Automatically Generating Text from Film Material – A Comparison of Three Models

Sebastian Korenič Tratnik
Jožef Stefan International Postgraduate School
Faculty of Computer and Information Science
Večna pot 113
Ljubljana, Slovenia

Erik Novak
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

The paper focuses on audio analysis and text generation using film material as an example. The proposed approach is done by using three different models (Wav2Vec2, HuBERT, S2T) to process the sound from different audio-visual units. A comparative analysis shows the strengths of different models and factors of different materials that determine the quality of text generation for functional film annotation applications.

KEYWORDS

Text generation, automated transcription, cinema, film, video

1 INTRODUCTION

Applications like automatic text captions for video materials have become more and popular and extensively used by users on different media, spanning from the computer, television, smartphones and other technologies that enable audio-visual consumption. However, even though these applications have to an extent already become a staple in our everyday lives, their performance often varies and still has not reached optimal functionality. There are many challenges when we work with text generation out of audio-visual materials. These span from the structure and quality, the type or category of sound, the age of the recordings and the models on which such translation is based on. The main goal of this paper is to provide a practical demonstration of a few basic models for automatic annotation. The goal is to take into account the currently most common procedures of such an endeavour and figure out how to minimize the loss function of the models to allow an optimal generation of text out of film or video more sufficiently.

The rest of this paper is organized in the following way. Section 2 provides a description of the problem in the context of contemporary consumption of audio-visual materials via most popular information and communication technologies. Section 3 delineates the methodology used and describes the approach used to tackle the problem in a concrete demonstration. Section 4 presents the models being used and describes our implementation of them, specifying the dynamics of the obtained results. A conclusion is reached in section 5, where the paper offers a discussion on the outcome and possible directions for future work.

2 PROBLEM DESCRIPTION

In recent years, audio-visual data has become as influent if not more influent as traditional text-based information. With this, the task of extracting information from the former and transforming it into the latter is becoming useful for different purposes [1, 2]. One example is that text annotations enable better comprehension in cases of bad sound quality or even allow the material to be understood in situations where sound consumption is impossible. Another one is a possible speed up of the video that the annotations provide due to their ability to keep the content integral in a clear graphic form. The consumption process can be made more time efficient with textual information compensating for the distortions of audio-visual quality that can be brought about with the manipulations of playing options. Furthermore, in a general sense, combining audio-visual material with text can solve many problems on different levels of film or video production. This can span from the preparing phases of pre-production such as writing the script, to the post-production phases where one needs good orientation over a vast quantity of material. Proper text generation can facilitate easier orientation in such work and allows for more efficient organization of the media materials.

In this paper, we will focus on those components that contribute to the quality of proper automated text generation as a prerequisite of such developmental strategies. The main contributions of this paper are: (1) an analysis of the factors that influence automatic transcription of film or video material (2) implementation and comparison of a few different models for sound annotation (3) reflection on how this process can be used for more complex tasks

3 METHODOLOGY

The problem we are solving is to take a piece of audio-visual material, convert it into a code that a model for automatic text generation can take as input and then generate output of text that matches the sound recording of the input in an optimal way. An optimal result should provide a close correspondence of the utterances in the film material and eventually identify different types and categories of sound such as dialogue, noise, music etc. We will do an analysis of the factors that influence the quality of automatically generated transcriptions in the following steps: 1) a comparison of different models for generating text from audio files, 2) an analysis of how the quality of transcriptions differs in relation to noise in the background (silence, music, dialogues), 3) an evaluation of how the clarity of speech influences the quality of transcriptions, and 4) an assessment to what extent it is more difficult to generate quality transcriptions from older audio-inscriptions (films).

Reflecting on the results of our procedure, we will think about how to improve the quality in cases when quality of transcriptions is bad. Aside from quality we will measure the time demands of models, that is how much time do the models need to generate transcriptions from the audio writing.

The following model were used:

1) Wav2Vec2 [4] is a framework for self-supervised representation learning from raw audio that was made open-source by Facebook. It is the first Automatic Speech recognition model included in Transformers as one of the central parts of Natural Language Processing. Figure 1 shows the model's architecture.

