
On abstractive and extractive summarization of instructional video transcripts using BERT

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The overflow of video content in the Internet (from YouTube, MOOCs, news
2 portals) necessitates automated summarizations of data. In our paper, we study
3 extractive and abstractive summarization for of instructional videos. Previously,
4 natural language processing efforts have been focused to meticulously curated
5 datasets far removed from textual inconsistencies that are inherent to videos. Our
6 work on text preprocessing allows to extend the approach summarization of auto-
7 generated amateur video transcripts. Next, we apply state-of-the-art pretrained
8 BERT transformer models to the problem and evaluate the efficiency of training
9 and fine tuning with datasets from WikiHow, How2 videos, and CNN. The results
10 are evaluated using ROUGE F1 and blind assessments by human experts.

11 1 Introduction

12 According to Forbes [*link TBA*], more than 500 million hours of videos are watched on YouTube
13 every day and a lot of time is wasted watching videos that are not useful. Video content is rapidly
14 growing and will remain the mainstream for sharing information in future. In this project YAVA
15 (“Your Active Virtual Audience”) we are aspiring to make online exchanges of information between
16 people via audio or video more efficient and enjoyable.

17 There have been a lot of research efforts recently focused on video summarization [e.g. see [Cai
18 et.al.], [Shemer et.al.], [Kaufman et.al.]]. The known methods work by extracting the most important
19 segments and concatenating them together. However, it has been demonstrated that a lot of the
20 time the result is not substantially better and sometimes even worse than random selection of video
21 fragments ([Mayu et.al.]).

22 Summarization in the area of multimodal video processing tackles the problem from a different angle.
23 Instead of producing summary by converting a long video into a short video, it extracts signals from
24 it speech-to-text, facial expressions, spectrogram of speaker’s voice; etc. (see [Samanth et.al.]). and
25 processes them separately produces a short text (an abstract of what it is about). This method has a
26 few advantages:

- 27 • We get access to a set of existing models for text summarization, substantially more mature
28 than those for videos (e.g. [Subramanian et.al.]).
- 29 • We can leverage existing text summarization datasets, which are more easily available, than
30 video datasets (e.g. [Mahnaz et.al.]).
- 31 • Processing texts during algorithm training takes less computational power than processing
32 videos.
- 33 • Arguably, a text summary of a long video is even better for the viewer than a short video,
34 especially from the perspective of a person who needs it to decide whether to watch the

35 full video. It doesn't consume the network bandwidth, doesn't require audio equipment
36 or noise-free environment, takes less device energy to reproduce (especially important for
37 mobile devices) , and the viewer can consume it at their own pace. You can skim the text in
38 any order, any time.

39 The models for these purposes have been developed better than for processing video as a whole (e.g.
40 see [Jaejin et.al.]), and that's why this approach referred to as "multimodal" summarization looks
41 very promising to us and has recently received a lot of attention from other researchers (e.g. see
42 [Palaskar et.al.], [Tripathi et.al.]).

43 The focus of our research is on how-to/instructions videos. According to [https://mediakix.com/
44 blog/most-popular-youtube-videos/](https://mediakix.com/blog/most-popular-youtube-videos/), this type of video is one of the most popular on youtube
45 these days. Also, viewers of such videos are interested in getting a tangible outcome, as compared to
46 viewers of entertainment or sports videos, therefore adding a summary will add more value. which
47 we will use for training purposes. Pioneering efforts in this area have been done by [Palaskar et.al.]
48 based on dataset of how2 videos [Sanabria et.al.]. We plan to improve on their results by taking
49 advantage of "WikiHow: A Large Scale Text Summarization Dataset" [Mahnaz et.al.], improving
50 the models, and applying more advanced techniques to evaluation of output. Why is it important /
51 challenging? We foresee many applications of this approach, especially in education and business,
52 where even minor improvements in information processing may make big differences when applied
53 at scale to online meetings, virtual classrooms and other forms of human interactions via video.

