Whitepaper
2018

AI Closed Captioning Services for Local and State Governments

IBM

This creates a clear situation where it is a best practice for governing bodies to begin captioning their material, as to not be perceived as discriminating against a specific audience.

The need for closed captions

Encouragingly, local and state legislatures have been adopting video into their processes. This includes streaming and archiving meetings, such as committee hearings or interim task force briefings, and making these available to the public. As a result, they have been fostering greater community engagement through enhanced involvement, while at the same time modernizing practices.

However, as government bodies embrace video, a dilemma around accessibility can emerge. In particular the notion of providing closed captions on content, making it more inclusive, which potentially can be costly. That said, there are strong motivating factors to caption video. One is to broaden how many people can watch, understand and embrace the content, such as those who are deaf or hard of hearing. Another is to cater to changing viewing habits, as online content can be watched muted with captions available.

A big motivator, though, is the concept that the legal landscape related to captioning may change for government entities. In 2016, the U.S. Department of Justice revised the Americans with Disabilities Act Title II regulations, introducing the possibility of accessibility requirements. In fact, the 2010 update stated that: "The Department intends to engage in additional rulemaking in the near future addressing accessibility in these areas and others, including next generation 9–1–1 and accessibility of Web sites operated by covered public entities and public accommodations." As a result, it was perceived that the 2016 update might incorporate this. It didn't, though, making it uncertain when such regulations might come into effect.

Some governing bodies, though, have been preemptive and now have strict accessibility legislation that requires captions or transcripts for all online video content. This practice is sound, too. Not only does it make content more widely accessible and caters to changing viewing habits, but it also works as a safeguard should regulations change. This quote from the Louisiana Law Blog[1] put it best:

"Despite the absence of clear regulatory guidance as to what standards may be required, the current enforcement position of the DOJ and of the "gotcha" plaintiffs is to rely on general regulations requirements under the ADA that goods and services be available and delivered to the public in a non-discriminatory manner."

This creates a clear situation where it is a best practice for governing bodies to begin captioning their material, as to not be perceived as discriminating against a specific audience.

[1] Boutwell, S.; "Website Accessibility and ADA Litigation", Louisiana Law Blog, 2016; https://www.louisianalawblog.com/business-and-corporate/website-accessibility-and-ada-litigation/

To use the $5 range, one is looking at roughly $300 per hour of captioned video.

Scaling caption generation

With the need for captions realized, the next step is to find an affordable, scalable way to implement them. Manual captioning is unfortunately very time consuming, as it can take 5-10 times the length of the video to caption it[2]. This means to caption an hour long committee meeting could take 5-10 hours, roughly an entire workday. Alternatively, paying for manual captioning can be costly. It's been stated that the "industry standard" cost is around $7-10 a minute[3], although a quick search can uncover ones priced in the $5 a minute range. To use the $5 range, one is looking at roughly $300 per hour of captioned video.

Time and cost have long been a reason why a lot of content isn't captioned today. However, as the need behind captioning mounts, so does the prospect of finding a solution. Thankfully, technology has begun to improve around automated captioning, moving to the forefront as an ideal solution to be able to scale caption generation.

ASR: how automated closed captioning works

To put it simply, automated closed captions are part of ASR (Automated Speech Recognition), created through a speech to text process. To make that concept a reality, several elements come into play that include:

- **Speech recognition**
  This involves being able to receive audio, from which it is then converted into a machine readable format, i.e. text.

- **Audio recognition**
  As technology was improved, audio recognition was introduced with the ability to separate sounds, like someone clapping, from actual speech when converting to a machine readable format.

- **Vocabulary**
  The ASR process will naturally try to match recognized speech against a large vocabulary list of terms. This is important to note, as the process doesn't try to phonetically create new words if it's not familiar with something. Instead it will try to match pronounced words with something similar within its vocabulary.

[2] "Accessible Video: Tips, Tricks, and Tools for YouTube (and Beyond)", MassMATCH: Massachusetts Rehabilitation Commission, 2010; http://www.massmatch.org/resources/accessiblevideo.php

[3] "Closed Caption Services", distribber; https://www.distribber.com/closed-caption-services

Thankfully work has been placed in reducing the accuracy gap through a key aspect: the introduction of artificial intelligence capabilities into the process.

The hurdle for ASR and caption generation

Technology around speech to text is not new. It has slowly gained in use, available now in smartphones and other devices. It has historically, though, not had the needed accuracy for many use cases. In fact it's not hard to think of the struggle someone has had in trying to use this technology by saying the same passage over and over again to be understood, trying to be clear, only for an "I'm sorry" message to be returned.

Caption generation has historically been one of those tricky areas, as if speech to text can struggle with someone deliberately trying to deliver a clear message, imagine the process trying to caption a faster paced committee meeting.

Thankfully work has been placed in reducing the accuracy gap through a key aspect: the introduction of artificial intelligence capabilities into the process.

AI and improving caption automation

Processes for automating caption generation have continued to improve, with a strong focus toward shrinking the accuracy gap that has previously existed. IBM has gone about doing this by infusing IBM Watson into the caption generation process. On the surface level, this has helped the speech recognition component as the AI can handle natural speech, accents and dialects better, although just like a normal human the accuracy will still not be as high as opposed to captioning simple speech that was spoken very clearly.

