Lecture Translator
Speech translation framework for simultaneous lecture translation

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger
Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani
Sebastian Stüker, Alex Waibel
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Adenauerring 2, 76131 Karlsruhe, Germany
m.mueller@kit.edu

Abstract

Foreign students at German universities often have difficulties following lectures as they are often held in German. Since human interpreters are too expensive for universities we are addressing this problem via speech translation technology deployed in KIT's lecture halls. Our simultaneous lecture translation system automatically translates lectures from German to English in real-time. Other supported language directions are English to Spanish, English to French, English to German and German to French. Automatic simultaneous translation is more than just the concatenation of automatic speech recognition and machine translation technology, as the input is an unsegmented, practically infinite stream of spontaneous speech. The lack of segmentation and the spontaneous nature of the speech makes it especially difficult to recognize and translate it with sufficient quality. In addition to quality, speed and latency are of the utmost importance in order for the system to enable students to follow lectures. In this paper we present our system that performs the task of simultaneous speech translation of university lectures by performing speech translation on a stream of audio in real-time and with low latency. The system features several techniques beyond the basic speech translation task, that make it fit for real-world use. Examples of these features are a continuous stream speech recognition without any prior segmentation of the input audio, punctuation prediction, run-on decoding and run-on translation with continuously updating displays in order to keep the latency as low as possible.

1 Introduction

The rapid development of communication technology nowadays makes it easier than ever before to communicate with other people independent of distance. With distances becoming irrelevant, one of the last barriers that hinders communications are different languages. Although English has become a lingua franca in large parts of the world, in many situations and for many people it is not an option. The different languages in the world also carry cultural heritage that needs to be protected. Forcing people to speak the same language will lead to a sever loss of cultural diversity.

There exist multiple possibilities to overcome this language divide. One possibility is to use interpreters for simultaneous interpretation. But since this is a very costly method, it is only possible in certain areas. One example is the European Parliament, where the demand for translation services is met by human interpreters.

Another area that can benefit from translation services are universities in non English speaking countries. Looking at the statistics, universities in English speaking countries have on average a higher percentage of students from abroad. One reason for this difference is the language barrier. While offering lectures in English might increase a university's attractiveness towards foreign students it is not desirable due to the loss in cultural identity and intellectual diversity that occurs when universities around the world stop teaching in their native language. Unlike the European Parliament, universities do not have the funds to employ sufficient amounts of human interpreters to simultaneously translate their lectures. Therefore, we developed a fully automatic translation solution that fits a university's budget and deployed it within the Karlsruhe Institute of Technology (KIT). By combining state-of-the-art automatic speech recognition (ASR) and machine translation (MT) with auxiliary technologies, such as resegmentation, punctuation prediction, and unsupervised speaker and domain adaptation we created a system that performs this task.

Developing systems for simultaneous translation poses several challenges. While the output should be of reasonable quality in order to being useful, the system is required to produce it in a timely fashion. Interactive scenarios like university lectures demand low latency. The delay of the output should be as low as possible in order to match the slides and the lecturers gestures. Due to reasons, such as multimodal channels for the consumer and the lack of a need of additional technology in the lecture hall, we display the translation result as captions in a web browser that students can view on their own devices, such as laptops, tablets and smart phones. Preliminary studies have shown that textual output is easier to digest than synthesized speech, especially if it does contain errors. Lately, we introduced various improvements in our setup to decrease the latency, e.g., by outputting preliminary captions fast and, if necessary, updating parts as both the transcription and translation hypotheses stabilize over time as more context is becoming available.

2 Related Work

The development of systems for speech translation started in the 90s. First systems were able to translate very domain specific and formalized dialogues. Later, systems supported greater variety in language, but were still built for specific domains (Stüker et al., 2007).

Despite a difference in the overall quality of the translations, MT systems suffer from not being able to anticipate context like human interpreters. MT systems are unable to do so because of the lack of background and context knowledge. This results in a higher delay of the translation. But there has been some research towards the reduction of the latency and the translation of incomplete utterances (Fügen and Kolss, 2007), (Sridhar et al., 2013), (Oda et al., 2015). The goal is to find the optimal threshold between quality and latency (Shavarani et al., 2015), (Yarmohammadi et al., 2013), (Oda et al., 2014).

With ongoing research and development, the systems have matured over the years. In order to assess whether our system helps students to better understand lectures, we have conducted a user study (Müller et al., 2016) (to appear). The outcome was that students actually benefit from our system.

