Incorporating Background Knowledge into Video Description Generation

Spencer Whitehead[1], **Heng Ji**[1], **Mohit Bansal**[2], **Shih-Fu Chang**[3], **Clare R. Voss**[4]
[1] Computer Science Department, Rensselaer Polytechnic Institute
{whites5,jih}@rpi.edu
[2] Computer Science Department, University of North Carolina at Chapel Hill
mbansal@cs.unc.edu
[3] Department of Electrical Engineering, Columbia University
sfchang@ee.columbia.edu
[4]US Army Research Laboratory
clare.r.voss.civ@mail.mil

Abstract

Most previous efforts toward video captioning focus on generating generic descriptions, such as, "A man is talking." We collect a news video dataset to generate enriched descriptions that include important background knowledge, such as named entities and related events, which allows the user to fully understand the video content. We develop an approach that uses video meta-data to retrieve topically related news documents for a video and extracts the events and named entities from these documents. Then, given the video as well as the extracted events and entities, we generate a description using a *Knowledge-aware Video Description* network. The model learns to incorporate entities found in the topically related documents into the description via an *entity pointer network* and the generation procedure is guided by the event and entity types from the topically related documents through a *knowledge gate*, which is a gating mechanism added to the model's decoder that takes a one-hot vector of these types. We evaluate our approach on the new dataset of news videos we have collected, establishing the first benchmark for this dataset as well as proposing a new metric to evaluate these descriptions.

1 Introduction

Video captioning is a challenging task that seeks to automatically generate a natural language description of the content of a video. Many video captioning efforts focus on learning video representations that model the spatial and temporal dynamics of the videos (Yao et al., 2015; Venugopalan et al., 2016; Yu et al., 2017). Although the language generation component within this task is of great importance, less work has been done to enhance the contextual knowledge conveyed by the descriptions. The descriptions generated by previous methods tend to be "generic", describing only what is evidently visible and lacking specific knowledge, like named entities and event participants, as shown in Figure 1a. In many situations, however, generic descriptions are uninformative as they do not provide contextual knowledge. For example, in Figure 1b, details such as *who is speaking* or *why they are speaking* are imperative to truly understanding the video, since contextual knowledge gives the surrounding circumstances or cause of the depicted events.

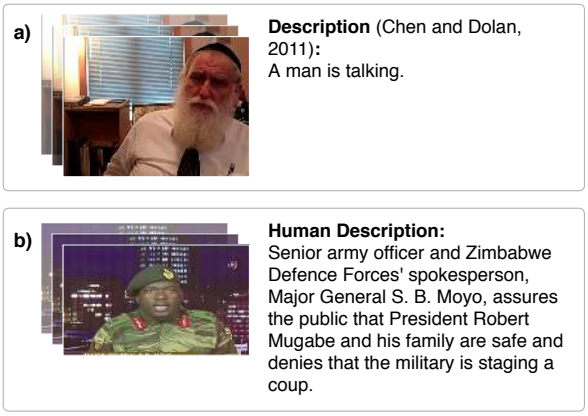

Figure 1: Comparison of machine (a) and human (b) generated descriptions.[1]

To address this problem, we collect a news video dataset, where each video is accompanied by meta-data (*e.g.*, tags and date) and a natural language description of the content in, and/or context around, the video. We create an approach to this task that is motivated by two observations.

First, the video content alone is insufficient to generate the description. Named entities or specific events are necessary to identify the participants, location, and/or cause of the video content. Although knowledge could potentially be mined from visual evidence (*e.g.*, recognizing the location), training such a system is exceedingly diffi-

[1](a) https://goo.gl/2StcD8, (b) https://goo.gl/VFR5nw

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3992–4001
Brussels, Belgium, October 31 - November 4, 2018.

cult (Tran et al., 2016). Further, not all the knowledge necessary for the description may appear in the video. In Figure 2a, the video depicts much of the description content, but knowledge of the speaker ("*Carles Puigdemont*") is unavailable if limited to the visual evidence because the speaker never appears in the video, making it intractable to incorporate this knowledge into the description.

Second, one may use a video's meta-data to retrieve topically related news documents that contain the named entities or events that appear in the video's description, but these may not be specific to the video content. For example, in Figure 2b, the video discusses the "*heightened security*" and does not depict the arrest directly. Topically related news documents capture background knowledge about the attack that led to the "*heightened security*" as well as the arrest, but they may not describe the actual video content, which displays some of the increased security measures.

