SIMULTANEOUS MACHINE TRANSLATION OF GERMAN LECTURES INTO ENGLISH: INVESTIGATING RESEARCH CHALLENGES FOR THE FUTURE

Matthias Wölfel[1]*, Muntsin Kolss*[1]*, Florian Kraft*[1]*, Jan Niehues*[1]*, Matthias Paulik*[1,2]*, and Alex Waibel*[1,2]

[1]Fakultät für Informatik, Universität Karlsruhe (TH), Germany
[2]Language Technologies Institute, Carnegie Mellon University, USA
{wolfel, kolss, fkraft, jniehues, paulik, waibel}@ira.uka.de

ABSTRACT

An increasingly globalized world fosters the exchange of students, researchers or employees. As a result, situations in which people of different native tongues are listening to the same lecture become more and more frequent. In many such situations, human interpreters are prohibitively expensive or simply not available. For this reason, and because first prototypes have already demonstrated the feasibility of such systems, automatic translation of lectures receives increasing attention. A large vocabulary and strong variations in speaking style make lecture translation a challenging, however *not* hopeless, task.

The scope of this paper is to investigate a variety of challenges and to highlight possible solutions in building a system for simultaneous translation of lectures from German to English. While some of the investigated challenges are more general, e.g. environment robustness, other challenges are more specific for this particular task, e.g. pronunciation of foreign words or sentence segmentation. We also report our progress in building an end-to-end system and analyze its performance in terms of objective and subjective measures.

Index Terms— automatic speech recognition, machine translation, speech-to-speech translation

1. INTRODUCTION

Lectures, seminars and oral presentations form a core part of knowledge dissemination and communication. For native speakers of other languages in the audience the opportunities for collaboration and exchange of information are severely reduced. Human translation services could be an option in a couple of cases to overcome the language barrier, but prohibitive costs limit their assignment to some exceptional circumstances. In all the other scenarios automatic translation systems could offer considerable practical benefits. For decades the underlying technologies have been developed independently of each other. A speech-to-speech translation system, however, is not only the cascade of its parts, but it requires additional system components and allows for a closed coupling by the exchange of word lattices or confusion networks. Since the goal is to produce output in a target language, the correctness of the intermediate components output can be of secondary concern. For example German compounds can be represented as separated words in the output of the automatic speech recognizer.

Building an automatic system for simultaneous translation of lectures poses many challenges, some of which are language-pair independent and some are language-pair dependent. In [1], a translation system for speeches and lectures was presented for the first time; this system translated English lectures into Spanish. Translating German lectures into English, as presented in this work, introduces additional difficulties not encountered in [1]. While we have presented a first system including a lot of technical details of the system and system structure in [2], this publication analyses and discusses challenges encountered in a German to English lecture translation system. While some challenges remain unsolved we propose possible solutions for others.

2. CHALLENGES

This section covers open and partly solved challenges faced by building a German to English lecture translation system.

2.1. Acoustic Environments

While body mounted microphones, if mounted correctly, clearly lead to the lowest word error rate, they are inconvenient and may generate too much distraction to the lecturer. Thus *automatic speech recognition* (ASR), which works reasonably well on recordings captured with mid- or far-field microphones, is essential to provide convenient and robust transcriptions of lectures. In addition ASR working on mid- or far-field microphones can also capture questions or comments from the audience. This information would be lost if only a single body mounted microphones would have been used.

From Figure 1 it is clear that a significant performance loss arises by moving the microphone away from the speaker's mouth. To reduce the quality gap of automatic transcriptions between close and distant speech capture additional technologies have to be developed which are able to estimate, track and compensate for the two distortions frequently encountered in mid- and far-field sound capture, namely *non-stationary noise* and *reverberation*. As shown in Figure 1, compensating for each of the two distortions in the acoustic feature space is already able to reduce the performance gap and a joint compensation reduces the gap further. A detailed description of the compensation framework used is presented in [3].

Since many lecturers are not professional speakers, the transcription of the speech signal is challenging on other aspects: the speaking styles can vary considerably and are much more conversational and informal than in prepared speeches or short utterances.

