White paper

Watson at work
Captioning goes cognitive: a new approach to an old challenge

IBM **Watson Media**

Cognitive systems can potentially inject greater context and improved accuracy at competitive costs

Issue:
Video captioning remains a challenge. New regulations may make it even more so.

Context:
Even with help from automated systems and third-party specialists, live and postproduction captioning is a relatively imperfect art that remains stubbornly resistant to breakthrough. New requirements for captioning of certain online video content will broaden demands for economical and accurate captioning.

Solution:
Cognitive systems have the potential to bring new capabilities to the captioning category, specifically in the areas of rapid ingest and the application of contextual, human-like understanding that can reduce errors and improve the comprehensibility of subtitles.

"Cognitive systems have a chance to succeed where previous captioning-automation platforms have fallen short."

The first commonly recognized instance of "closed-caption" subtitling was WGBH-TV's 1972 broadcast of the iconic cooking show "The French Chef" featuring the late Julia Child. Almost ever since, there has been a parade of initiatives introduced to automate or at least ease the painstaking process of translating spoken words and sounds to textual descriptions and subtitles.

External captioning services that offloaded the burden from broadcasters and networks were followed by the introduction of software that aimed to recognize sounds and speech. For more urgent demands, real-time captioning came onto the scene in 1982 when the National Captioning Institute inaugurated a service that promised to return textual interpretations of words and sounds within four to five seconds after its airing. Its secret ingredient: hordes of fleet-fingered court reporters wearing headphones and churning out nearly-live text within four to five seconds of its airing.

All of these entrants elevated the state of the art in captioning. None were ideal. "Live" captioning has produced numerous faux pas as typists struggled to understand and convert unfamiliar words, names and expressions to immediate output. Software solutions historically have been imperfect as well, and often for the same reasons. Misspellings, "inaudible" designations and, especially, difficulty in interpreting the surrounding context of spoken words have tried the patience of television industry customers who are forced to review and correct flawed output before applying captions to their finished video assets.

As a result, it's easy to sympathize with a certain "heard-it-before" sense of fatigue among video industry professionals who have been promised failsafe captioning solutions only to find the craft remains what it has always been: a manual-intensive and stubbornly imperfect process.

Cognitive systems breakthrough
Today, however, there's a new possibility on the captioning scene that has the potential for a genuine breakthrough. Cognitive systems combine earlier capabilities of video processing intelligence with something new – the ability to analyze, understand and "learn" the surrounding context of video content in much the same way that humans do. Thus, the word "fault" in the context of a tennis match is treated differently than it would be in the context of a daytime soap opera episode. And the resulting interpretation and display of words and descriptions that both precede and follow it are rendered more accurately and with much improved contextual presentation. Cognitive systems have a chance to succeed where previous captioning-automation platforms have fallen short because they create output that more closely tracks the intention, purpose and verbatim assemblage of words and sounds tied to video content. Because of their ability to examine and interpret, they work in almost the same way human transcription specialists do. Except they're faster.

“Self-learning from every correction, recognition accuracy improves with every use. ”

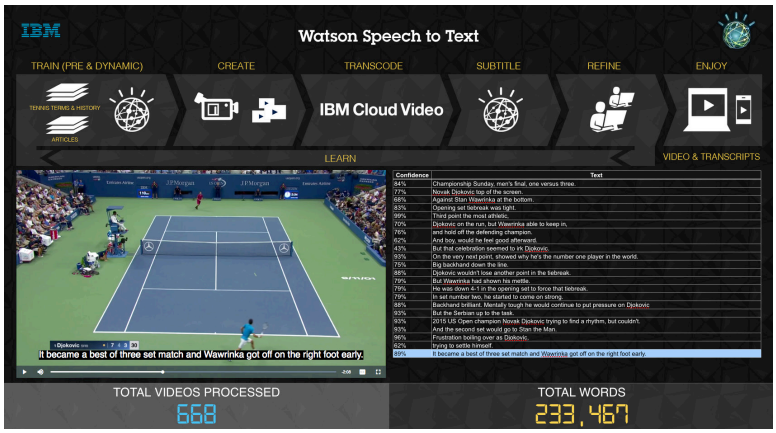
The key enabling technologies Watson brings to bear for captioning purposes are:

- **Language model customization** creates domain-specific language models to increase recognition accuracy
- **Custom corpora** extend vocabulary with words in context
- **Custom words** extend vocabulary with words and their phonetic form
- **Custom acoustic models** improve recognition accuracy for videos with specific audio conditions (background noise, special accents and so on)