Another way AI is improving caption automation is around the realm of context. Homophones, words that sound the same but actually carry different meanings, make the process of caption generation hard, even for humans. The best way to demonstrate this is an example, so let's pick on "band" and "banned", two words that sound more or less the same. Now let's take a context example:

He was "banned" from the task force committee meeting.

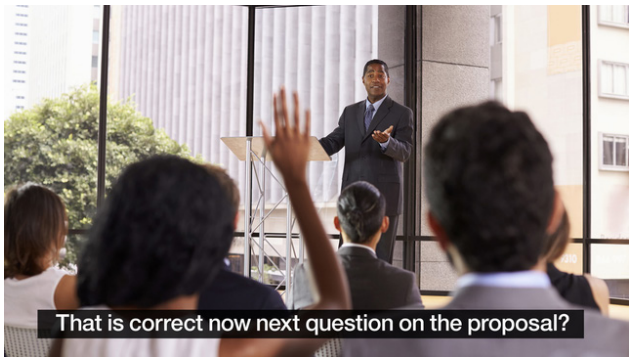
Given the context, it should be easy for a human to tell the difference. For an automated caption process that wasn't trained for context, though, it could be confused to use either band or banned. Now let's pick something even harder, let's go for "sensor" or "censor" in this context:

Unfortunately the problem is the way the "censor" did their job.

Now in that limited context, it might be clear that it's referring to someone whose job it is to censor material. However, swapping out "their" to "its" and suddenly it's more likely to be talking about a sensor instead.

Being able to train AI plays a key role in its ability to generate increasingly accurate closed captions.

In this case, there is a lot of value in being able to teach an AI about context. Let it learn by example, which can be done through allowing the AI to read over past examples, spotting trends so it knows not just homophones but also when someone slurs their speech what they might have actually meant. This is one way that AI is improving caption generation, the other is by being trainable.



Training AI to improve accuracy

Being able to train AI plays a key role in its ability to generate increasingly accurate closed captions. This is achieved through leaning on the idea that concepts and names will repeat across content from the same government body. For example, the name and specific spelling of a state senator might come up many times during meetings and other video content. Being able to teach the AI and add this specific spelling to the vocabulary will go a long way to producing captions that are more accurate from the start.

For IBM's implementation, this is done in a couple of ways. One is through adding words to the available vocabulary, in essence teaching it new words. For example, imagine that a government official's name is Peter Cartwell. Now without training the AI, it's very likely to take this and try to caption it as "Peter Cartwheel". As a result, the AI can be trained that when something sounds like "Peter Cartwheel" it's actually "Peter Cartwell". The same can be applied to local company names as well as landmarks or even street names. For example if a street is called "Livengood Road" it could be taught that "Living good" could be "Livengood".

Now this is one method, the other is to allow the AI to learn based on looking over a corpus, a collection of written texts. This functions by letting the artificial intelligence learn by example and, consequently, make better and more informed decisions when selecting how to caption something. A great, local example of this could be if a neighborhood is referred to as the "Hyde neighborhood". Now without context, "Hyde" would likely be captioned as "hide". This is where the corpus comes in, helping to showcase scenarios of when to use Hyde versus hide. A sample would be maybe community stated in context would be a give away to caption the phrase as "the Hyde Community", rather than "the hide community".

IBM offers methods to caption live or previously recorded video, including ways to train IBM Watson, and expanding its vocabulary by using an online dashboard.

Separating AI instances

Once you get into the practice of training the AI, and in effect "biasing" the results, it also introduces the notion of the importance of having separate instances for each government body. While this does preserve individual work, the main reason for this is to maintain accuracy, and make sure what's being taught is relevant. For example, while the "Hyde Community" might be relevant to your instance, it won't be to another government body where that word wouldn't come up. Similarly, imagine another government body has a "Wayzata Blvd" that was taught in their instance which would not come up in your results as well.

By separating these instances, it reduces the vocabulary words introduced, so that the AI doesn't have to decipher if someone just said "Wayzata" or if they slightly slurred and said "why's that?". Unfortunately, all words will rarely be spoken with crystal clear clarity, and as a result the fewer irrelevant words that the AI has to possibly insert into a passage, the better.

Implementation

Government bodies looking to introduce an automated closed captioning solution have options for both on-demand and live content. An ideal roll out should depend on something that is trainable, especially on the topic of live content where you can't edit material after the fact for accuracy.

IBM offers methods to caption live or previously recorded video, including ways to train IBM Watson, and expanding its vocabulary by using an online dashboard. For on-demand assets, the technology can also learn based on any corrections made. For example, if a new government official speaks for the first time and the spelling of their name needs to be altered, the AI learns how to spell the name and can apply the spelling correctly the next time. Corpus, for better context, can also be uploaded through providing completed caption files, in the SRT or VTT format, or through providing a custom document that contains a single sentence per line. Both need to be completely error free before submission, as Watson will use this to learn from.

For the purposes of live captioning, a script or any sort of additional context before a committee meeting or overview, like a schedule, can be used to help aid the direction the AI takes. For the task of live captioning, this functions differently from on-demand captions which will generate a caption file associated with the asset. For live content the technology begins by generating a transcript using speech to text. This is then fed over IP or serial to a closed caption encoder to be delivered as 608 captions during the broadcast, allowing end viewers to watch closed captioned content live.

To learn more about the AI based captioning solutions from IBM and how they can work for your state or local government use case, request a demo.

Produced in the United States of America
November 2018