **Multimedia and multimodal retrieval**.

## KEYWORDS

movie summarization; video summarization; video understanding; multimedia; multimodal interaction; information retrieval

### ACM Reference Format:

George Awad and Keith Curtis. 2023. A New Approach for Evaluating Movie Summarization. In *Proceedings of the 2nd Workshop on User-centric Narrative Summarization of Long Videos (NarSUM '23)*, October 29, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3607540.3617137>

## 1 INTRODUCTION

In recent years, the exponential growth of multimedia content, particularly movies and videos, has posed significant challenges for content consumption and comprehension. The vast amount of available audiovisual data necessitates efficient and effective methods to extract and present concise yet informative summaries. Movie summarization, a subfield of natural language processing (NLP) and computer vision, aims to address this challenge by generating

brief synopses or highlights that capture the essence of a movie's plot, themes, and key moments.

The advancements in deep learning, natural language processing, and computer vision techniques have led to substantial progress in automatic movie summarization technology. Such technologies have found applications in various domains, including video indexing, recommendation systems, and content understanding. However, the effectiveness and quality of these automated summarization methods vary significantly depending on the algorithms, features, and datasets used for evaluation.

While numerous movie summarization algorithms have been proposed [3, 8, 8, 14], the lack of standardized evaluation benchmarks and metrics hinders fair and meaningful comparisons among these methods. Existing evaluation approaches often lack comprehensiveness, making it challenging to discern the strengths and limitations of each technique. Consequently, this research gap emphasizes the need to establish a comprehensive benchmark that facilitates rigorous evaluation, comparison, and advancement of movie summarization technology.

The primary objective of this paper is to introduce an evaluation benchmark for movie summarization technology as applied at TRECVID (The annual video retrieval evaluation benchmark). TRECVID runs every year a set of content-based video evaluation tasks such as video search, activity detection, video captioning, etc. Our proposed evaluation benchmark incorporates a diverse dataset, encompassing various genres and styles of movies. By including real-world complexity, our benchmark aims to provide a reliable assessment of the generalization capability of the evaluated summarization methods.

This work makes the following contributions:

- Proposal of a novel and comprehensive evaluation benchmark for movie summarization technology.
- a new dataset with diverse and representative movie content, that enhances the benchmark's realism and robustness.
- Introduction of a set of performance metrics designed to capture both the quality (e.g. tempo, redundancy, readability, and contextuality) and informativeness of generated summaries.

The rest of this paper is organized as follows: Section 2 provides an overview of the task, data, and summarization queries. In Section 3, we describe the proposed evaluation benchmark and utilized metrics. Section 4 and 5 present the evaluated system approach and results respectively. Finally, Section 6 concludes the paper.

### 1.1 Related Work

There are several academic works related to automatic movie summarization. In [6] the authors propose an automatic approach to mine and summarize movie reviews using supervised learning and

<sup>1</sup><https://www.kinolorber.com/>

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

NarSUM '23, October 29, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0277-8/23/10...\$15.00  
<https://doi.org/10.1145/3607540.3617137>

graph-based ranking algorithms. Another work [14] proposes a human attention-based movie summarization dataset and model that can automatically edit a condensed version of a movie by selecting keyframes. A comprehensive overview of automatic text summarization techniques is presented in [12], which includes a discussion of summarization approaches, techniques used, standard datasets, and evaluation metrics. A method for summarizing full-length movies via sparse graph construction was presented in [8]. Additionally, [3] reviews some of the works in the automatic video summarization field, with a special focus on movie summarization. A video summarization track was previously introduced at TRECVID [1] which evaluated video summaries of the major life events of specified characters during a given time frame in BBC Eastenders TV series.

[11] introduced a novel dataset providing detailed graph-based annotations of social situations along with a thorough analysis of the dataset. MovieNet was introduced by [5]. This dataset contained 1100 movies with trailers, photo's, plot descriptions. It provides tags of cinematic style, description sentences, scene boundaries, etc. [2] proposed a dataset for learning from key scenes of a movie rather than the entire movie. They created the Condensed Movie Dataset (CMD) which consisted of key scenes from over 3,000 movies, accompanied by high-level semantic descriptions of the scenes, character face-tracks and metadata.