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation and by the National Science Foundation (NFS) under the project STR-DUST (Grant number IIS-0325905). The authors would like to thank Christian Fügen for providing the system implementation and Susanne Burger, Edda Heeren, Oliver Langewitz, Stephan Settekorn and Marleen Stollen for transcribing and translating the lecture development and evaluation data.

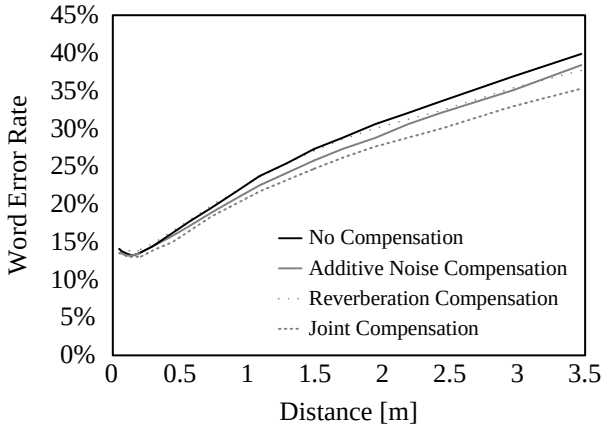

Fig. 1. Word error rates without compensation and with different compensation approaches for different speaker to microphone distances of a German speaker.

2.2. Compounds

German language contains a large amount of compound[1] words and morphological forms which cause additional difficulties for automatic transcription as well as translation.

In the standard case a recognizer's vocabulary is static and tuned to the domain of interest a-priori. In German it is quite common to form a new word by prefixing a noun with further nouns describing the main noun in a specific context. This dynamic language process contradicts a static compiled vocabulary and produces difficulties like

- a bad vocabulary coverage (if truncated) or
- an underestimation of language model weights (for infrequently or never seen words) and
- a significant increase in search space complexity.

To reduce the number of words in the vocabulary of the recognition and translation system and to reduce, at the same time, the *out of vocabulary* (OOV) words the compound words have to be split into their parts. A rigorous splitting of compounds, however, reduces the effective context represented in the language model. To keep those contexts where it would be most useful, namely in-domain data, we split only those compounds which are not represented in those data. Splitting only those compounds provides a suited generalization capability for unseen compounds which have not been observed during training. By splitting the compound words, as described, the OOV rate dropped from 2.3% to 1.5% for our test set.

Handling of compounds is also important in *machine translation* (MT), as a word alignment based on word-to-word correspondences is usually the first step in training a system. Without compound splitting, the alignment might choose just one of the English constituents as counterpart of a German compound, resulting in dropping content during translation.

We found that the optimal compound split for ASR is *not* the best for MT. Splitting more aggressively, for example using the frequency-based method described in [4], leads to better translation quality even though this may generate much shorter and partially erroneous word fragments on the German input side. Machine translation systems do not impose language modeling on the input word sequence and thus do not suffer from a compound split dependent context order decrease on the input side, and "incorrect" splits on the German side are easily bridged by using phrase translations if necessary.

In an integrated speech translation system, the recognizer output must be matched to the translation system. We apply the second, more aggressive compound splitting on top of the splitting already performed in speech recognition, both to the training data and to the speech recognizer output during translation.

[1]Compounds are comprised of at least two nouns and thus an innumerable number of possible words can be formed.

2.3. Modeling Pronunciation of Foreign Words

An analysis of technical lectures given at Universität Karlruhe (TH) yielded that German scholars embed a high number (2% of the words) of technical English terms into their presentations. For example typical English words in a German speech recognition lecture are "hidden Markov model" or "minimum variance distortionless response". Because German and English share the same heritage, German contains a high number of words which are identical to English. In the investigated technical lectures 21% of the words are represented in *both*, in the German as well as in the English languages. Fillers and words which could neither be assigned to German nor English appeared 13% of the time.

Depending on the word the pronunciation (and even the meaning) might vary or the original pronunciation is used (the meaning however might vary). Typical examples of the former are "bald", "kind" or "man". Typical examples of the latter are "handy" (in German used for cell-phone) or "brief" (in German used for letter). Additional difficulties arise by the fact that some words share the same pronunciation, however have different transcription (and meaning) in English and German. An example is "eagle" (in German it would be written "Igel" with the meaning hedgehog). In those cases one has to rely on the context represented in the language model.

