MOOC Translation for openHPI

Automated Translation of E-learning content and community created content

Motivation

The audience of openHPI courses in English language often has a very international background. Only a minority of the participants are native English speakers. For many of them, subtitles in their native language are a valid and important enhancement of the course materials. Transcribing and translating the courses manually would require a major financial investment.

Another use case emerges on OpenWHO, the MOOC platform of the World Health Organization (which is powered by the same software as openHPI). Often, the population in those countries that are frequently affected by epidemics doesn’t speak English but rather a variety of local dialects. Producing separate courses (on the same topic) in all required languages would be very cost- and time-intensive.

Automated transcription and translation is a very effective tool to reduce the costs for these procedures. We already have several components at hand that we are using in this context. However, the whole process still requires many time-intensive manual steps that could be replaced by an appropriate tooling.

openHPI Forum talk by Christian Willems and Markus Egg: https://www.tele-task.de/lecture/video/6574

Fig. 1: draft proposal for openHPI subtitle transcription and translation

Challenges

Translation of learning content is a very complex workflow (see Fig. 1), that consists of a number of very different steps. A large proportion of the workflow can be automated, some can be outsourced to the crowd. However, some steps may need editorial supervision or manual quality assurance.
A workflow for providing translated video subtitles roughly goes through the following steps:

I. Video Transcription
   A. Automated Speech Recognition (Speech to text)
   B. Sentence Boundary Detection (Determine proper punctuation)
   C. Cueing (Maintain timing information for subtitles)
   D. Quality Assurance (Important: mistakes from transcription will “grow exponentially” in the subsequent steps)

II. Subtitle Translation
   A. Pre-processing (Prepare text content for translation; must be full, single sentences)
   B. Machine Translation
   C. Post-processing (Re-create subtitle cues with proper timing; format conversion)

III. Connect subtitle tracks to video resources

Additional content types in the context of MOOCs that are only subject to translation, but not to transcription, would be PowerPoint Slides, (rich-)text content, quizzes, or forum discussions.

Project Objectives

There are a number of tools for a part of the workflow steps available from research projects, EU projects or technology partners, including:

- Automated Subtitle Generation Toolkit (ASGT) - developed at the HPI by Xiaoyin Che et al. in an earlier Master’s Project
- Transcription: Universidad Politecnica de Valencia (https://politrans.upv.es/)
- Translation Engines: TraMOOC, SAP Leonardo translation engine, DeepL

Other tools exist on the market or as Open Source initiatives, some tools might yet have to be built.

The actual project objectives are:

- Rethink the proposed translation workflows (see Fig. 1)
- Evaluate existing tools and choose a technology stack
- Evaluate the use of open standards for machine-translation APIs
- Try to avoid as many manual steps as possible
- Implement the “missing links”, establish a consistent workflow
- Create a dashboard for content editors allowing to supervise and control the workflow
- Design a solution that can be integrated with various content sources (e.g. other LMS)
- Optional: Speech Synthesis (Text to speech) for translated subtitles?

Project Partner

World Health Organization / OpenWHO

Contact

Prof. Dr. Christoph Meinel  
meinel@hpi.de  
H-1.37

Tom Staubitz  
thomas(staubitz@hpi.de  

Christian Willems  
christian.willems@hpi.de  
H-1.37