Evaluation of the KIT Lecture Translation System

Markus Müller*, Sarah Fünfer†, Sebastian Stüker*, Alex Waibel*
*Karlsruhe Institute of Technology, Adenauerring 2, 76131 Karlsruhe, Germany
†Freelancer, Essenweinstr. 33, 76131 Karlsruhe, Germany
m.mueller@kit.edu, europeersarah@gmx.de, sebastian.stueker@kit.edu, waibel@kit.edu

Abstract

To attract many foreign students is among the goals of the Karlsruhe Institute of Technology (KIT). One obstacle to achieving this goal is the fact that lectures at KIT are usually held in German which many foreign students are not sufficiently proficient in, as, e.g., opposed to English. While the students from abroad are learning German during their stay at KIT, it is very challenging to become proficient enough in it in order to follow such a complex communication situation as a lecture. As a solution to this problem we offer our automatic simultaneous lecture translation at KIT's lecture halls which automatically translates the German lectures into English in real time for the students. While not as good as human interpreters, the system is available at a price that KIT can afford in order to offer it in potentially all lectures. In order to assess whether the quality of the system is high enough in order to be of use to our foreign students at KIT we have conducted a user study on the benefit of the system to its users over the course of two terms. In this paper we present this study, the way it was conducted and its results. As it turns out the results indicate that the quality of the system has passed a threshold as to be able to support students in their studies. The detailed feedback participants to the study have helped to identify the most crucial weaknesses of the systems and has guided which development steps to take next.

Keywords: automatic speech recognition, machine translation, user study

1. Introduction

Many students studying at KIT are from abroad. While they usually speak English fluently, they are often not sufficiently proficient in German in order to follow the content of German lectures. Since most lectures at KIT are given in German, this means that students from abroad struggle with the language barrier during their stay at KIT, most prominently in their academic studies. Universities in English speaking countries do not suffer from this problem, hence they have a higher percentage of foreign students. One of the goals of KIT's internationalization strategy is to become more attractive for students from abroad, among others by lowering the language barrier while at the same time still teaching in German due to considerations of cultural and scientific diversity.

In this the use of human interpreters for translating lectures at KIT is not a feasible option, as unlike the European Parliament, KIT does not have the financial resources for employing sufficient amounts of human interpreters. Instead, we have started to offer an automatic simultaneous translation system at several of KIT's lecture halls (Cho et al., 2013). The system translates the German lectures into English in real time and displays the results of the automatic speech recognition and machine translation as subtitles on a web page. The web page can then be viewed on the personal devices of the students, e.g., smart phones or laptop computers.

Translating university lectures is a challenging task. There exist different constraints that have to be met in order to build a system that is of actual use to the students. The two key aspects are the quality of the output of the system itself, as well as the time it takes to create it. Translations have to be output in a timely fashion without large delays. Our system first transcribes the audio by the use of a speech recognition system and then automatically translates the transcribed text. In order for the students to access the transcription and translation using their own electronic device we have implemented this system as a web-service. This way, the students can use their laptops, tablets or smart phones to access our service. In each supported lecture hall, the system is integrated into the PA. Due to this seamless integration, there is no need for the lecturer to carry an additional microphone. The system is automatically started and stopped at predefined times.

We first introduced this system during summer term 2012. Since its inauguration, we offer this service to the students on a regular basis. It is installed in our main lecture hall (Audimax) and in multiple other lecture halls. With the system being available in more lecture halls, we are able to offer this service to a wider audience.

In this paper we now present the results of a user study on how the students actually benefit from the system, and which aspects of the systems should be improved with the highest priority. We carried out the study over the course of two terms.

The rest of the paper is organized as follows: In the next chapter we give a brief overview of work done in this field. In Section 3 we describe the evaluation procedure. Following that, we present the results in Section 4. We conclude in Section 5 with an outlook where we show next possible steps to improve our system.

2. Related Work

Multiple evaluations of translation systems have been published in the past. Early work in the field includes an end-to-end evaluation of a speech-to-speech translation system (Gates et al., 1996). First systems were built for for a certain domain. More recent systems are no longer limited in their domain (Fügen et al., 2006). There exist systems for transcribing (Stüker et al., 2007) or translating (Fügen et al., 2007) lectures and speeches. To the best of our knowledge, the most recent end-to-end evaluation in simultaneous translation is (Hamon et al., 2009). In this work, we evaluate the system described in (Cho et al., 2013).

