Proceedings of Machine Translation Summit XVIII

https://mtsummit2021.amtaweb.org

1st Workshop on Automatic Spoken Language Translation in Real-World Settings

Organizers:
Claudio Fantinuoli and Marco Turchi

1st Automatic Spoken Language Translation in Real-World Settings

Organizers
Claudio Fantinuoli fantinuoli@uni-mainz.de
Mainz University/KUDO Inc.
Marco Turchi turchi@fbk.eu
Fondazione Bruno Kessler

1 Aim of the conference

To date, the production of audio-video content may have exceeded that of written texts. The need to make such content available across language barriers has increased the interest in spoken language translation (SLT), opening up new opportunities for the use of speech translation applications in different settings and for different scopes, such as live translation at international conferences, automatic subtitling for video accessibility, automatic or human-in-the-loop respeaking, or as a support system for human interpreters, to name just a few. Furthermore, specific needs are emerging in terms of user profiles, e.g. people with different abilities, and user experiences, e.g. use on mobile devices.

Against this backdrop, the Spoken Language Translation in Real-World Settings workshop aims to bring together researchers in the areas of computer science, translation, and interpreting, as well as users of SLT applications, such as international organizations, businesses, broadcasters, content media creators, to discuss the latest advances in speech translation technologies from both the perspective of the Computer Science and the Humanities, raising awareness on topics such as the challenges in evaluating current technologies in real-life scenarios, customization tools to improve performance, ethical issues, human-machine interaction, and so forth.

2 Invited Speakers

2.1 Marcello Federico, Amazon AI

Recent Efforts on Automatic Dubbing

Automatic dubbing (AD) is an extension of automatic speech-to-speech translation such that the target artificial speech is carefully aligned in terms of duration, lip movements, timbre, emotion, and prosody of the original speaker in order to achieve audiovisual coherence. Dubbing quality strongly depends on isochrony, i.e., arranging the target speech utterances to exactly match the duration of the original speech utterances. In my talk, I will overview ongoing research on AD at Amazon, while focusing on the following aspects: verbosity of machine translation and prosodic alignment. Controlling the output length of MT is crucial in order to generate utterances of the same duration of the original speech. The goal of prosodic alignment is instead to segment the translation of a source sentence into phrases, so that isochrony is achieved without negatively impacting on the speaking rate of the synthetic speech. Along my talk, I will present experimental results and demo videos on four dubbing directions – English to French, Italian, German and Spanish.

Bio

Marcello Federico is a Principal Applied Scientist at Amazon AI, USA, since 2018. He received the Laurea degree in Information Sciences, summa cum laude, from the University of Milan, Italy, in 1987. At Amazon, he leads a research project on automatic dubbing and oversees the science work behind the Amazon Translate service. His research expertise is in automatic dubbing, machine translation, speech translation, language modeling, information retrieval, and speech recognition. In these areas, he co-authored 225 scientific publications, contributed in 20 international and national projects, mostly as scientific leader, and co-developed open source software packages for machine translation (Moses) and language modeling (IRSTLM) used worldwide by research and industry. He has served on the program committees of all major international conferences in the field of human language technology. Since 2004, he is on the steering committee of the International Conference on Spoken Language Translation (IWSLT) series. He has also been editor-in-chief of the ACM Transactions on Audio, Speech and Language Processing; associate editor for Foundations and Trends in Information Retrieval, and a senior editor for the IEEE/ACM Transactions on Audio, Speech, and Language Processing. He has been a board member of the Cross Lingual Information Forum and the European Association for Machine Translation (chair of EAMT 2012), founding officer of the ACL SIG on Machine Translation. He is currently President of the ACL SIG on Spoken Language Translation and associate editor of the Journal of Artificial Intelligence Research. He is a senior member of the IEEE and of the ACM.