## 2 TASK OVERVIEW

Two main subtasks were supported in movie summarization track:

- (1) Video Summary
- (2) Text Summary

**Video Summary** Given a movie, a character, and image / video examples of that character, generate a video summary highlighting major key-fact events about the character. Video summaries are limited by a maximum summary length.

**Text Summary** Given a movie, a character, and image / video examples of that character, generate a textual summary to include key-fact events about the character role in the movie. Textual summaries are limited by a maximum number of sentences.

### key-fact events definition

- Any events that are important and critical in the character's story-line.
- They should cover the character's role from the beginning to the end of the movie.
- Example - Character Jeremy's key-fact events in the short movie "Super Hero":
  - Charlie bullies Jeremy.
  - Charlie and Jeremy fight at the playground.
  - Jeremy's mother reveals to the principal that Jeremy has a terminal illness.
  - Jeremy gets admitted to the hospital.
  - Jeremy passes away.

A key-fact event regarding a character does not necessarily require that character to be visible in the scene. In the above example *Super Hero*<sup>2</sup>, in one scene Jeremy's mother revealed to the principal that Jeremy had a terminal illness. This would clearly count as key-fact

| Movie           | Genre             | Duration |
|-----------------|-------------------|----------|
| Calloused Hands | Drama             | 92 min   |
| Liberty Kid     | Drama             | 88 min   |
| Like Me         | Horror / Thriller | 79 min   |
| Losing Ground   | Comedy / Drama    | 81 min   |
| Memphis         | Drama             | 79 min   |

Table 1: MSUM training dataset

| Movie            | Genre          | Duration |
|------------------|----------------|----------|
| Archipelago      | Drama          | 110 min  |
| Bonneville       | Drama          | 93 min   |
| Chained For Life | Comedy / Drama | 88 min   |
| Heart Machine    | Drama          | 84 min   |
| Littlerock       | Drama          | 82 min   |

Table 2: MSUM testing dataset

regarding Jeremy even though he was not present in that scene. This is available for viewing on the link provided in the footnote.

The purpose of this task is to summarize the important key-facts for a character. As such, this task is different from producing a movie trailer. Key events should appear in the order in which they become apparent in the movie, and should ideally capture that character's story-line.

The number of allowed key-facts is limited per movie and character to the number of key-facts provided by pre-task annotators. This is to prevent an unlimited summary situation where a 'summary' could be generated to an exceedingly long length in order to include as many clips or sentences as possible, thereby negating the entire purpose of video summarization and of this task. For video summaries a time length of 10 seconds is allowed per key-fact. One of the major challenges of the task is to separate major key-facts from non-consequential events. With a non-consequential event being one that has little to no impact on that characters overall story-line. One example could be: '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'.

### 2.1 Dataset

A set of 10 movies from Kinolorberedu was licensed and the organizers were able to distribute it to participants after signing a data agreement. A set of 5 movies were reserved for training (7 hrs) and another 5 movies were used for the testing set (7.6 hrs). Number of scenes across the full 10 movies ranged between 28 and 58. The selected movies were meant to be diverse in genre with self contained story-line and reasonable overall quality. Further information about the movies' genres and duration are provided below in Tables 1 and 2.

<sup>2</sup><https://www-nlpir.nist.gov/projects/tv2022/SuperHero.webm>

**Table 3: Movie summarization queries and specifics**

| Movie            | Character | Max. Length | Max. Lines | Key-facts | scenes |
|------------------|-----------|-------------|------------|-----------|--------|
| Archipelago      | Cynthia   | 190 seconds | 38         | 19        | 57     |
| Archipelago      | Edward    | 140 seconds | 28         | 14        | 57     |
| Bonneville       | Arvilla   | 190 seconds | 38         | 19        | 41     |
| Chained for Life | Mabel     | 130 seconds | 26         | 13        | 38     |
| Heart Machine    | Cody      | 160 seconds | 32         | 16        | 28     |
| Heart Machine    | Virginia  | 110 seconds | 22         | 11        | 28     |
| Littlerock       | Atsuko    | 190 seconds | 38         | 19        | 39     |
| Littlerock       | Cory      | 160 seconds | 32         | 16        | 39     |

## 2.2 Summarization Queries

Each query consisted 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 maximum summary time (in seconds) was specified for each character. While for text summaries, the maximum number of sentences was specified for each character as well. A sentence for text summary could be either a key-fact (the focus of the task), or a filler sentence. The maximum number of sentences a run could submit for a given character includes all key-facts and filler sentences.