3. Evaluation of the System

Measuring the perceived quality of the lecture translation (LT) system is a difficult task. It is influenced by many factors. In order to evaluate the LT system in the field, we asked ourselves the following questions: Do students using the LT comprehend the lecture better? Does the LT help students? We also wanted to know whether they like the user interface and can handle it with ease. For our analysis, we decided on four evaluation methods to help us answering these questions and get a global and comprehensive view of the actual system performance: (a) Automatic measurements in form of web access and duration of stay on the website, (b) individual interviews, (c) short surveys and (d) a comprehensive questionnaire.

The first part of our study was conducted during the summer term 2014. Automatic metrics, e.g., the number of people using the LT and the average duration of use of the LT per session per person have been collected throughout the term. Small surveys were sent around every 3–4 weeks to users that had shown interest in the project. Towards the end of the term, we conducted personal interviews and a survey among students who had used the system.

We conducted the second part of our study towards the end of the winter term 2014. Automatically collected data was once more gathered throughout the term. Two short surveys were sent around. Unfortunately, it was not possible to get interview partners in the winter term as (foreign) students were not willing to participate.

3.1. Frequency of Use

In order to learn more about the amount of people using the LT per session, we collected usage data in anonymized from. We measured the average time one person stayed logged in and the number of students using the LT during one session. However, for an exact documentation of the time students stayed logged in, they needed to actively close the web page. As this was not always the case, the exact duration of use could not always be determined.

3.2. Short Surveys

We conducted two types of short surveys: Exit polls and short questionnaires that were sent to the students via mail. For technical reasons, no exit poll could be integrated into the system during the first part of the test. At the beginning of each term, we presented the lecture translator in all lectures where this service was offered. We also offered a mailing list to which students could subscribe. This list was used to distribute online surveys. We sent around anonymous short surveys with four questions on a regular basis (every 3–4 weeks) during the lecture period. In July 2014 and in January 2015 we used a large questionnaire and therefore did not distribute any short surveys.

3.3. Large Questionnaire

In order to gain more detailed feedback, we designed a comprehensive questionnaire. It contained questions regarding the background of the users, a system evaluation, an evaluation of the components "automatic speech recognition" and "machine translation". In addition to that, we included a section where the students could express ideas and identify problems with the system. The questionnaire covered three A4 pages. It was distributed to all students that claimed to have used the LT at least once in the summer and/or in the winter term.

In order to increase engagement in the winter term and to reach as many students as possible, we made this survey available online. We asked the lecturers to publish the link with their lecture notes. Thus, even foreign students who work from home were able to participate. We also distributed the link via our mailing list and published it on the LT-homepage.

For rating questions, we provided a 5-point Likert scale ranging from 1 (worst option) to 5 (best option). The average is denoted by 3. An additional field n/a was provided in case the question could not be answered or did not apply. We evaluated the answered questionnaires in three different groups: The entire group of students, students with German as mother tongue and students with a different mother tongue.

3.4. Interview

By conducting personal interviews based on a standardized set of questions, we wanted to get a more detailed view about the impression the users had from the LT. It was also a possibility to get for more detailed answers.

The interviews were divided into several parts, resembling those of the questionnaire. The first part asked about general information about the interviewee, including his/her mother tongue and language knowledge. The second part asked general questions about the system, including its usefulness, the user behaviour and whether it was easier to follow the lecture. We also wanted to know in which situations the LT helped the most. The third respectively forth part was about the ASR-component and the MT-component. The questions were targeted towards the evaluation of the usefulness of the system and how disruptive the errors were. In the last part, we also wanted to learn more about useful features, ideas, suggestions or specific problems.

4. Results

The results section is divided into multiple parts, reflecting the different modalities used for the evaluation. We begin by presenting the results from our analysis of the frequency of use. Then, we outline the results from our two different types of surveys. This section concludes with an analysis of the interviews we conducted.

4.1. Frequency of Use

We introduced our system at the beginning of the term in the various lectures. Usually, most of the students are attending the lectures during this period. It was also this period when most of the activity was registered. This indicates that students were curious about the project. In the summer term, an active usage of the LT in all seven lectures could be observed. However, in three lectures this was only the case at the day of the presentation of the system and for a short amount of time (about 10 minutes). In two lectures, there were 4 activity spikes with an average of 2.4 users per session and an average stay of about 2 minutes. Relatively high activity was observed in one lecture with an average

of ten users per session and an average stay per user per session varying from 2 to 50 minutes. This was probably due to the fact that we were actively promoting the system throughout the term as the lecture was held by our lab.

In the winter term the highest level of activity was recorded at the day of the presentation of the system: Between 3 and 40 students per lecture accessed the LT. In 6 lectures in the fields of computer science, mathematics and economic sciences, 10 or even more people showed initial interest. During the term, 29 more events (log-in with a duration of stay over 1 minute) were registered. In 5 lectures an average of 2.9 persons were active for 2 or 3 more sessions. In one lecture, we recorded activities of 1 to 4 persons on a regular basis throughout the term (12 more times after the initial presentation). This is attributed to the fact that we actively promoted the LT in this lecture.