54 Summarizing content is challenging even for a human. The rules of identifying what's important and
55 what can be omitted are subjective, changeable and very hard to formalize. While watching a long
56 video conference, participants often get tired and lose attention. Finally, a lot depends on the context.
57 Yet, as hard as it is, most people get it, and this skill improves through a lot of learning and practice.
58 It gives us hope that training machines to help facilitate this process is both possible and useful.

59 Also, evaluating the quality of summaries and obtaining benchmarks is another problem. As shown in
60 research [Mayu et.al.], engaging human experts for evaluation of results is expensive, while automated
61 techniques lack depth. We will use a combination of both techniques to maximize the quality of
62 results.

63 In our work, we are exploring transferability of modern text summarization techniques to instructional
64 videos scripts on large annotated data sets that we created by preprocessing YouTube videos and data
65 from other authors. We discuss heuristics that were discovered on this data, impacts on the quality of
66 generated summaries, and propose different ways of improving summarization process to deal with
67 these issues. Finally, we identify promising directions for future research.

68 2 Prior work

69 2.1 Text Summarization

70 Text summarization is the task of generating shorter versions of documents while maintaining
71 important information [need link]. This area of research in the natural language processing community
72 has grown rapidly over the past several years due to its practical applications among various industries
73 such as news, reviews, education. Summarization systems take two general approaches: extractive and
74 abstractive. Extractive summarization provides users with textual summaries that have been copied
75 and concatenated from important parts of a document. It is a reliable task capable of maintaining
76 sentence structure and factual correctness. Abstract summarization generates a summary with content
77 that is not always found in the underlying text. It is a complex task that mimics human summarization
78 by generalizing and paraphrasing key points made in the document.

79 Prior to 2014, summarization was centered on extracting lines from single documents using statistical
80 models and neural networks had limited success[6, 7]. Sutskever et al. and Cho et al work on
81 sequence to sequence models opened up new possibilities for neural networks in natural language
82 processing. From 2014 to 2015, LSTMs (variety of RNN) became the dominant approach that
83 achieved state of the art results. They became successful in tasks such as speech recognition, machine
84 translation, parsing, image captioning, etc. It paved the way for abstractive summarization, which
85 began to score competitively against extractive summarization. In 2017, Attention is all you need
86 [8] provided a solution to the 'fixed length vector' problem, enabling neural networks to focus on

87 important parts of the input for prediction tasks. Transformers with attention became more dominant
88 for certain tasks [9].

89 2.2 Multi-modal Summarization

90 Research surrounding multimedia has improved greatly to bridge the gaps between multi-modal
91 content such as speech, visuals, and text. Summarization has been used in meeting records [10],
92 sports videos [11], news [12], each encapsulating synchronized speech, videos, and subtitles. Video
93 summaries consist of cutting important frames out of the video to create a succinct compact version.
94 More recently, research around multimodal summarization, which combines the textual and visual
95 modalities to align with the video content, have reached an early benchmark [13 - shruti's work].
96 The How2Dataset [5] is a collection of 2,000 hours of instructional videos with English subtitles and
97 crowdsourced Portuguese translations. It covers different how-to domains such as sports, cooking,
98 and education. The dataset has been created to be used as a benchmark for multimodal natural
99 language tasks, used in various competitions and research settings. This How2Dataset precedes
100 more recent work constructing data from instructional web videos in the How2100M [14] dataset.
101 The dataset is large-scale and has 136 million video clips and transcripts of humans performing or
102 describing various tasks, but there are no human annotated summaries.

