Stories of Love and Violence: Zero-Shot Interesting Events Classification for Unsupervised TV Series Summarization

Alison Reboud[1*], Ismail Harrando[1], Pasquale Lisena[1] and Raphaël Troncy[1]

[1]EURECOM, Sophia Antipolis, France.

*Corresponding author(s). E-mail(s): alison.reboud@eurecom.fr;
Contributing authors: ismail.harrando@eurecom.fr;
pasquale.lisena@eurecom.fr; raphael.troncy@eurecom.fr;

Abstract

In this paper, we propose an unsupervised approach to generate TV series summaries using screenplays that are composed of dialogue and scenic textual descriptions. In the last years, the creation of large language models has enabled Zero-Shot text classification to perform effectively in some conditions. We explore if and how such models can be used for TV series summarization by conducting experiments with varying text inputs. Our main hypothesis being that interesting moments in narratives are related to the presence of interesting events, we choose candidate labels to be events representative of two genres (crime and soap opera) and we obtain competitive results with respect to the state-of-the art baseline.

Keywords: Summarization, Moment Detection, Zero-Shot Classification, Knowledge Graphs, Face recognition

1 Introduction

With an ever-growing number of videos uploaded daily over a multitude of channels[1] and a large variety of genres and topics, comes an increasingly pressing need for efficient management of multimedia content. The entertainment sector, which includes movies and TV series, constitutes a particularly rich collection of videos and a good target for video summarization. It is indeed more and more via streaming platforms that the public discovers new audiovisual content and it becomes interesting for them to be able to display key segments of a program in order to facilitate the user search and browsing experience.

While there is a high interest for general-purpose video summarization methods [1], a line of work around genre-specific (video summarization) also exists, as outlined in Sreeja et al. [2]. These authors underline that if specific actions play a major role for sport videos, the presence of the main characters in a video segment might instead be important for movies. Our method proposes to leverage on the specifics of the entertainment domain. Namely, we exploit the fact that TV series episodes are often associated to transcripts and/or screenplays. The complex narrative of this type of material is an interesting case study from a computational linguistics point of view, and we argue that their summarization can benefit from the progress made in natural language processing in the last years. For text summarization as well, many techniques leverage domain-specific knowledge [3]. For example, the best approaches aiming at summarizing news articles are based on the observation that the main points of an article are presented at the beginning of the document [4]. Similarly, summarizing scientific articles is best done when taking into account the very specific structure of this genre of document [5].

In this paper, we tackle the task of TV series summarization which aims to produce shorter summaries covering the episodes' most interesting scenes, by proposing a text-based unsupervised method, using screenplays or transcripts previously segmented into scenes or shots. We evaluate our approach on two different genres: crime (from the *CSI: Crime Scene Investigation* series [6, 7]) and soap opera (from *BBC EastEnders* series). We show that it is possible to rely on a very general unsupervised model (zero-shot text classification), using the right label instead of focusing on the architecture of the model. One consideration we take into account while building our approach is that, due to a time consuming annotation process [8], labelled data for TV series summarization is scarce.[2] We therefore believe it is crucial to develop unsupervised approaches for this task and we establish this criterion as a requirement for our model. At first, the usage of text classification may seem counter intuitive for summarization as, in many settings, we do not know the semantic content of a text beforehand. However, because some topics, events and words often appear together, there is a long tradition of classifying movies and TV series

[1] https://www.forbes.com/sites/tjmccue/2020/02/05/looking-deep-into-the-state-of-online-video-for-2020/

[2] Trailers can not qualify as good proxies for this task because they precisely avoid spoilers, which are often the key events that we instead wish to include in our summaries.

into genres [9]. We follow Ben-Ahmed et al. [10] in their hypothesis that the most interesting moments of a series episode should be semantically close to its genre or to events recurrent in the considered genre. Our work also leverages on the fact that large language models boosted the performance of Zero-Shot classification which is the task of classifying textual inputs using only the label information without seeing any training examples of that label [11]. Therefore, we consider zero-shot classification models to be a good opportunity to test our hypothesis about the importance of genre with an unsupervised model that can easily be used for other genres in the future. To the best of our knowledge, this method has not been yet explored for the task of TV series summarization.

Screenplays containing mixed information (dialogues and scenic information describing what the spectator sees and hears), we ask ourselves what is the most relevant text type for a text classification approach based on genre? This work also aims at answering the following questions: Can TV series summarization benefit from zero-shot classification methods? How can we determine the relevant classification labels for the task of summarization? Do different zero-shot models yield different results? When coupled with existing approaches, does this method provide complementary information?

Our main contribution consists in showing that, with the right label, it is possible to obtain results on par with other state of the art approaches, at the task of unsupervised TV series summarization, with an 'out-of-the-box' tool. We show how to determine that label and we observe that our method yields even better results when ensembled with centrality measurements developed by Papalampidi et al. [7]. Since we test our general approach on two different genres and datasets with complementary evaluation methods, the specifics of our methods vary with the dataset. The remainder of the paper is therefore structured as follows: we first present some related work (Section 2). In Section 3, we present our general approach. In Section 4, we detail our experiments and discuss the results on the CSI dataset, while we present our experiments on the BBC EastEnders dataset in Section 5. After discussing limitations and generalisation issues in Section 6, we conclude and outline some future work in Section 7.

2 Related Work

We present some of the research in the field of automatic video summarization, with a special focus on TV series and movies as the notion of interestingness remains dependent on the domain and use-case. We also present the related field of spoiler detection. Since our method is based on zero-shot text classification, we review the most recent advances in this field.

2.1 TV Series Summarization and Related Tasks

To be summarized, videos are usually split into segments, which are then classified as being interesting or not. A plethora of methods for general-purpose video summarization rely solely on visual cues to predict visual interestingness.

A comprehensive review can be found in Apostolidis et al. [1]. As movies and TV series are generally accompanied by speech and metadata, approaches for this domain are usually multimodal or textual [2]. One inspiring line of work from general-purpose video summarization, for our particular use case, is multimodal and semantic or category-driven methods. Several works [12, 13] indeed aim to increase the similarity between the semantics of the summary and of the associated metadata, action or video category[3] with a reward system. Instead of video categories, our approach aims to create summaries semantically close to some named events.

Adapted to movies and TV series, some early approaches attempted to generate trailers, relying on a combination of multimodal low-level features such as motion, contrast, statistical rhythm [14] spatio-temporal saliency, AM-FM speech and part of speech tagging [15], with the goal to draw a multimodal saliency curve. The MediaEval benchmarking initiative [16], for which higher-level features were also used, helped fostering the research in the field. Interestingness in movies has been approached with the help of related concepts such as movie genre classification [10, 17] or emotional resonance [18, 19]. However, in this paper, instead of extracting salient moments (like in movie trailers), we wish to build summaries which cover the whole narrative arc with its major events. For this task, some have considered important characters identification [20, 21], while our proposed approach is rather event-centric. Papalampidi et al. [22] took upon the challenge of formalising narrative structure. Based on expert knowledge on narratives, they considered that movie scripts contain five turning points (Opportunity, Change of Plans, Point of no Return, Major Setback and Climax) and showed that its is feasible to automatically identify them from screenplays.[4]. The authors also released the so-called TRIPOD dataset that contains movie screenplays and Turning Points annotations. On a follow-up work, they proposed a sparse movie graph which indicates the similarity between scenes using multimodal information [23], while Lee et al. [24] identified these turning points with a supervised transformer approach. As we want to develop a method which can be applied to TV series episodes that are not self-contained, and therefore do not necessarily follow such a defined structure, we do not use the five turning points identification as a proxy task for TV series summarization. Instead, closer to our work is Papalampidi et al. [7] which demonstrated that these turning points can also be used as a latent representation when gold standard TV series summaries are available. We compare our unsupervised approach to theirs, using the same metrics.

Another important contribution in the field is the TRECVID VSUM challenge for which participants had to develop unsupervised methods to summarize episodes of the BBC Eastenders TV series. For this challenge, one team

[3]Video categories include: Changing Vehicle Tire, Getting Vehicle Unstuck, Groom Animal, Making Sandwich, Park-our, Parade, Flash Mob Gathering, BeeKeeping, Bike Tricks, Dog Show, Base Jump, Bike Polo, Eiffel Tower, Excavator River Crossing, Kids Playing in Leaves, MLB, NFL, Notre Dame Cathedral, Statue of Liberty, and Surfing

[4]The authors report a 17.33% Partial Agreement score on the percentage of turning points where there is an overlap of at least one scene between the prediction and the ground truth

proposed a purely visual approach [25]. With the exception of our 2021 submission, the other teams have all based their methods on fan-written text [26–28]. On the contrary, our current approach is not dependant on the availability of such external data. Addressing the summarization task from a slightly different perspective, some works generated text summaries from movies or TV series (abstractive summarization) [29, 30].

Finally, despite being mostly interested in user-generated content on social media and review sites, another line of work related to our task is spoiler detection [31, 32]. Such work include a model [32] based on the writing style of the online comments (tense, degree of objectivity) and on named entity recognition. Closest to our work is a deep neural spoiler detection model with a genre-aware attention mechanism approach [33]. The authors also conducted a spoiler characteristics analysis where they extracted semantic frames from spoiler sentences in the dataset. They found frames associated with "*killing*" to be frequent in thriller spoilers, while romance had more frames linked to personal relationships. We directly use these results to define our text classification candidate labels.

2.2 Zero-Shot Text Classification

Recently, approaches relying on large pre-trained generic language models have proven very successful for a wide range of NLP downstream tasks. Text classification, which consists in associating the correct label to a text segment for any domain and for aspects such as topic, emotion or event is one of them. If the fine-tuning step on the downstream task still increases performance for this task, such models have also proven efficient without this additional step [11]. Zero-shot classification is particularly powerful because it can be used for labels partially or fully unseen during the model development, making it a ready-to-use tool without the need for task-specific datasets.

