White Paper

Automation in Captions and Subtitles Workflows

Sana Afsar
Interra Systems, Inc.

INTRODUCTION

Captions display the audio portion of media content as text, identify speakers, and describe other relevant sounds like clapping and applause, primarily for those who are deaf or hard of hearing. Captions are categorized as either open or closed. Open captions are always in view, as they are burnt into the video, whereas closed captions can be turned on and off by viewers. Special hardware or software decoders are required to view closed captions, and consumers need to learn how to turn captions on or off in order for them to be utilized. Despite these shortcomings, closed captions are preferred over open captions. Since closed captions exist as a separate text stream, they can be removed, and video can be archived and indexed. The same video can be relayed with audio and captions in multiple languages. Furthermore, unlike open captions, they are not subject to loss of quality when the encoded video is compressed. That's the reason the majority of the captions are closed captions. Subtitles are similar to captions; however, subtitles only contain the dialog portion of the audio. Subtitles were created for viewers that can hear but cannot understand the language or accent, or the speech is not entirely clear. Although subtitles are used interchangeably with captions, the technical equivalent of captions is Subtitles for Deaf or Hard-of-hearing (SDH), which contains additional descriptions for non-dialog audio. Outside of the U.S., the term subtitle is only used to refer to captions. The term SDH is used only when a distinction is required.

Captions and subtitles are mandated by all major broadcasters and regulated by laws because they help make content accessible to millions of viewers with hearing impairments and they make content more accessible to viewers across the world.
This paper will discuss the importance of content delivery with good quality captions, legal requirements related to accessibility, and it will examine the challenges broadcasters face when delivering captions. In addition, it will discuss new and affordable solutions for ensuring captions quality and thus ensuring full content monetization.

BENEFITS OF CLOSED CAPTIONS

Captions provide viewers with access to entertainment and important information. Beyond benefitting viewers that are hard of hearing, they also help non-native speakers. Many viewers without hearing impairments choose to use captions to ensure that they don't miss a single word of dialog. Moreover, a significant portion of viewers take advantage of captioning at public places like the gym, restaurants, airports, hospitals and libraries, where audio is low or muted. Captioning also aids in comprehension and memory retention for viewers, especially when the content is technical in nature.

Today, the majority of social media users watch promotional videos without sound. With applications like Facebook enabling video auto-play in silent mode, viewers can only fully understand the videos if they are accompanied by captions. This has been validated by internal tests at Facebook, which found that captioned video ads saw a 12 percent increase in engagement. A similar trend is observed for promotional videos on LinkedIn platform.

The availability of subtitles in local languages has made it possible to watch media content that is available only in a foreign language. Subtitles also aid in learning new languages.

The presence of text alongside video makes content more discoverable to search engines and ranked higher in search results, thus helping it achieve more views. Experiments have shown that captions and subtitles boost literacy skills as well. Increased exposure to words creates more fluent and proficient readers and helps in improving vocabulary.

REGULATORY COMPLIANCE

Various types of regulatory compliance for captions have been enacted to ensure equal opportunities to people with hearing disabilities. The Federal Communications

Commission (FCC) has created guidelines for ensuring caption quality and a good viewing experience.

The FCC rulings require closed captioning of TV programs and content reshown on the internet. Since most premium content that is delivered online in the U.S. also appears on television, it must have captions in order to be compliant. The FCC rules for caption quality require that captions be:

Accurate: Captions must match the spoken words in the dialog and convey background noises and other sounds.
Synchronous: Captions must coincide with their corresponding spoken words and sounds and must be displayed on the screen at a speed that can be read by viewers.
Complete: Captions must run from the beginning to the end of the program.
Properly placed: Captions should not block other important visual content on the screen, overlap one another or run off the edge of the video screen.

Captions also need to adhere to certain guidelines, such as TV content ratings in the United States, for profanity and content advisory. The FCC has made the transmission of content advisory information mandatory using eXtended Data Service (XDS) of closed captions. Channel blocking (e.g., muting the program audio, rendering the video black or otherwise indecipherable, eliminating program-related captions) occurs as soon as a program rating packet with the appropriate content advisory rating level is received.

Similarly in UK, the Office of Communications (Ofcom) was created in 2002 and requires UK broadcasters to provide television access services, like subtitling, sign language, and audio descriptors. Additionally, the Authority for Television On Demand (ATVOD), an independent co-regulator for video on demand in the UK, provides comprehensive guidelines for subtitling.