3 Speech Translation Framework

The Speech Translation Framework used for the lecture translation system is a component based architecture. It is designed to be flexible and distributed. There are 3 types of components: A central server, called the "mediator", "workers" for performing different tasks and clients that request certain services. Our setup has 3 different kinds of workers: ASR systems, punctuation predictors and MT systems. But the communication protocol itself does not distinguish between these different types and does not limit the types of work be to performed.

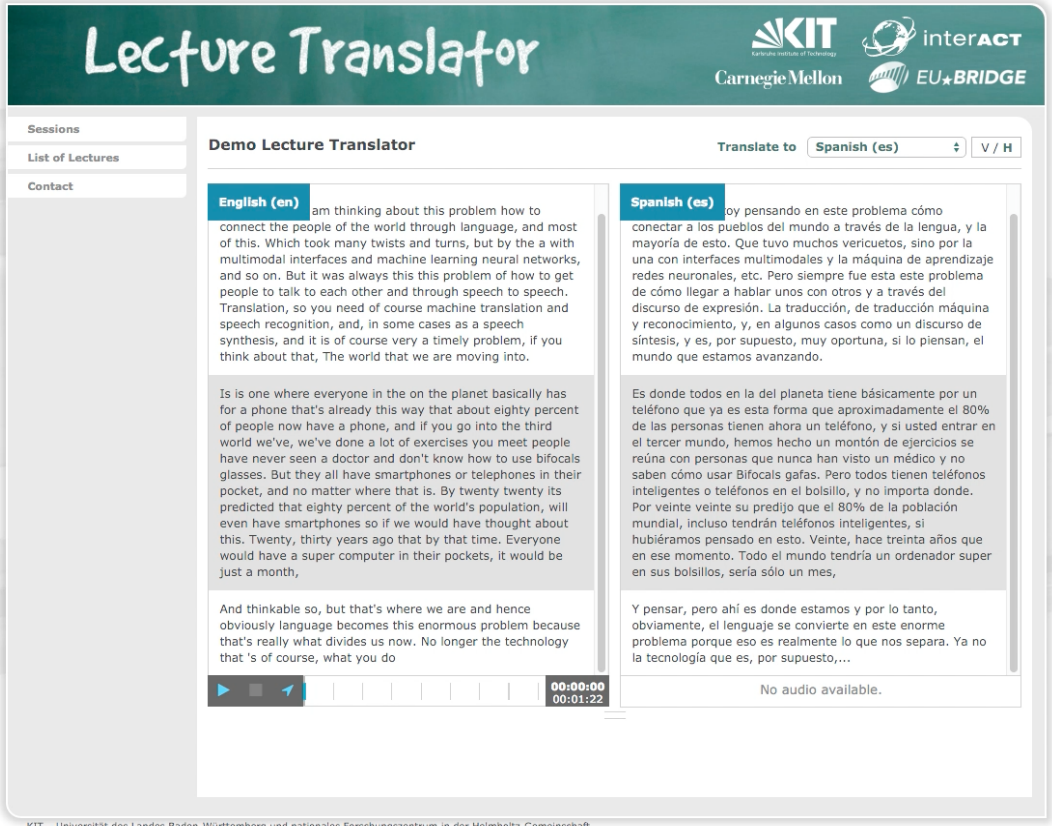
Each worker registers on the central mediator, providing a "fingerprint" and a name the mediator. The fingerprint tells the mediator which type of service the worker provides. Based on these fingerprints, the mediator selects the appropriate chain of workers to perform the requested task. E.g., if a client asks for a Spanish transcription of English audio, the mediator would first select an English ASR worker and would then route the output through a segmenter for English Text and finally run the output through the MT to translate the English text into Spanish.

4 Lecture Translator

4.1 System Description

The Lecture Translator (LT) at KIT was implemented based on the speech translation framework described above (Cho et al., 2013). We developed all workers in-house. The audio is being transcribed using the Janus Recognition Toolkit (JRTk) (Woszczyna et al., 1994), which features the IBIS single-pass decoder (Soltau et al., 2001). The acoustic model was trained using several hundred hours of recordings from lectures and talks.

Figure 1: *User interface of the Lecture Translator showing an ongoing session*



am thinking about this problem how to connect the people of the world through language, and most of this. Which took many twists and turns, but by the a with multimodal interfaces and machine learning neural networks, and so on. But it was always this this problem of how to get people to talk to each other and through speech to speech. Translation, so you need of course machine translation and speech recognition, and, in some cases as a speech synthesis, and it is of course very a timely problem, if you think about that, The world that we are moving into.