A major contribution to the field has been brought by Yin et al,. who tackled three main problems with zero-shot text classification: its restrictive focus on topic categorization, the treatment of labels as indices rather than as words with a meaning and an disparate evaluation with various datasets and evaluation setups [11]. They proposed a standardized evaluation on "conceptually different and diverse aspects": topics, emotions and situations. In particular, they showed that beyond the *restrictive* version of zero-shot classification (in which during a training phase, the classifier is allowed to see similar data with their labels), zero-shot classification also handles the *wild* version where the classifier does not see any examples of the labeled data. The later version is primordial for our task as we do not know if it saw words close to the candidate labels that we propose during training.

Several Transformer-based frameworks [34–36] have been developed lately. They make use of the pre-trained knowledge and representational ability of such models in the context of zero shot classification. In this paper, we will use the Entail method developed by Yin et al. [11], as one of our baselines since this is the representative of this family of frameworks. Entail uses a pre-trained

transformer architecture that is trained on the task of Textual Entailment, which represents the task of classifying the logical relation between two given sentences, the *premise* and the *hypothesis*: *entailment* (agreement), *contradiction*, or *neutrality*. The task of zero-shot text classification is then formulated as a logical entailment inference. We also select ZeSTE (Zero-Shot Topic Extraction)[5], a zero-shot text classification model using common-sense knowledge as another method [37]. In short, this model relies on the presence of explicit knowledge in ConceptNet [38] to score and classify documents into a given set of labels. This offers an alternative approach that is explainable as we can visualize the graph of connections between the label and the content of the document, and that has state-of-the-art performance on several text classification datasets [37]. Both approaches are explained in more details in Section 3.2. We use these two models throughout the paper to test our hypothesis of the potential of zero-shot text classification for TV content summarization. While previous works have used zero-shot approaches for abstractive summarization [39], to the best of our knowledge, TV series summarization for audiovisual data through zero-shot classification is a novel direction of work.

3 Approach

In this section, we present our zero-shot classification method. Screenplays contain mixed information: dialogues and scenic information describing what is visually happening. Dialogues are a transcription of the speech and scenic information are visual instruction in the screenplay indicating the movement, position, or tone of an actor, or the sound effects and lighting. For the CSI series, thanks to an homogeneous formatting across the episodes screenplays, we are able to write a script which separates these two types of texts. Our main motivation for this step, is that while dialogue can be automatically obtained with Automatic Speech Recognition techniques, scenic information cannot. Getting insights into which type of data is the most relevant to the task allows to somehow assess how automatic the methods is and how it would perform if we would only have access to the raw audio and video material. We ultimately use three types of text inputs: dialogue only, scenic information only and original screenplay (mixed information). For each text input and every scene, our approach consists in obtaining a score denoting the probability that it belongs to the candidate label of interest. We then select the scenes with the highest confidence as the ones that we predict to be part of the summary.

3.1 Candidate Labels

One of our hypothesis being that the scenes included in a summary are representative of a TV series or movie genre, we select different ways to choose candidate labels related to a genre:

[5] https://github.com/D2KLab/ZeSTE

- **Genre-based method**: The candidate labels chosen correspond to the name of the series genre(s).
- **Event-based method**: The candidate labels chosen correspond to named events.

Beyond the genre name, the idea of this method is to obtain candidate labels that are representative of events often happening in a specific genre. As mentioned in Section 2, Chang et al. [33] conducted an analysis that provides genre specific words for the Romance and Thriller genres in order to develop supervised genre-aware spoiler detection models. More precisely, they use FrameNet [40], a tool built on the semantic frame theory, for sentences semantic role labeling where sentences are parsed and associated to semantic frames according to their structure. Semantic frames are descriptions of a type of event, relation, or entity and the participants in it. For the sentence 'John drowned Martha', it would for example tag 'John' as 'killer', 'drowned' as 'killing' and 'Martha' as 'victim'. The authors used the SEMAFOR parser to extract semantic frames from spoiler sentences for different genres including Thriller and computed their normalised frame frequency (NFF = count of each frame divided by the total number of frames). Figure 1 shows the difference of NFFs for each frame and the most contrasting 10 frames for the two genres Thriller and Romance.

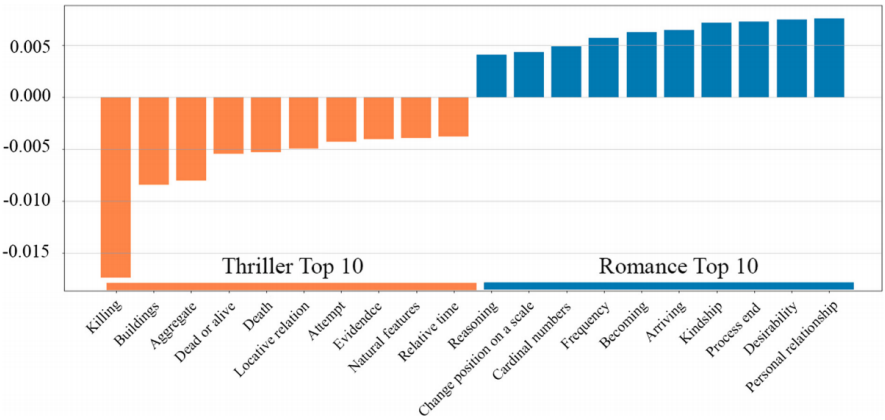

Fig. 1 Top 10 frames between Thriller and Romance (NFF of Thriller frames - NFF of Romance frames) from Chang et al. [33]

For our approach, as we are interested in making summaries that capture the key events of a narrative, we select as candidate labels the frames names describing an event, among the 10 frames displayed. Hence, for the genre Thriller, we select the labels "killing", "death" and "attempt". The authors interpret the contrast in the distribution of the frames as a significant relationship between the genre and contents of a spoiler sentence. As ultimately the key scenes we want to extract could probably qualify as spoilers, these results

also give more empirical grounding to our hypothesis that genre could be used for summary scenes retrieval.

3.2 Models

To tackle the task of key narrative event extraction, we choose two state-of-the-art approaches for zero-shot text classification that use two different sources of knowledge: latent knowledge from pre-trained language models (Entail method) and explicit common sense knowledge about genres through ConceptNet (ZeSTE method). Both approaches perform well on several text classification benchmarks, and are freely available and open-source. The goal of our paper is to illustrate the potential of zero-shot classification for TV series summarization. We therefore limit our investigation and comparison to these two models to make our analysis more focused and concise and we leave as future work the study of alternative models.

3.2.1 Entail

Given a sentence acting as a *premise*, the task of Natural Language Inference (NLI) aims at determining its relation with an *hypothesis* sentence as either true (entailment), false (contradiction), or undetermined (neutral). NLI datasets consist of sequence-pairs that are generally approached by a transformer architecture such as BERT [41]. Both the premise and the hypothesis are the inputs of a model which classification head predicts one of the following labels: contradiction, neutral, entailment. The method developed by Yin et al. [11] consists in using a model pre-trained on that task as zero-shot text classifier. More precisely, the text to be labeled is the *premise* and the candidate labels are injected in the sentence "This text is about" + label, to form an *hypothesis*.

The confidence with which the Entail model predicts the hypothesis to be entailed by the premise is interpreted as the confidence of the label to be true. While, in the original paper, the label-weighted F1 obtained was 37.9% on Yahoo Answers with the smallest version of BERT, fine-tuned on the multi-genre NLI (MNLI) corpus [42], we use the HuggingFace implementation[6] which reports a F1 of 53.7% by using the BART model pre-trained on MNLI [43].

3.2.2 ZeSTE

ZeSTE is a different approach based on the assumption that a document about a topic such as "crime" will probably also mention other words from the same lexical field such as "victim" or "perpetrator". ConceptNet is used to produce a "topic neighbourhood", which is a list of candidate words related to the candidate labels. More specifically, the topic neighborhood is created by querying every node that is N hops away from the label node. Using ConceptNet Numberbatch (ConceptNet's graph embeddings), a cosine similarity score is computed between each node and the candidate label. This similarity score

[6] https://huggingface.co/zero-shot/

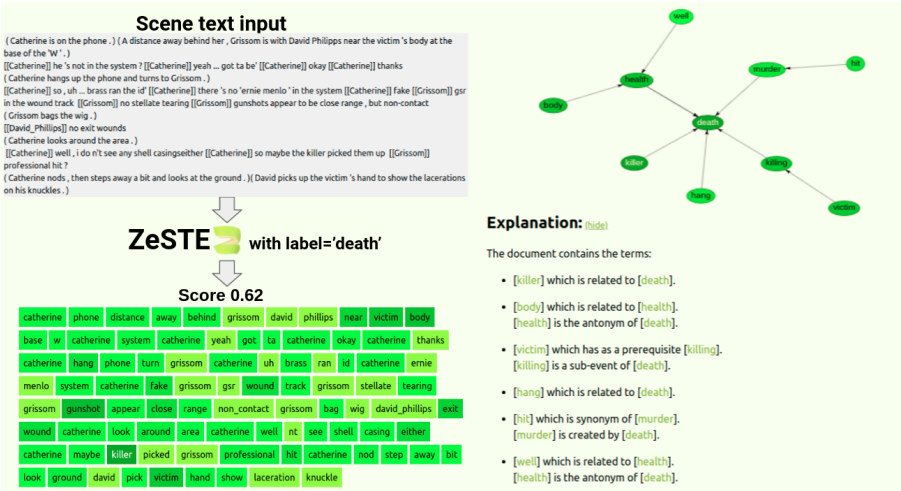

Fig. 2 Text and explanation of a scene classified by ZeSTE as 'death' as the label with the highest confidence

(Equation 1) represents the relevance of any term in the neighborhood to the candidate label. Next, the documents to classify are assigned a score following the same method (the Numberbatch concept embedding for word w is denoted by nb_w).

$$numberbatch\_score(doc, label) = \sum_{token \in doc \cap LN(label)} sim(nb_{token}, nb_{label}) \quad (1)$$

The original authors tried different ways to choose a neighborhood. Following their evaluation procedure on the BBC News dataset, we select their best configuration: 3-hops neighborhoods (for the best performance/computation power ratio), using all the relations (47 relations defined in ConceptNet). Finally, as shown in the example in Figure 2, all predicted document labels can be explained by the model by showing the path between the nodes. The darker the colour of the child node, the highest the similarity with the parent node.

