7th International Symposium on Live Subtitling and Accessibility
Universitat Autònoma de Barcelona
5-6 November 2020 (online)

KEYNOTE

“Automatic Speech Recognition for Captioning Lectures”

Tatsuya Kawahara (Kyoto University)

VIDEO PRESENTATION https://youtu.be/wXkFKdoMY4A

ABSTRACT

As live streaming and video sharing of many events such as lectures and meetings have become prevailing, the demand for captioning is ever increasing. In this context, we have investigated automatic speech recognition (ASR) technology for captioning lectures. While re-speaking to commercial dictation software is often adopted, it still requires much skill and training. Therefore, we focus on the system that automatically transcribes the speech of lecturers. There are two scenarios: one is video captioning and the other is live captioning. Live captioning requires small latency or real-time processing while video captioning deals with already recorded material without time constraint, and thus demands high quality. Note that there are still differences between video captioning and verbatim records as video captioning has more weights on faithfulness, and verbatim records have priority on readability. These two factors of faithfulness and readability are both important but in trade-off relations.

First, we have conducted captioning video lectures of the Open University of Japan, which provides 300 courses via TV and radio programs, and the majority are also broadcasted via the Internet. The lectures are recorded in a studio, providing a good acoustic condition, but topics and vocabulary are technical and not covered by most of the commercial ASR engines. Thus, we adapt the ASR system to each course. Typically, by using textbook the accuracy gets 90%, and when a script is prepared, though this is not true in many lectures, the accuracy reaches 95%.

We have also developed live captioning software. While steno-type is used by professional people in parliaments, courts and broadcasts, the layman volunteers at school and town events use normal PCs with software that allows for collaborative PC captioning. It typically requires 3 or 4 persons in one lecture. Similarly, dedicated ASR systems have been developed for parliaments, courts and broadcasts, while layman volunteers use commercial or free software. Since ASR output essentially includes errors, we need a human editor to make corrections. It translates that the ASR system and a human make a collaboration, and the same collaborative captioning software can be used. In Japan, free software named IPtalk is most widely-used for live captioning by volunteers. We have developed an ASR plug-in for this software, so ASR results can be post-edited easily.

However, ASR-based systems are not yet prevailing for captioning. For usable level, accuracy of 85-90% is desired, and if below 80% the system is not usable because correction is not possible in real-time. To achieve the usable accuracy, it is necessary to ensure (1) fluent speaking, (2) clean recording by tuning a microphone and an amplifier, and (3) coverage of technical terms by customizing the lexicon.

There are several issues in captions. First, in terms of the amount of text, too many texts are not easy to read. Second, in terms of faithfulness, too verbatim texts are not easy to read. Third, with regard to timing, perfect real-time is not friendly. Moreover, these depend on the user. The most controversial issue is the verbatim caption vs. summarized caption. The verbatim output keeps the speaking style and speakers' characters and liked by some hard-of-hearing people. ASR is suitable for this purpose. On the other hand, summarized caption like movie subtitles only keeps the content of speech, and liked by many deaf people. It is only possible by human editors who have summarization skills.