In order to annotate the key-fact events, human assessors watched movies from the test set and marked out all key-facts for either 1 or 2 main characters from each movie. Key-facts were later double-checked and spell checked to remove any clear and obvious errors. Full queries including the number of key-facts in the ground truth, number of scenes per movie, maximum allowed video summary length, and the maximum allowed text summary sentences are listed in Table 3.

## 3 EVALUATION

The following subsections describe the processes that human assessors followed for the evaluations of both video and text summaries. Both use objective and subjective ratings as described.

### 3.1 Video Summaries

For the evaluation of video summaries, human assessors watched the submitted summaries. Assessors were also provided with a list of key-facts as provided by a different set of earlier assessors for this task. For any matching key-facts, assessors marked the matching key-fact and provided the clip number from the submitted summary for it to be independently verified later. Subjective ratings for tempo, contextuality, and redundancy were also provided. Figure 1 shows a screenshot of the tool used for the assessments of video summaries. The tool provides a way to play the summary from the provided list of video clips and/or text summaries and mark the inclusion of found key-facts.

It is important to point that submitted key-facts do not need to exactly match previous key-facts provided by assessors, in which a simple automatic comparison of sentences would have sufficed, and a degree of subjectivity must be part of such video summary evaluations. It is the task of assessors to decide if submitted key-facts *can be said to* closely enough match ground-truth key-facts from their perspective. This is particularly important in regard to

textual summary evaluations as many submissions can be quite poor and would likely all achieve zero scores if dependant on an automatic exact match to be deemed accurate.

### 3.2 Textual Summaries

For the evaluation of text summaries, human assessors first read a list of marked key-fact sentences in the submitted test summary. The format of submitted runs allowed systems to distinguish between a submitted key-fact and a filler sentence. These were compared with the list key-facts for each query provided by a different set of earlier assessors. Matching key-facts were marked by assessors and the matching submitted sentence was provided so that these could be later verified. Following this, assessors then read the entire submission comprised of both key-facts and filler sentences. Subjective ratings for readability, contextuality, and redundancy were then provided. Figure 2 shows a screenshot of the tool used for the assessments of text summaries.

As with all video summary evaluations, a degree of subjectivity needs to be part of the evaluation process and the task of final assessors is to decide if submitted video shots closely enough match key-facts provided by earlier assessors to be awarded points.

### 3.3 Metrics

The Scores for this task were calculated by assigning marks to correct key-facts and to subjective attributes such as readability and contextuality. The objective-all rating refers to the percentage of correct key-facts in the submitted textual or video summary. For example, 5 key-facts retrieved out of a possible 13 would give an objective-all score of 0.385. Precision refers to the number of correctly retrieved key-facts divided by correct key-facts + incorrectly retrieved key-facts, i.e.  $p = (\text{correct} / (\text{correct} + \text{false}))$ .

Tempo is a video specific subjective score which rates how well a video summary comes together. Readability is the textual specific alternative which rates the overall readability of a text summary. Contextuality measures how well context is provided for important information within the summary. Redundancy measures content considered unnecessary or superfluous. All three subjective ratings are measured on a scale of 1 to 7, with 7 being best except for redundancy. All three subjective ratings are combined into an overall subjective-all score. We acknowledge that many available movie datasets don't include the original videos which limits the use of subjective metrics as proposed in this framework.

## Movie Video Summarization

**Task:** You are given a video summary of a movie character, and list of textual key-facts about that character. Please watch the summary and select the key-facts that are shown in the video summary. Write the clip number that this key-fact appears at (**clip numbers are shown in the top left corner of the video while playing**).