3.3 Evaluation

We first evaluate our results with F1 scores, a metric which 'has been adopted by the vast majority of the state-of-the-art works on video summarization' [1], including SUMMER, the method we compare ourselves to on the CSI dataset. Besides this metric, several works also include a human-evaluation [7, 22]. Following Awad et al. [44] who investigated different aspects of human evaluation for the task of video summarization, for the soap opera genre, we assess temporability, contextuality, redundancy and the number of questions a method allowed to answer. For the genre of crime we also investigate the number of crime aspects covered by our method.

4 Summarizing Crime TV Series Episodes

In this section, we evaluate our genre-based summarization approach, on the CSI dataset [6, 7], which is, according to the authors, associated to the crime genre. The experiments presented in this section can be reproduced using the code published at https://github.com/alisonreboud/screenplay_summarization.

4.1 Dataset

The Crime Scene Investigation (CSI) dataset contains 39 CSI video episodes together with their screenplays segmented into scenes, each one being associated to a binary label denoting whether the scene should be part of the summary or not.[7] It also contains word-level labels indicating if the perpetrator is mentioned in the dialogue. An episode scene contains in average 21 sentences and 335 tokens. For the scenes chosen for the summary, the three human annotators had to indicate whether they selected the scene based on one/more or none of the following six reasons to justify why a scene is important: i) it revealed the victim, ii) the cause of death, iii) an autopsy report, iv) crucial evidence, v) the perpetrator, and vi) the motive/relation between perpetrator and victim. The dataset creators considered these reasons to be aspects that should be covered by crime series summaries. An episode contains in average 40 scenes from which 30% are labelled positively. Although only 3 episodes (out of 39) contain a second investigation case (instead of just one), we followed the authors in assuming no such prior knowledge considering that TV series and movies often contain sub-plots.

4.2 Experiment

We perform the text classification on every scene. In order to compare ourselves with the original SUMMER approach [7], we configure our model to include 30 percents of the scenes in the episode summaries. Applying the genre-based method, the candidate labels are “thriller” and its sub-genre “crime” (as described in the dataset). For the event-based method, the candidate labels are “killing”, “death” and “attempt” (see Section 3).

To assess whether our approach yields complementary results to the SUMMER ones (obtained on the mixed information, not separating dialogues from scenic information), we also combined our results. As explained in Section 2, SUMMER is an approach that computes centrality measures between scenes to identify turning points and that chooses the scenes with top centrality measures. After min-max scaling these scores, we average them with our Zero-Shot Classification (ZSC) scores (2).

[7] https://github.com/EdinburghNLP/csi-corpus

$$ensemble\_score(scene) = \sum_{scene} 0.5 \times ZSCscore_{scene} + 0.5 \times SUMMERscore_{scene} \quad (2)$$

4.3 Results and Discussion

Table 1 presents the results of our experiments on the CSI dataset, where SUMMER corresponds to the state of the art results on this dataset.

Table 1 F1 for different text inputs (ZSC = Zero-Shot Classification, SI = Scenic Information, MI = Mixed Information)

		ZSC			ZSC+SUMMER		
		Dialogue	SI	MI	Dialogue	SI	MI
Genre-based method							
crime	Entail	37.32	39.13	38.01	38.75	42.074	41.09
thriller	Entail	39.53	35.91	36.76	40.00	40.84	38.24
crime	ZeSTE	37.44	36.61	40.98	44.14	45.20	44.11
thriller	ZeSTE	36.98	40.52	41.20	45.36	45.08	45.013
Event-based method							
killing	Entail	41.53	**45.49**	41.03	46.34	**48.55**	45.089
death	Entail	40.92	44.77	40.80	45.30	48.97	47.013
attempt	Entail	26.71	32.69	25.45	33.28	40.52	30.89
killing	ZeSTE	40.14	39.17	43.66	46.43	45.14	47.95
death	ZeSTE	43.67	43.25	**46.21**	47.74	46.28	**48.59**
attempt	ZeSTE	37.22	36.95	38.49	43.72	43.44	44.19
	SUMMER			44.70			

First, comparing the results obtained for the genre-based method with the ones obtained for the event-based method, we observe that for both the Entail and the ZeSTE models, the results obtained with the genre-based method are inferior, suggesting that the name of the genre is not the best candidate label for the summarization via text classification. The F1 scores of the genre-based method reaches a maximum of 41.21% which is under the SUMMER performance. When combined with SUMMER results, the results outperform SUMMER alone in four out of six cases for the ZeSTE model. For the genre-based method, ZeSTE slightly outperforms the Entail model.

For the event-based method, our approach yields the highest mean F1 with the label "killing" using Scenic Information and the Entail model (F1 = 45.49%) and for the label "death" with mixed information and the ZeSTE model (F1 = 46.21%). These labels are semantically close to each other and are the two most representative of the event frames of the genre Thriller. On the other hand, the label "attempt" performs the worst of all keywords, across methods Entail and ZeSTE. This could probably be explained by the fact that "attempt" is the least domain-specific word among the labels we tried. In a CSI episode context, the word is probably to be understood as "murder attempt", but the two general zero-shot classification models we use miss the information that our interest only lies in this specific context.

The fact that the words "killing" and "death" are successful labels for crime cases summarization makes intuitive sense from a human point of view. Indeed, this type of crime cases we try to summarize has also be called 'Whodunit'[8] where the word "it" stands precisely for a killing or murder. For these two labels, it is also always the case, for mixed information input types and models, that the combination of our approach and SUMMER obtains a higher F1 mean than SUMMER and zero shot classification alone, reaching a F1 score up to 48.59%.

In order to assess the statistical significance of our results, we perform a t-test (1) between the F1 scores obtained by SUMMER and the F1 scores of our best approach (ZeSTE with label 'death') (2) between the F1 scores obtained by SUMMER and the ZeSTE (label 'death') + SUMMER approach. Our null hypothesis is that the two distributions are identical. We respectively obtain p-values of 0.626 and 0.098, which are both above a significance level of 0.05. For such a significance level, these results do not allow us to reject the null hypothesis and we therefore consider our approaches to be on par with the state-of-the-art.

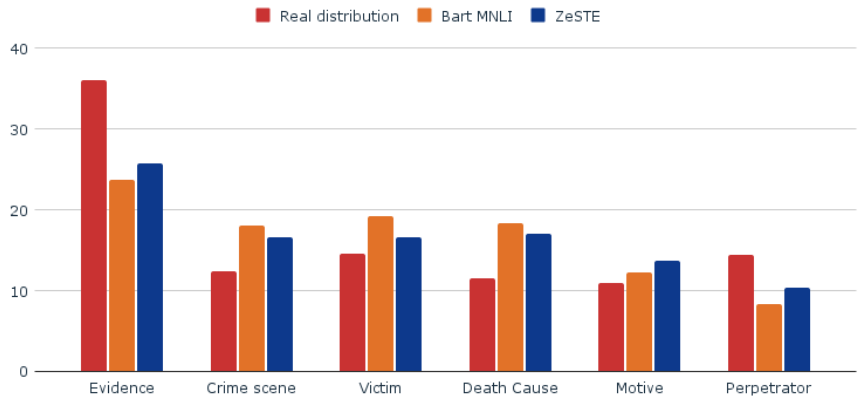

Fig. 3 Average composition of the scenes correctly predicted as being part of the CSI summary by the best performing Entail and ZeSTE models

In terms of models, there is no clear winner between ZeSTE and Entail. However, they do present differences in terms of the text input it deals with the best. We observe that for Entail, scenic information systematically outperform the other text types with mixed information performing the worst. For ZeSTE, instead, mixed information always yields the best results. A possible reason to explain why the performance is reduced when both dialogue and scenic information are used is the fact that BART truncates long texts while ZeSTE does not. This truncation might cause some useful information to be excluded when both information are concatenated. The results also suggest that the

[8] https://en.wikipedia.org/wiki/Whodunit

role played by visual (scenic descriptions) and audio (dialogue) information is an axis worth investigating and that there is a point in isolating texts of different nature, describing two different modalities. We observe that both models are quite sensitive to the label choice and the label ranking (in terms of performance) is quite similar for both models.

Since our goal is to produce informative summaries and given that the SUMMER dataset creators gave some cues about what they consider to be a good summary for this genre – a summary that covers different crime-related aspects which they define to be Evidence, Crime scene, Victim, Death Cause, Motive, Perpetrator of an episode – we compare in Figure 3 the distribution of aspects for the scenes chosen by our method with the true distribution of the dataset. We choose to plot the best performing labels for Entail and ZeSTE, which are respectively "killing" and "death". First, we observe that the distribution of aspects obtained for ZeSTE and Entail are quite similar. In the ground truth summaries (real distribution), the aspect 'Evidence' is twice more represented than the other aspects. While 'Evidence' is also the most frequent aspect in the two models predictions, the frequency of aspects is more evenly distributed with the other aspects. This shows that the summaries created with the approach presented are diverse, covering different aspects of crime plots.

