Transcript Analysis

Lazykh’s Example Transcript Analysis

In this analysis, we aim to gain insights into how Cary Kaiming Huang, a content creator on YouTube with channels carykh and lazykh, creates his videos. We will focus on the example transcript provided in the lazykh repository, which is an open-source software developed by Cary to automate the production of animated videos. Our ultimate goal is to leverage the lazykh software as a starting point for our project to automate the process of turning text into video.

As we only have access to one example transcript, our analysis will be limited in scope. However, we will still be able to learn from the structure and content of the transcript. Additionally, we will examine any specific tags or markers used within the text, to inform our automated video creation process.

Once we have a solid understanding of the transcript's structure and content, we will explore the lazykh software source code to see how it utilizes the information from the transcript to generate videos. This exploration will help us identify potential improvements or modifications necessary to better automate the process of turning text into video using the lazykh software. By combining our insights from the transcript analysis and our understanding of the software, we aim to create a robust pipeline for converting text into animated videos.

Example Transcript Overview

Before we dive into the analysis, let's take a look at the example transcript provided in the lazykh repository. This will give us a better understanding of its structure, formatting, and content. Here's the complete transcript as it appears in the repository:

```
<explain> So.
I up loaded a video show casing my [polished planets]
<confused> cause I also added
<sad> [five] other planets
<explain> in twenty twelve!
And although it went under the [radar] for quite a few years
<explain> Yeah, eventually, it became my [most viewed video]!
<rq> but why am I returning
to this topic, eight years later?

<explain> Well, back then,
<sad> I didn't really want, or know, how to share my world files [online],
so I thought the video, alone, was enough.
<explain> Now, with a new perspective,
and wanting to be more collaborative, in the you tube online world,
I'm coming to the realization that,
<happy> it would be really cool, to see other Mine craft players
playing around, with my giant Earth.
<explain> in their own creative ways.
So, I'll put links to download all the world files, in the description.
So you guys can play with it!
<happy> Remember, it's got four times the blocks, six times the planets, and
a hundred
times the caring es than that other world.

<sad> And, uh, I I know that Pep in f t asses one to one Earth replica
will blow my earth out of the water when it's done,
<explain> but they're not quite done yet!

I'll also put the yo sam itty and men juror sponge world files online,
<sad> but not the pock will bells Cannon in Mine craft, or
slope Desert worlds, since those rely on mod, that are defunct now.

<happy> I also want to say thanks, to moe hang and Notch for making world
files
backward compatible,
<sad> because I was worried for the longest time
that my world files from Mine craft one point two,
would be super difficult to convert into the modern Mine craft one point
sixteen.
<happy> But in truth, it couldn't have been easier!

<explain> Back on the topic of the planets world.
In the last week,
I've also manually added
maybe fifteen to twenty new objects to the size comparison arena,
<rq> since I am the Scale of the Universe guy, I guess.
<explain> So I thought it might be fun to use this video to show them off to
you!
```

```
<happy> So let's jump back in to mine craft.
```

From the example transcript, we can observe the following:

1. **Emotion tags:** The transcript uses emotion tags enclosed in angle brackets, such as <happy>, <sad>, and <explain>. These tags provide an indication of the speaker's emotional state during different parts of the video, which may be useful for generating appropriate animations or visuals to match the content.
2. **Dialogue:** The transcript is mainly composed of dialogue, which is the spoken content of the video. The dialogue is divided into lines or paragraphs, often separated by emotion tags or other markers.
3. **Markers:** The transcript may contain additional markers, such as [...], which could be used to highlight specific words or phrases within the dialogue. These markers may represent important keywords or concepts that can be emphasized during the animation process.
4. **Questions:** The transcript may include questions, denoted by the <rq> tag, which could represent rhetorical questions or prompts for further discussion. These questions can be used to structure the video content and guide the viewer through the narrative.
5. **Repetition:** The transcript contains instances of repeated words or phrases, such as "I I." This repetition could be a result of the speaker's natural speech pattern or an emphasis on certain points. When converting the text into an animated video, it might be useful to consider whether to keep these repetitions to make the speech pattern sound more natural or adjust them for smoother narration.
6. **Punctuation:** There are instances in the transcript where punctuation, such as commas, may be used more frequently than expected. This could indicate a specific speech pattern, pauses, or emphasis on certain parts of the content. This could also be due to the implementation of the software that picks a different pose for the same emotion.
7. **Unconventional spelling:** The transcript includes instances of unconventional spelling or abbreviations of words, which could be intentional or a result of transcription errors. When analyzing the content and converting it into a video, it is crucial to consider these variations and adjust them as needed to ensure proper pronunciation and understanding.