2.2 Prof. Silvia Hansen-Schirra, Mainz University

CompAsS - Computer-Assisted Subtitling

With growing research interest and advances in automatic speech recognition (ASR) and neural machine translation (NMT) and their increasing application particularly in the captioning of massive open online resources, implementing these technologies in the domain of TV subtitling is becoming more and more interesting. The CompAsS project aims at researching and optimizing the overall multilingual subtitling process for offline public TV programmes by developing a multimodal subtitling platform leveraging state-of-the-art ASR, NMT and cutting-edge translation management tools. Driven by scientific interest and professional experience, the outcome will reduce resources required to re-purpose high-quality creative content for new languages, allowing subtitling companies and content producers to be more competitive in the international market. Human and machine input will be combined to make the process of creating interlingual subtitles as efficient and fit for purpose as possible from uploading the original video until burning in the final subtitles. Post-editing of written texts is standard in the translation industry, but is typically not used for subtitles. By post-editing subtitles, the project hopes to make significant gains in productivity while maintaining acceptable quality standards. The planned pipeline foresees the use of ASR as a first step for automatic film transcript extraction, followed by human input, which converts the ASR texts into monolingual subtitles. These subtitles are then translated via NMT into English as relay language and several target languages (e.g., German) and finally post-edited. From a scientific perspective, the CompAsS project evaluates the multimodal text processing of movie transcription with automatic-speech recognition and neural machine translation. Applying well-established methods from translation process research, such as keylogging, eye tracking, and questionnaires, this study provides the basis for the interface design of the CompAsS subtitling tool. We investigate how professional subtitlers and translation students work under eight different conditions: two transcription, three translation and three post-editing tasks. We use established measures based on gaze and typing data (i.e. fixations, pauses, editing time, and subjective ratings) in order to analyze the impact of ASR and NMT on cognitive load, split attention and efficiency.

Bio

Silvia Hansen-Schirra is Professor for English Linguistics and Translation Studies and Director of the Translation Cognition (TraCo) Center at Johannes Gutenberg University Mainz in Germersheim. She is the co-editor of the book series "Translation and Multilingual Natural Language Processing" and "Easy – Plain – Accessible". Her research interests include machine translation, accessible communication and translation process research.

2.3 Juan Pino, Facebook AI

End-to-end Speech Translation at Facebook

End-to-end speech translation, the task of directly modeling translation from audio in one language to text or speech in another language, presents advantages such as lower inference latency but faces a data scarcity challenge, including for high resource languages. In this talk, various data and modeling solutions are presented in order to overcome this challenge. Similar to the textual machine translation case, multilingual speech translation provides maintainability and quality improvements for lower resource language pairs. We present our initial efforts on this topic. As simultaneous speech translation is a prerequisite for practical applications such as simultaneous interpretation, we also give an overview of our investigations into end-to-end simultaneous speech translation. Finally, we describe initial work on speech translation modeling for speech output.

Bio Juan Pino is a Research Scientist at Facebook, currently working on speech translation. He received his PhD in machine translation from the University of Cambridge under the supervision of Prof. Bill Byrne.

2.4 Prof. Bart Defrancq, Ghent University

Will it take another 19 years? Cognitive Ergonomics of Computer-Assisted Interpreting (CAI)