**Tempo and Rhythm:** (1 - 7, high is best)

Value: 4

**Contextuality:** (1 - 7, high is best)

Value: 4

**Redundancy:** (1 - 7, low is best)

Value: 4

**video summaries:**

- littlerock-cory-1
- littlerock-cory-2
- littlerock-cory-3
- littlerock-cory-4
- littlerock-cory-5
- littlerock-cory-6
- littlerock-cory-7
- littlerock-cory-8
- littlerock-cory-9
- littlerock-cory-10
- littlerock-cory-11

**Key-facts:**

- ☐ Cory meets Atsuko and her brother Rintaro at a party in a motel room.
- ☐ Cory tells Atsuko that his mother is dead.
- ☐ Cory wants to be a model and an actor and go to Hollywood.
- ☐ Cory owes Brody \$200, but only has \$50 to give him now.
- ☐ Brody gives Cory one hour to get the \$150 he owes him.
- ☐ Cory work for his father and does not like it.
- ☐ Cory was worried when Atsuko did not come back to the motel last night.
- ☐ Cory asks his father if Atsuko could stay with them since her brother left.
- ☐ Cory takes money out of the till at Mica's Restaurant where he works.
- ☐ Cory takes Atsuko to a July 4th parade.
- ☐ Cory says he loves Atsuko.
- ☐ Brody comes to the restaurant looking for Cory but is told he is not there.
- ☐ Cory tries to kiss Atsuko, who pushes him away.
- ☐ Cory asks to borrow a couple hundred dollars from his father, who refuses to give it to him.
- ☐ Cory confesses that Brody gave him drugs to sell and Cory smoked them, and cannot pay Brody back.
- ☐ Cory is observing his work in the art opening and is happy.

[Submit your key-facts assessments](#)

Figure 1: Video summary assessment tool

## Movie Text Summarization

**Task:** You are given a text summary of a movie character, and list of textual key-facts about that character. Please read the summary and select the key-facts that are written/mentioned/inferred in the text summary. When you find a matching key-fact, please copy and paste the part (text segment) in the summary that matches it.

**Readability:** (1 - 7, high is best)

Value: 4

**Contextuality:** (1 - 7, high is best)

Value: 4

**Redundancy:** (1 - 7, low is best)

Value: 4

**Text Summaries:**

- little\_rock.cory.1
- little\_rock.cory.2
- little\_rock.cory.3
- little\_rock.cory.4
- little\_rock.cory.5
- little\_rock.cory.6
- little\_rock.cory.7
- little\_rock.cory.8
- little\_rock.cory.9
- little\_rock.cory.10
- little\_rock.cory.11

**Key-facts:**

- ☐ Cory meets Atsuko and her brother Rintaro at a party in a motel room.
- ☐ Cory tells Atsuko that his mother is dead.
- ☐ Cory wants to be a model and an actor and go to Hollywood.
- ☐ Cory owes Brody \$200, but only has \$50 to give him now.
- ☐ Brody gives Cory one hour to get the \$150 he owes him.
- ☐ Cory work for his father and does not like it.
- ☐ Cory was worried when Atsuko did not come back to the motel last night.
- ☐ Cory asks his father if Atsuko could stay with them since her brother left.
- ☐ Cory takes money out of the till at Mica's Restaurant where he works.
- ☐ Cory takes Atsuko to a July 4th parade.
- ☐ Cory says he loves Atsuko.
- ☐ Brody comes to the restaurant looking for Cory but is told he is not there.
- ☐ Cory tries to kiss Atsuko, who pushes him away.
- ☐ Cory asks to borrow a couple hundred dollars from his father, who refuses to give it to him.
- ☐ Cory confesses that Brody gave him drugs to sell and Cory smoked them, and cannot pay Brody back.
- ☐ Cory is observing his work in the art opening and is happy.

