Zero-Shot Classification of Events for Character-Centric Video Summarization

ALISON REBOUD, ISMAIL HARRANDO, PASQUALE LISENA, and RAPHAËL TRONCY, EURECOM, France

This paper describes an event classification and character-centered approach proposed by the D2KLab team at EURECOM for the 2021 TRECVID Video Summarization Task [Awad et al. 2020]. Our approach relies on defining a list of typical important events in a soap opera and using this list of named events as candidate labels for a zero-shot text classification method. This additional data source is used together with the provided videos, scripts and master shot boundaries. We also use BBC EastEnders characters' images crawled from the Google search engine in order to train a face recognition system. All our runs use the same general method, but with varying constraints regarding the number of shots and the maximum duration of the summary. The runs submitted are as follows:

- EURECOM1: 5 shots with highest similarity scores and the total duration of the summary is < 150 sec;
- EURECOM2: 10 shots with highest similarity scores and the total duration of the summary is < 300 sec;
- EURECOM3: 15 shots with highest similarity scores and the total duration of the summary is < 450 sec;
- EURECOM4: 20 shots with highest similarity scores and the total duration of the summary is < 600 sec.

ACM Reference Format:
Alison Reboud, Ismail Harrando, Pasquale Lisena, and Raphaël Troncy. 2021. Zero-Shot Classification of Events for Character-Centric Video Summarization. In *Proceedings of TRECVID 2021, International Workshop on Video Retrieval Evaluation (TRECVID 2021).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Considering video summarization as an important task for digital content retrieval and reuse, the TRECVID Video Summarization Task (VSUM) 2021 aims at fostering the research in this field by asking its participants to automatically summarize "the major life events of specific characters over a number of weeks of programming on the BBC EastEnders TV series"[1] [Awad et al. 2020]. More precisely, for three different characters of the series, the participants have to submit 4 summaries with respectively 5, 10, 15 and 20 automatically selected shots. These generated summaries are evaluated by the assessors according to their tempo, contextuality and redundancy as well as with regards to how well they contain answers to a set of questions unknown to the participants before submission. In addition to the videos, the episodes transcripts are provided by the organizers. In 2020, we have addressed the VSUM task by matching fan-written synopsis to transcripts using as hypothesis that each paragraph mentioned in these synopsis correspond to important moments to include in a summary [Harrando et al. 2020]. However, such synopsis are not always available. This year, we propose a new approach based on zero-shot classification of named events.

[1] https://www-nlpir.nist.gov/projects/tv2021/vsum.html

TRECVID 2021, December 7–10, 2021, Virtual Conference

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 APPROACH FOR THE MAIN TASK

Figure 1 illustrates our general approach for the main task composed of three main steps: transcript classification, face recognition and shot selection.

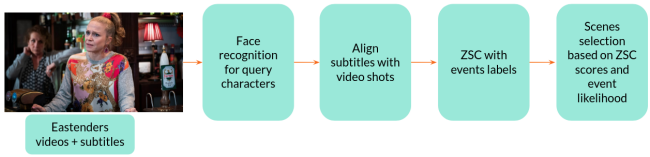

Fig. 1. Fan-driven and character centered approach

2.1 Face Recognition

The dataset considered for the task consists of 10 video episodes which amount to approximately 19 000 shots. The summarization task aims to produce shorter videos of 5 to 20 shots (which is respectively 0.02% and 0.10% of the original episode duration).

The compression rate being high, we discard all the scenes where the character of interest in not present in the scene. In order to do so, we extract and recognize faces using the Face Celebrity Recognition library [Lisena et al. 2021], a method which uses images gathered from crawling the web with the character's name as the keyword query. We also added the phrase "EastEnders" to the names to avoid including images of people with the same name. The faces are first detected with an MTCNN. Each detected face then gets associated with a FaceNet embedding. We empirically define a threshold of standard deviation 0.24 for cosine similarity under which we consider that the faces are outliers and we eliminate them. Finally, a multi-class SVM classifier outputs the final prediction.

We also align the provided XML transcripts with the given shot segmentation. If a sentence encompasses multiple shots, we select all the shots as we expect a good summary to avoid including scenes with cut utterances. However, this increases the noise of our summaries and diminishes the number of distinct moments. We believe this constraint is a limitation of the shot segmentation and that a scene segmentation would be more relevant to the task.

2.2 Shot Transcript Classification

The instructions for VSUM state that the method developed for the task should be able to differentiate between meaningful and trivial events, choosing for example 'the birth of a child rather than a

short illness'. Therefore, we tackle this task by trying to define what could be such events, hypothesising that soap opera episodes are repetitive enough that the type of major events of an episode can be defined in advance, without having watched the series. We use the results of a research work which investigates if soap opera viewers' perceptions of the likeliness of some life events differ from the non-viewers [Seese 1987]. In this work, the authors defined events which they thought often happen in soap operas (Table 1). We construct our model with the hypothesis that the least likely events are also the most interesting ones and should probably be included in the summary. For instance, if the scene contains a 'suicide attempt', it should be more interesting than a 'happily married' scene. For that reason, we take the inverse of the perceived likelihood (on a scale from 1 to 5) of an event as its weight (Table 1). We do not assume the evaluation team to be specifically composed of soap opera viewers and hence select the likelihood scores reported for the non-viewers group. The weight of the event gets further multiplied by the confidence score obtained from the zero-shot classifier (which was normalized for each class with RobustScaler[2]). Finally, because we wish to extract informative scenes which should therefore be long enough, the score per scene gets further multiplied by the log of the length of the shot dialogue (Equation 1).

$$score(shot\_i) = max\_l \in labels(zsc(trans\_i, l) * weight(l) * log(len(trans\_i)) \quad (1)$$

where $shot_i$ is the unique id of the shot, $trans_i$ is its corresponding transcript, *labels* is the list of events, with their importance expressed with *weight*().