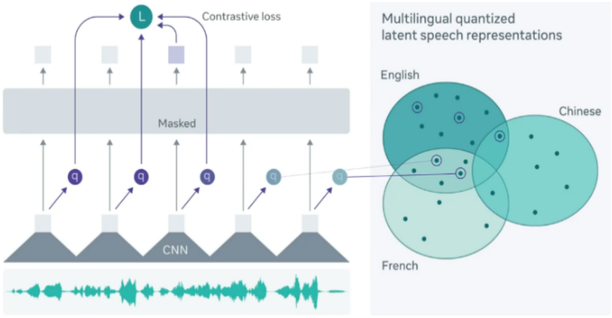

Figure 1. Wav2Vec2 learns speech units from multiple languages using cross-lingual training [4].

The model starts by processing the raw waveform with a multilayer convolutional neural network. This yields latent audio representations of 25ms that are fed into a quantizer and a transformer. From an inventory of learned units, the quantizer chooses appropriate ones, while half of the representations are masked before being used. The transformer then adds information from the whole of the audio sequence and with the output leads to solving the contrasting task with the model identifying the correct quantized speech units for the masked positions.

2) HuBERT [3] (Hidden-Unit BERT) is an approach for self-supervised speech representation that uses masking in a similar way and in addition adds an offline clustering step that provides aligned target labels for a prediction loss. This prediction loss is applied over the masked regions, which leads the model to learn a combined language and acoustic model over the continuous inputs. By focusing on the consistency of the unsupervised clustering step rather than the intrinsic quality of the assigned cluster labels, HuBERT can either match or improve the Wav2Vec2 model. Figure 2 shows the model's architecture.

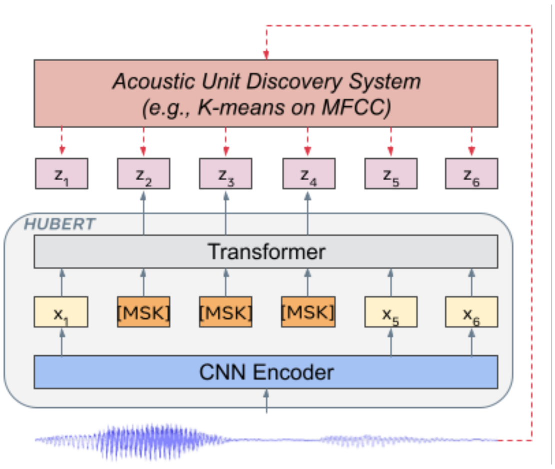

Figure 2. HuBERT predicts hidden clusters assignments using masked frames (y2, y3, y4 in the figure) generated by one or more iterations of k-means clustering [7].

3) S2T [5] (Speech2Text) is a transformer-based encoder-decoder (seq2seq) model that uses a convolutional downsampler to dramatically reduce the length of audio inputs over one half before they are fed into the encoder. It generates the transcripts autoregressively and is trained with standard autoregressive cross-entropy loss.

4 EXPERIMENT SETTING

4.1 Evaluation metric

We have used WER (Word error rate) as the metric of the performance of the models which computes the error rate on the comparison of substitutions, deletions, insertions and correct words. Original text was used for each of the model and each film example, removing the punctuation.

$$\mathbf{WER} = \frac{\mathbf{S+D+I}}{\mathbf{N}}$$

where...
S = number of substitutions
D = number of deletions
I = number of insertions
N = number of words in the reference

4.2 Data set

The dataset was formed with clips of different films. The films used were classics of world cinema (*The Godfather, 2001: A Space Odyssey, Star Wars, Frankenstein, Fight Club, Paris, Texas, Scent of A Woman, Tomorrow and Tomorrow and Tomorrow*). 14 clips of sizes spanning from 5 to 30 seconds were used with the lengthier ones incorporating different sound contents (like speech, shouting, whispering etc.). The first step was to prepare the audio in such a format that the models will be able to read it, so the clips were changed from mp4 to wav. An online converter, **cloudconvert** [https://cloudconvert.com], was used as the clips were fairly short and the results could be directly added to the Kaggle dataset from the browser itself.

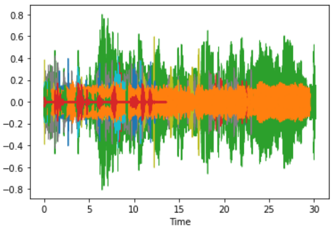

Figure 3: A superposition of waveform graphs of all the examples.

4.3 Implementation details

Programming was done on Kaggle, where code was written in Python and after the experiments were set up, and the GPU was activated for faster computation. The general process using each of the models is the following. First, an encoder takes raw data and puts it in the model. In our demonstration, tokenizers were used at the start, but as S2T tokenizers was not equipped to get the audio, it had to be changed to a processor. To retain consistency, the same step was applied to the other two models as well. Once data gets in the model, the model predicts particular syllables for each sound with certain probabilities and then in an additional step selects those with the highest probability based in the context of the semantic whole of the sentence. In the final step, the decoder (again the tokenizers / the processors) takes the output of the model and transforms it into text.