[Submit your key-facts assessments](#)

Figure 2: Textual summary assessment tool

## 4 PARTICIPATING SYSTEM

Although 8 teams signed up to participate in the task, only 1 team [7] from the National Institute of Informatics, Japan and University of

Information Technology, VNU-HCMC, Vietnam was able to submit 4 runs with slight variations in each of them.

To summarize the system applied by the finishing team, they first segmented the movie into shots using TransNet\_v2 [10]. They

then passed shots through a pipeline to calculate a ‘face similarity ranking score’ and ‘text similarity ranking score’. Scores were then fused to generate a single importance score per shot to determine final shots. To calculate the ‘face similarity ranking score’, MTCNN [13] was used for face detection and VGGFace2 [4] was used for face representation. Cosine similarity was then used to match faces in the shot and the input face query. For calculation of ‘text similarity ranking score’, two sources were used: ‘Audio to Text Similarity Score’ and ‘Video Captioning Similarity Score’.

For video captioning, shots were first passed through a Video Captioning Network, generated captions were then compared with key-facts from the training set. Bert was used for feature extraction and cosine distance was used for text similarity measurement. For the audio to text, shots were passed through an audio to text system, generated audio captions then went through the same process as the video caption. A time score was used to penalize shots which were too long before final fusion of scores. For more details we refer the reader to the system paper [7].

## 5 RESULTS

### 5.1 Video Summaries

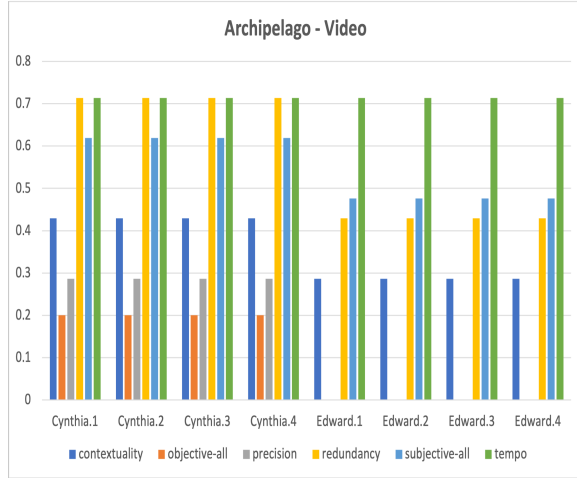


Figure 3: MSUM video summarization: Archipelago

Figures 3 to 7 show the video summarization task results for all testing set movies. Each figure plots the performance of the summarization system’s 4 runs across all corresponding characters and movies for all metrics. As a general observation, for the movie *Archipelago* best results are seen for the character Cynthia. Subjective scores for tempo and redundancy are quite good, however objective scores of 0.2 are quite poor. There were no objective scores achieved for the character Edward. This means that no correct key-facts were retrieved by the summarization system for this movie and character and the system is said to have performed poorly on this query.

For the movie *Bonneville*, there was only one character for summarization in this movie, Arvilla. Generally, both subjective and objective scores for this movie and character are quite poor. The