Web Content Accessibility Guidelines (WCAG) is another internationally recognized digital accessibility standard set by the World Wide Web Consortium (W3C) and is referred to in the accessibility policy of many European universities and elsewhere.

Besides, most broadcasters and OTT platforms have specific guidelines regarding the format of captions and subtitles.

CURRENT SITUATION AND CHALLENGES

Captions and subtitles have undeniable benefits. However, until recently they were more of a luxury rather than an integral part of media. Many video providers are still struggling to add quality captions and subtitles to their content and meet regulations due to the technical and financial challenges and lack of trained professionals.

Adding words to the video is not as simple as it seems. To start with, one needs a transcript and needs to perform intelligent segmentation taking care of all captioning rules. The intelligent segmentation phase relies on language understanding to ensure that the transcript is divided into semantically coherent units and each unit can be displayed properly on the video frame. Each caption text should not contain more than two or three lines to avoid covering too much of video. A line should not contain more than 42 characters or it will run out of space. Segmentation needs to take into account natural pauses, clause boundaries, compound words, and more, to ensure that the meaning of a sentence is not lost. Scene change detection can be employed at this step to ensure that captions do not run over the scene change boundaries. Then the segments need to be synchronized, or time stamped so that each segment can be displayed at the right time. In the past, this task was labor intensive and time consuming. Now tools are available to automate this process. Captioning of live telecasts, like news, sports events, and other live shows is especially challenging, as all these actions need to be done on the fly.

Catering to the character sets used in different languages and using correct encoding is also an error-prone task and needs to be done carefully to avoid any issues during display. The reviewed transcripts are then encoded as captions in media. This is also a challenging task, since a variety of encoding methods are used at the time of authoring, interchange, and delivery and in different geographies. Any slight mistake in encoding can also make captions unusable. Just as for audio and video, different methods and standards have evolved to meet specific requirements for captions. While the DTV 608-708 format is popular in the U.S., the European region widely uses teletext. These two can be encoded as a separate track within media or a side-car file while editing and exchange. Alternatively, they can be encoded within video frames for TV transmissions. There are different ways in which they can be encoded with video frames for different video codecs. They can also accompany the media as a separate TTML, STL or WebVTT file for internet delivery or a CineCanvas or DCDM file for a movie theater release.

Further captioning requires special attention while editing, as many times captioning is lost or distorted during editing/sub-clipping captioned video and format conversion. There is always a possibility of losing captions during transcoding if the operator misses the option to enable extracting and re-encoding captions in a new format or if the transcoding software itself is not capable of extracting and re-encoding captions. Sometimes during editing, metadata information to display captions properly onscreen is lost, rendering the captions in the edited portion unusable.

Changes in frame rate are also difficult to tackle while maintaining proper caption quality. To maintain caption sync with video, captions need to be encoded at the same frame rate as that of the video. And if the video frame rate changes, it can make captions out of sync, unless it is corrected.

Different encoding points may introduce shift or drift in captions, which become noticeable only after few minutes of video. Shift usually happens when captions are inserted before or after the intended audio. Sometimes this misalignment is found throughout the file (i.e., there is an overall shift, early or late, between audio and captions).

For a synchronization drift issue, captions start out with the correct timing, but slowly become earlier and earlier (or later and later) over the duration of the video. This can happen when the video and caption frame rate is different.

A key requirement today is to produce good quality captions and subtitles in all of the required languages in an increasingly short turnaround time, using cost-effective methods. Automated solutions have emerged as an efficient, affordable way to support these new workflows.

QUALITY CHECKING OF CAPTIONS AND SUBTITLES

Providing captions and subtitles with media boosts viewer engagement and expands its reach. However, any reduction in quality can ruin the viewing experience and can even incur regulatory fines and penalties. Enhancements in digital technology have helped improve the captioning process, but it's still evolving and requires comprehensive quality checks starting with transcription to transmission.

Format Detection

Since closed captions are "closed," aka hidden inside media, confirming its presence is not a trivial task. Captions and subtitles come in various forms. Since special hardware and software decoders are required to display specific formats, it is important that captions and subtitles in the intended format accompany the media. Otherwise the target decoder (i.e., TV set or media player) might not display them at all.