Is is one where everyone in the on the planet basically has for a phone that's already this way that about eighty percent of people now have a phone, and if you go into the third world we've, we've done a lot of exercises you meet people have never seen a doctor and don't know how to use bifocals glasses. But they all have smartphones or telephones in their pocket, and no matter where that is. By twenty twenty its predicted that eighty percent of the world's population, will even have smartphones so if we would have thought about this. Twenty, thirty years ago that by that time. Everyone would have a super computer in their pockets, it would be just a month,

And thinkable so, but that's where we are and hence obviously language becomes this enormous problem because that's really what divides us now. No longer the technology that 's of course, what you do

toy pensando en este problema cómo conectar a los pueblos del mundo a través de la lengua, y la mayoría de esto. Que tuvo muchos vericuetos, sino por la una con interfaces multimodales y la máquina de aprendizaje redes neuronales, etc. Pero siempre fue esta este problema de cómo llegar a hablar unos con otros y a través del discurso de expresión. La traducción, de traducción máquina y reconocimiento, y, en algunos casos como un discurso de síntesis, y es, por supuesto, muy oportuna, si lo piensan, el mundo que estamos avanzando.

Es donde todos en la del planeta tiene básicamente por un teléfono que ya es esta forma que aproximadamente el 80% de las personas tienen ahora un teléfono, y si usted entrar en el tercer mundo, hemos hecho un montón de ejercicios se reúna con personas que nunca han visto un médico y no saben cómo usar Bifocals gafas. Pero todos tienen teléfonos inteligentes o teléfonos en el bolsillo, y no importa donde. Por veinte veinte su predijo que el 80% de la población mundial, incluso tendrán teléfonos inteligentes, si hubiéramos pensado en esto. Veinte, hace treinta años que en ese momento. Todo el mundo tendría un ordenador super en sus bolsillos, sería sólo un mes,

Y pensar, pero ahí es donde estamos y por lo tanto, obviamente, el lenguaje se convierte en este enorme problema porque eso es realmente lo que nos separa. Ya no la tecnología que es, por supuesto,...

For translation, we used a phrase-based decoder (Vogel, 2003). It uses advanced models for domain adaptation, bilingual and cluster language models in addition to Discriminative Word Lexica for producing the translation. We use POS-based word reordering (Rottmann and Vogel, 2007; Niehues and Kolss, 2009). The translation model was trained on 1.8 million sentences of parallel data. It includes data from various sources and in-domain data.

4.2 System Operation

The LT is in regular use for multiple years now and currently translates approx. 10 different lectures per term. We have installed this system in multiple lecture halls, among them KIT's largest hall, called "Audimax".

In each hall, the system is tightly integrated in the PA to ensure smooth operation. The audio is captured via the PA from the microphone that the lecturer uses to address the audience. The operation of the system itself is time controlled: It starts at the time when the lecture begins and runs until the lecture is finished. The workers of the system run distributed over multiple servers. This ensures overall system stability as it allows for fail-overs in case of server failure. There are multiple instances of each worker running in order to translate multiple lectures in parallel.

During the every day operation the LT does not require any special preparations from the lecturer prior to each lecture because of the integration into the PA and the time controlled operation. But the quality of the output can be improved if slides or lecture notes are being made available beforehand. This way, the system is able to adapt to the specific domain of a lecture by covering any terms or named entities special to this lecture. The second advantage that we use is that the same lectures are usually given repeatedly in different terms. This way, we can use several iterations of the same lecture to improve the performance. Using the collected data, we adapt the ASR to certain speakers and ASR and MT to certain topics.

As the goal is to provide the service as cost efficient as possible, we decided to use the devices that the students already own to display the output. The Lecture Translator is therefore a web based service. Listeners wanting to see the transcription can go to the website of the service[1] to see a list of currently running sessions. Depending on the permissions from the lecturer, the output can be displayed either only to people who know the password or viewers from within KIT or globally. A screen-shot from the user interface running an active session is shown in Figure 1. The transcription is displayed on the left part of the window while the translation is shown on the right. The user has the choice of various target languages, depending on the source language.

Our system currently supports the translation from German audio into English and French text. Using English as input language, the system is able to produce French, German and Spanish output.