Analysis of the Example Transcript

Emotion Tags

The emotion tags in the transcript may not directly reflect the emotion or sentiment present in the text itself, but rather serve as instructions on how the dialogue should be delivered, similar to how a movie director would guide an actor. We can use the emotion tags to adjust the 2D cartoons poses and visuals to match the intended delivery of the dialogue. By doing this we can create a more engaging and dynamic video that the intended audience would want to watch.

It might prove difficult to train a ML Model which can envision how a line of text should be delivered i.e., which pose the character *should* have. However, we can use sentiment analysis on the lines of text as a starting point, and incrementally improve how engaging and dynamic the cartoon character delivers the lines of text by using more sophisticated methods.

Dialogue

The dialogue in the example transcript can be categorized as conversational and is written from the point of view of Cary. The cartoon character and Cary are the same person for all intents and purposes. In our case this will be quite different. A manager or executive will write a piece of text, most likely an email, that is not written for the purpose of being a transcript of a video and it is not written as dialogue. So, when making the videos we will have to try and see if such a piece of text translates well to being delivered by a cartoon character.

Furthermore, the structure of the example transcript is different than everyday writing, in which the structure is arranged into sentences and paragraphs. The transcript is structured based on what billboard is shown on screen. The billboard is the visual (image or text) shown on screen besides the character. Furthermore, a new line is needed when we want to change the pose of the character. So, we should break sentences into phrases by adding new lines, based on what pose we want the character to have, and based on what billboards we want to show for the duration of a line of text.

Markers

The example transcript contains words between square brackets ([...]). This indicates the topic of the line of text. The lazykh software has a feature which allows the user to draw a billboard for every line of text. The words between the square brackets are called the topic of the line of text and are displayed as a title beneath the billboard. When the topic of a line of text is not present, then the whole line of text becomes the topic, and the whole line of text would also be displayed beneath the billboard as the title of the billboard.

Questions

If rhetorical questions or prompts for further discussion are present in the transcript, they are denoted by the <rq> tag. The creator of lazykh mentions that this type of utterance is often included in his scripts, and that this is the reason for it being here. I think that this type of pose can be helpful in our case, since we want to change the behavior of people, this could be a useful tactic to employ when we want the audience to think about a certain topic. However, this could be exceedingly difficult to implement, as you would need to know that the author of the input text wants the audience to think about something. Furthermore, including the rhetorical question poses is not necessary to create a minimum viable product.

Repetition

The use of repetition of certain words is to make the speech more natural sounding. This is something not necessary to include in our scripts. If we included it, we would include it as an improvement on the minimum viable product. However, there may be other ways to make the speech sound more natural. Making the speech sound natural is not one of our priorities.

Punctuation

The example transcript contains more comma’s than one would typically use when writing prose. The reason for this is that the lazykh software is implemented in such a way that the pose changes when a comma or full stop is encountered. Controlling the poses of the cartoon character is one of our main objectives, but this level of control may not be necessary for the minimum viable product. This is something we can investigate after we have finished that.

Unconventional Spelling

The unconventional spelling and abbreviation of words is intentional. One of the dependencies of lazykh (Gentle) might not recognize certain words, due to the dictionary of words that is used. Which would result in the animation not being properly synchronized to the speech. To work around this problem, Cary has used alternative spelling, so that Gentle can recognize the words by breaking up the words into parts that are present in the dictionary. This can be a problem in our case too.

Analysis' Conclusion

The emotion tags are important to make the video more engaging and dynamic to watch. Our starting point will be to perform emotion classification and/or sentiment analysis on the text to assign the emotion tags. The structure of the annotated transcript does not follow the structure of everyday writing. Sentences must ideally split on change of billboards and/or emotions. The markers present in the example transcript are not relevant for the MVP but may be worth revisiting when we have finished building the MVP and aim to improve it. The rhetorical questions tags follow the same trajectory. These tags can be utilized to further enhance how engaging the video will be but is not part of the MVP. The repetition of certain words and use punctuation is also a further enhancement, respectively, they would make the speech more natural sounding, and more engaging by having multiple poses during the same emotion. Finally, the issue of unconventional spelling is something that we should address in the MVP. If the MVP cannot synchronize certain words and sounds, then this would not be considered an enhancement, but it should be fixed as part of the MVP.