103 3 Problem Statement

104 In our work we set the following goals:

- 105 • Curate and publish a single source of truth data set of text and summaries aggregated and
106 formatted from WikiHow articles, How2 videos, and CNN stories
- 107 • Apply existing BERT-based text summarization models to make them applicable to auto-
108 generated scripts from instructional videos and generalize them to work on instructional
109 videos
- 110 • Augment ROUGE metrics [Chin-Yew Lin] for evaluation of the results with a framework
111 for formalized expert assessment based on our research and criteria proposed by previous
112 works

113 For our confidence about the feasibility of the project, we first ran a series of manual experiment by
114 dumping a few auto-generated scripts YouTube scripts and running them through online summariza-
115 tion services. The first results were very disappointing. However, we noticed that auto-generated
116 scripts don't have punctuation and line breaks don't necessarily correspond to the logical ends of
117 sentences. After fixing these issues, we got meaningful summaries and proceeded to generalizing the
118 approach as follows.

119 4 Methodology

120 From the initial exploration and data analysis we saw that in the process of applying existing
121 summarization models to Youtube video scripts we will deal with challenges imposed by parsing
122 speech-to-text output add more complexity to text summarization. For example, in one of the sample
123 videos in our test data set closed captioning confuses the speaker's words "*how you get a text from*
124 *a YouTube video*" for "*how you get attacks from a YouTube video*". So, our work includes several
125 iterations of the process described below:

- 126 • Collection and aggregation of data from multiple sources (HowTo video scripts, WikiHow,
127 CNN stories, YouTube)
- 128 • Preprocessing of video scripts to make them fit the text summarization models (e.g. errors in
129 word recognition, lack of punctuation in closed captioning, getting rid of special characters
130 etc., aligning inputs aggregated from multiple sources to common format)
- 131 • Text summarization models: selection, deployment, training, and fine-tuning
- 132 • Experiments: applying models to the data and evaluation of the outputs using ROUGE
133 metrics and human expert judgements

134 4.1 Collection

135 We believe that sentence fluency and generalization is best captured in the larger corpus of news,
136 instructional texts, and youtube videos. For this reason, we combined the following three datasets:

- 137 • **CNN/Daily Mail dataset** provided by Hermann et. al 2015, the How2 Dataset, and Wikihow.
138 The datasets illustrate different summary styles that range from one sentence long phrases
139 to short paragraph summaries. CNN/Daily mail includes a combination of news articles and
140 story highlights written in various lengths.
- 141 • **Wikihow dataset**, a large scale text summarization containing over 200,000 single document
142 summaries. We included it to increase performance and generalizability, we included the
143 Wikihow is a variety of ‘How To’ instructional texts compiled from wikihow.com, ranging
144 from topics such as ‘How to deal with coronavirus anxiety’ to ‘How to play Uno.’ Similar
145 to CNN news articles, the articles inside the dataset vary in size and topic but are structured
146 to drive across direct messages / instructions to the user.
- 147 • **How2 Dataset** of 8,000 videos (approximately 2,000 hours). This dataset was constructed
148 from ‘How To’ YouTube videos that averaged 90 seconds long and 291 word long transcripts.
149 It also includes human generated sentence summaries written to generate interest in the
150 viewer. Summaries were two to three sentences in length with an average length of 33
151 words.

152 As part of this research, we are exploring different combinations of data during training of summa-
153 rization models and evaluate how they perform on instructional video scripts in any domain.

154 4.2 Preprocessing

155 The format of CNN /Daily Mail stories, wikiHow articles, and howTo scripts is different. We invested
156 substantial efforts into converting them to a format that can be used. For the convenience of other
157 researchers who may want to use similar methodology, we shared the results of aligning them to the
158 same format that can be training.

159 Another stream of work we have done at this stage is based on the heuristics observed during
160 evaluation of results. We expected the differences in conversational style of the video scripts and
161 writtent text of CNN stories (on which the models were pretrained) will impact quality of the output.
162 In our first experiments, it manifested in a very distinct way. The model considered the first one-two
163 sentences to be very important for summaries, and we ended up with getting many summaries looking
164 like "hi!" and "hello, this is <first and last name>". It inspired us for implementing an improvement
165 by using entity detection spacy and nltk to remove introduction from the text that we feed to
166 summarization model.

