
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, Content F1, and blind assessments by human experts.

11 1 Introduction

12 According to Forbes, more than 500 million hours of videos are watched on YouTube every day and
13 a lot of time is wasted watching videos that are not useful. Video content is rapidly growing and will
14 remain the mainstream for sharing information in future. In this project YAVA (“Your Active Virtual
15 Audience”) we are aspiring to make online exchanges of information between people via audio or
16 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 Our approach tackles the problem from a different angle. Instead of producing summary by converting
23 a long video into a short video, we will convert the video into a short text (an abstract of what it is
24 about) automatically generated based on the script of speech. This method has a few advantages:

- 25 • We get access to a set of existing models for text summarization, substantially more mature
26 than those for videos (e.g. [Subramanian et.al.]).
- 27 • We can leverage existing text summarization datasets, which are more easily available, than
28 video datasets (e.g. [Mahnaz et.al.]).
- 29 • Processing texts during algorithm training takes less computational power than processing
30 videos.
- 31 • Arguably, a text summary of a long video is even better for the viewer than a short video,
32 especially from the perspective of a person who needs it to decide whether to watch the
33 full video. It doesn’t consume the network bandwidth, doesn’t require audio equipment
34 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.

We understand that the approach also has limitations, e.g. it will perform best on videos where the majority of information is conveyed via words. However, at later stages of the research we can add other separately extracted signals, such as spectrogram of speaker’s voice; emotions on people’s faces, illustrative pictures, etc. (see [Samanth et.al.]). 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.

From the initial exploration and data analysis we saw that in the process of applying the models of summarization to videos we will deal with challenges imposed by parsing speech-to-text output add more complexity to text summarization (e.g. errors in word recognition, lack of punctuation in closed captioning, etc.). 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”.

Finally, 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.

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.

1.1 Style

Papers to be to NeurIPS 2020 must be prepared according to the instructions presented here. Papers may only be up to eight pages long, including figures. Additional pages *containing only a section on the broader impact, acknowledgments and/or cited references* are allowed. Papers that exceed

87 eight pages of content will not be reviewed, or in any other way considered for presentation at the
88 conference.

89 The margins in 2020 are the same as those in 2007, which allow for $\sim 15\%$ more words in the paper
90 compared to earlier years.

91 Authors are required to use the NeurIPS \LaTeX style files obtainable at the NeurIPS website as
92 indicated below. Please make sure you use the current files and not previous versions. Tweaking the
93 style files may be grounds for rejection.

94 1.2 Retrieval of style files

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

96 <http://www.neurips.cc/>

97 The file `neurips_2020.pdf` contains these instructions and illustrates the various formatting re-
98 quirements your NeurIPS paper must satisfy.

99 The only supported style file for NeurIPS 2020 is `neurips_2020.sty`, rewritten for $\LaTeX 2_{\epsilon}$.
100 **Previous style files for \LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

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

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

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

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

114 The formatting instructions contained in these style files are summarized in Sections 2, 3, and 4
115 below.

116 2 General formatting instructions

117 The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long.
118 The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points.
119 Times New Roman is the preferred typeface throughout, and will be selected for you by default.
120 Paragraphs are separated by $\frac{1}{2}$ line space (5.5 points), with no indentation.

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

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

129 Please pay special attention to the instructions in Section 4 regarding figures, tables, acknowledgments,
130 and references.

131 **3 Headings: first level**

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

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

134 **3.1 Headings: second level**

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

136 **3.1.1 Headings: third level**

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

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

140 **4 Citations, figures, tables, references**

141 These instructions apply to everyone.

142 **4.1 Citations within the text**

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

146 The documentation for `natbib` may be found at

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

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

150 `\citet{hasselmo}` investigated\dots

151 produces

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

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

155 `\PassOptionsToPackage{options}{natbib}`

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

158 `\usepackage[nonatbib]{neurips_2020}`

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

163 **4.2 Footnotes**

164 Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number¹
165 in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote
166 with a horizontal rule of 2 inches (12 picas).

¹Sample of the first footnote.

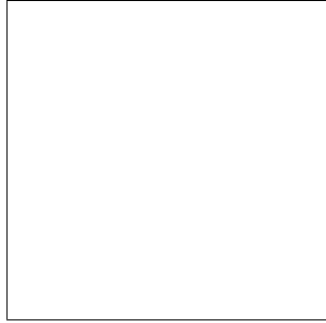


Figure 1: Sample figure caption.

Table 1: Sample table title

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

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

168 4.3 Figures

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

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

175 4.4 Tables

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

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

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

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

184 This package was used to typeset Table 1.

185 5 Final instructions

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

²As in this example.

6 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 file uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdf fonts` 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}{\mathbb{R}} %real numbers
\newcommand{\Nat}{\mathbb{N}} %natural numbers
\newcommand{\CC}{\mathbb{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.

6.1 Margins in L^AT_EX

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 L^AT_EX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Broader Impact

Authors are required to include a statement of the broader impact of their work, including its ethical aspects and future societal consequences. Authors should discuss both positive and negative outcomes, if any. For instance, authors should discuss a) who may benefit from this research, b) who may be put at disadvantage from this research, c) what are the consequences of failure of the system, and d) whether the task/method leverages biases in the data. If authors believe this is not applicable to them, authors can simply state this.

Use unnumbered first level headings for this section, which should go at the end of the paper. **Note that this section does not count towards the eight pages of content that are allowed.**

230 References

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

235 [1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In
236 G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp.
237 609–616. Cambridge, MA: MIT Press.

238 [2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the*
239 *GENeral NEural Simulation System*. New York: TELOS/Springer-Verlag.

240 [3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent
241 synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

242 @articleDBLP:journals/corr/abs-1810-09305, author = Mahnaz Koupaee and William Yang Wang, title =
243 WikiHow: A Large Scale Text Summarization Dataset, journal = CoRR, volume = abs/1810.09305, year = 2018,
244 url = <http://arxiv.org/abs/1810.09305>, archivePrefix = arXiv, eprint = 1810.09305, timestamp = Wed, 31 Oct
245 2018 14:24:29 +0100, biburl = <https://dblp.org/rec/journals/corr/abs-1810-09305.bib>, bibsource = dblp computer
246 science bibliography, <https://dblp.org>