Finally, a small exploration of the scenes wrongly included in the summaries by our method revealed some examples where the error does actually not come from the classification itself: we observe that the included scene is indeed strongly associated to the label from a human point of view. Figure 2 illustrates such a case. This particular example is an autopsy scene that ZeSTE (rightly) associates strongly to the keyword 'death' because it contains among others, the words 'body', 'victim' and 'killer' which all are in the ConceptNet neighborhood of 'death' via the relations mentioned in Figure 2. This association to the label is, however, not sufficient to make the scene relevant enough to be included in the summary. We include more examples of errors in Appendix A.

5 Summarizing Soap Opera TV Series Episodes

In this section, we further evaluate the robustness of our approach by testing it on an different genre, a soap opera TV series, while adapting the evaluation method. We present the results obtained for the summarization of the BBC EastEnders series with a human evaluation on the criteria of tempo, contextuality, redundancy and the model's capacity to answer a set of questions about the plot. The experiments presented in this section can be reproduced using the code published at https://github.com/MeMAD-project/trecvid-vsum.

5.1 Dataset

The TREC Video Retrieval Evaluation (TRECVID) campaign aims at fostering the research in content-based exploitation and retrieval of information

from digital video via open metrics-based evaluation [44]. One of the tasks proposed in 2020[9] and 2021[10] is the Video Summarization Task (VSUM). The participants have to automatically summarize "the major life events of specific characters over a number of weeks of programming on the BBC EastEnders TV series". The dataset consists in 244 video episodes (464 hours) segmented into 471.527 shots, together with their transcripts. For the 2021 edition of the challenge, for five different characters of the series, the participants had to submit 4 summaries with 5, 10, 15 and 20 automatically selected shots over 10 episodes with a maximum duration of respectively 150, 300, 450 and 600 seconds. 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. We believe this type of evaluation which includes more subjective metrics, complements well the F1 evaluation performed for the CSI dataset. The plot questions evaluation can be compared with the qualitative analysis performed on the CSI dataset which aimed at discovering if our model was able to cover the different aspects that the dataset creators thought should be included in a crime series.

5.2 Experiment

As the task focuses on some specific characters and does not provide a transcript-shot alignment, we enhance our general approach described in Section 3 with additional preprocessing steps that we describe below. Furthermore, as we were only allowed to submit one method for evaluation, we reduced the number of experiments we could do: we select the Entail model, using the dialogue text (the full screenplay of this TV series is not made available by TRECVID) and we focus on the event labels (event-based method in Section 3) as our first experiments show better results than just the genre label.

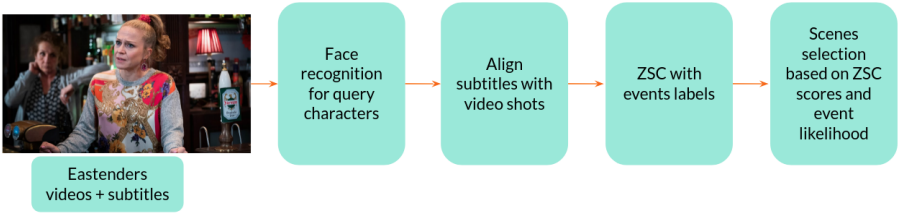

Fig. 4 Our approach for the VSUM challenge (ZSC = Zero-Shot Classification)

[9] https://www-nlpir.nist.gov/projects/tv2020/vsum.html
[10] https://www-nlpir.nist.gov/projects/tv2021/vsum.html

5.2.1 Recognizing Character Faces in Videos

The portion of the dataset considered for the challenge contains 10 episodes, that is approximately 19 000 shots, which must be reduced to either 5 or 20 shots (respectively 0.02 or 0.10 percent of the original episodes), while for the previous experiments on the CSI dataset, 30 percents of the scenes were to be selected in the summary. Because of this important compression, we wish to filter out irrelevant scenes and to reduce further the noise. We therefore consider that for a shot to be important for a character, the character needs to be present in this shot. For that, we use the Face Celebrity Recognition library [45], a model that relies on pictures obtained from search engines with the actor's name as the keyword query. In this particular case, we added the word "EastEnders" to the character names to be certain that we do not add pictures corresponding to different people with a similar name. For each video frame, faces are first detected with an MTCNN [46] algorithm. The system then extracts face embeddings with FaceNet [47]. Based on the hypothesis that most of the faces obtained with the search do depict the actors of interest, other faces (such as people appearing together with the actor) are automatically removed by eliminating outliers to reach a cosine similarity of face embeddings below a standard deviation of 0.24 (this threshold was empirically defined). After this filtering step, a multi-class SVM classifier is trained. To increase the recognition consistency between video frames, the system also uses the Simple Online and Real time Tracking algorithm (SORT) [48] which returns groups of detection of the same person in consecutive video frames. In our experiment, we keep the shots with any of the five characters when the confidence score is more than 0.5. Another required preprocessing step is to transform the given XML transcripts into timestamped text and aligned it with the provided shot segmentation. When a sentence spreads over two shots we include it in both shots as a good summary should probably avoid having scenes with cut utterances. However, this might lead to some noise.

Even if cheaper to produce, we believe that for a better coherence and smoothness of the summary, a segmentation into scenes (like in the CSI dataset) is more appropriate than one into shots. A shot is indeed simply the continuous sequence between two edits or cuts[11], while a scene is the basic unit in a screenplay, usually associated to one main story element [49].

5.2.2 Selecting Events for the Soap Opera Genre

Following the semantic parsing method explained in Section 3, we first extract the important events semantic frames for the romance genre. We obtain the following frames: 'process end', 'arriving', 'becoming', 'change position on a scale' and 'reasoning'. All these frames depict quite general events that could refer to many potential sub-events: 'process end' could be the end of a relationship, of a work contract or of an education but also of minor events such as finishing the dinner. For the thriller genre, the event frames extracted ('killing','death'

[11] https://en.wikipedia.org/wiki/Shot_(filmmaking)

and 'attempt' (murder)) were merely synonyms depicting a single event. The extracted frames for romance rather suggest that this genre encompasses a wider diversity of major events and that we should therefore adapt our method, offering it the flexibility of considering a combination of different event labels rather than a unique one.

Furthermore, as stated in Section 2, the task of summarization, even if narrowed down to the specific type of narratives, remains very dependent on the instructions given for the annotation and/or evaluation. For this challenge, it is specifically stated that the model 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'. As the challenge settings does not allow to submit different methods, we anticipate some of the limitations of our semantic parsing approach as a way to extract important events in soap operas and adapt our method to find mode precise soap opera events labels. Specifically, we focus on human knowledge, using the results of a study which aimed to investigate whether soap opera viewers' perceptions of the likelihood of some life events differ from the non-viewers [50]. In this study, the authors select events which they believe are typically happening in soap operas (Table 2). Our assumption is that the most likely an event, the least important it is for a summary. For example, we assume that a scene depicting the event 'happily married' is less interesting for a summary than one showing a 'suicide attempt'. We therefore assign to each event a weight equal to the inverse of their perceived likelihood (on a scale from 1 to 5). As we can not assume that the evaluators are especially soap opera viewers, we choose to use the likelihood scores given by the non-viewers group. To score each shot, we multiply its confidence score from the zero-shot classifier (which we first normalize for each class using RobustScaler[12]) with the weight of the class (inverse of the perceived likelihood in Table 2). Furthermore, to avoid extracting short scenes, and therefore very few information, we further multiply this score with the log of the length of the shot transcript content (Equation 3).

$$score(shot_i) = max\_l \in labels(zsc(trans_i, l) * weight(l) * log(len(trans_i)) \quad (3)$$

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 $weight(l)$ for l in labels.

Finally, we select the top N shots for each character based on the max score on all classes as a summary. Because of the constraint on the length of the summary, if the selected shots are too long, we push out the longest scene from the top N and replace it with the N+1th one, and so on until we get a total runtime that fits the summary length requirement.

Table 2 Life events labels, their perceived likelihood for non-viewers (scale from 1 to 5 higher is more likely) and their associated weight (inverse of the likelihood) [50]

Label	Likelihood	Weight
extramarital affair	1.98	0.51
get divorced	1.96	0.51
illegitimate child	1.45	0.69
institutionalized for emotional problem	1.43	0.70
happily married	4.05	0.25
serious accident	2.96	0.34
murdered	1.81	0.55
suicide attempt	1.26	0.79
blackmailed	1.86	0.54
unfaithful spouse	2.23	0.45
sexually assaulted	2.60	0.38
abortion	1.41	0.70

5.2.3 Baselines

A first comparison we make is with the 2020 edition results presented in Table 3. Despite focusing on other characters and episodes, the two editions of the task should have a similar level of difficulty. In 2020, we had proposed a method which relied on data augmentation ("Ours_i" method on Table 3). As explained in Section 2, we had found and scraped one fandom synopsis per episode and we had assumed that each sentence of the synopsis was related to an important moment. After using the same face recognition step as presented in section 5, our approach computes the similarity between the synopsis and the episodes transcripts associated to each shot. We defined similarity as being the sum of TF-IDF weights (computed on the transcript) for each word appearing in both the synopsis and the transcript [26]. NII_UIT, the other participant team in the 2020 edition, relied purely on visual features [25]. Their importance score is an average of a face recognition (of the characters of interest) and a representation score. The later is obtained from the sequence to sequence model VASNet [52] which takes as input features extracted from GoogLeNet [53] trained on ImageNet [54].