167 The CNN/Daily Mail dataset has been preprocessed to remove news anchor introductions. For
168 our Wikihow and How2 transcripts, we split sentences using the Stanford Core NLP toolkit and
169 preprocessed the data in the same method used by (See et. al.).

170 4.3 Summarization models

171 For our first large-scale experiment, we used an out-of-box PreSumm abstractive model created by
172 Yang trained on CNN and Daily Mail [Yang] . Next, we used the PreSumm abstractive model trained
173 on 5,000 samples from the How2 dataset, 3,097 samples from Wikihow with a 100,000 step size. In
174 our most recent, but not final, experiment we also trained the PreSumm extractive model on 13,907
175 Wikihow dataset, 5,000 How2 dataset with 50,000 steps. These models are very demanding in terms
176 of both memory and computational resources. Training has to be run on very powerful machines. We
177 leveraged VMs in GCP and Azure, thanks to the free limits allocated to students by both. We started
178 with trying to deploy on TPU, but after several days of trying and porting the code we got blocked on
179 transient memory failures and defaulted to using VMs with multiple GPUs and a big SSD to allow
180 for storing checkpoints of the models. Compared to a CPU-based machine, we got several orders of
181 magnitude improvements. [to be added: which models we use]

Table 1: Comparison of results

| Experiment | | Rouge-1 | Rouge-L |
|-------------------------------------|-------------------------------------|---------|---------|
| Model | Pretraining Data | | |
| PreSum | CNN and Daily Mail | 18.08 | 18.01 |
| PreSum with preprocessing | CNN and Daily Mail | 20.51 | 18.86 |
| PreSum with pre- and postprocessing | CNN and Daily Mail | 22.47 | 20.07 |
| PreSum | How-To, WikiHow, CNN and Daily Mail | 24.4 | 21.45 |
| PreSum with postprocessing | How-To, WikiHow, CNN and Daily Mail | 26.32 | 22.47 |

4.4 Evaluation

The PreSumm model created by Yang trained on CNN and Daily Mail [Yang] resulted in SOTA rouge scores when applied to samples from those datasets. However, when tested on our How2 Test dataset, it gave very poor performance and a lack of generalization in the model (see Table 1). Looking at the data, we found that the model tends to pick the first one or two sentences for the summary. This can be explained by the fact that the first paragraph of a news article often captures the gits of it, which the model learned. However, in the case of our instructional videos, the first sentences would be a non-informative introduction, such as "Hi there! My name is ...". Based on that, we hypothesized that removing introductions from the text will help improve ROUGE scores. Indeed, we got a few points better after applying preprocessing described in the Section 4.2 above. Yet another improvement in the score was accomplished by taking advantage of one more observation: most curated summaries follow a template that starts with "Learn how ...". So, we added these two words in the beginning of the summary at post-processing stage. With all that, we still couldn't get higher than 22.5 ROUGE-1 F1 and 20 ROUGE-L F1. Reviewing scores and texts of individual summaries showed that the model is doing better on some topics, such as medicine, and worse on others, such as sports. Again, this makes sense for a model that is trained on news: it isn't reasonable for it to be good with yoga-specific terminology, while news about health care are very common.

So, in our next series of experiments, we used our own dataset for training. We were able to push the scores higher: by 4 for ROUGE-1 and 2.5 ROUGE-L F1 on the results with and without preprocessing, compared to the CNN-trained model. However, there is still a lot of room for improvement, as more specialized model by [Shruti et.al.] claims to go above 40 ROUGE score.

