KIT Lecture Translator: Multilingual Speech Translation with One-Shot Learning

Florian Dessloch, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai-Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Thomas Zenkel, Alexander Waibel
Karlsruhe Institute of Technology, Karlsruhe, Germany; Carnegie Mellon University, Pittsburgh, PA, USA
firstname.lastname@kit.edu

Abstract

In today's globalized world we have the ability to communicate with people around the world. However, in many situations the language barrier still presents a major issue. For example, many foreign students studying at KIT are initially unable to follow a lecture in German. Therefore, we offer an automatic simultaneous interpretation service for students.

To fulfill this task, we have developed a low-latency translation system adapted to the lecture domain which covers several language pairs. While the switch from traditional statistical machine translation to neural machine translation (NMT) significantly improved performance, to integrate NMT into the speech translation framework required several adjustments. We have addressed the run-time constraints and different types of input. Furthermore, we utilized one-shot learning to easily add new topic-specific terms to the system. In addition to better performance, NMT also enabled us increase our covered languages through the use of multilingual models. Combining these techniques, we are able to provide an adapted speech translation system for several European languages.

1 Introduction

In today's globalized world we have the opportunity to communicate with people all over the world. But, often the language barrier still poses a challenge and prevents communication. At KIT, there are many international students from around the world. To deal with the language barrier and support foreign students in lectures, KIT offers an automatic lecture translation (LT) service in many lecture halls. When a lecture begins, a recording client is triggered which records the lecturer's speech and presentation screen, and sends them to our simultaneous LT system which returns both the transcription and translation in real-time via a web interface.

Starting from the initial version of lecture translation (Fügen et al., 2006), our system has continuously developed (Kolss et al., 2008; Cho et al., 2013). In 2012, the LT system was first operated in several lecture halls in KIT with limited coverage; German was the primary spoken language, translated into English. We now support both German and English as input languages with three additional target languages: French, Spanish, and Italian. Furthermore, a preliminary multilingual system for 24 languages is also available. In order to provide efficient recognition and translation services to the students, we address the following research areas: 1) **Low-latency**: Transcription and translation needs to be synchronized with the speech of the lecturer as much as possible. How can we provide systems with very low latency? 2) **Multilingualism**: How can we minimize the effort and maintenance needed to train and support many languages? 3) **Adaptation**: Which adaptation techniques are applicable for online and low-latency speech translation?

2 Low-latency Speech Translation Framework

Speech is simultaneously recorded by a recording client and sent to a server. There, the three main components of the system, automatic speech recognition (ASR), segmentation, and machine translation

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 89–93
Santa Fe, New Mexico, USA, August 20-26, 2018.

(MT), transcribe and translate the audio, which is shown to the user in an interface. The segmentation system is a monolingual translation system that adds case and punctuation information to the ASR output, and segments it into appropriate sentences for the translation system. While the main use-case is online translation, where the user can follow the lecture concurrently on his smart phone or laptop, we also offer a web-based archive for viewing previously recorded lectures.

One of the main weaknesses in earlier versions of our speech translation framework was the latency of the system. Since MT systems are usually trained on the sentence-level, the translation would only be displayed if the whole sentence was recognized. In order to overcome this drawback, we extended our framework to handle intermediate outputs (Niehues et al., 2016). This allows us to display a translation for a partly recognized sentence, and later update it with the full sentence translation. The same technique is also applied to display intermediate hypotheses from the speech recognition which is described in Section 3.

3 Automatic Speech Recognition

We utilize the DNN-HMM model to build our ASR component. We trained a deep neural network with several lectures' audio to model many thousands of context-dependent phonemes. We also utilized lecturers' materials such as lecture notes and reading materials to build adapted vocabulary and language models for the scheduled lectures. While using phoneme-based acoustic modeling is stable for many different languages, the automatic adaptation of vocabulary and language model allows us to significantly improve transcription quality based on information from the lectures of the same course and lecturer. A further advantage to the DNN-HMM model for our use case is that it is a very efficient model for building low-latency ASR systems. The latency of our ASR system has to be very low to keep transcription and translation as synchronized with the speech of the lecturer as possible.

Low-latency ASR By using a dynamic decoding framework for ASR, we can avoid the detection of audio segments, and incrementally perform decoding as soon as a fraction of speech is recorded. This so-called run-on incremental recognition helps us avoid the latency caused by waiting for the end of the current segment. Normally, only at the end of an utterance is the most probable hypothesis determined. However, since waiting until the end of the utterance leads to a high latency, we detect when a part of the hypothesis becomes stable and can be kept.