Thus, we propose to retrieve topically related news documents from which we seek to extract named entities (Pan et al., 2017) and events (Li et al., 2013) likely relevant to the video. We then propose to use this knowledge in the generation process through an *entity pointer network*, which learns to dynamically incorporate extracted entities into the description, and through a new *knowledge gate*, which conditions the generator on the extracted event and entity types. We include the video content in the generation by learning video representations using a spatio-temporal hierarchical attention that spatially attends to regions of each frame and temporally attends to different frames. We call the combination of these generation components the *Knowledge-aware Video Description* (KaVD) network. The contributions of this paper are as follows:

- We create a knowledge-rich video captioning dataset, which can serve as a new benchmark for future work.
- We propose a new *Knowledge-aware Video Description* network that can generate descriptions using the video and background knowledge mined from topically related documents.
- We present a knowledge reconstruction based metric, using entity and event F1 scores, to evaluate the correctness of the knowledge conveyed in the generated descriptions.

Figure 2: Examples from the news video dataset (video, meta-data, and description) with some retrieved topically related documents.[2]

2 Approach

Figure 3 shows our overall approach. We first retrieve topically related news documents using tags from the video meta-data. Next, we apply entity discovery and linking as well as event extraction methods to the documents, which yields a set of entities and events relevant to the video. We represent this background knowledge in two ways: 1) we encode the entities through entity embeddings and 2) we encode the event and entity typing information into a *knowledge gate vector*, which is a one-hot vector where each entry represents an entity or event type. Finally, with the video and these representations of the background knowledge, we employ our KaVD network, an encoder-decoder (Cho et al., 2014) style model, to generate the description.

2.1 Document Retrieval and Knowledge Extraction

We gather topically related news documents as a source of background knowledge using the video meta-data. For each video, we use the corresponding tags to perform a keyword search on documents from a number of popular news outlet web-

[2](a) https://goo.gl/3cF1oU, (b) https://goo.gl/NkwHvJ

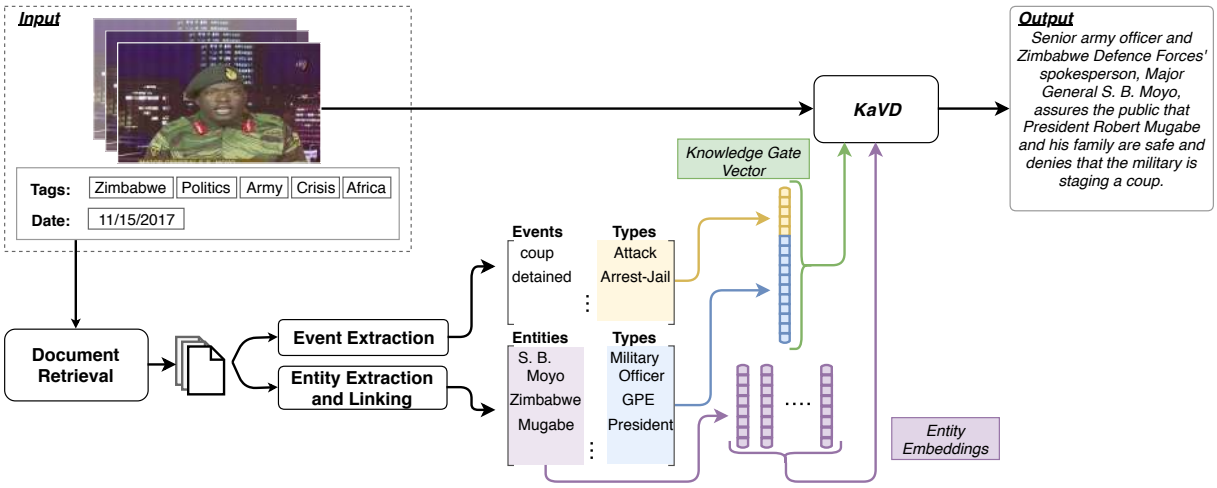

Figure 3: Overall pipeline of our approach.

sites.[3] We filter these documents by the date associated with video, only keeping documents that are written within d days before and after the video upload date.[4] The keyword search gathers documents that are at least somewhat topically relevant and filtering by date increases the likelihood that the documents reference the specific events and entities of the video, since the occurrences of entity and event mentions across news documents tend to be temporally correlated. We retrieve an average of 3.1 articles per video and find that on average 68.8% of the event types and 70.6% of the entities in the ground truth description also appear in corresponding news articles. In Figure 3, the retrieved background documents include the entity "*Mugabe*" and the event "*detained*", which are relevant to the video description.