5 EXPERIMENT RESULTS

The ground rules for our project were that each model had a particular function that took sound as input and produced text as output with each audio having the text extracted separately. Subsequently different models were compared according to the accuracy of the results according to different criteria and a variety of scenarios (noise, music, number of characters, tempo of speech etc.). We will illustrate the obtained results via a concrete example. We will take a clip with relatively clear sound from the film *A Few Good Men* (1992), a digitized version of a well preserved celluloid film. The sound is clear and the dialogue takes places in a court practically in complete silence of the surroundings with the speech changing from normal tone to screaming. The clip is 22 seconds long and its waveform is shown in Figure 4. The original text is as following:

A: Did you order the Code Red?!
B: You don't have to answer that question!
C: I'll answer the question. You want answers?
A: I think I'm entitled!
C: You want answers!?
A: I want the truth!
C: You can't handle the truth! Son, we live in a world that has walls, and those walls have to be guarded by men with guns. Who's gonna do it? You? You, Lieutenant Weinberg?

The produced transcriptions are as follows:

Wav2Vec2:
YOU WAR THE CORA YOU DON'T HAVE TO ANSWER THE QU
ESTION I'LL ANSWER THE QUESTION YOU WANT ANSWERS I
THINK I'M ENTITLE YOU WANT ANT A AT THE TRUE YOU CA
N'T HANDLE THE TRUTH SON WE LIVE IN A WORLD THAT H
AS WALLS AND THOSE WALLS HAVE TO BE GUARDED BY ME
N WITH GUNS WHO'S GON TO DO IT YOU YOU LIEUTENANT
WINEBERG

HuBERT:
OMARTER TE CORET YOU DON'T HAVE TO ANSWER THAT Q
UESTION I'LL ANSWER THE QUESTION YOU WANT ANSWERS
I THINK I'M ENTITLED YOU WANT ANSWERRTHE TRUTH YO
U CAN'T HANDLE THE TRUTH SON WE LIVE IN A WORLD TH
AT HAS WALLS AND THOSE WALLS HAVE TO BE GUARDED B
Y MEN WITH GUNS WHO'S GOING TO DO IT YOU YOU LIEUT
ENANT WINBURG

S2T:
DEAR LORD THE CORRET YOU DON'T HAVE THE ANSWER T
HAT QUESTION I'LL ANSWER THE QUESTION YOU WANT AN
SWERS BUT THEY CAN'T ENTITLE YOU ONE AND THE TRUTH
YOU CAN'T HANDLE THE TRUTH SOME WE LIVE IN A WORL
D THAT HAS WALLS AND THOSE WALLS HAVE TO BE GUARD
ED BY MEN WITH GUNS WHOSE TENANT DO IT YOU LIEUTE
NANT WINEBURG THOSE HAVE TO BE GUARDED BY MEN WI
TH GUNS WHOSE CANNON DO IT YOU YOU LIEUTENANT WI
NEBURG YOU LIEUTENANT WINEBURG

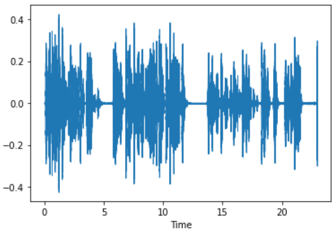

Figure 4. A scene from A Few Good Men (1992), a still and waveform graph from the used sequence.

The lower the WER number, the better the results. The models did not have a noticeable variation of speed, while the quality of their performance varied due to different factors. Hubert gave overall the best results from the point of view of readability. According to the rate of correspondence between input audio and output text, HuBERT comparably gave the better rate of the transcription in case of videos with poor audio quality from Wav2Vec2, i.e. that from older or damaged films, while Wav2Vec2 gave better performance in case of background music, but had the tendency of adding too much insertions. S2T had the tendency to produce mistakes, seen in peaking numbers over 1.0. The overall results are given in Table 1.

It is important to note that the average given does not reflect the better overall accuracy, but is the sum of different factors. So the models can be good at transcribing particular words, but can add or drop extra words in the process and therefore make the overall text less comprehensible. An important factor is the way the original text that is used for comparison is written – omitting punctuations and properly writing the words even if they are mispronounced will improve the results. Finally, it is crucial that all the texts are in caps lock, or the comparison won't work and will produce misleading results.