Lecture Dictionary Adaptation The web interface allows lecturers to upload lecture materials such as slides and reading materials that will be accessible for download by students. We make additional use of these materials by automatically extracting out-of-vocabulary (OOV) words which are not recognizable by the default ASR system. We generate automatic pronunciations for these word, and map them to a common word to obtain language model probabilities. This is based on the intuition that these words are likely to occur in the lecture and should possess higher probability. Adaptation is performed on a per-lecture basis so that each lecture has its own specialized vocabulary.

4 Neural Machine Translation

The main advancement over previous lecture translation systems is the switch from SMT to NMT, and the necessary adaptations to do so. In order to use NMT in our framework, we had to develop several adaptations. First, we improved the run-time for the monolingual translation system by using a dedicated target encoding. Secondly, we used multi-task learning to improve the performance on translating the partial sentences necessary in low-latency translation. Finally, we developed methods to easily integrate topic-specific terms. But the switch also allowed us to significantly increase our language coverage.

Monolingual MT Automatic speech recognition (ASR) systems typically do not generate punctuation marks or reliable casing. To create segments and better match typical MT training conditions, we use a monolingual NMT system to add sentence boundaries, insert proper punctuation, and add case where appropriate before translating (Cho et al., 2017). To train, we create parallel data where the source is the lowercased sentence with all punctuation removed, and the target is features indicating case with punctuation attached. The output vocabulary is then quite small; less than 100. Rare source words are replaced with POS tags. The training data is randomly segmented so that segment boundaries and punctuation types are well-distributed throughout the corpus. At test time, we follow the sliding window technique describe in (Cho et al., 2017), and always keep the previous l_w words as context.

Adaption to Speech Since we are using the low-latency framework described in Section 2, the system does not only need to translate complete sentences, but also partial sentences. In phrase-based MT, this did not pose a problem. But if the NMT system is only trained on complete sentences, it learns to always generate complete sentences. Therefore, it will fantasize an ending for an incomplete sentence. We address this problem by additionally training the system to translate partial sentences. Accordingly, we first generate artificial training data. To improve corrections while maintaining performance, we use multi-task learning and train the model to perform both tasks, the translation of partial sentences and the translation of full sentences (Niehues et al., 2018).

One-Shot Learning In addition to overall translation quality, we identify the importance of translating rare events which do not appear many times in the training data but are critical to individual lectures They can be difficult to translate using NMT, but it is crucial for the system to translate them consistently. In order to incorporate external translations into the system, we designed a framework that allows the model to dynamically interact with external knowledge bases via both data augmentation and modeling (Pham et al., 2018). During training, we pre-train phrase-tables with the parallel corpora, and use them to annotate possible translations for the rare-words that appear less than 3 times in the training data. We consider word-splitting methods such as BPE crucial efficiently represent words that do not appear in the training data, and therefore allow proper annotation. By using the COPY-NET the model is able to learn a bias towards the annotation, which might otherwise have be assigned very small probabilities by the NMT softmax function. Finally, we use reinforcement learning to guide the search operation to encourage copying the annotation into the generated sequence.

Multilingual MT In order to build a single neural translation model able to translate into more than twenty European languages, we follow the approach described in (Pham et al., 2017). Our goal is to keep the neural architecture as compact as possible while still maintaining parity with the translation quality of systems trained on individual language pairs on the same data. Fundamentally, we our system shares its main components across languages: the encoder, the decoder and the attention layer, but employs different softmax output layers and word embedding layers for different target languages based on their vocabularies. In this way, the system does not need to calculate over all the words from all target languages.

5 Results

WERs and Latency of the ASR In Table 1 , we present the performance of our multilingual speech recognition component in term of word error rates (WER) and word latency. The word latency is measured as the difference between the time a word is spoken and the time when its transcription is available at the display component. Since words span a duration we use their end time. Each test set consists of about 20-30 lecture talks. Typically, recognized words will appear in the display client about 1 second after real-time. The archived WERs without adaption are below 20.0% for all languages.

#	Language	WER (%)	Word Latency (s)
1	English	15.2	0.84
2	German	19.4	1.03
3	Spanish	14.1	0.79
4	French	19.3	1.11
5	Italian	17.5	0.94

Table 1: WER and Latency of The ASR

Input	Das binäre *Zahlensystem* ist ...
Baseline:	The binary *payment system* is ...
One-Shot:	The binary *numeral system* is ...