The automatic logging of events provided some insight into the usage of the LT. Having students using it on a regular basis is an indication that the LT was appreciated and helpful.

4.2. Short Surveys

As further measure, we sent short surveys to students that showed interest in the system. In the summer term, 24 students from four different lectures showed interest, 15 of them were foreign students. In the winter term, 66 students showed interest, with 45 of them being foreign students. The return rate of the questionnaire in both periods was low. In the summer term, we got 4 answers for the first survey in May and 3 for the second one in June. During the winter term, we received 4 answers for the first survey in November and only one for the second one in December.

In the summer term, 4 out of 7 students used the LT 3 times or more. Considering the fact that most students probably only had one lecture per week in the Audimax with the LT running, that number is quite high. One of the students used it two times, two only one time. Out of 7 students, 5 considered the LT useful, one did not answer that question. Especially the transcript was considered helpful in general. Negative aspects mentioned included the latency of the LT, difficulties when logging in, the fact that it is hard to follow the slides and the LT at the same time as well as that the LT was not available in all lectures.

In the winter term, 4 students participated in the first survey, but only one in the second. Although this return rate is low, the answers were helpful. 3 of the 4 students actually used the LT. The one who had not used it explained that he understood German and therefore did not need the LT. Interestingly enough, he said he would have loved to use the system one year ago. This entails that it is especially useful for foreign students that begin to study in Germany.

Opinions on the quality of the LT were mixed. While one student considered the speech recognition worked rather poorly, two described it as "good" or "very good". One person also mentioned the machine translation which she/he considered "very good". Two said they were going to continue to use the LT. One wanted to try it from time to time and one person said she/he would not use it again.

4.3. Large Questionnaire

The analysis of the large questionnaire is divided into several subsections In the first subsection, we provide an overview of the general questions we asked. In the next two sections, we analyse the evaluation of both ASR and MT components. This part concludes with the last part of the questionnaire about ideas or encountered problems.

4.3.1. General

As shown in Table 1, 22 students from 5 lectures answered the questionnaire, 2 of them having Chinese as mother tongue, 2 of them Spanish, 1 Russian. The level of German of the foreign students was quite high, varying from B2 to C2. Their English level was a bit lower, between B2 and C1. All participants had a high level of English, ranking from B1 to C2.

Looking at the usage, 10 students only used the LT once, 10 students 2 to five 5 and 2 students 6 to 10 times. The majority (14) of the participants in the study stated they would use the system again. The students were mainly male and studied business engineering or computer science. 5 of them have been studying in Germany for less than one year and 5 are studying more than 4 years.

General Information	**All**	**NG**	**G**
Male	20	4	16
Female	2	1	1
Years of studying in Germany			
less than 1	5	2	3
1-2	7	2	5
3-4	3	1	2
more than 4	5	0	5
n/a	2	0	2
How often did you use the LT?			
1	10	2	8
2-5	10	3	7
6-10	2	0	2
Would you use the LT again?			
yes	14	3	11
no	8	2	6

Table 1: General information of participating students, **NG** non-German, **G** German

4.3.2. System Evaluation

The results from the system evaluation are shown in Table 2. The general impression of the LT was rather positive, with 3.21 points on a scale from 1 to 5 for both groups. Germans rated the system slightly better than foreign students. It was also considered useful, with 3.23 of 5 points.

When asked in more detail about the perceived usefulness, particularly foreign students explained that it improved the understanding of the lectures and they would benefit from having it in other lectures, too. However, they were not so sure about the effect of their performance in studying and whether the LT made it easier to follow the lectures. The latter phenomenon was explained by some students.

They sometimes considered it difficult to switch between the lecturer, the slides and the output of the LT. The ease of use was also rated positively and got 3.27 points. Especially the layout of the user interface was considered clear and got the highest marks (4.27).

4.3.3. ASR Component

The evaluation results of the ASR component are shown in Table 3. The general impression of the ASR component was positive, especially among foreign students. They also considered the transcription useful. The overall quality was evaluated with 3.01 points. The largest difference in the results between German and non-German students was observed in the category of transcription errors. Those were considered less distracting by foreign students than by German students. The lowest mark in this section got the delay of the transcription, with 2.62 points. The highest mark scored the transcription of general terms. Foreign students considered the recognition quality of technical terms better. The overall usefulness of ASR received mark of 3.13. Foreign students answered that it helped to improve their performance in the subject and their comprehension of the lecture. As for the transcript, not all participants stated that it did made it easier to follow the lecture.