As the used example shows, it is mostly clarity of speech that will determine how the models perform. As the models were pre-trained and were not trained according to the specific data used, they were in general surprisingly efficient. The discrepancies in different treatments of the same audio are visible, but in general as long as the dialogue was clear, the results were comparable. Music seemed to cause bigger problems for the model than background noise, while additional speech in the background proved most problematic. Emotional influences on speech did not prove that problematic and even affective utterances were transcribed comparably with neutral speech if the sound data was of high quality.

Table 1. The WER scores for each model. The bold values represent the best performances on the given clip. The best performing model is HuBERT.

Clip number	Wav2Vec2	HuBERT	S2T
1	69%	**53%**	91%
2	100%	**0%**	100%
3	100%	**95%**	**95%**
4	**27%**	30%	36%
5	**17%**	**17%**	**17%**
6	39%	**18%**	43%
7	**28%**	**28%**	64%
8	70%	**46%**	55%
9	50%	**25%**	100%
10	57%	**37%**	73%
11	62%	**38%**	51%
12	100%	**95%**	100%
13	60%	**33%**	73%
14	9%	**4%**	9%
Average	56%	**37%**	65%

The WER usually shows the results in a metric between 0 and 1, however in case the annotation results were extremely unsuccessful, the higher extreme may surpass the limit. In our case, up to 1.6 was reached, however in the chart, it was limited down to 1.0 for purposes of clarity.

5 DISCUSSION AND FURTHER WORK

So as a general principle, when taking clips from films, the main factor that can potentially influence the quality of the generated text in a negative way is the background noise. As one can expect, the model will work best when nothing is in the background and worst when people are talking in the background. Ideally, to improve the quality one would train the models for the specific material, using a similar type of material and accordingly doing a pre-classification according to the main categories of sound analysis (ie. monologue, dialogue, background noise, music, echo, normal speech, loud speech, shouting, whispering etc.) - especially when using older or less preserved material, which drastically differs in sound data from newer or more preserved works.

In our research we expanded on and adapted existing work on automated text generation models, providing an analysis of the factors that determine the quality of such results from film material. As an example, we applied our approach on different film material, ranging in the quality and age of the clips and the structure of the sound data.

A useful strategy for the future from the perspective of film practice would be to find ways to link transcriptions with a script. A precondition of such an endeavour would be to implement an algorithm for recognizing the person speaking and identifying the source with descriptions ("person A is speaking, then person B, then person A has a long monologue, person C answers" etc.). Another important task would be identifying the sounds of different categories and providing fitting audio-signs (sound of squeaking steps, playing of music etc.). From these steps one could eventually at least to some extent automatically generate scripts for films or find ways to develop tools for easier text-based classification of audio-visual material.

CONCLUSIONS

In this paper we explored ways to generate text out of audio information presented in film and video material. We used three different models to evaluate various film units, Wav2Vec2, HuBERT, and S2T. We found that the model HuBERT achieved best results, while the remaining two methods performed similarly.

ACKNOWLEDGMENTS

The research described in this paper was supported by International Postgraduate School Inštitut Jožef Štefan, Ljubljana, Slovenia in the class Textual/ Multimedia mining and semantic technologies held by dr. Dunja Mladenič under the mentorship of Erik Novak. We also thank Besher Massri, Aljoša Rakita and Martin Abram for additional feedback.

REFERENCES

1 A. Ramani, A. Rao, V. Vidya and V. B. Prasad, "Automatic Subtitle Generation for Videos," *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 132-135, doi: 10.1109/ICACCS48705.2020.9074180.

2 Rustam Shadiev, Yueh-Min Huang, Facilitating cross-cultural understanding with learning activities supported by speech-to-text recognition and computer-aided translation, Computers & Education, Volume 98, 2016, Pages 130-141

3 Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. [arXiv:2106.07447v1 [cs.CL], Submitted on 14 Jun 2021].

4 Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.. [arXiv:2006.11477v3, Submitted on 20 Jun 2020 (v1), last revised 22 Oct 2020 (this version, v3)].

5 Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, Juan Pino. fairseq S2T: Fast Speech-to-Text Modeling with fairseq. [arXiv:2010.05171v1, Submitted on 11 Oct 2020] .

6 Wav2vec2.0: Learning the structure of speech from raw audio. [https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio Submitted on 24 Sep 2020, Access. 9.1.2022

7 Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 3451-3460.