Table 2: An example of one-shot learning

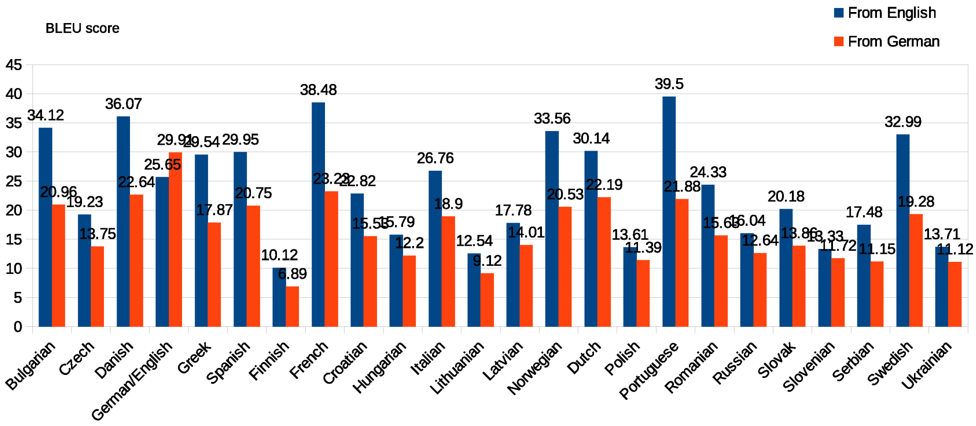

Figure 1: MT Performance when translating from English or German to 24 European languages

Machine Translation Figure 1 shows the results of the multilingual system, translating from English and German to 24 European languages using a single model trained on the multilingual data. Compared to a standard bilingual system trained on the same data, it achieves better performance: for English↔German, we see 25.65 BLEU as compared to 24.92 translating into German, and 29.91 BLEU as compared to 28.74 translating into English. The results confirm our assumption that multilingual information helps to improve low-resourced translation systems trained individually.

This system achieves its best BLEU scores translating from English to Portuguese and German to English. This is reasonable, as there are adequate amounts of data in those directions and there are related languages which can assist by providing additional context. At the other end of the spectrum, the system obtains its worst results when translating into Finnish as there is not much parallel training data, and Finnish is the most morphologically-rich language in our set, further impoverishing the data condition.

When translating in specific domains, words which are generally rare can be incredibly important to translate correctly. For example, if we consider a lecture about the binary numeral system or *Zahlensystem*, it is necessary to translate this term or the meaning of the lecture is lost. One-shot learning allows us to do so, as shown in Table 2. Without one-shot learning, we have not seen this term before. Using byte-pair encoding, the system is generate a translation for *Zahlensystem*, but it incorrectly generates the translation *payment system* for the similar German word *Zahlsystem*. By adding the phrase {*Zahlensystem # numeral system*} to our memory, we are able to correctly translate this word in context.

6 Conclusion

This paper describes recent advancements for low-latency speech-to-text translation. Using several techniques, we were able to use fully neural methods for the machine translation component of our system. Further, by using multi-task and reinforcement learning, we were able to use NMT in a low-latency framework that can be easily adapted to new topics. These neural methods have allowed us to significantly increase our covered languages. Our multilingual model is able to translate from two source languages to 24 target languages, while fitting in memory on a moderate-size GPU.

References

Eunah Cho, Christian Fügen, Teresa Herrmann, Kevin Kilgour, Mohammed Mediani, Christian Mohr, et al. 2013. A real-world system for simultaneous translation of german lectures. In *INTERSPEECH*.

Eunah Cho, Jan Niehues, and Alex Waibel. 2017. Nmt-based segmentation and punctuation insertion for real-time spoken language translation. *Proc. Interspeech 2017*, pages 2645–2649.

Christian Fügen, Muntsin Kolss, Dietmar Bernreuther, Matthias Paulik, Sebastian Stüker, Stephan Vogel, and Alex Waibel. 2006. Open domain speech recognition & translation: Lectures and speeches. In *ICASSP*.

Muntsin Kolss, Matthias Wölfel, Florian Kraft, Jan Niehues, Matthias Paulik, and Alex Waibel. 2008. Simultaneous german-english lecture translation. In *IWSLT 2008*, pages 174–181.

Jan Niehues, Thai-Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alexander Waibel. 2016. Dynamic transcription for low-latency speech translation.

J. Niehues, N-Q Pham, T-L Ha, M. Sperber, and A. Waibel. 2018. Low-latency neural speech translation. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018)*, Hyderabad, India.

Ngoc-Quan Pham, Matthias Sperber, Elizabeth Salesky, Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. KIT's Multilingual Neural Machine Translation systems for IWSLT 2017. *IWSLT 2017*.

Ngoc-Quan Pham, Jan Niehues, and Alex Waibel. 2018. Towards one-shot learning for rare-word translation with external experts. In *Proceedings of the Second Workshop on Neural Machine Translation*. Association for Computational Linguistics.