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

Table 3 Average score for each run and team (Entail [26] and NIIUIT [25]) in TRECVID VSUM 2020

TeamRun	Percentage
Ours_1	31%
Ours_2	31%
Ours_3	**35**%
Ours_4	32%
NIIUIT_1	9%
NIIUIT_2	8%
NIIUIT_3	8%
NIIUIT_4	6%

We also compare our results with the ones obtained by the other challenge participants for the 2021 edition (Table 4). Both NII_UIT [28] and ADAPT [27] proposed a method relying on some external data (respectively Wikipedia articles and fan-written synopsis) about the BBC Eastenders show. More precisely, NII_UIT derives an importance score for each shot which is a combination of a face recognition score, a co-appearance face score, a wiki and a virtual event score. The wiki score is obtained from comparing the similarity between the transcripts and the ten most interesting sentences (manually selected) from the characters Wiki page. The virtual event score is a visual one, obtained from an EfficientNet B4 [55] network, which was trained to detect social events in video frames [56]. The ADAPT approach also relies on a face recognition step and a matching with keywords manually extracted from scraped BBC Eastenders video metadata and fan-based web sites.

In Table 4, we also include the results we obtained on a subtask of the challenge, where the evaluation questions are revealed to the participants before submission. These results allow us to assess the gain in performance when the specific important moments are known. For this subtask, we used a longformer language model pre-trained on a QA dataset (Squad-v2)[13]. NII_UIT used the same importance score they computed for their main method, to which they added a question score. This score is obtained after concatenating and embedding the questions (with the Universal Sentences Encoder [57]) to obtain a similarity score. ADAPT relied on the same approach as for the main task, only modifying the list of keywords of interest.

5.3 Results and Discussion

Table 4 shows the overall results (combining evaluation metrics and characters) for the following constraints:

[13] https://huggingface.co/mrm8488/longformer-base-4096-finetuned-squadv2

Table 4 Average score for each run and team (Entail [51], NIIUIT [28], ADAPT [27]) in TRECVID VSUM 2021

Team_Run	Main Task	Subtask
ADAPT_1	31.20%	15.60%
ADAPT_2	34.20%	11.40%
ADAPT_3	27.40%	17%
ADAPT_4	27.80%	25%
Entail_1	17.40%	32.20%
Entail_2	30.40%	31.80%
Entail_3	32.80%	30.80%
Entail_4	**37.60%**	34.60%
NII_UIT_1	7.40%	19.60%
NII_UIT_2	12.20%	22.40%
NII_UIT_3	29.60%	28.20%
NII_UIT_4	22.80%	**49.20%**

- Entail_1: 5 shots with highest scores and the total duration of the summary is <150 sec;
- Entail_2: 10 shots with highest scores and the total duration of the summary is < 300 sec;
- Entail_3: 15 shots with highest scores and the total duration of the summary is < 450 sec;
- Entail_4: 20 shots with highest scores and the total duration of the summary is < 600 sec.

Our run 4 with 20 shots reached 37.60%, the best score across all teams. Given that the approach from the ADAPT and NII_UIT teams relied both on the necessary condition of gathering fan-written material specific to the series and on manually selecting important keywords or sentences from them, the results obtained by our approach are encouraging. Indeed, despite not being obtained automatically either, our candidate labels are not specific to the episodes nor to the BBC Eastenders show but could potentially generalise to other series of the soap-opera genre. Our approach is also more minimalistic than the NII_UIT team who combined 5 different scores. We find it interesting that every team relied on different types of text sources (labels, synopsis of the episodes, wikipedia pages of the characters) to extract the most interesting shots. However, since all methods relied on some additional components, which differ between teams (similarity measures, text embeddings, face recognition pipeline), it is impossible to isolate the text source element and to conclude about their individual relevance. Harmonizing the other components would be an interesting experiment for future work. Contrary to the ADAPT team, the smaller the compression, the better our results. Interestingly, for the subtask where queries were known in advance, except from the run NII_UIT_4 which obtained 49.20%, no run obtained better results than the best run for the main task. While these results suggest that answering this type of question is still a very challenging task, it might also be that soap opera events are good enough of a proxy for a complete and informed question about the character. Table 3 shows the results obtained with the same constraints for the 2020 edition. The fact that our best results from the 2020 edition (35%), which ranked first [26] did not outperform our zero-shot classification method proposed in the 2021 edition (37.6%), despite being guided by a fandom synopsis for each episode, might be another cue speaking in favour of genre-events being a good guide to find interesting moments in soap opera episodes.

We present in Table 5 all detailed results for the TRECVID VSUM 2021 edition. In particular, the last two rows show the mean for the temporability, contextuality and redundancy metrics.

- Temporability refers to 'how well do the video shots flow together? Do shots cut mid-sentence? Do they flow together nicely so it would not be obvious that this is an automatically generated summary'.

Table 5 All results (T=Tempo, C=Context, R=Redundancy)

Team_Run_Query	T	C	R	Q1	Q2	Q3	Q4	Q5	final_score
ADAPT_1_Archie	5	3	2	Yes	No	Yes	No	Yes	62%
ADAPT_2_Archie	6	5	4	Yes	Yes	Yes	No	Yes	79%
ADAPT_3_Archie	4	6	4	No	Yes	No	No	No	30%
ADAPT_4_Archie	5	5	3	No	Yes	No	No	No	31%
Entail_1_Archie	3	4	5	No	Yes	No	No	No	26%
Entail_2_Archie	3	4	4	Yes	Yes	No	No	Yes	59%
Entail_3_Archie	3	5	5	Yes	Yes	No	No	Yes	59%
Entail _4_Archie	**3**	**5**	**4**	**Yes**	**Yes**	**No**	**No**	**Yes**	**60%**
NII_UIT_1_Archie	3	2	7	No	No	No	No	No	6%
NII_UIT_2_Archie	3	3	5	No	Yes	No	No	No	9%
NII_UIT_3_Archie	4	3	4	No	No	No	Yes	No	27%
NII_UIT_4_Archie	2	2	6	No	No	No	No	No	6%
ADAPT_1_Jack	6	5	2	No	No	No	No	No	17%
ADAPT_2_Jack	6	4	2	No	No	No	No	No	16%
ADAPT_3_Jack	5	5	4	No	No	No	Yes	No	30%
ADAPT_4_Jack	4	5	3	No	No	No	No	No	14%
Entail_1_Jack	6	3	3	No	No	No	No	No	14%
Entail_2_Jack	5	5	4	No	No	No	No	Yes	30%
Entail_3_Jack	4	4	2	No	No	No	No	Yes	30%
Entail_4_Jack	**5**	**4**	**2**	**No**	**No**	**No**	**No**	**Yes**	**31%**
NII_UIT_1_Jack	2	2	5	No	No	No	No	No	7%
NII_UIT_2_Jack	3	2	6	No	No	No	No	No	7%
NII_UIT_3_Jack	4	3	5	No	No	No	Yes	No	26%
NII_UIT_4_Jack	6	4	4	No	No	No	Yes	No	30%
ADAPT_1_Max	3	3	3	No	Yes	No	No	No	27%
ADAPT_2_Max	2	3	5	No	No	No	No	No	8%
ADAPT_3_Max	2	4	4	No	No	No	No	No	8%
ADAPT_4_Max	3	3	4	No	No	No	No	No	10%
Entail_1_Max	4	3	3	No	No	No	No	No	12%
Entail_2_Max	4	3	3	No	No	Yes	No	No	28%
Entail_3_Max	4	3	3	No	Yes	Yes	No	No	44%
Entail_4_Max	**4**	**3**	**4**	**No**	**Yes**	**Yes**	**No**	**No**	**43%**
NII_UIT_1_Max	3	3	4	No	No	No	No	No	10%
NII_UIT_2_Max	3	3	4	No	No	No	No	No	10%
NII_UIT_3_Max	3	3	4	No	Yes	No	No	No	26%
NII_UIT_4_Max	3	3	4	No	Yes	No	No	No	26%
ADAPT_1_Peggy	2	3	3	No	Yes	No	No	No	26%
ADAPT_2_Peggy	2	3	3	No	Yes	No	No	No	26%
ADAPT_3_Peggy	2	3	4	No	No	Yes	No	No	25%
ADAPT_4_Peggy	2	3	3	No	No	Yes	No	Yes	42%
Entail_1_Peggy	3	3	3	No	No	No	No	No	11%
Entail_2_Peggy	3	3	4	No	No	No	No	Yes	10%
Entail_3_Peggy	3	3	5	No	No	No	No	Yes	9%
Entail_4_Peggy	**3**	**3**	**4**	**No**	**No**	**No**	**No**	**Yes**	**10%**
NII_UIT_1_Peggy	2	3	3	No	No	No	No	No	10%
NII_UIT_2_Peggy	3	3	4	No	No	No	No	No	10%
NII_UIT_3_Peggy	3	3	4	No	No	Yes	No	No	26%
NII_UIT_4_Peggy	2	3	4	No	No	No	No	No	9%
ADAPT_1_Tanya	3	2	5	No	Yes	No	No	No	24%
ADAPT_2_Tanya	4	4	5	No	No	No	Yes	Yes	43%
ADAPT_3_Tanya	4	4	4	No	Yes	Yes	No	No	44%
ADAPT_4_Tanya	3	4	5	No	Yes	No	No	Yes	42%
Entail_1_Tanya	4	2	6	Yes	No	No	No	No	24%
Entail_2_Tanya	2	4	5	Yes	No	No	No	No	25%
Entail_3_Tanya	2	2	6	Yes	No	No	No	No	22%
Entail_4_Tanya	**5**	**4**	**5**	**Yes**	**Yes**	**No**	**No**	**No**	**44%**
NII_UIT_1_Tanya	2	1	7	No	No	No	No	No	4%
NII_UIT_2_Tanya	3	3	5	No	Yes	No	No	No	25%
NII_UIT_3_Tanya	4	4	5	No	Yes	Yes	No	No	43%
NII_UIT_4_Tanya	4	4	5	No	Yes	Yes	No	No	43%
Mean	**3.47**	**3.4**	**4.12**						
Entail_mean	**3.65**	**3.5**	**4**						

- Contextuality is defined as: 'Does the content provide the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed?'.
- Redundancy is defined as: 'Does the video contain content considered to be unnecessary or superfluous?' [44].