Language	German	English	Both	Average
Compensation	Word Error Rate			
None	10.7%	60.0%	15.4%	13.8%
Mapping	11.1%	34.6%	13,8%	12.7%
Parallel	11.4%	26.4%	14.7%	13.4%

Table 1. Word error rate per language for the baseline system (none) and two different compensation techniques. Results from [5].

An analysis of the recognition errors per class, see first line in Table 1, shows that the recognition performance of the English words is significantly lower than the German words and thus causes a significant increase in average WER. This can be explained by the fact that only a couple of English phonemes are represented in the German phoneme set and that rule based pronunciation dictionaries for German do usually not handle the pronunciation of foreign words and thus generate a wrong (following the German rules) mapping of the phoneme sequence.

In order to reduce the performance gap between the German words and the English words in a German speech recognition system one could employ both the German as well as the English phoneme set (parallel) or map the English pronunciation dictionary to German phonemes (mapping). Both methods have been successfully applied in [5]. Table 1 compares the different approaches. Both methods are able to significantly reduce the WER for English words, however introduce additional errors to the German words. In the best case it is possible to reduce the average WER by more than one percent.

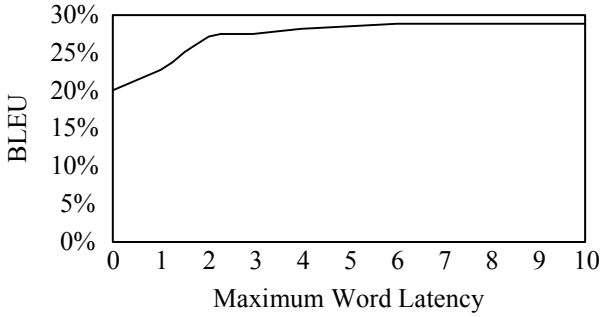

Fig. 2. BLEU score vs maximum word latency.

2.4. Sentence Boundaries and Latency

Another challenge in lecture translation is the absence of sentence boundaries in the spoken input. Because there is a wide variety in speaking style and ability among lecturers, this difficulty is often compounded by ungrammatical language containing disfluencies such as repetitions and false starts.

Machine translation systems, however, perform best when operating on well-formed sentence or utterance units. In integrated speech translation systems, a common approach is loose coupling between speech recognition and machine translation, i.e., using an additional intermediate component to segment the unstructured first-best ASR hypothesis into shorter sentence-like units prior to passing them on to the following machine translation component. For best translation performance, the chosen segment boundaries will correspond to semantic or syntactic boundaries [6, 7, 8]. Segmentation can be done based on a hidden-event language model, or pauses in the speech signal, since pauses often correspond to punctuation marks or other semantic boundaries. Generally, longer segments on average mean fewer segmentation errors and higher translation quality. Segment lengths of 10-40 words are usually required to come near the potential performance of the underlying translation system.

A lecture translation system must run in real-time since the speaker will not wait for the system to catch up. While the actual processing speed is not the main limitation on current standard hardware, longer utterance segments directly increase the *latency* of the system, i.e., the delay between words being uttered by the speaker and the output of the translation of these words. A low latency, however, is desirable because a long delay between the original utterance, during which the speaker may navigate with a light-pointer over a slide, and the output of the translation will make it more difficult for the listener to follow the lecture, and could severely impact the understanding of the presented material. The translation quality must therefore be compromised by using shorter segment lengths to avoid unacceptably long latencies.

An alternative to the standard segmentation approach is directly processing the continuous input word stream from the speech recognizer in real-time, as proposed in [2]. Using a continuously maintained rotating translation lattice, such a stream decoder design, makes decisions when to generate translation output independently from a fixed input segmentation with hard boundaries. At the same time, it allows for full word reordering within a sliding local window that also defines the maximum output delay. When most or all required word reorderings lie within this window (ranging between two to four words), this approach reaches the translation performance of the standard input segmentation model at greatly reduced latency values and is able to reach its potential at very low latencies, as can be seen in Figure 2.