We apply a high-performing, publicly available entity discovery and linking system (Pan et al., 2017) to extract named entities and their types. This system is able to discover entities and link them to rich knowledge bases that provide fine-grained types that we can exploit to better discern between entities in the news documents (*e.g.*, "`President`" versus "`Military Officer`").[5] Additionally, we use a high-performing event extraction system (Li et al., 2013) to extract events and their arguments. For example, in Figure 3, we get entities "*S. B. Moyo*", "*Zimbabwe*", and "*Mugabe*" with their respective types, "`Military Officer`", "`GPE`", and "`President`". Likewise, we obtain events "*coup*" and "*detained*" with their respective types, "`Attack`" and "`Arrest-Jail`". The entities and events along with their types provide valuable insight into the context of the video and can bias the decoder to generate the correct event mentions and incorporate the proper entities.

We encode the entities and events into representations that can be fed to the model. First, we obtain an entity embedding, $\mathbf{e}_m$, for each entity by averaging the embeddings of the words in the entity mention. Second, we encode the entity and event types into a one-hot knowledge gate vector, $\mathbf{k}_0$. Each element of $\mathbf{k}_0$ corresponds to an event or entity type (*e.g.*, "`Arrest-Jail`" event type or "`President`" entity type), so the j^{th} element, $k^{(j)}$, is 1 if the entity or event type is found in the related documents and 0 otherwise. $\mathbf{k}_0$ serves as the initial knowledge gate vector of the decoder (Section 2.2). The entity embeddings give the model access to semantic representations of the entities, while the knowledge gate vector aids the generation process by providing the model with the event and entity types.

2.2 KaVD Network

Our model learns video representations using hierarchical, or multi-level, attention (Yang et al., 2016; Qin et al., 2017). The encoder is comprised of a spatial attention (Xu et al., 2015) and bidirectional Long Short-Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997) temporal encoder. The spatial attention allows the model to attend to different locations of each frame (Figure 4), yielding frame representations

[3] BBC, CNN, and New York Times.

[4] $d = 3$ in our experiments.

[5] We only use types that appear in the training data and are within 4 steps from the top of the 7,309 type hierarchy *here*.

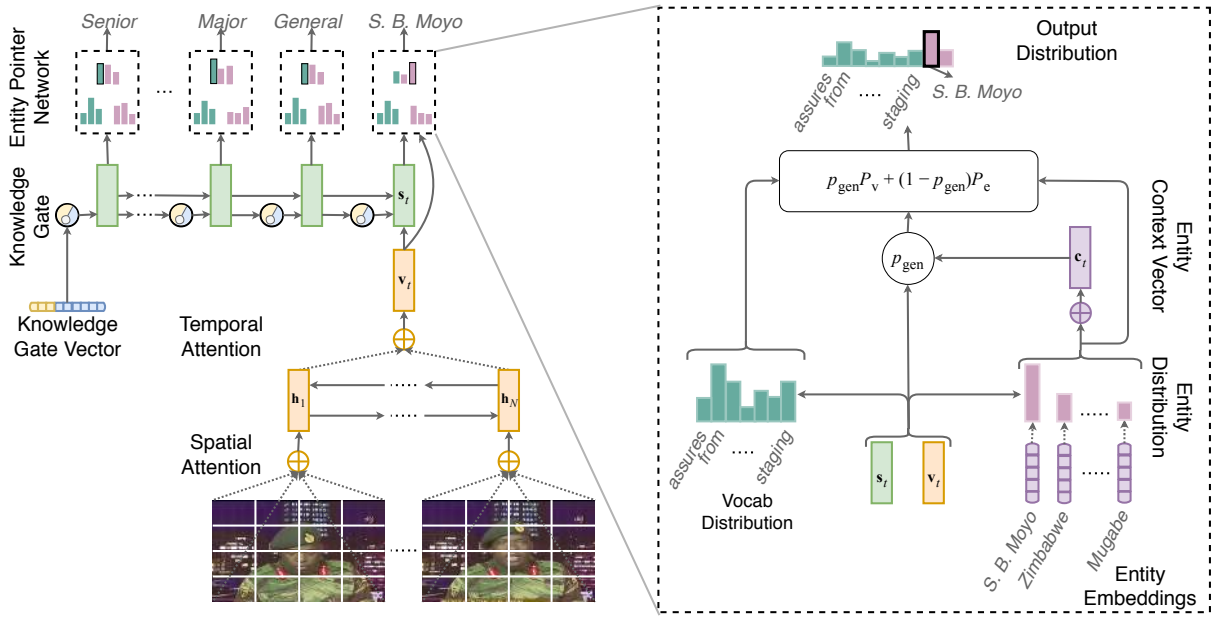

Figure 4: KaVD Network. At each decoder time step, the model computes p_{gen} to determine whether to emit a vocabulary word or a named entity from the topically related documents.