Our results across characters and runs are above the mean except for redundancy for which we are slightly under the mean. In general, the results obtained by all teams across all evaluation metrics show that the task remains a challenging one.

Table 6 Evaluation questions used by assessors in TRECVID VSUM 2021. The questions our model was capable to answer are in bold.

Archie:
What happens when Phil throws Archie in to a pit?
What happens after Danielle reveals to Archie that Ronnie is her mother?
Where do Peggy and Archie get married?
What happens when Archie arrives at the pub after Peggy invited him?
What happens when Archie is kidnapped?
Jack:
What happens when police break in the door of Jack and Tanya's home?
Where are Max and Jack during the violent confrontation between them when a gun is drawn?
Who does Jack offer to pay in order to withdraw their statement to the police?
Why is Jack a suspect in the hit and run on Max?
What does Jack reveal to Tanya about his dodgy past?
Max:
What were the cause of Max's serious injuries which left him in hospital?
What is/was the relationship between Max and Tanya?
What kind of weapon does Max obtain from Phil?
Where are Max and Jack during the violent confrontation between them when a gun is drawn?
Who is responsible, or who does Max believe is responsible, for the serious injuries which left him in hospital?
Peggy:
Who does Peggy ask to kill Archie?
Where do Peggy and Archie get married?
Show one of the challenges which Peggy faces in her election run.
What does Peggy overhear Archie saying, which causes their marriage to be over?
What is Janine doing to irritate or anger Peggy?
Tanya:
What does Tanya reveal to the police while being interviewed at the station?
What is/was the relationship between Max and Tanya?
What does Jack reveal to Tanya about his dodgy past?
What does Tanya discover in the sink and on Jack's clothes?
What big move were Tanya and Jack planning for the future?

In Table 6, we report the evaluation questions for each character. The questions for which our best model (run 4) was able to answer are marked in bold. We observe that the majority of the evaluation questions are 'What' questions (16 out of 25), most of them being about events. From these 16

questions, our approach is able to answer 9 of them. On the contrary, our system did not allow to answer any of the other types of questions, namely the 'Who', 'Why', 'Where' questions and one instruction. These results both speak in favour of events/actions as being the first important aspect of a summary but also suggest that our model would benefit from covering other aspects such as locations and persons, which might also be genre specific. It is difficult to draw conclusions regarding the type of events that our system is able to capture. However, the questions provide with interesting cues about typical events happening in soap operas. The events we used for candidate labels are all related to love or violence. This is also mainly the case for the events in the evaluation questions. We find a marriage, a kidnapping, a police break, injuries leading to an hospitalization, an attempt murder, etc. These questions cover some events that were not exactly in our candidate list but are nevertheless quite close to them. If we link these results with the ones obtained for the thriller genre where 'killing' was the main event, we can see that despite the fact that romance and thriller are the most distinctive genres according to the semantic parsing experiments [33], in both cases, stories are all about violence and (a little bit of) love. The results are interesting because they point towards a new research question: are love and violent events always the most interesting ones for narratives across genre?

6 Discussion

6.1 Limitations

In this paper, we realised that while the semantic frame extraction step was useful for the experiments on the crime series, it did not generalise well to the soap opera genre. Instead, for that genre, we used a list of events we found in the literature [50]. The fact that our approach is yet missing an automated way to obtain these named events candidates probably constitutes its current biggest limitation and a direction for future work. Our error analysis also suggested that semantic similarity with the candidate labels is not always a sufficient condition to be included in the summary. As suggested by the increase in performance when averaging the ZSC and SUMMER scores, our method would probably benefit in integrating a step which aims to minimise the redundancy of our summaries, for example by computing a centrality score *a la* SUMMER. Then, in this work, we have dealt with two well-defined genres. However, as TV series become more complex, we would need to evaluate our approach on series which genre is not as a clear cut as in the CSI and BBC Eastenders episodes. Finally, in terms of evaluation, more automatic metrics could be investigated in the future. In a classification setting, with a F1 score, a scene is either considered as being important or not. However, it might be worth introducing some nuances. Let's consider two scenes which are not included in the ground truth summary: one, despite not being the one scene chosen in the summary, relate to important event mentioned in the summary,

while the other scene is not linked to any important events. An appropriate evaluation should probably account for such a difference. For example, we have seen that the VSUM TRECVID human evaluation, instead of using some specific scenes as ground truth, interrogates whether the selected scenes can answer a specific set of questions. As such, in the future, we plan to evaluate our summaries with automatic metrics which assess question-answering capabilities [58]. Following [59], another direction could be to draw inspiration from the text summarization field in which the content similarity between a generated and a ground-truth summary is usually measured. Popular similarity metrics include N-gram based metrics such as ROUGE [60], BLEU [61] or METEOR [62] or neural ones such as BERTscore [63].

6.2 Generalisation to Other Genres

A limitation of our work is that we only tested it on two different genres. However, we expect that our method can generalise to other genres. First, we believe that one strength of our approach is the flexibility it offers, to adapt to the particular application and need by adapting the keywords (as we did for the soap opera experiment). We give guidelines on how to find these themes and we think they could serve as a stepping stone for other genres.

- **Generalisation capabilities to other known genres**: We do not expect our method to perform worse when applied to other genres, as we did not find reasons or papers supporting that soap operas and crime are less complex than other genres. On the contrary, methods that aim to automatically predict the genre of movies obtain the best scores for Western and War movies, which suggests that other (if not all) genres have markers that can be detected given a proper understanding of them. In fact, we suspect that genres such as Horror and Action have recurrent themes that can also be used to guide our method.
- **Generalisation capabilities to arbitrary genres**: Most media content comes with metadata concerning its genre, and in the absence of this, a model to predict the genre can be used [64, 65]).
- **Generalisation capabilities to mixed genres**: In this paper, we concluded that across both genres, important terms seem to be dramatic events about love and violence (hence, the title of the article). The specific events expected in a TV series most likely differ according to the genre but in future work, deriving a complete list of the 'dramatic events' across genres could help generalisation capabilities. Training a 'dramatic event classifier' is also a possibility. We also observe that people seem to believe that movies tend to be predictable across genres[14]. Some movies seem to follow the same narrative arcs, even if from very different genres[15] (e.g. Dances with Wolves,

[14] https://www.quora.com/Have-movie-plots-gotten-more-predictable-and-less-developed-or-are-humans-just-so-much-smarter-now-than-50-years-ago-that-we-anticipate-the-ending-and-story-development-quicker

[15] https://www.businessinsider.com/movies-with-the-same-plot-2013-4

Avatar). There are then reasons to believe that using event labels, even with mixed genres, could be relevant.

7 Conclusion and Future Work

We have proposed a new method for unsupervised summarization, and we have demonstrated the effectiveness of zero-shot classification with events representative of a genre as candidate labels for crime series and soap operas. When provided with a screenplay, we were able to observe that the Entail model performs best when handling only visual information data. We think our approach is helping to push interpretability: contrary to modelling interestingness without proxies, this approach allows to justify the choice of summary scenes by their closeness to non subjective labels. Another major strength of this approach is its flexibility. Realising that video summarization is subjective, some recent work are interested in producing personalised query based video summaries [66]. We have provided all the code to reproduce the results presented in this paper at https://github.com/alisonreboud/screenplay_summarization and https://github.com/MeMAD-project/trecvid-vsum.

In the future, we would like to be able to test how zero-shot classification performs when a user is interested in extracting emotionally interesting scenes or other different concepts related to interestingness. The Entail model is also especially interesting for testing query-based approaches as the pre-training of the model with an 'hypothesis' sentence offer possibilities that go way beyond the sentence we used for classification. The fact that for soap opera and crime, which are two very different genres, important moments were about dramatic events, makes us wonder if an approach based on classification of dramatic events could perform well across genres. While trying to design an approach to find events candidates, we realised that there is a gap in the literature when it comes to classifying events between dramatic and trivial or describing the most common events of a movie genre. In a future work, we plan to close this gap, potentially by relying on human annotation.

Acknowledgement

This work has been partially supported by the European Union's Horizon 2020 research and innovation program within the MeMAD (grant agreement No. 780069) project, and by the French National Research Agency (ANR) within the kFLOW project (Grant n°ANR-21-CE23-0028).

A Appendix

This annex complements Figure 2 in presenting the scenes that were wrongly labeled by ZeSTE for season 4 episode 22, when using the label 'death'. The words that are identified by ZeSTE as being the more closely related to the label are highlighted.

A.1 Examples of false positives on the CSI dataset

Scene 4

(ROBBINS examines ERNIE MENLO'S feet as WARRICK watches.)
Robbins: for what it's worth, these bruises correspond to the holes in his sock
(ROBBINS walks around the body toward the head-side of the table.)
Warrick: well, he's been worked over pretty good. He's got a nice fat lip. Robbins: yeah. there was a good clot in the wound, and the tissues were contused. I'd say it occurred at least an hour or two before death. I teased out a couple of small-caliber projectiles from his brain.
(He holds out the bullets.) (He gives them to WARRICK.)
Robbins: one was embedded in the right frontal cortex. The other lodged in the first cervical vertebra' Warrick: it's copper-washed lead. must be a .22.Robbins: you know, historically, .22s were the hit man's bullet of choice
(Quick CGI POV: The gun shot sounds and the bullet hits the brain and swishes around inside.)
Robbins: they have the energy to enter into the cranial vault, but not enough to exit, so they just ricochet around inside, shredding the gray matter until they stop.
(End of CGI POV.)(Resume to present.)
Warrick: nice.Robbins: there's also extensive crush injury to both hands, with fractures of the metacarpals and the phalanges
(ROBBINS picks up the body's wrist to show WARRICK the knuckles)
Robbins: bruises appear perimortem. Warrick: any idea what might have caused that kind of damage?
(ROBBINS indicates the x-rays up on the view box behind WARRICK who turns around to look at them.)
Robbins given the fracture pattern, i'd guess it was some sort of blunt object. Warrick: maybe a ball peen hammer. Robbins: what gets you to that? Warrick: they used to tell me back in the days, the first time you got caught cheating, they'd give you a couple whacks on the hand with a ball peen hammer. Robbins: ow Warrick: the second time, you'd lose a limb. Robbins: third time? Warrick: a long walk in the desert with a shovel.