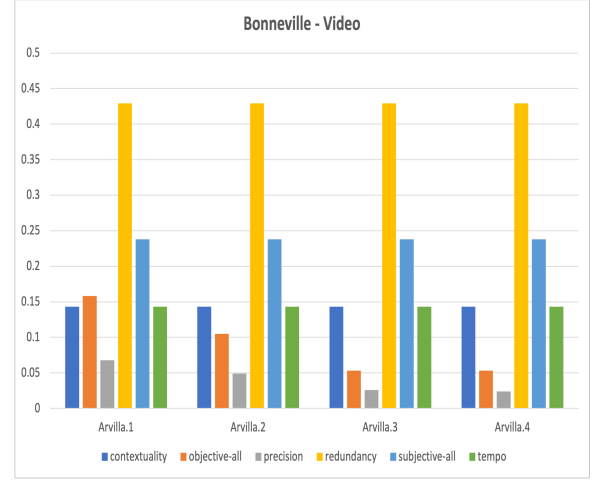


Figure 4: MSUM video summarization: Bonneville

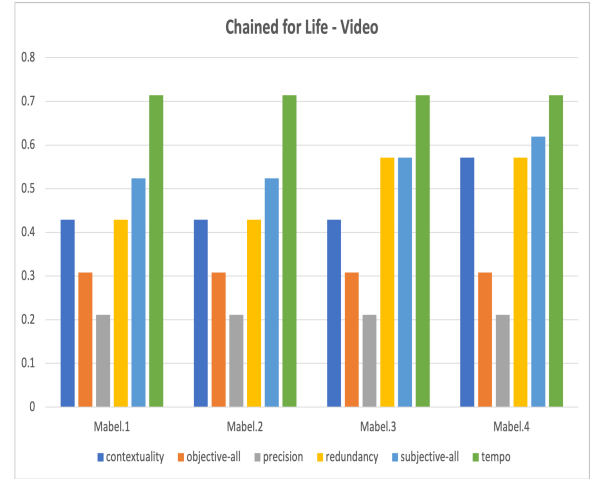


Figure 5: MSUM video summarization: Chained for Life

movie *Chained for Life* contained just one character for summarization, Mabel. Subjective scores were not as high as for Archipelago - Cynthia, however objective scores are slightly better.

In the movie *Heart Machine*, two characters for summarization, Cody and Virginia were considered. Slightly better scores were achieved for Cody than for Virginia, however there is not much difference between characters. Perfect redundancy scores were achieved on Cody for three of the four runs. Objective scores are not great for either character, however these scored slightly better on Cody.

For the last movie titled *Littlerock*, contained two characters for summarization, Atsuko and Cory. Best objective scores were achieved on Atsuko while best subjective measures were achieved for Cory. Scores for objective measures were generally better in this movie. Very good all around subjective scores were achieved