Sync With Audio and Video

Once caption or subtitle text is extracted, it should be checked for sync with audio and video. It's also crucial to check the positioning of captions over video to ensure that they are not covering or hiding a critical part of the video such as burnt-in text. Caption alignment can only be verified when compared with audio. A misalignment of more than a second or two may make captions unreadable, cause viewer confusion, and result in the final captions of a show segment or commercial running into the next program.

Accuracy and Completeness

Captions should be checked to see if they match the spoken dialog verbatim. Many times simplifying text or using alternate text does not convey the same meaning and is very frustrating to lip readers. Additionally, broadcasters need to ensure that there isn’t any drop in captions intermittently within video.

Languages Check

An asset might have individual caption files for different languages. Broadcasters need to ensure that content is being distributed in the right languages with the correct captions at the right times.

Reading Speed, Row and Column Count Check

Captions need to have different reading speed for different types of content, for example, slow reading speed for kid's program. It needs to have different row and column count for different display resolutions.

Segmentation Quality check

The segmentation needs to be evaluated to ensure that the meaning of a sentence is not lost while breaking a long sentence into smaller displayable segments. It needs to adhere to different guidelines proposed by OTT platfoms like Netflix Timed Text guidelines or BBC guidelines.

Presence of Undesirable Text

Captions need to be checked for presence of profanity to meet content advisory regulations. Sometimes the presence of ads and website links are also undesirable.

Spell Check

The caption text might contain some misspellings introduced during transcription or encoding and needs to be spell checked.

Metadata Conformance

The caption file needs to be checked for adherence to metadata and conformance related to frame rate, positioning, character set and time codes to ensure proper decoding.

Checking the quality of captions must be done at every encoding point in the content creation path, which can occur many times in a workflow. Doing this manually every time requires many skilled professionals, linguistic experts, and lots of time. It would increase turnaround time and make captions economically burdensome. The good news is that automated QC of captions has become more advanced. Using the right set of QC tools broadcasters can ensure a positive consumer experience and compliance of closed captioning with government regulations.

WORKFLOW FOR CAPTIONS AND SUBTITLES

A state-of-the-art Captions QC and Repurposing system should include extensive subtitles and closed caption data verification and correction. The system should ensure that content is being created, edited, distributed, and received in the right languages, with the correct captions and subtitles, at the right times. It can be used to check and correct captions at any point in the content lifecycle — from ingest through editing — and verify versions for delivery to OTT platforms and other partners. The system must ensure captions and subtitles are correct and in-sync at every step, in the proper format, and that the correct languages appear on the correct track, all without human intervention.

An automated QC and Repurposing system should accurately extract captions from all of the industry supported formats and perform quality checks and corrections on them. Such a system that uses speech-to-text and character recognition technologies can enable video providers and distributers to comply with industry regulations by making sure that captions and subtitles are in sync with audio and video. There should also be checks for profanity and other undesirable words, misspelled words, display duration, reading speed, safe title area and various other attributes of text required for a good viewing experience.

INTERRA SYSTEMS' SOLUTION

Interra Systems offers a leading machine learning (ML) driven live and file-based Captioning solutions. It provides comprehensive quality checks and correction efficiency in a flexible environment for several languages.

Interra Systems' BATON Captions meets all of the extensive requirements for high-quality subtitles and closed caption creation, verification as well as correction. The software puts all of the captions encoded in the content in an easy-to-read format with timing information, simplifying the review process. Users can review the text and errors, along with the frame-accurate audio and video playback in the BATON Media Player.

CONCLUSION

From entertainment to education, video is everywhere. Not providing accurate captions makes video content inaccessible to millions of users with hearing impairments. Subtitles take this a step further and open up content to additional viewers across the world. Global popularity of many television shows is solely attributed to captions and subtitles. Besides accessibility, captions and subtitles improve literacy and reading comprehension, help in learning new languages and boost video discoverability in search engines. That is why they are mandated by all major broadcasters and legally regulated.

Thanks to new ML solutions for text detection and speech to text conversion, many automated workflows for captioning — which may have seemed impossible or too expensive just a few years ago — are now a reality. The global amount of video content is set to grow and so are captions and subtitles. It is better to be prepared today with automated Captions solutions than to face a flood of caption-and subtitle-related challenges later. By deploying a complete solution for ingest to delivery, broadcasters can ensure content is delivered with superior quality captions and subtitles to meet legal requirements and to fully monetize their content.