Scene 6

(BRASS and GRISSOM interview SAM BRAUN with his LAWYER next to him.)
Lawyer: as much as my client appreciates your flair for the dramatic, the show's over, gentlemen. what do you have? Grissom: the tire patterns at the scene of teddy keller's murder are consistent with the wheel base and turning radius of your client's limousine Lawyer: as well as every other limo in vegas. we also found neon glass embedded in all four tires Lawyer: the whole town's a construction site Lawyer: it's a tenuous link, at best. well, then ... how did his blood end up in the back of your client's limousine?
(THE LAWYER doesn't say anything.)
Brass: you waited until teddy cleared the security cameras
(Quick flashback to: ['BACKSEAT'OF'LIMO'] The door opens and SAM pulls TEDDY KELLER into the back seat with him.) (The door shuts behind him.) (SAM back hands TEDDY KELLER in the face causing his nose to bleed.)(A large splotch of blood falls to the seat.)
Sam'Braun: we're not through talking, kid
(End of flashback.) (Resume to present.)
Brass: and then you took him for a ride... vegas style. Just like the old days, huh

(Quick flashback to: ['NEON'GRAVEYARD'-'NIGHT'](They pull TEDDY KELLER out of the limo's backseat.)
Teddy'Keller: please, please, let me go
(The limo door shuts and they lead TEDDY along the neon graveyard.)
Teddy'Keller: please. please
(TEDDY'S shirt is opened and the fat suit he's wearing is revealed.)
Sam'Braun: let me show you what i do to cheaters Teddy'Keller: no, no !
(They pull TEDDY over to the sign and he shoots him twice in the back of the head.) (End of flashback.) (Resume to present.) (SAM leans over and whispers something to his LAWYER.) (When done, the LAWYER turns, looks, and smiles at BRASS and GRISSOM.)
Lawyer: my client offered the young man a ride home Lawyer: they stopped briefly at the neon graveyard, where they held a private conversation regarding the ethics of defrauding a casino
(BRASS chuckles.)
Brass: that must have been some chat Brass: we know he left the casino with the money Lawyer: the young man returned the money as a sign of respect for my client and his position in the community. Brass: i'm sure he did
(GRISSOM and SAM stare at each other.)
Brass: so, what next? you gonna tell me you're being set up? it happens to you a lot, huh, sam?

Scene 9

(SARA, WARRICK and DAVID PHILLIP work on the victim's body.) (Camera view down on ERNIE MENLO'S body on the autopsy table.) (He's still in the fat suit.) (SARA picks up something off of his forehead and puts it in a clean envelope.)(DAVID PHILLIPS puts the victim's clothes in a package.) (WARRICK works on the victim's lacerated hand.) (SARA removes the rolex watch.) (She looks at it.)(It's 9:41 am.) Sara: no ticks.it's authentic (She flips the watch over and looks at the back.) Sara: logo sticker is n't worn down. watch could be new. Warrick: guy hits the jackpot, has to celebrate. goes and buys some bling-bling to impress the strippers with (DAVID lifts up the body's foot and sees the holes in the sock's heels.) Warrick: what have you got?
(WARRICK and SARA both look at the feet.)
Warrick: air conditioned socks'

A.2 Examples of false negatives on the CSI dataset

Scene 12

(Camera swoops down to show ERNIE MENLO'S dead body at the base of the'W' in the WHISKEYTOWN letter sign.) (BRASS, GRISSOM and CATHERINE stand around the body.)
Brass: two shots to the back of the head Brass: double tap
(GRISSOM shines his flashlight on the wound at the back of the victim's neck.)
Grissom: he's wearing a wig and a fat suit. it's not halloween, is it? Catherine: in this town, it's always halloween
(BRASS picks up the NEVADA DRIVER'S LICENSE.) (It reads:)
Brass:'ernie menlo' Brass: well, he was n't carrying a very'fat' wad Catherine: rolex is still on his wrist Catherine: probably rules out robbery Catherine: what do you think? Grissom: i do n't know
(GRISSOM turns around and looks at the various signs abandoned and thrown away littering the area.)

Grissom: i'm looking for a sign

Scene 16

(DAVID HODGES explains the composition of the glass as GRISSOM looks through the scope at the shards.)

David˙Hodges: the glass fragments you found at the apartment building are primarily lead-based. Different curvatures and textures with traces of florescent powder, phosphorous and mercury. Grissom: neon glass David˙Hodges: i checked out that graveyard once. David˙Hodges: pretty interesting. Grissom: the comparison? David˙Hodges: your sample's consistent with the glass collected from the first crime scene. Grissom: see? that connects the two murders. we've got a timeline.

Scene 32

(Sirens wail in the distance.) (ERNIE MENLO sits in the chair in the center of the darkened room.) (In front of him stands two men - one holding a bright light on him, the other interrogates him.)

Interrogator: I'm going to make this really simple. who are you working with? Ernie˙Menlo: i'm, uh, unemployed at the moment. Interrogator: you got any idea what we did to chumps like you back in the day? Ernie˙Menlo: uh, no. look, could you put the a.c. on in here or somethin'? that, or just, uh, let me go. i mean, you ca n't keep me in here. it's against the law Interrogator: there's no law in this room

(He looks at both his interrogators.)

Ernie˙Menlo: you can't touch me

A.3 Error Analysis

Analyzing the scenes which were wrongly selected for the summary of episode 22 season 4 with our method, we see that we have the same type of error as observed in Figure 2: the error does not come from the classification itself. Namely, scenes 4 and 9 are autopsy scenes containing words such as 'body', 'autopsy', 'victim', 'bullet or 'brain'. Similarly, scene 6 is an interrogation which gives information about a murder. The scenes are, hence, strongly associated to the label 'death'. However, the relation with the label is not a sufficient condition for the scene to be selected in the summary: autopsy or interrogations scenes seem to be quite common in the CSI episodes but a summary should only include the most relevant scenes for the plot.

Analyzing the three summary scenes which were not retrieved by our method, we can see that only scene 32 is not semantically close to the label 'death'. It is worth noting that this particular scene was retrieved by the SUMMER method as well as by the method which averages the ZeSTE and SUMMER scores (the other two scenes were not). While we can assume that a reason to have included scene 16 in the summary might be its last utterance ('see? that connects the two murders. we've got a timeline'), which reveals that the plot is about two connected murders, we find it generally difficult, when considering the scenes independently, to justify why these scenes are more relevant than the false positive ones. This speaks in favour of an approach, such as SUMMER, which rather put scenes in a more global perspective, computing

a centrality score for each scene. It might also partly explain why, on average, we obtain better results when averaging our event classification scores with SUMMER scores.

References

[1] Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Video Summarization Using Deep Neural Networks: A Survey. Proceedings of the IEEE **109**(11), 1838–1863 (2021)

[2] Sreeja, M., Kovoor, B.C.: Towards genre-specific frameworks for video summarisation: A survey. Journal of Visual Communication and Image Representation **62**, 340–358 (2019)

[3] Kryściński, W., Keskar, N.S., McCann, B., Xiong, C., Socher, R.: Neural Text Summarization: A Critical Evaluation. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 540–551. ACL, Hong Kong, China (2019)

[4] Zhang, H., Liu, H.: Visualizing structural "inverted pyramids" in english news discourse across levels. Text & Talk **36**(1), 89–110 (2016)

[5] Altmami, N.I., Menai, M.E.B.: Automatic summarization of scientific articles: A survey. Journal of King Saud University-Computer and Information Sciences (2020)

[6] Frermann, L., Cohen, S.B., Lapata, M.: Whodunnit? Crime Drama as a Case for Natural Language Understanding. Transactions of the Association for Computational Linguistics **6**, 1–15 (2018)

[7] Papalampidi, P., Keller, F., Frermann, L., Lapata, M.: Screenplay Summarization Using Latent Narrative Structure. In: 58th Annual Meeting of the Association for Computational Linguistics, pp. 1920–1933. ACL, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.174

[8] Rochan, M., Wang, Y.: Video summarization by learning from unpaired data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7894–7903 (2019). https://doi.org/10.1109/CVPR.2019.00809

[9] Bazin, A.: What Is Cinema? Volume I, 1st edn. University of California Press, Oakland, CA, USA (2005). http://www.jstor.org/stable/10.1525/j.ctt5hjhmc

[10] Ben-Ahmed, O., Huet, B.: Deep Multimodal Features for Movie Genre and Interestingness Prediction. In: International Conference on Content-Based

Multimedia Indexing (CBMI), pp. 1–6 (2018)

[11] Yin, W., Hay, J., Roth, D.: Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3914–3923. ACL, Hong Kong, China (2019)

[12] Zhou, K., Xiang, T., Cavallaro, A.: Video Summarisation by Classification with Deep Reinforcement Learning. British Machine Vision Conference (BMVC) (2018)