Finally, to get one score per shot (and not per candidate event label), we select the max score on all event labels. To generate the submissions, we keep the N shots with the highest score. To respect the summary length requirement, in case the generated summary is too long, we un-select the longest scene from the top N and replace it with the N+1th one, recursively until the summary length constraint is met.

Table 1. Life events labels and their perceived likelihood (scale from 1 to 5) according to [Seese 1987]

Label	Likelihood
extramarital affair	1.98
get divorced	1.96
illegitimate child	1.45
institutionalized for emotional problem	1.43
happily married	4.05
serious accident	2.96
murdered	1.81
suicide attempt	1.26
blackmailed	1.86
unfaithful spouse	2.23
sexually assaulted	2.60
abortion	1.41

[2]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

Table 2. Overall results

Team	Main task	Subtask
ADAPT	30.15%	17.25%
EURECOM	29.55%	30.10%
NII_UIT	18%	29.85%

3 APPROACH FOR THE QUERIES SUBTASK

The goal of the substask is similar to the main one, except that the queries used for the evaluation by the task organisers are revealed for the subtask (after submission to the main task). Our approach considers this task to be similar to a Question-Answering task where the goal is to predict where the answer to the question lies in the text. We use HuggingFace's Transformer QA pipeline (using longformer as a base model, pretrained on Squad-v2 QA task, or longformer-squadv2) to score each line in the script as a potential answer to the question for each character. We then rerank the top 10 answers using Sentence-BERT (paraphrase-mpnet-base-v2), scoring each by cosine similarity to the question. This tends to push answers that are more similar to the question to the top run. To avoid having long runs, we drop scenes that are too long. These scenes get picked consistently because they contain a lot of words and thus are likely to match with the questions somehow. In this submission, we limit shot length to 20s.

4 RESULTS

Table 2 shows our and the other teams results. We ranked second for the main task and first for the substask. For both tasks, our results are close to 30% which was also the type score we obtained in 2020 [Harrando et al. 2020] with an approach which was relying on the provision of fan made synopsis, contrary to this year. For the substask (where queries are known), it is somehow surprising to see that non of the teams achieved results better than the score of the best team for the main task. Tables 3 and 4 display respectively the characters for which we obtained the best and worst results in the main task. We obtained the best score across characters with the run 4 (37.60%). Interestingly, for this run, our event classification method allowed to answer 9 of the 16 'What' questions and zero of the remaining 9 'Who', 'Why', etc. questions. These results could indicate that events/actions are the first important facts of a summary but also suggest that our model could gain from covering other aspects such as persons and locations.

Table 3. Detailed results for the queries about Archie with 20 shots included in the summary

Query	Main task	Subtask
What happens when Phil throws Archie in to a pit?	Yes	No
What happens after Danielle reveals to Archie that Ronnie is her mother?	Yes	No
Where do Peggy and Archie get married?	No	No
What happens when Archie arrives at the pub after Peggy invited him?	No	No
What happens when Archie is kidnapped?	Yes	No

Table 4. Detailed results for the queries about Peggy with 20 shots included in the summary

Query	Main task	Subtask
Who does Peggy ask to kill Archie?	No	No
Where do Peggy and Archie get married?	No	No
Show one of the challenges which Peggy faces in her election run.	No	Yes
What does Peggy overhear Archie saying, which causes their marriage to be over?	No	No
What is Janine doing to irritate or anger Peggy?	Yes	Yes

REFERENCES

George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Lukas Diduch, Jeffrey Liu, Alan F. Smeaton, Yvette Graham, Gareth J. F. Jones, Wessel Kraaij, and Georges Quénot. 2020. TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *TRECVID 2020 Workshop*. NIST, Gaithersburg, MD, USA.

Ismail Harrando, Alison Reboud, Pasquale Lisena, Raphael Troncy, Jorma Laaksonen, Anja Virkkunen, and Mikko Kurimo. 2020. Using Fan-Made Content, Subtitles and Face Recognition for Character-Centric Video Summarization. In *TRECVID 2020 Workshop*. NIST, Gaithersburg, MD, USA.

Pasquale Lisena, Jorma Laaksonen, and Raphaël Troncy. 2021. FaceRec: An interactive framework for face recognition in video archives. In *2nd International Workshop on Data-driven Personalisation of Television (DataTV)*, ACM (Ed.). New-York.

Gayle Seese. 1987. *Soap opera viewers' perceptions of the real world.* Master's thesis. University of Central Florida.