4.3.4. Machine Translation Component

In part 4, students were asked to evaluate the machine translation (MT) component. The results are shown in Table 4. The all students combined described the general impression of the system with 3.19 points better than the one of the ASR-component. Foreign students alone, however, got a better impression from the ASR component (3.4 ASR versus 2.67 MT). Although they appreciated the translation quality, the usefulness was rated below 3 by the foreign students. The score of the overall quality of MT was slightly higher than the one of the ASR. Foreign students, however, rated the ASR component higher.

They also rated the delay of the translation with the lowest score, as it was considered rather high. The foreign students also perceived the translation quality to be fluctuating. During the term, we added several improvements to the system. The students noticed the improvement of the quality throughout the term.

The scores of the translation quality of general terms received the highest rating in this section by both groups. But the amount of questions answered with "n/a" was quite high in this part: More than 30% of the answers did not include a rating. 6 people, including one foreign student, did not have an opinion about the MT quality. Compared to ASR and to the quality aspects, the usefulness of the MT component was rated a bit lower, attaining 2.87 points. Nearly 40% of the answers to the questions were rated n/a: Eight people did not express their opinion on the usefulness of MT at all. The differences between German speaking and foreign students were rather distinct, with more than 1 point of difference. It might be that, as the lectures were in German, the ASR component added value for the German speaking students, whereas the English output was an additional language for them to process.

4.3.5. Ideas / Problems

The last part of the questionnaire offered the students the possibility to propose suggestions and name things to improve. Ten students responded by putting in one or more comments. The most common suggestions were to create an archive of lectures to be downloaded as well as offering a service for the translation of the presentation slides. In addition to that, offering the translation from German to Chinese was proposed.

The students also suggested to improve issues they encountered. They proposed to reduce of the time lag of the LT, as well as to improve the accuracy of the translation. Especially the translations of technical terms appeared to be sub-optimal in the current setup. There were also comments that students considered the project interesting and that they were interested in its future development and improvements.

4.4. Interview

To offer a more interactive way of providing feedback, we also offered to do interviews. Two foreign students, a female business-studies student from China and a male computer science student from Ecuador accepted being interviewed after two lectures during which they had used the system. The Chinese girl considered the LT little useful as of her very good knowledge and level of understanding of German. However, she got the impression that the system might be useful for foreign students. She considered the speech recognition to be working very well, the resulting transcript being of great use for students.

Nevertheless, she perceived it hard to follow the slides / the lecturer and the LT simultaneously. Especially during periods when the lecturer was explaining charts or a picture she considered it difficult to switch back and forth. The other interviewee also considered this to be a potential problem. In this context, the time lag was considered too large, being even bigger for the translation. This resulted in the translation being less useful. Both mentioned mistakes in transcripts and translations, but it seemed less important to them. One of them also mentioned technical problems during one lecture, but those were due to the unstable Internet connection in the lecture hall at that time.

When asked for improvements, they expressed a multitude of ideas. One idea was to add subtitles to videos that are recorded and published by some lecturers. Those videos are published on the Internet. By adding subtitles and translation to those videos, it would be possible to reach out to even more potential users. The two interviewees were not sure whether they preferred an offline archive to the live function. Although an archive might be useful for later studying and the preparation for the final exam, the live function allows for more situational context information. Moreover, the Chinese girl told us that a lot of her friends did not actually attend lectures but study from home.

5. Conclusion

The different methods allowed us to identify the strengths and weaknesses of our system. We learned about some difficulties and got suggestions and ideas for the future. The

		All	NG	G
General Impression		3.21	3.00	3.27
The service is...	terrible – wonderful	3.23	3.20	3.24
The experience is...	frustrating – satisfying	3.18	2.80	3.29
The system is...	not useful – useful	3.23	3.00	3.29
Perceived usefulness		3.26	2.8	3.47
Using the LT improves my performance in studying for this subject.	disagree – agree	3.13	2.20	3.55
Using the LT increases my understanding of the lecture.	disagree – agree	3.13	3.20	3.09
Using the LT makes it easier to follow the lecture.	disagree – agree	2.81	2.00	3.18
I would find the LT useful in other lectures.	disagree – agree	3.94	3.80	4.00
Perceived ease of use		3.27	2.65	3.45
I enjoy using the LT.	disagree – agree	3.19	2.00	3.56
The service works as expected.	disagree – agree	2.73	2.00	2.94
The features provided are sufficient.	disagree – agree	2.86	2.20	3.06
The layout of the user interface is clear.	disagree – agree	4.27	4.40	4.24