In order to calculate ROUGE metrics, we used py-rouge package and initialized evaluator with a 100-word limit penalty as follows:

```

#nltk.download("punkt")
rouge_evaluator = rouge.Rouge(
    metrics=["rouge-n", "rouge-l"],
    max_n=4,
    limit_length=True,
    length_limit=100,
    length_limit_type="words",
    apply_avg=True,
    apply_best=False,
    alpha=0.5, # Default F1_score
    weight_factor=1.2,
    stemming=True,
)

```

```

*****
Reference: now that you have spent the time cleaning your oven learn how to keep it clean with expert tips in this free h
ow to video on how to better clean your oven

Hypothesis: make sure your oven is clean .<q>clean your oven .<q>make sure you want to clean the oven with a towel .<q>ge
t your food .<q>put your food in your baking soda and water .<q>do n't go to the kitchen .

rouge-1:      P: 29.55      R: 40.62      F1: 34.21
rouge-2:      P: 6.98      R: 9.68      F1: 8.11
rouge-3:      P: 2.38      R: 3.33      F1: 2.78
rouge-4:      P: 0.00      R: 0.00      F1: 0.00
rouge-l:      P: 24.16      R: 31.50      F1: 27.34
rouge-w:      P: 14.23      R: 9.78      F1: 11.59
*****

```

Figure 1: An example where ROUGE metric is confusing.

We have observed examples of bad summaries with high ROUGE score, such as in Figure ??, and good summaries with low ROUGE score. We believe that ROUGE is fine as a starting point for comparison, but the real evaluation of the output quality still requires human experts.

5 Conclusion

We are continuing to work on improving summarization for instructional videos, as measured by both ROUGE and human experts. By the end of the project, we hope to accomplish scores that are comparable to current SOTA, but more generalizable. We also plan to provide a more detailed analysis on correlations between features of a video (e.g. topic, length, number of likes) and the quality of summaries produced on our experiments, as well as a more detailed description of our expert evaluation process.

Broader Impact

The contribution of our research is three-fold:

- We created and published a data set of how-to videos with time-tagged scripts, machine-generated summaries¹
- We generalized existing text summarization models to the scripts extracted from the videos [Sanabria et.al.]
- We augmented ROUGE metrics [Chin-Yew Lin] for evaluation of the results with a framework for formalized expert assessment based on our research and criteria proposed by previous works *[that's in work]*

At a high level, we hope that our analysis of transferability of summarization techniques from text to videos will have both practical and theoretical impacts by helping identify promising directions for future research.

References

We will align the formatting of references for the final submission. Current list is accurate, but not standardized.

- [1] Yang Liu, Mirella Lapata. Text Summarization with Pretrained Encoders. (2019) URL. <https://arxiv.org/abs/1908.08345v2>
- [2] Abigail See, Peter J. Liu, and Christopher D. Manning. (2017) Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073-1083.
- [3] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Neural Information Processing Systems*, 2014.

¹<https://github.com/alebryvas/berk266/> - it's not public repository yet, but we can provide access upon request

- [4] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. (2017). Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102. Association for Computational Linguistics.
- [5] Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018. URL. <https://arxiv.org/abs/1811.00347>
- [6] Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of AAAI 2005*, Pittsburgh, USA.
- [7] Svore, K., Vanderwende, L., and Burges, C. (2007). Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of the EMNLP-CoNLL*, pages 448–457. [7, 8]
- [8] Yu-Hsiang Huang. Attention is all you need - pytorch. <https://github.com/jadore801120/attention-is-all-you-need-pytorch>, 2018.
- [9] Nima Sanjabi. Abstractive text summarization with attention-based mechanism. Master’s thesis, Universitat Politècnica de Catalunya, July 2018.
- [10] Berna Erol, D-S Lee, and Jonathan Hull. 2003. Multimodal summarization of meeting recordings. In *Multimedia and Expo, 2003. ICME’03. Proceedings. 2003 International Conference on*, volume 3, pages III–25. IEEE.
- [11] Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. Multi-modal summarization of key events and top players in sports tournament videos. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 471–478. IEEE
- [12] Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.
- [13] Shruti Palaskar, Jindrich Libovicky, Spandana Gella, Florian Metze. 2019. Multimodal Abstractive Summarization for How2 Videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596. Association for Computational Linguistics.
- [14] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV 2019*. <https://arxiv.org/abs/1906.03327>.