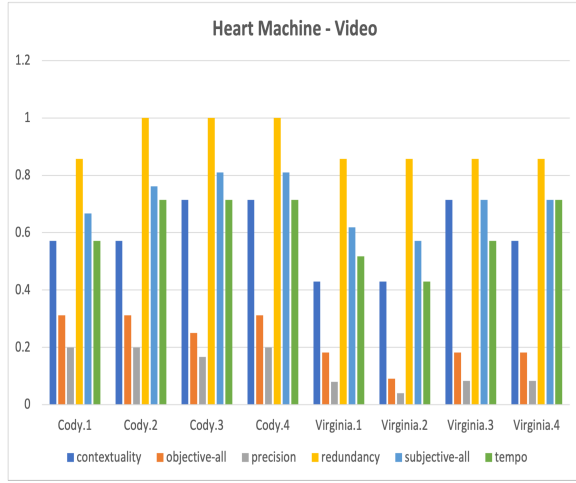


Figure 6: MSUM video summarization: Heart Machine

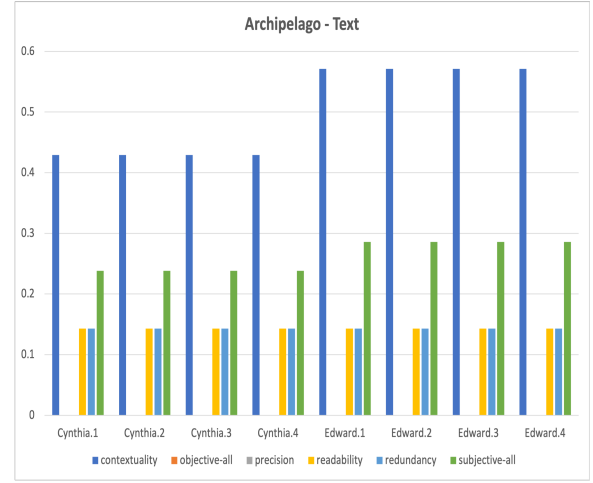


Figure 8: MSUM text summarization: Archipelago

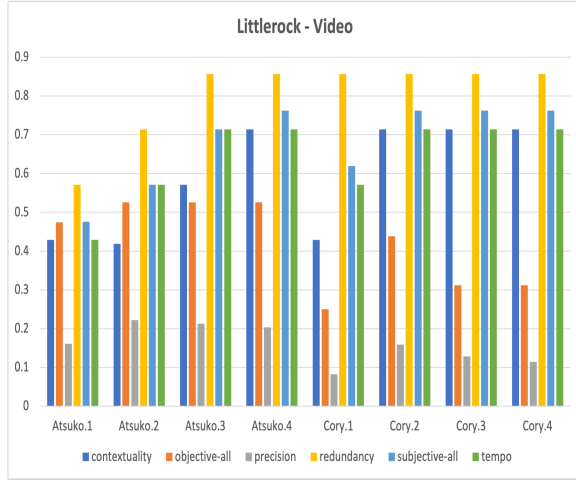


Figure 7: MSUM video summarization: Littlerock

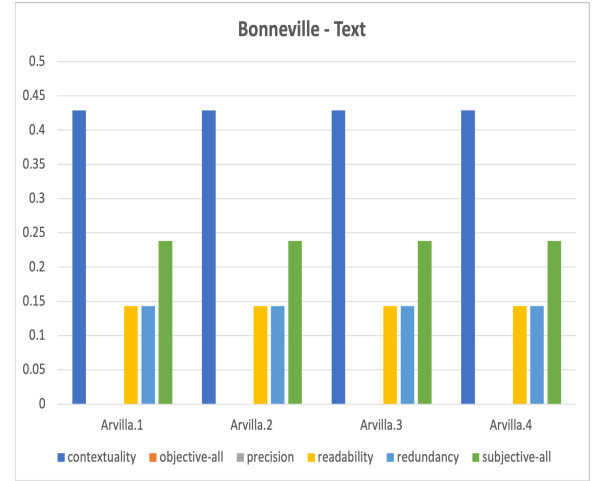


Figure 9: MSUM text summarization: Bonneville

for Cory, while these also underachieved on objective scores in comparison to Atsuko.

## 5.2 Text Summaries

Figures 8 to 12 show the text summarization task results. Each figure plots the performance of the summarization system across all corresponding characters and movies for all metrics. In the movie *Archipelago*, best results are seen for the character Edward, however these are only for subjective scores. No objective scores were achieved for either character.

For the movie *Bonneville*, both subjective and objective scores for this movie and character are quite poor. While for the movie *Chained for Life* subjective scores were higher than for *Archipelago* and *Bonneville* above, while some objective scores were also achieved in this movie.

In the movie *Heart Machine*, higher objective scores were achieved for Virginia than for Cody, while there is very little difference between characters for subjective scores. The achieved objective scores for the characters of this movie were higher compared to the observed scores for the characters of the previously mentioned movies, on this task.

Finally for the movie *Littlerock*, better results were achieved for Cory than for Atsuko on both objective and subjective metrics. Zero marks on objective metrics were achieved for Atsuko, while at least some marks on objective metrics were achieved on Cory.

## 6 CONCLUSION

In this paper we proposed a new evaluation framework for movie summarization based on key-fact events of selected characters to highlight the most important events affecting the story-line of their role. The framework was tested on a new licensed movie dataset