Table 2: Overall system evaluation, **NG** non-German, **G** German, with different categories

		All	NG	G
General		2.95	3.40	2.82
The transcription quality is...	unsatisfying – satisfying	2.86	3.40	2.71
The usefulness of the transcription is...	low – high	3.05	3.40	2.94
Quality		3.01	3.13	2.97
The errors of the transcription were...	distracting – not distracting	2.75	3.40	2.53
The delay in transcription was...	high – low	2.62	2.80	2.56
The transcription was...	disfluent – fluent	2.90	2.60	3.00
During lectures, transcription quality was...	fluctuating – consistent	3.06	3.20	3.00
During the term, transcription improved...	not at all – clearly	3.15	3.00	3.20
The transcription of general terms was...	bad – good	3.44	3.40	3.46
The transcription of technical terms was...	bad – good	3.28	3.50	3.21
Usefulness		3.13	2.93	3.22
The transcription helped me improve my performance in studying for this subject.	disagree – agree	3.19	3.00	3.27
The transcription made it easier to follow the lecture.	disagree – agree	3.06	2.80	3.18
The transcription made it easier to comprehend the content of the lecture.	disagree – agree	3.13	3.00	3.20

Table 3: Evaluation ASR component, **NG** non-German, **G** German, with different categories

amount of feedback is rather limited, it contained interesting aspects.

The feedback from people using it was rather positive. They considered it useful and appreciated this service being offered. Yet, there were some features missing. We did implement some of them since the time when we conducted this study. Examples are a reduced latency in both the transcription and translation.

As to the questions we asked at the beginning, "Does the tool help students?", "Is it easier to follow the lectures?": Based on the feedback, the LT does help (foreign) students. They especially appreciated the automatic transcription. In its current form, however, it is sometimes difficult to switch between the lecturer, the presentation and the LT while keeping track of the situation. Thus, especially when a lecturer is explaining charts, pictures or is taking notes it might be a bit confusing. Nevertheless, suggestions like the integration of the slides, an archive of the lectures and/or the possibility to annotate or highlight text would be of great help in this respect.

6. Bibliographical References

Cho, E., Fügen, C., Herrmann, T., Kilgour, K., Mediani, M., Mohr, C., Niehues, J., Rottmann, K., Saam, C., Stüker, S., et al. (2013). A real-world system for si-

		All	NG	G
General		3.19	2.67	3.36
The translation quality is...	unsatisfying – satisfying	3.20	3.00	3.27
The usefulness of the translation is...	low – high	3.18	2.25	3.46
Quality		3.13	2.68	3.28
The errors of the translation were...	distracting – not distracting	2.86	2.50	3.00
The delay in translation was...	high – low	2.87	2.25	3.09
The translation was...	disfluent – fluent	3.00	2.50	3.17
During a lecture, translation quality was...	fluctuating – consistent	2.93	2.25	3.18
During the term, translation improved...	not at all – clearly	3.29	3.00	3.36
The translation of general terms was...	bad – good	3.63	3.50	3.67
The translation of technical terms was...	bad – good	3.36	3.00	3.42
Usefulness		2.83	2.08	3.14
The translation helped me improve my performance in studying for this subject.	disagree – agree	2.71	2.00	3.00
The translation made it easier to follow the lecture.	disagree – agree	2.77	2.00	3.11
The translation made it easier to comprehend the content of the lecture.	disagree – agree	3.00	2.25	3.33

Table 4: Evaluation MT Component, **NG** non-German, **G** German, with different categories

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

Fügen, C., Kolss, M., Paulik, M., and Waibel, A. (2006). Open domain speech translation: from seminars and speeches to lectures. In *TC-STAR workshop on speech to speech translation, Barcelona, Spain*, pages 81–86.

Fügen, C., Waibel, A., and Kolss, M. (2007). Simultaneous translation of lectures and speeches. *Machine Translation*, 21(4):209–252.

Gates, D., Lavie, A., Levin, L., Waibel, A., Gavaldà, M., Mayfield, L., Woszczyna, M., and Zhan, P. (1996). End-to-end evaluation in JANUS: a speech-to-speech translation system. In *Dialogue Processing in Spoken Language Systems*, pages 195–206. Springer.

Hamon, O., Fügen, C., Mostefa, D., Arranz, V., Kolss, M., Waibel, A., and Choukri, K. (2009). End-to-end evaluation in simultaneous translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–353. Association for Computational Linguistics.

Stüker, S., Fügen, C., Kraft, F., and Wölfel, M. (2007). The isl 2007 english speech transcription system for european parliament speeches. In *INTERSPEECH*, pages 2609–2612.