that emphasize the most important regions of each frame. The temporal encoder incorporates motion into the frame representations by encoding information from the preceding and subsequent frames (Yao et al., 2015). We use a LSTM decoder, which applies a temporal attention (Bahdanau et al., 2015) to the frame representations at each step. To generate each word, the decoder computes its hidden state, adjusts this hidden state with the *knowledge gate* output at the current time step, and determines the most probable word by utilizing the *entity pointer network* to decide whether to generate a named entity or vocabulary word. Pointer networks are effective at incorporating out-of-vocabulary (OOV) words in output sequences (Miao and Blunsom, 2016; See et al., 2017). In previous research, OOV words may appear in the input sequence, in which case they are copied into the output. Analogously, in our approach, named entities can be considered as OOV words that are from a separate set instead of the input sequence. In the following equations, where appropriate, we omit bias terms for brevity.

Encoder. The input to the encoder is a sequence of video frames, $\{F_1, ..., F_N\}$. First, we extract frame-level features by applying a Convolutional Neural Network (CNN) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Ioffe and Szegedy, 2015; Szegedy et al., 2015, 2017) to each frame, F_i, and obtaining the response of a convolutional layer, $\{\mathbf{a}_{i,1}, ..., \mathbf{a}_{i,L}\}$, where $\mathbf{a}_{i,l}$ is a D-dimensional representation of the l^{th} location of the i^{th} frame (*e.g.*, the top left box of the first frame in Figure 4). We apply a spatial attention to these location representations, given by

$$\alpha_{i,l} = a_{\text{space}}\left(\mathbf{a}_{i,l}\right) \quad (1)$$

$$\xi_{i,l} = \text{softmax}\left(\alpha_{i,l}\right) \quad (2)$$

$$\mathbf{z}_i = \sum_{l=1}^{L} \xi_{i,l}\mathbf{a}_{i,l} \quad (3)$$

where a_{space} is a scoring function (Bahdanau et al., 2015). Frame representations $\{\mathbf{z}_1, ..., \mathbf{z}_N\}$ are input to a bi-directional LSTM, producing temporally encoded frame representations $\{\mathbf{h}_1, ..., \mathbf{h}_N\}$.

Decoder. The decoder is an attentive LSTM cell with the addition of a *knowledge gate* and *entity pointer network*. At each decoder step t, we apply a temporal attention to the frame representations,

$$\beta_{t,i} = a_{\text{time}}\left(\mathbf{h}_i, \mathbf{s}_{t-1}\right) \quad (4)$$

$$\eta_{t,i} = \text{softmax}\left(\beta_{t,i}\right) \quad (5)$$

$$\mathbf{v}_t = \sum_{i=1}^{N} \eta_{t,i}\mathbf{h}_i \quad (6)$$

where $\mathbf{s}_{t-1}$ is the previous decoder hidden state and a_{time} is another scoring function. This yields a single, spatio-temporally attentive video representation, $\mathbf{v}_t$. We then compute an intermediate hidden state, $\hat{\mathbf{s}}_t$, by applying the decoder LSTM

to $\mathbf{s}_{t-1}$, $\mathbf{v}_t$, and previous word embedding, $\mathbf{x}_{t-1}$. The final decoder hidden state is determined after the knowledge gate computation.

The motivation for the knowledge gate is that it biases the model to generate sentences that contain specific knowledge relevant to the video and topically related documents, acting as a kind of coverage mechanism (Tu et al., 2016). For example, given the retrieved event types in Figure 3, the knowledge gate encourages the decoder to generate the event trigger "*coup*" due to the presence of the "`Attack`" event type. Inspired by the gating mechanisms from natural language generation (Wen et al., 2015; Tran and Nguyen, 2017), the knowledge gate, $\mathbf{g}_t$, is given by

$$\mathbf{g}_t = \sigma\left(\mathbf{W}_{g,v}[\mathbf{x}_{t-1}, \mathbf{v}_t] + \mathbf{W}_{g,s}\hat{\mathbf{s}}_t\right) \quad (7)$$

$$\mathbf{k}_t = \mathbf{g}_t \odot \mathbf{k}_{t-1} \quad (8)$$

where all $\mathbf{W}$ are learned parameters and $[\mathbf{x}_{t-1}, \mathbf{v}_t]$ is the concatenation of these two vectors. This gating step determines the amount of the entity and event type features contained in $\mathbf{k}_{t-1}$ to carry to the next step. With the updated $\mathbf{k}_t$, we compute the decoder hidden state, $\mathbf{s}_t$, as

$$\mathbf{s}_t = \hat{\mathbf{s}}_t + \left(\mathbf{o}_t \odot \tanh\left(\mathbf{W}_{s,k}\mathbf{k}_t\right)\right) \quad (9)$$

where $\mathbf{o}_t$ is the output gate of the LSTM and $\mathbf{W}_{s,k}$ is a learned parameter.