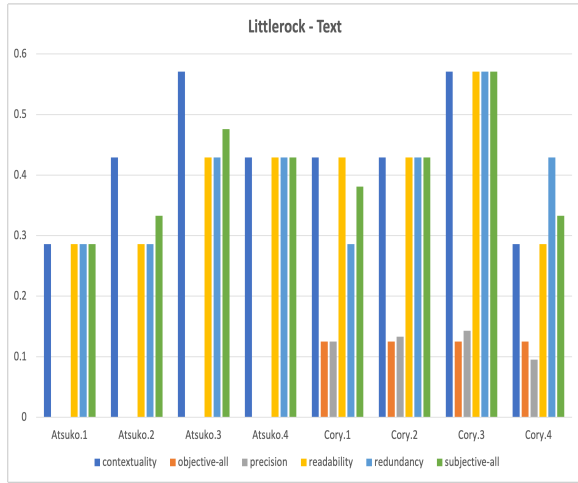


Figure 12: MSUM text summarization: Littlerock

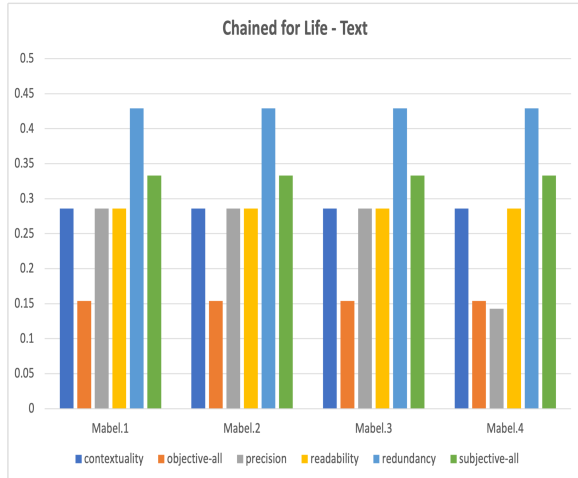


Figure 10: MSUM text summarization: Chained for Life

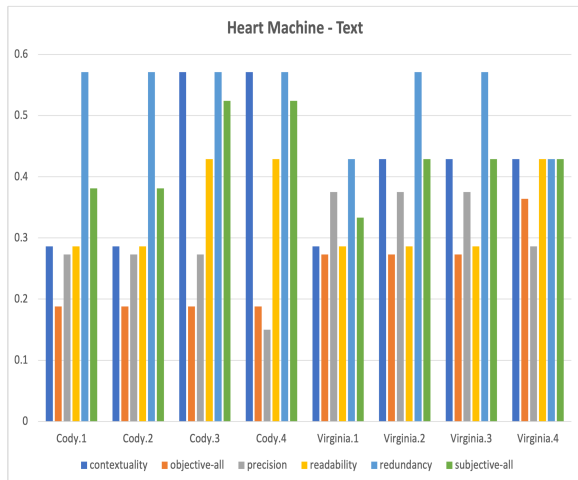


Figure 11: MSUM text summarization: Heart Machine

available for the public. Two sub-tasks were introduced, namely text summarization and video summarization.

Objective and subjective metrics were applied in both summarization sub-tasks. In general, text summarization of movie characters are more challenging compared to video summarization. The participation of just one team in the task limits the conclusions that can be drawn.

*Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST, or the U.S. Government.*

## REFERENCES

- [1] George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al. 2021. TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. *arXiv preprint arXiv:2104.13473* (2021).
- [2] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*.
- [3] Xavier Bost. 2016. *A storytelling machine?: automatic video summarization: the case of TV series*. Ph. D. Dissertation. Université d'Avignon.
- [4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.
- [5] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 709–727.
- [6] Atif Khan, Muhammad Adnan Gul, Mahdi Zareei, RR Biswal, Asim Zeb, Muhammad Naeem, Yousaf Saeed, and Naomie Salim. 2020. Movie review summarization using supervised learning and graph-based ranking algorithm. *Computational intelligence and neuroscience* 2020 (2020).
- [7] Nam Nguyen, Tien Hung Nguyen, Cong Nguyen Thanh, Hao Vo, Khiem Le, Tien Do, Tien-Dung Mai, and An Pham Nguyen. 2022. NII\_UT at TRECVID 2022: Movie Summarization Task. (2022).
- [8] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2021. Movie summarization via sparse graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13631–13639.
- [9] Alan F Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVID. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 321–330.
- [10] Tomáš Souček and Jakub Lokoč. 2020. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838* (2020).
- [11] Paul Vicol, Makarand Tapaswi, Lluís Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8581–8590.
- [12] Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. 2022. Automatic text summarization methods: A comprehensive review. *arXiv preprint arXiv:2204.01849* (2022).
- [13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.
- [14] Defang Zhao, Dandan Zhu, Xiongkuo Min, Jiaomin Yue, Kaiwei Zhang, Qiangqiang Zhou, Guangtao Zhai, and Xiaokang Yang. 2023. Human attention based movie summarization: Dataset and baseline model. *Neurocomputing* 534 (2023), 106–118.