On abstractive and extractive summarization of instructional video transcripts using BERT

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

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

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

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

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

To be Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

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

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

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

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

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

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

2 Prior work

2.1 Text Summarization

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

Prior to 2014, summarization was centered on extracting lines from single documents using statistical models and neural networks had limited success[6, 7]. Sutskever et al. and Cho et al work on sequence to sequence models opened up new possibilities for neural networks in natural language processing. From 2014 to 2015, LSTMs (variety of RNN) became the dominant approach that achieved state of the art results. They became successful in tasks such as speech recognition, machine translation, parsing, image captioning, etc. It paved the way for abstractive summarization, which began to score competitively against extractive summarization. In 2017, Attention is all you need [8] provided a solution to the 'fixed length vector' problem, enabling neural networks to focus on important parts of the input for prediction tasks. Transformers with attention became more dominant for certain tasks [9].

2.2 Multi-modal Summarization

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

3 Problem Statement

In our work we set the following goals:

- Curate and publish a single source of truth data set of text and summaries aggregated and formatted from WikiHow articles, How2 videos, and CNN stories
- Apply existing BERT-based text summarization models to make them applicable to auto-generated scripts from instructional videos and generalize them to worj on instructional videos
- Augment 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

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

4 Methodology

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

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

4.1 Collection

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

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

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

4.2 Preprocessing

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

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

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

4.3 Summarization models

For our experiments, we trained the PreSumm abstractive model on 5,000 samples from the How2 dataset, 3,097 samples from Wikihow with a 100,000 step size. We also trained the PreSumm extractive model on 13,907 Wikihow dataset, 5,000 How2 dataset with 50,000 steps.

4.4 Evaluation

Alexandra or Bryan

5 Conclusion

The style files for NeurIPS and other conference information are available on the World Wide Web at

`http://www.neurips.cc/`

The file `neurips_2020.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

The only supported style file for NeurIPS 2020 is `neurips_2020.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

The LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

Preprint option If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text "Preprint. Work in progress." in the footer. This version may be distributed as you see fit. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2020.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 6, 7, and 8 below.

6 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 8 regarding figures, tables, acknowledgments, and references.

7 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

7.1 Headings: second level

Second-level headings should be in 10-point type.

7.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

8 Citations, figures, tables, references

These instructions apply to everyone.

8.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

```
http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf
```

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2020` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2020}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

8.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

8.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

8.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

[1] Sample of the first footnote.

[2] As in this example.

Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

9 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

10 Preparing PDF files

Please prepare submission files with paper size “US Letter,” and not, for example, “A4.”

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

```
\usepackage{amsfonts}
```

followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{I\!\!R} %real numbers
\newcommand{\Nat}{I\!\!N} %natural numbers
\newcommand{\CC}{I\!\!\!\!C} %complex numbers
```

Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

10.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

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 summaries
- 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

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

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Note that the Reference section does not count towards the eight pages of content that are allowed.**

@articleDBLP:journals/corr/abs-1810-09305, author = Mahnaz Koupaee and William Yang Wang, title = WikiHow: A Large Scale Text Summarization Dataset, journal = CoRR, volume = abs/1810.09305, year = 2018, url = http://arxiv.org/abs/1810.09305, archivePrefix = arXiv, eprint = 1810.09305, timestamp = Wed, 31 Oct 2018 14:24:29 +0100, biburl = https://dblp.org/rec/journals/corr/abs-1810-09305.bib, bibsource = dblp computer science bibliography, https://dblp.org

———

[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.

[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 Meting 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.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.