Our next step is to generate the next word. The model needs to produce named entities (*e.g.*, "*S. B. Moyo*" and "*Robert Mugabe*") throughout the generation process. These named entities tend to occur rarely if at all in many datasets, including ours. We overcome this issue by using the entity embeddings from the topically related documents as potential entities to incorporate in the description. We adopt a soft switch pointer network (See et al., 2017), as our entity pointer network, to perform the selection of generating words or entities.

For our entity pointer network to predict the next word, we first predict a vocabulary distribution, $P_{\text{v}} = \psi\left(\mathbf{s}_t, \mathbf{v}_t\right)$, where $\psi(\cdot)$ is a softmax output layer. $P_{\text{v}}(w)$ is the probability of generating word w from the decoder vocabulary. Next, we compute an entity context vector, $\mathbf{c}_t$, using a soft attention mechanism:

$$\gamma_{t,m} = a_{\text{entity}}\left(\mathbf{e}_m, \mathbf{s}_t, \mathbf{v}_t\right) \quad (10)$$

$$\epsilon_{t,m} = \text{softmax}\left(\gamma_{t,m}\right) \quad (11)$$

$$\mathbf{c}_t = \sum_{m=1}^{M} \epsilon_{t,m}\mathbf{e}_m \quad (12)$$

Here, a_{entity} is yet another scoring function. We use the scalars $\epsilon_{t,m}$ as our entity probability distribution, P_{e}, where $P_{\text{e}}\left(E_m\right) = \epsilon_{t,m}$ is the probability of generating entity mention E_m. We compute the probability of generating a word from the vocabulary, p_{gen}, as

$$p_{\text{gen}} = \sigma(\mathbf{w}_c^\top \mathbf{c}_t + \mathbf{w}_s^\top \mathbf{s}_t + \mathbf{w}_x^\top \mathbf{x}_{t-1} + \mathbf{w}_v^\top \mathbf{v}_t) \quad (13)$$

where all $\mathbf{w}$ are learned parameters. Finally, we predict the probability of word w by

$$P(w) = p_{\text{gen}}P_{\text{v}}(w) + (1 - p_{\text{gen}})P_{\text{e}}(w) \quad (14)$$

and select the word of maximum probability. In Equation 14, $P_{\text{e}}(w)$ is 0 when w is not a named entity. Likewise, P_{v} is 0 when w is an OOV word. For the example in Figure 4, the vocabulary distribution, P_{v}, has the word "*from*" as the most probable word and the entity distribution, P_{e}, has the entity "*S. B. Moyo*" as the most probable entity. However, by combining these two distribution using p_{gen}, the model switches to the entity distribution and correctly generates "*S. B. Moyo*".

3 News Video Dataset

Current datasets for video description generation focus on specific (Rohrbach et al., 2014) and general (Chen and Dolan, 2011; Xu et al., 2016) domains, but do not contain a large proportion of descriptions with specific knowledge like named entities as shown in Table 1. In our news video dataset, the descriptions are replete with important knowledge that is both necessary and challenging to incorporate into the generated descriptions.

Our news video dataset contains AFP international news videos from YouTube.[6] These videos are from October, 2015 to November, 2017 and cover a variety of topics, such as protests, attacks, natural disasters, trials, and political movements. The videos are "on-the-scene" and contain some depiction of the content in the description. For each video, we take the YouTube descriptions given by AFP News as the ground-truth descriptions we wish to generate. We collect the tags and meta-data (*e.g.*, upload date). We filter videos by length, with a cutoff of 2 minutes, and remove videos which are videographics or animations. For preprocessing, we tokenize each sentence, remove punctuation characters other than

[6] https://www.youtube.com/user/AFP

Dataset	Domain	#Videos	#Sentences	Vocab Size	Named Entities/Sentence
TACos M-L (Rohrbach et al., 2014)	Cooking	14,105	52,593	2,864	0.1×10^{-4}
MSVD (Chen and Dolan, 2011)	Multi-category	1,970	70,028†	13,010	0.4×10^{-2}
MSR-VTT-10K (Xu et al., 2016)	20 categories	10,000	200,000†	29,316	1.4×10^{-1}
News Video (Ours)	News	2,883	3,302	9,179	2.1

Table 1: Comparison of our news video dataset to other datasets. † indicates that the dataset has multiple, single-sentence reference descriptions for each video.

periods, commas, and apostrophes, and replace numerical quantities and dates/times with special tokens. We sample frames at a rate of $1fps$. We randomly select 400 videos for testing, 80 for validation, and 2,403 for training. We make the dataset publicly available: https://goo.gl/2jScKk.