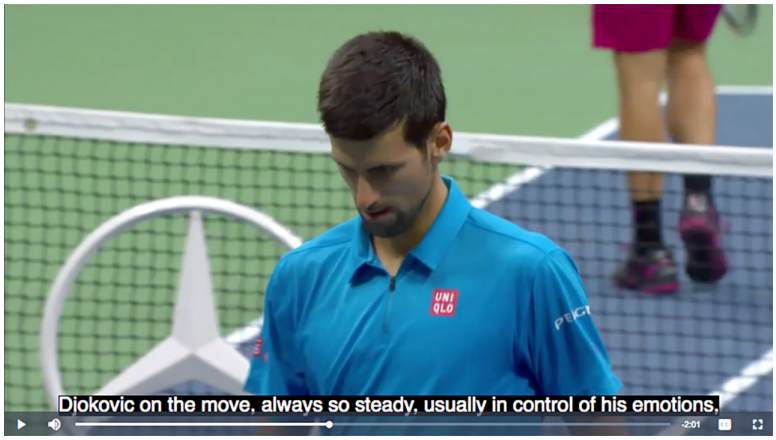
Captions powered by Watson

IBM Watson® uses automated speech-recognition capabilities to ingest spoken and audial elements of video assets. It then applies a range of cognitive functionality to assess and act on the interpreted data. Additionally, Watson enables customized captioning solutions by leveraging features such as corpus, vocabulary and custom audio models to further enhance the accuracy of first-run caption scripts.

Watson automatically generates captions for ingested videos using the Watson Speech to Text API. The caption editor feature of the API is designed to review and correct the automatically generated captions. The editor interface is designed for both experts and non-professionals and is optimized for maximum efficiency. Self-learning from every correction, recognition accuracy improves with every use. Names and proper nouns are automatically extracted from reviewed captions and used as glossary words, helping to ensure they will be recognized and spelled properly in subsequent uses. Caption generation is done by using a smart layout algorithm. Using this algorithm, Watson automatically segments caption cues at natural breaking points, which can result in greater readability.



The Watson + US Open demo shows how captions are automatically generated using the Watson Speech to Text API



Watson captions the US Open (demo environment)

The money question

Ultimately, of course, cognitive systems will have to prove their worth on an economic scale by producing usable output at an equal or lower cost than the prevailing industry norm. At the upper end of the range, video captioning expenditures for live content can range from USD 3 - 5 per minute, inclusive of both auto-generated scripts and human editing; the lower range for non-live and video-on-demand content is closer to USD 1 - 2 per minute. These are the boundaries cognitive systems must match or surpass to gain any sort of market traction.

The good news on this front is that cognitive systems are participants in a rising market for automatic content recognition (or ACR) at large, with positive implications for scale economics. The global market research firm MarketsandMarkets projects a compounded annual growth rate of 27.2% through 2021 for ACR technology in the media and entertainment sector, driven by a confluence of applications tied to audio fingerprinting, watermarking, music recognition, and music discovery. Essentially, the rising demand for better ways to find and discover personalized media content will lend itself to greater cost-sharing of ACR technologies across more sectors, potentially resulting in cost improvements for captioning users (among many other adopters). Also contributing to a greater scale is the broadening of the market for video captioning itself. New U.S. captioning rules that became effective in July 2017 require that online video content which was originally “shown on TV” appear with captions available. This requirement again has the potential to expand the overall market size for captioning, potentially providing a larger pool of investment that may help to bring costs down.

Watson uses automated speech-recognition capabilities to ingest spoken and audial elements of video assets

"...transcripts presented to editors for final review are more accurate and closer to market-ready than predecessor, technologies have allowed."

Accuracy, accuracy, accuracy

Of course, the other big consideration point for video industry players is compliance with federal rules. In the U.S., the FCC requires that captioning be:

- **Accurate**: Captions must match the spoken words in the dialogue and convey background noises and other sounds to the fullest extent possible.
- **Synchronous**: Captions must coincide with their corresponding spoken words and sounds to the greatest extent possible 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 to the fullest extent possible.
- **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.

Cognitive systems can contribute to most of these cornerstone requirements because of their ability to couple audio recognition with broader contextual understanding, so that the transcripts presented to editors for final review are more accurate and closer to market-ready than predecessor technologies have allowed.

Implementation scenarios

Video industry participants that want to understand the contribution cognitive systems may play in captioning efforts going forward may want to experiment with early-state implementations that provide a testing ground. Circumstances that may warrant initial trials may include any of these considerations:

- Where the total cost is similar or lower than available solutions
- Where an in-house solution is preferred
- Where turnaround time is a key factor
- For content where no regulations apply

The role of cognitive systems in television industry captioning practices is in its early stages. But there's clearly keen interest in adopting a solution that shows promise for transcending the limitations and boundaries of legacy approaches. The powerful combination of automated content recognition with cognitive/learning capabilities will bring new capability sets to a longstanding television industry practice. In the end, cognition may be exactly what captioning has wanted, ever since Julia Child showed the world how to cook like a French master chef.

IBM Watson Media
505 Howard Street
San Francisco, CA 94105

Produced in the United States of America
February 2020