In 1926 the first experiments were held where interpreters were required to interpret diplomatic speeches (semi)- simultaneously. Different experimental setups were put to the test to study interpreters' performances and simultaneous interpreting was successfully carried on from 1928 on in different diplomatic contexts (Baigorri-Jalón 2014). However, the real breakthrough only came in 1945 with the Nüremberg trials, where simultaneous interpreting was offered for weeks in a row and served as a model for the organisation of standing diplomatic conferences. Recent years have seen the development of the first usable CAI-tools for simultaneous interpreters, based on automatic speech recognition (ASR) technologies. These tools provide interpreters not with full transcripts of speeches but rather with lists of specific target items that pose problems, such as numbers, terms and named entities. Full transcripts are of little use for simultaneous interpreters as they are working with extremely narrow time frames with regard to the source text and combine several cognitive, language-related tasks. Adding the (language-related) task of consulting a running transcript of the source speech would probably over-burden cognitive processing in interpreters. Experiments with simulated ASR and ASR prototypes have shown that the provision of targeted information improves interpreters' performances on the accuracy dimension with regard to the rendition of the target items (Desmet et al. 2018, Fantinuoli Defrancq 2021). The first analyses of cognitive load associated with consulting ASR while interpreting suggest that no additional cognitive load is involved with the use of the prototype ASR. However, all aforementioned studies were conducted in quasi-experimental settings, with carefully presented speeches by native and near-native speakers, in physical interpreting booths and using prototypes whose features are based on intuition rather than on ergonomic analysis. There is a real risk that in the absence of systematic ergonomic analysis, CAI-tools will face the same fate as simultaneous interpreting technology. In my contribution I will apply Cañas' (2008) principles of cognitive ergonomics to the integration of ASR in interpreting booths or

remote simultaneous interpreting (RSI) platforms. According to Cañas, successful integration of software in the human workflow relies on 4 requirements: it should (1) shorten the time to accomplish interaction tasks; (2) reduce the number of mistakes made by humans; (3) reduce learning time; and (4) improve people's satisfaction with a system. Cognitive ergonomics seeks improvement in those areas to make the execution of the overall task assigned to what is called the "Joint Cognitive System", i.e. the joint processing by humans and devices involved in that task (Woods Hollnagel 2006), more successful. I will argue that although the first research results based on data from physical booths are encouraging, the integration of ASR in the interpreters' workflow on RSI platforms will face particular challenges.

References

Baigorri-Jalón, J. (2014). From Paris to Nuremberg. The Birth of Conference Interpreting. Amsterdam: Benjamins. Cañas, J. (2008). Cognitive Ergonomics in Interface Development Evaluation. Journal of Universal Computer Science, 14 (16): 2630-2649.

Defrancq, B., Fantinuoli, C. (2020). Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. Target 33(1): 73-102.

Desmet, B., Vandierendonck, M., Defrancq, B. (2018). Simultaneous interpretation of numbers and the impact of technological support. In Interpreting and technology (pp. 13–27). Language Science Press.

Woods, D. Hollnager, E. (2006). Joint Cognitive Systems: Patterns in Cognitive Systems Engineering. Boca Raton: CRC Press.

Bio

Born in 1970, studied Romance Philology at Ghent University (1987-1991) and was granted a PhD in Linguistics at the same University in 2002. Worked at the College of Europe as a French lecturer from 1992 until 1995, as a researcher at Ghent University from 1995 until 2007, as a visiting professor at the Université Catholique de Louvain-la-Neuve from 2004 until 2009 and as a postdoctoral researcher at Hogeschool from 2007 until 2010. Trained as a conference interpreter in 2010 and was appointed as an assistant professor of interpreting and translation the same year. Has been head of interpreter training both at the masters' and at the postgraduate levels since 2010, both at Hogeschool Gent and University Ghent (since 2013, when the department was moved from the Hogeschool to the University in the framework of an institutional reform). Is a member of the Department Board, the Faculty Board, the Research Commission of the alpha-Faculties, the Doctoral School Board and of the CIUTI Board.

3 Scientific committee

- Alina Karakanta Fondazione Bruno Kessler
- Evgeny Matusov AppTek
- Jan Niehues University of Maastricht
- Sara Papi Fondazione Bruno Kessler
- Franz Pöchhacker University of Vienna
- Bianca Prandi Johannes Gutenberg-Universität Mainz
- Pablo Romero-Fresco Universidade de Vigo
- Juan Pino Facebook
- Claudio Russello UNINT - Rome
- Matthias Sperber Apple
- Sebastian Stueker Karlsruhe Institute of Technology S
- hinji Watanabe Johns Hopkins University

Contents