4 Experiments

4.1 Model Comparisons

We test our method against the following baselines: **Article-only.** We use the summarization model of See et al. (2017) to generate the description by summarizing the topically related documents. **Video-only (VD).** We train a model that does not receive any background knowledge and generates the description directly from the video. **VD with knowledge gate only (VD+Knowledge Gate), VD with entity pointer network only (VD+Entity Pointer), and no-video (Entity Pointer+Knowledge Gate).** These test the effects of the knowledge gate, entity pointer network, and video encoder in isolation.

Each model uses a cross entropy loss. Video-based models are trained using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0002 and have a hidden state size of 512 as well as an embedding size of 300. We use Google News pre-trained word embeddings (Mikolov et al., 2013) to initialize our word embeddings and compute entity embeddings. For visual features, we use the Conv3-512 layer response of VGGNet (Simonyan and Zisserman, 2014) pre-trained on ImageNet (Deng et al., 2009).

4.2 Evaluations

METEOR (Denkowski and Lavie, 2014) and ROUGE-L (Lin, 2004) are adopted as metrics for evaluating the generated descriptions. We choose METEOR because we only have one reference description per video and this metric accounts for stemming and synonym matching. We also use ROUGE-L for comparison to summarization work. These capture the coherence and relevance of the generated descriptions to the ground truth.

Generating these descriptions is concerned with not only generating fluent text, but also the amount of knowledge conveyed and the accuracy of the knowledge elements (*e.g.*, named entities or event structures). Previous work in natural language generation and summarization (Nenkova and Passonneau, 2004; Novikova et al., 2017; Wiseman et al., 2017; Pasunuru and Bansal, 2018) scores and/or assigns weights to overlapping text, salient phrases, or information units (*e.g.*, entity relations (Wiseman et al., 2017)). However, knowledge elements cannot be simply represented as a set of isolated information units since they are inherently interconnected through some structure.

Therefore, for this knowledge-centric generation task, we compute F1 scores on event and entity extraction results from the generated descriptions against the extraction results on the ground truth. For entities, we measure the F1 score of the named entities in the generated description compared to the ground truth. For events, given a generated description, w^s, and the ground truth description, w^c, we extract a set of event structures, $\mathcal{Y}^s$ and $\mathcal{Y}^c$, for both descriptions such that $\mathcal{Y} = \{(t_k, r_{k,1}, a_{k,1}, ..., r_{k,m}, a_{k,m})\}_{k=1}^{K}$ where there are K events extracted from the description, t_k is the k^{th} event type, $r_{k,m}$ is the m^{th} argument role of t_k, and $a_{k,m}$ is the m^{th} argument of t_k. For the description in Figure 2a, one may obtain:

$\mathcal{Y} = \{($`Demonstrate`, `Entity`, "*Pro-independence supporters*", `Place`, "*Barcelona*"$)\}$

Next, we form event type, argument role, and argument triples $(t_k^s, r_{k,m}^s, a_{k,m}^s)$ and $(t_j^c, r_{j,m}^c, a_{j,m}^c)$ for each event structure in $\mathcal{Y}^s$ and $\mathcal{Y}^c$, respectively. We compute the F1 score of the triples, considering a triple correct if and only if it appears in

the ground truth triples.[7] This metric enables us to evaluate how well a generated description captures the overall events, while still giving credit to partially correct event structures. We compute these F1 scores on 50 descriptions based on manually annotated event structures. We also perform automatic F1 score evaluation on the entire test set using the entity and event extraction systems of Pan et al. (2017) and Li et al. (2013), respectively. The manual evaluations offer accurate comparisons and control for correctness, while the automated evaluations explore the viability of using automated IE tools to measure performance, which is desirable for scaling to larger datasets for which manual evaluations are too expensive.