5 Intermediate Output

One of the main problems of earlier versions of our speech translation framework was the latency of the system. Since machine translation systems are usually trained on sentence level, the translation can only be displayed if the whole sentence is recognized. In order to overcome this drawback, we extended our framework to handle intermediate outputs. This allows us to display a translation of a partly recognized sentence and later update it with the translation of the whole sentence. The same technique is also be applied to the to display intermediate hypotheses from the speech recognition that are later updated.

In the framework, each message has properties defining the time span to which its content relates. For example, if the MT component generates a new translation, it will generate a message with the start and end time of the translation and the translation itself. In the baseline system, the start time has to be equal or greater than the end time of all previous messages.

In order to limit the complexity, we only allow to update the most recent messages. Every time a message with a new starting time is received, this implicitly will mark all messages prior to this starting time as final and no updates to the content of these messages is allowed. Allowing updates for every message would be too complex, as we also allow to change the time span of the updated messages. This would lead to difficulties for all messages except the most recent one. Furthermore, in this case the different components would need to store information about the whole session instead of only information about the non-final sections.

In order to facilitate the new possibilities of the framework, each component was extended in order to handle intermediate output and input. On the input side, the content of the new message can no longer be simply attached to the previous output, but it might also overwrite part of the stored content. Therefore, additional bookkeeping is necessary. On the output side, we can now already output preliminary results and later update them with better hypotheses.

When generating new messages we have to make sure that we do not mark content as final by using a new start time for the next message although the input for this text has not been marked final by the previous component.

6 Conclusion

In this paper we presented our automatic simultaneous translation system for university lectures. The lecture translator is installed in four lecture halls at KIT and has been running for several years now. The system features several techniques that are specifically tailored at the needs of a simultaneous system processing an unsegmented stream of continuous speech. Feedback from the students and a systematic user study have shown that the system helps students to better follow the lectures if they are not (yet) completely fluent in German. Currently we are increasing the number of lecture halls at KIT that the system is installed in and are working with other universities that are also interested in deploying the system.

References

Eunah Cho, Christian Fügen, Teresa Herrmann, Kevin Kilgour, Mohammed Mediani, Christian Mohr, Jan Niehues, Kay Rottmann, Christian Saam, Sebastian Stüker, et al. 2013. A real-world system for simul-

[1] http://lecture-translator.kit.edu

taneous translation of german lectures. In *INTERSPEECH*, pages 3473–3477.

Christian Fügen and Muntsin Kolss. 2007. The influence of utterance chunking on machine translation performance. In *Proceedings of the eighth Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, pages 2837–2840, Antwerp, Belgium.

Markus Müller, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2016. Evaluation of the KIT Lecture Translation System. In *Language Resources and Evaluation Conference (LREC)*, Portoroz, Slovenia, May.

Jan Niehues and Muntsin Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. In *Proceedings of the Workshop on Statistical Machine Translation*, WMT 2009, Athens, Greece.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Optimizing segmentation strategies for simultaneous speech translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, Maryland, USA.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Syntax-based simultaneous translation through prediction of unseen syntactic constituents. In *The 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Beijing, China, July.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skövde, Sweden.

Hassan S. Shavarani, Maryam Siahbani, Rantim M. Seraj, and Anoop Sarkar. 2015. Learning segmentations that balance latency versus quality in spoken language translation. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

Hagen Soltau, Florian Metze, Christian Fugen, and Alex Waibel. 2001. A one-pass decoder based on polymorphic linguistic context assignment. In *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pages 214–217. IEEE.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 230–238, Atlanta, Georgia, USA.

Sebastian Stüker, Christian Fügen, Florian Kraft, and Matthias Wölfel. 2007. The isl 2007 english speech transcription system for european parliament speeches. In *INTERSPEECH*, pages 2609–2612.

Stephan Vogel. 2003. SMT Decoder Dissected: Word Reordering. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China.

Monika Woszczyna, N. Aoki-Waibel, Finn Dag Buø, Noah Coccaro, Keiko Horiguchi, Thomas Kemp, Alon Lavie, Arthur McNair, Thomas Polzin, Ivica Rogina, Carolyn Rose, Tanja Schultz, Bernhard Suhm, M. Tomita, and Alex Waibel. 1994. Janus 93: Towards spontaneous speech translation. In *International Conference on Acoustics, Speech, and Signal Processing 1994*, Adelaide, Australia.

Mahsa Yarmohammadi, Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Baskaran Sankaran. 2013. Incremental segmentation and decoding strategies for simultaneous translation. In *IJCNLP*, pages 1032–1036.