The long-distance word reordering requirements of the German-English language pair interfere with the desire for short latency. Producing timely translation will lead to severe translation errors when the necessary context has been destroyed by a segment boundary or lies outside the reordering horizon. Unfortunately, even tolerating long delays for the sake of better translations will not yield much better results with today's systems because good models for long-distance word reordering itself, even for translation of perfectly grammatical text in batch mode, have been notoriously hard to come by.

2.5. Word Reordering

Although German and English are related languages the word order is very different, especially for the position of the verb: in a German subordinate clause the verb is at the last position, whereas it is at the second position in the English sentence. Furthermore, German verbs might consist of two parts where the auxiliary verb is at the second position of the sentence while the main part is at the end. Consequently, reordering over a rather limited distance does not lead to a reasonable translation quality, as the readability of the translation is strongly affected by the awkward word order.

To tackle this problem we introduced rule-based reordering [9] in which a word lattice, with different word orders, is created as a pre-processing step to translation. The rules to create the lattice are automatically learned from data tagged with *part-of-speech* (POS). In the training, POS based reordering patterns were extracted from word alignment information. The context in which a reordering pattern is seen was used as an additional feature. At decoding time, a lattice structure is built with the previously extracted rules for each source utterance as input for the decoder.

Figure 3 shows two examples, one in which rule-based reordering helps and one in which word reordering is still a problem. Rule-based reordering is able to improve the BLEU score in a range between 3% and 5% relative.

3. END-TO-END SYSTEM EVALUATION

An end-to-end lecture translation system can be evaluated with objective error measures such as the WER for speech recognition and the BLEU score for MT, which are the primary metrics we used for developing our prototype German-English system. A detailed technical description of this system can be found in [10].

Lecture translation from German to English is a difficult task, and given the significant challenges described in this paper, it is not surprising that the absolute translation quality lags behind the quality of systems for more limited tasks and better controlled conditions.

A sufficient ASR performance is the obstacle since translation quality scales almost linearly with the WER of the input hypotheses. From our own experience we feel that a WER of 15% is quite readable as stand alone, while useful translation of lectures become possible for WER below 10%. This difference is due to the error compensation capacity of the human. The misrecognition of words which are acoustically similar, but semantically different, can be easily compensated by the human. This information of acoustic similarity is lost after translation and thus a compensation becomes impossible. Take for example the two example pairs "zu geben" (to give) and "zugeben" (to admit) or "Reden" (speeches) and "Reben" (vine). The former is an error in word boundaries while the latter is a phoneme error.

Our current system, which has an overall WER of 12.5% under real-time conditions on standard hardware, closely fails this thresh-

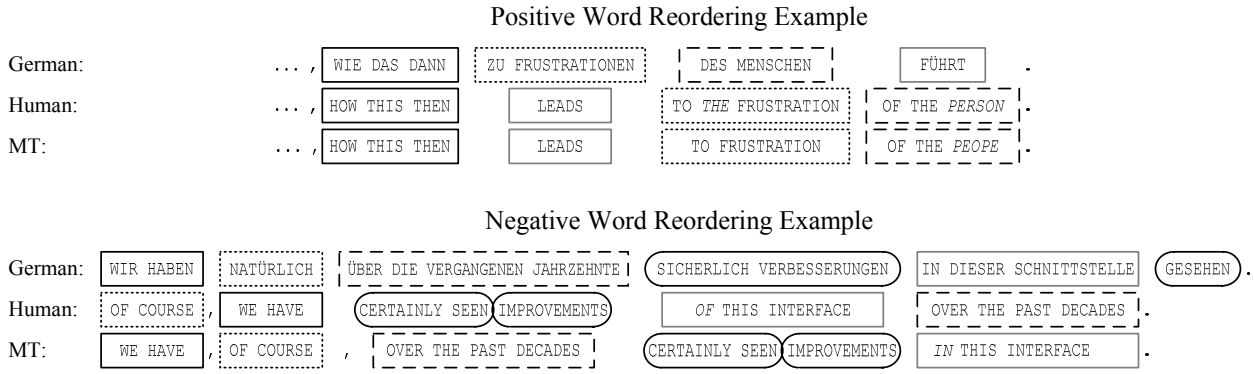

Fig. 3. A positive and a negative example of automatic word reordering. The different boxes show which parts are translations of each other and do not need to correspond to the phrase pairs used in the translation.

old, but an analysis of the recognition errors reveals that German words are correctly recognized 89.5% of the time and thus nearly hits this boundary. While on the reference segmentation and transcription our system achieves a BLEU score of 33.8 when scored against one reference transcription, the errors introduced by automatic sentence generation and speech recognition lower this to 24.2.