5 Results and Analysis

The KaVD network outperforms almost all of the baselines, as shown in Table 2, achieving statistically significant improvements in METEOR and ROUGE-L w.r.t. all other models besides the no-video model ($p < 0.05$).[8] The additions of the entity pointer network and knowledge gate are complementary and greatly improve the entity incorporation performance, increasing the entity F1 scores by at least 6% in both the manual and automatic evaluations. In Figure 5a, the entity pointer network is able to incorporate the entity "*Abdiaziz Abu Musab*", who is a leader of the group responsible for the attack. We find that the entity and event type features from the knowledge gate help generate more precise entities. However, noise in the article retrieval process and entity extraction system limits our entity incorporation capabilities, since on average only 70.6% of the entities in the ground truth description are retrieved from the articles. Lastly, the video encoder helps generate the correct events and offers qualitative benefits, such as allowing the model to generate more concise and diverse descriptions, though it negatively affects the entity incorporation performance.

The video alone is insufficient to generate the correct entities (Table 2). In Figure 5a, the VD baseline generates the correct event, but generates the incorrect location "*Kabul*". We observe that when the visual evidence is ambiguous, this model may fail to generate the correct events and entities. For example, if a video depicts the destruction of buildings after a hurricane, then the VD baseline may mistakenly describe the video as an explosion since the visual evidence is similar.

The article-only baseline tends to mention the correct entities as shown in Figure 5a, where the description is generally on topic but provides some irrelevant information. Indeed, this model can generate descriptions unrelated to the video itself. In Figure 5b, the article-only baseline's description contains some correct entities (*e.g.,* "*Colombia*"), but is not focused on the announcement depicted in the video. As See et al. (2017) discuss, this model can be more extractive than abstractive, copying many sequences from the documents. This can lead to irrelevant descriptions as the articles may not be specific to the video.

Our entity and event F1 score based metrics correlate well with the correctness of the knowledge conveyed in the generated description. The consistency in model rankings between the manual and automatic entity metrics shows the potential of using automated entity extraction approaches to evaluate with this metric. We observe discrepancies between the manual and automatic event metrics, in part, due to errors in the automated extraction and the addition of more test points. For example, in the generated sentence, "*Hundreds of people are to take to the streets of...*", the event extraction system mistakenly assigns a "`Transport`" event type instead of the correct "`Demonstrate`" event type. In contrast, such mistakes do not appear in the manual evaluations.

6 Related Work

Most previous video captioning efforts focus on learning video representations through different encoding techniques (Venugopalan et al., 2015a,b), using spatial or temporal attentions (Yao et al., 2015; Pan et al., 2016; Yu et al., 2016; Zanfir et al., 2016), using 3D CNN features (Tran et al., 2015; Yao et al., 2015; Pan et al., 2016), or easing the learning process via multi-task learning or reinforcement rewards (Pasunuru and Bansal, 2017a,b). Compared to other hierarchical models (Pan et al., 2016; Yu et al., 2016), each level of our hierarchy encodes a different dimension of the video, leveraging global temporal features and local spatial features, which are shown to be effective for different tasks (Ballas et al., 2015; Xu et al., 2015; Yu et al., 2017).

We move towards using datasets with captions that have specific knowledge rather than generic

[7] This criterion is used for computing precision and recall.
[8] Found via paired bootstrap resampling (Koehn, 2004).

Model	METEOR	ROUGE-L	Entity F1	Auto-Entity F1	Event F1	Auto-Event F1
Article-only	8.6	13.2	8.7	8.5	1.9	3.6
VD	9.1	17.9	2.5	1.5	1.0	7.3
VD+Entity Pointer	9.7	18.1	15.3	13.6	5.7	7.0
VD+Knowledge Gate	9.8	18.5	10.2	10.7	6.7	8.3
Entity Pointer+Knowledge Gate	10.1	18.7	**23.7**	**20.9**	2.2	**9.9**
KaVD	**10.2**	**18.9**	22.1	19.7	**9.6**	8.9

Table 2: METEOR, ROUGE-L, and manual/automated entity (Entity F1/Auto-Entity F1) and event (Event F1/Auto-Event F1) F1 score results of the baselines and KaVD network on our news video dataset.

a) 

Title: Deadly Shabaab attack rocks Somali capital

b) 

Title: Santos: 'Green light' for referendum on Colombia peace deal

Model	Description
Article-only	somali capital mogadishu on saturday. at least 276 people have died and the government news agency sonna says only 111 of them have been identified. a turkish military but instead witnessed her burial. no group has yet said it was behind on instead he attended her burial. "anfa'a said she had spoken to her sister 20 minutes before on
VD	a suicide bomber killed # people in a bus carrying # people killed in a bus in central kabul.
VD+Entity Pointer	A suicide bomber killed # people were killed in a bus near the northern city of Mogadishu, police said.
VD +Knowledge Gate	At least # people were killed and # wounded when a busy bus station in Kabul, killing at least # people dead and others who died in the rubble of the deadliest attack in the country.
EntityPointer +Knowledge Gate	At least # people were killed in a suicide car bomb attack on a suicide car bomb attack on a police vehicle in Mogadishu, police said.
KaVD	A suicide bombing claimed by the Abdiaziz Abu Musab group time killed # people in Somalia's capital Mogadishu, killing # people, officials said.