[13] Lei, J., Luan, Q., Song, X., Liu, X., Tao, D., Song, M.: Action Parsing-Driven Video Summarization Based on Reinforcement Learning. IEEE Transactions on Circuits and Systems for Video Technology **29**(7), 2126–2137 (2018)

[14] You, J., Liu, G., Sun, L., Li, H.: A Multiple Visual Models Based Perceptive Analysis Framework for Multilevel Video Summarization. IEEE Transactions on Circuits and Systems for Video Technology **17**(3), 273–285 (2007)

[15] Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., Avrithis, Y.: Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention. IEEE Transactions on Multimedia **15**(7), 1553–1568 (2013)

[16] Demarty, C.-H., Sjöberg, M., Ionescu, B., Do, T.-T., Gygli, M., Duong, N.: MediaEval 2017 Predicting Media Interestingness Task. In: MediaEval Workshop (2017)

[17] Hesham, M., Hani, B., Fouad, N., Amer, E.: Smart Trailer: Automatic generation of movie trailer using only subtitles. In: First International Workshop on Deep and Representation Learning (IWDRL), pp. 26–30 (2018). IEEE

[18] Xu, M., Jin, J.S., Luo, S., Duan, L.: Hierarchical Movie Affective Content Analysis Based on Arousal and Valence Features. In: 16th ACM International Conference on Multimedia. MM '08, pp. 677–680. Association for Computing Machinery, New York, NY, USA (2008)

[19] Haq, I.U., Muhammad, K., Hussain, T., Del Ser, J., Sajjad, M., Baik, S.W.: QuickLook: Movie Summarization Using Scene-Based Leading Characters with Psychological Cues Fusion. Information Fusion **76**, 24–35 (2021)

[20] Sang, J., Xu, C.: Character-Based Movie Summarization. In: 18th ACM

International Conference on Multimedia, pp. 855–858 (2010)

[21] Bost, X., Gueye, S., Labatut, V., Larson, M., Linarès, G., Malinas, D., Roth, R.: Remembering winter was coming. Multimedia Tools and Applications **78**(24), 35373–35399 (2019). https://doi.org/10.1007/s11042-019-07969-4

[22] Papalampidi, P., Keller, F., Lapata, M.: Movie Plot Analysis via Turning Point Identification. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1707–1717. ACL, Hong Kong, China (2019)

[23] Papalampidi, P., Keller, F., Lapata, M.: Movie Summarization via Sparse Graph Construction. In: Thirty-Fifth AAAI Conference on Artificial Intelligence (2020). AAAI Press

[24] Lee, M., Kwon, H., Shin, J., Lee, W., Jung, B., Lee, J.-H.: Transformer-based Screenplay Summarization Using Augmented Learning Representation with Dialogue Information. In: Third Workshop on Narrative Understanding, pp. 56–61. ACL, Online (2021). https://aclanthology.org/2021.nuse-1.6

[25] Le, D., Vo, H., Nguyen, T., Do, T., Pham, T., Vo, T., Nguyen, T., Nguyen, V., Ngo, T.: NII˙UIT AT TRECVID 2020. In: TRECVID 2020 Workshop (2020)

[26] Harrando, I., Reboud, A., Lisena, P., Troncy, R., Laaksonen, J., Virkkunen, A., Kurimo, M., *et al.*: Using Fan-Made Content, Subtitles and Face Recognition for Character-Centric Video Summarization. In: TRECVID 2020 Workshop (2020)

[27] Potyagalova, A., Jones, G.J.F.: DCU ADAPT at TRECVID 2021: Video Summarization - Keeping It Simple. In: TRECVID 2021 Workshop (2021)

[28] Tran, K.D., Quang, N.P.L., Do.T, Mai.T, Truong, A.P.N.: NII˙UIT AT TRECVID 2021: Video Summarization Task. In: TRECVID 2021 Workshop (2021)

[29] Aparício, M., Figueiredo, P., Raposo, F., de Matos, D.M., Ribeiro, R., Marujo, L.: Summarization of Films and Documentaries Based on Subtitles and Scripts. Pattern Recognition Letters **73**, 7–12 (2016)

[30] Yu, T., Liu, Z., Fung, P.: AdaptSum: Towards Low-Resource Domain Adaptation for Abstractive Summarization. In: 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pp. 5892–5904. ACL,

Online (2021)

[31] Chang, B., Lee, I., Kim, H., Kang, J.: "Killing Me" Is Not a Spoiler: Spoiler Detection Model using Graph Neural Networks with Dependency Relation-Aware Attention Mechanism. In: EACL, pp. 3613–3617 (2021)

[32] Jeon, S., Kim, S., Yu, H.: Spoiler detection in TV program tweets. Information Sciences **329**, 220–235 (2016)

[33] Chang, B., Kim, H., Kim, R., Kim, D., Kang, J.: A Deep Neural Spoiler Detection Model Using a Genre-Aware Attention Mechanism. In: 22nd Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD), pp. 183–195 (2018)

[34] Zhang, J., Lertvittayakumjorn, P., Guo, Y.: Integrating Semantic Knowledge to Tackle Zero-shot Text Classification. In: NAACL-HLT (2019)

[35] Ye, Z., Geng, Y., Chen, J., Chen, J., Xu, X., Zheng, S., Wang, F., Zhang, J., Chen, H.: Zero-shot Text Classification via Reinforced Self-training. In: ACL (2020)

[36] Weller, O., Lourie, N., Gardner, M., Peters, M.: Learning from task descriptions. In: EMNLP, pp. 1361–1375 (2020)

[37] Harrando, I., Troncy, R.: Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph. In: 3^{rd} Conference on Language, Data and Knowledge (LDK), Zaragoza, Spain (2021)

[38] Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In: Thirty-First AAAI Conference on Artificial Intelligence, pp. 4444–4451. AAAI Press, ??? (2017)

[39] Liu, P.J., Chung, Y.-a., Ren, J.J.: SummAE: Zero-Shot Abstractive Text Summarization using Length-Agnostic Auto-Encoders (2019)

[40] Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1, pp. 86–90. ACL, Montreal, Quebec, Canada (1998). https://doi.org/10.3115/980845.980860

[41] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL, pp. 4171–4186. ACL, Minneapolis, MN, USA (2019)

[42] Williams, A., Nangia, N., Bowman, S.: A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In:

NAACL, pp. 1112–1122. ACL, New Orleans, Louisiana, USA (2018). https://aclanthology.org/N18-1101

[43] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: ACL, pp. 7871–7880 (2020)

[44] Awad, G., Butt, A.A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., Delgado, A., Zhang, J., Godard, E., Diduch, L., Liu, J., Smeaton, A.F., Graham, Y., Jones, G.J.F., Kraaij, W., Quénot, G.: TRECVID 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In: TRECVID 2020 Workshop. NIST, Gaithersburg, MD, USA (2020)

[45] Lisena, P., Laaksonen, J., Troncy, R.: FaceRec: An Interactive Framework for Face Recognition in Video Archives. In: ACM (ed.) 2nd International Workshop on Data-driven Personalisation of Television (DataTV) (2021)

[46] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE signal processing letters **23**(10), 1499–1503 (2016)

[47] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823. IEEE Computer Society, Boston, MA, USA (2015)

[48] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple Online and Realtime Tracking. In: IEEE International Conference on Image Processing (ICIP), pp. 3464–3468. IEEE Computer Society, Phoenix, AZ, USA (2016)

[49] Liu, C., Shmilovici, A., Last, M.: Towards story-based classification of movie scenes. PloS one **15**(2), 0228579 (2020)

[50] Seese, G.: Soap opera viewers' perceptions of the real world. Master's thesis, University of Central Florida (1987)

[51] Reboud, A., Harrando, I., Lisena, P., Troncy, R.: Zero-Shot Classification of Events for Character-Centric Video Summarization. In: TRECVID 2021 Workshop (2021)

[52] Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., Remagnino, P.: Summarizing Videos with Attention. In: Asian Conference on Computer Vision, pp. 39–54 (2018). Springer

[53] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

[54] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., *et al.*: ImageNet Large Scale Visual Recognition Challenge. ImageNet Large Scale Visual Recognition Challenge **115**(3), 211–252 (2015)

[55] Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[56] Ahmad, K., Conci, N., Boato, G., De Natale, F.G.: USED: a large-scale social event detection dataset. In: 7th International Conference on Multimedia Systems, pp. 1–6 (2016)

[57] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., *et al.*: Universal sentence encoder for English. In: Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 169–174 (2018)

[58] Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., Gallinari, P.: QuestEval: Summarization asks for fact-based evaluation. In: Conference on Empirical Methods in Natural Language Processing, pp. 6594–6604. ACL, Punta Cana, Dominican Republic (2021)

[59] Plummer, B.A., Brown, M., Lazebnik, S.: Enhancing video summarization via vision-language embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5781–5789 (2017)

[60] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-1013

[61] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. ACL, USA (2002)

[62] Lavie, A., Agarwal, A.: METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Second Workshop on Statistical Machine Translation, pp. 228–231 (2007)

[63] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: 8th International Conference on Learning Representations (ICLR). OpenReview.net, Addis Ababa, Ethiopia (2020). https://openreview.net/forum?id=SkeHuCVFDr

[64] Matthews, P., Glitre, K.: Genre analysis of movies using a topic model of plot summaries. Journal of the Association for Information Science and Technology **72**(12), 1511–1527 (2021)

[65] Hoang, Q.: Predicting Movie Genres Based on Plot Summaries. CoRR **abs/1801.04813** (2018). https://doi.org/10.48550/arXiv.1801.04813

[66] Huang, J.-H., Worring, M.: Query-controllable Video Summarization. In: 2020 International Conference on Multimedia Retrieval. ICMR '20, pp. 242–250. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3372278.3390695