Human evaluation of our system reveals that the generated English simultaneous translation provides a rough idea of what the German lecturer said. The probably most striking problem for humans is the rather awkward word order which significantly reduces the readability of the English translation and increases the cognitive workload even in those cases where all words have been translated correctly. Furthermore, the different word order in German and English leads to another problem. Since the word alignment of the German and English verb is quite difficult, in some sentences the German verb is not aligned. This causes the problem that there are phrase pairs, which remove the German verb in the translation. In those cases, even if everything else is translated correctly, the sentence is hard to understand, since the verb determines the meaning.

A third important objective measure is the latency of the end-to-end system. For human interpreters, a latency of half a sentence seems common. An analysis of our test set yielded an average segment length of approximately 19 words on manual segmentation and 7 words on automatic segmentation. Taking into account that the different components in the automatic system introduce additional latency it becomes obvious that the theoretical lower bound of 4 words in our system can not be reached. Instead we observe a latency between one and two sentences which is due to the different system components. Such a long latency can be quite irritating, in particular if the lecturer has already switched to a new slide.

4. CONCLUSION AND OUTLOOK

We have identified and described five particular research challenges to provide simultaneous MT of German lectures into English. We believe that those research challenges need to be addressed in order to significantly improve the comfort for the lecturer (no body-mounted microphone required), the quality of the translation (e.g. word reordering) as well as the user satisfaction (e.g. low latency).

It has been found by Hamon et al. [11], by comparing human and computer speech-to-speech translations, from English into Spanish, that the transmission of content already reaches an understandable level which is not far away from that of human translations. This rises our hopes that reasonably good simultaneous MT of German lectures into English lies within reach and we look with hope on the solutions to come.

5. REFERENCES

[1] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stüker, S. Vogel, and A. Waibel, "Open domain speech recognition & translation: Lectures and speeches," *Proc. of ICASSP*, 2006.

[2] M. Kolss, S. Vogel, and A. Waibel, "Stream decoding for simultaneous spoken language translation.," *Proc. of Interspeech*, 2008.

[3] M. Wölfel, "A joint particle filter and multi-step linear prediction framework to provide enhanced speech features prior to automatic recognition," *Proc. of Hands-free Speech Communication and Microphone Arrays*, 2008.

[4] P. Koehn and K. Knight, "Empirical methods for compound splitting," *Proc. of European Chapter of the Association for Computational Linguistics*, 2003.

[5] S. Ochs, M. Wölfel, and S. Stüker, "Verbesserung der automatischen Transkription von englischen Wörtern in deutschen Vorlesungen," *Proc. of Elektronische Sprachsignalverarbeitung*, 2008.

[6] E. Matusov, D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, M. Ostendorf, and H. Ney, "Improving speech translation by automatic boundary prediction.," *Proc. of Interspeech*, 2007.

[7] C. Fügen and M. Kolss, "The influence of utterance chunking on machine translation performance.," *Proc. of Interspeech*, 2007.

[8] M. Paulik, S. Rao, I. Lane, S. Vogel, and T. Schultz, "Sentence segmentation and punctuation recovery for spoken language translation," *Proc. of ICASSP*, 2008.

[9] K. Rottman and S. Vogel, "Word reordering in statistical machine translation with a pos-based distortion model," *Proc. of Theoretical and Methodological Issues in Machine Translation*, 2007.

[10] M. Kolss, M. Wölfel, F. Kraft1, J. Niehues1, M. Paulik, and A. Waibel, "Simultaneous german-english lecture translation," *Proc. of International Workshop on Spoken Language Translation*, 2008.

[11] O. Hamon, D. Mostefa, and K. Choukri, "End-to-end evaluation of a speech-to-speech translation system in TC-STAR," *Proc. of Machine Translation Summit XI*, 2007.