Model	Description
Article-only	colombia's marxist rebels against her family. and last year, when given the leg of helena gonzlez's nephew years ago is still fresh the as pope francis arrived in colombia on wednesday for a six-day the
VD	president donald trump says that he will be talks to be to be talks to be talks in the country's country to be talks, saying he says he would be no evidence's state and kerry says.
VD+Entity Pointer	President Maduro says the FARC president warns that the ceasefire to Prime Minister says that he will be ready to help President Maduro says that he is no evidence of President Bashar talks in Bogota.
VD +Knowledge Gate	US Secretary of State John Kerry, who will not any maintain in Syria, after a ceasefire in Syria, saying that the United Nations says, it will not to be into a speech in its interview.
EntityPointer +Knowledge Gate	Venezuela's President FARC envoy to Colombia is a definitive ceasefire in the FARC conflict, with FARC rebels, the FARC rebels.
KaVD	Colombia's government, signed the peace agreement with the FARC peace accord in the FARC rebels.

Figure 5: Comparison of generated descriptions. The KaVD network generates the correct entities and correct events, while other models may contain some wrong entities or wrong events.

captions as in previous work (Chen and Dolan, 2011; Rohrbach et al., 2014; Xu et al., 2016). There are efforts in image captioning to personalize captions (Park et al., 2017), incorporate novel objects into captions (Venugopalan et al., 2016), and perform open domain captioning (Tran et al., 2016). To the best of our knowledge, our dataset is the first of its kind and offers challenges in entity and activity recognition as well as the generation low probability words. Datasets with captions rich in knowledge elements, like those in our dataset, take a necessary step towards increasing the utility of video captioning systems.

We employ similar approaches to those in automatic summarization, where pointer networks (Vinyals et al., 2015) and copy mechanisms (Gu et al., 2016) are used (Gulcehre et al., 2016; Nallapati et al., 2016; Miao and Blunsom, 2016; See et al., 2017), and natural language generation for dialogue systems (Wen et al., 2015; Tran and Nguyen, 2017). The KaVD network combines the copying capabilities of pointer networks (See et al., 2017) and semantic control of gating mechanisms (Wen et al., 2015; Tran and Nguyen, 2017) in a complementary fashion to address a new, multi-modal task.

7 Conclusions and Future Work

We collect a news video dataset with knowledge-rich descriptions and present a multi-modal approach to this task that uses a novel Knowledge-aware Video Description network, which can utilize background knowledge mined from topically related documents. We offer a new metric to measure a model's ability to incorporate named entities and specific events into the descriptions. We show the effectiveness of our approach and set a new benchmark for this dataset. In future work, we are increasing the size of dataset and exploring other knowledge-centric metrics for this task.

Acknowledgments

This work was supported by the U.S. DARPA AIDA Program No. FA8750-18-2-0014 and U.S. ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. 2015. Delving deeper into convolutional networks for learning video representations. In *ICLR*.

David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *WMT*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *ACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *ACL*.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out: ACL workshop*.

Yishu Miao and Phil Blunsom. 2016. Language as a latent variable: Discrete generative models for sentence compression. In *EMNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshops at ICLR*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *NAACL*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *EMNLP*.

Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *CVPR*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *ACL*.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*.

Ramakanth Pasunuru and Mohit Bansal. 2017a. Multi-task video captioning with video and entailment generation. In *ACL*.

Ramakanth Pasunuru and Mohit Bansal. 2017b. Reinforced video captioning with entailment rewards. In *EMNLP*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *NAACL*.

Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. In *IJCAI*.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *GCPR*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In *CVPR Workshops*.

Van-Khanh Tran and Le-Minh Nguyen. 2017. Natural language generation for spoken dialogue system using rnn encoder-decoder networks. In *CoNLL*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *ACL*.

Subhashini Venugopalan, Lisa Anne Hendricks, Raymond Mooney, and Kate Saenko. 2016. Improving lstm-based video description with linguistic knowledge mined from text. In *EMNLP*.

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015a. Sequence to sequence – video to text. In *ICCV*.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015b. Translating videos to natural language using deep recurrent neural networks. In *NAACL*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *NIPS*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *EMNLP*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *ICCV*.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*.

Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. 2016. Spatio-temporal attention models for grounded video captioning. In *ACCV*.