SignSpeak

Scientific understanding and vision-based technological development for continuous sign language recognition and translation

Grant Agreement Number 231424

Small or medium-scale focused research project (STREP)
FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics

Project start date: 1 April 2009
Project duration: 36 months

Evaluation results of SignSpeak technology and GUI demonstrators

Major Deliverable D.7.4 – PUBLIC – M36

Release version: V1.0 – May 25th, 2012

Author: EUD, RWTH and CRIC.
Reviewers: CRIC.

SEVENTH FRAMEWORK PROGRAMME

Introduction

The purpose of this deliverable was to obtain a human evaluation of the software and technology developed in the SignSpeak project, as well as asking the evaluators for possible applications of the developed technology. To this end, the evaluation was conducted in cooperation with the "Deaf and Sign Language Research Team Aachen" (DESIRE) at RWTH Aachen University. A total of 6 Deaf and 3 hearing evaluators performed the evaluation.
The design of the human evaluation was done by EUD, TID, RWTH and CRIC. CRIC carried out the implementation by developing the web interface described in D6.3, which is available at http://www.signspeak.eu/demonstrator/, and RWTH worked on the evaluation design by selecting and grouping the videos and analysing the results.

Evaluation methodology

We performed a user evaluation on the test set of the RWTH-Phoenix Weather corpus, using the single signer setup with Signer03. Since the evaluation of signed videos turned out to be time consuming and laborious for the evaluators, after speaking to the DESIRE team and the EUD we decided to evaluate only a subset of the test set and presented a total of 18 video segments (with an average duration from 5 to 10 seconds) and their translations to the evaluators. The subset of the videos was manually selected and equally divided into three groups of six videos according to their subjective translation quality. The evaluators were asked to evaluate each video in the group and consequently to evaluate the group as a whole. They were not told that the videos were grouped according to translation quality. The three groups had neutral names Group 1, Group 2, and Group 3, and we chose to show first medium translations, then poor translations and in the end good translations. In the following, the three groups are referred to as "poor", "medium" and "good" instead of their neutral names.
The evaluation was conducted according to two criteria as is common in machine translation research (see [Koehn 09, Chapter 8][1] for details):

- The *adequacy* measures whether the translated text conveys the meaning of the signed video.
- The *fluency* measures whether the translations are grammatically correct and fluent.

For each evaluation, the evaluators applied a score from 1 ("poor") to 5 ("excellent"). One goal of the evaluation was to find out the impact of the different parts of the pipeline on the translation quality. To measure this, we used translations obtained from three different setups:

- *recognition*: For eight videos, the whole pipeline of recognition and translation was run.
- *transcription*: For eight videos, the human annotated glosses were translated by the machine translation system, i.e. no recognition was performed.
- *reference*: For two videos, the original text spoken by the announcer was presented. The reference translations were used to test which scores the evaluators would give to human translations.

When comparing the first two setups, we can see the impact of recognition errors on the translation quality. When comparing the latter two setups, we can see the impact of the translation system on the overall system output.

[1] P. Koehn: Statistical Machine Translation. Cambridge University Press, 2009.

Adequacy

The adequacy scores of the individual videos (for each of the three groups) are presented in Figure 1. The bars indicate the average scores given by the evaluators; the error bars indicate the variance of the scores among the different evaluators. The scores indicate that the manual placement in the groups was quite accurate, as the best scores were given for group 3, followed by group 1 and group 2.

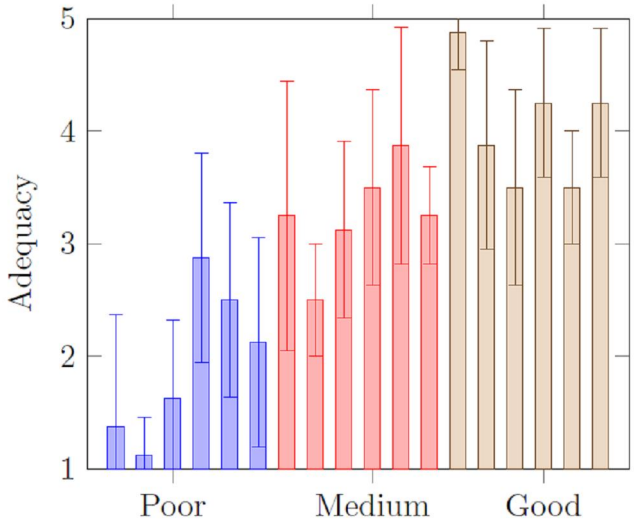

Figure 1. Adequacy scores of individual videos

Figure 2 compares the adequacy scores given to the individual videos with the scores given to the whole group of videos. The results indicate that the evaluators tended towards the medium score "3" when evaluating the adequacy of the whole group.

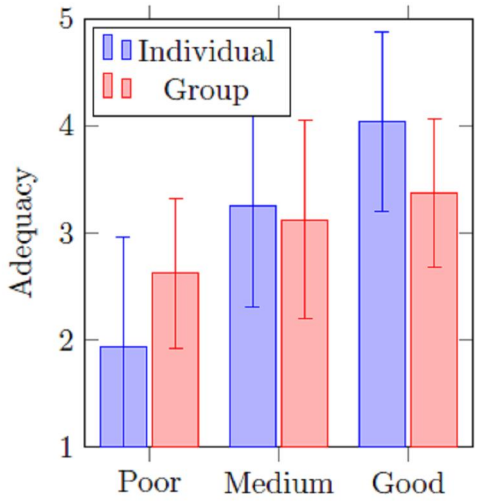

Figure 2. Adequacy: Individual videos vs. whole groups

Figure 3 shows the scores grouped according to the different kinds of setups. In the group "recognition", the sign language recognition system was applied to the video, and the output, a sequence of glosses, was translated into spoken German by the sign language translation system. In the group "transcription", the manually transcribed glosses were translated into spoken German, that is, the recognition step was skipped. In the group "reference", the original text spoken by the announcer was presented to the evaluators for comparison purposes. Figure 4 shows the average adequacy scores for the three groups. As was expected, the score for the whole pipeline is worse than for the translation system alone, since recognition errors inevitably propagate and lead to translation

errors. The degragation of the scores when using machine translation instead of the reference text is smaller than the degragation caused by using the sign language recognition output for translation. This indicates that the recognition system leads to more issues in the whole pipeline than the translation system. Note also that the reference text spoken by the announcer did not receive perfect scores. The reason is that in the RWTH-Phoenix Weather corpus, the human interpreters have to interpret the text spoken by the announcer under real time conditions and therefore do not always give an exact interpretation of the spoken text, but sometimes leave out some information.

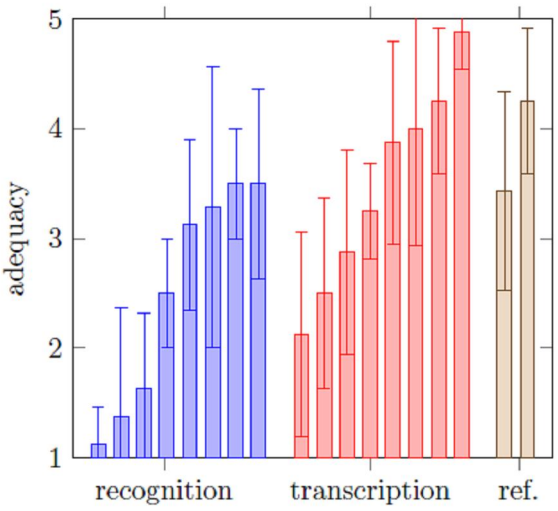

Figure 3. Adequacy: videos using recognition, transcription, reference

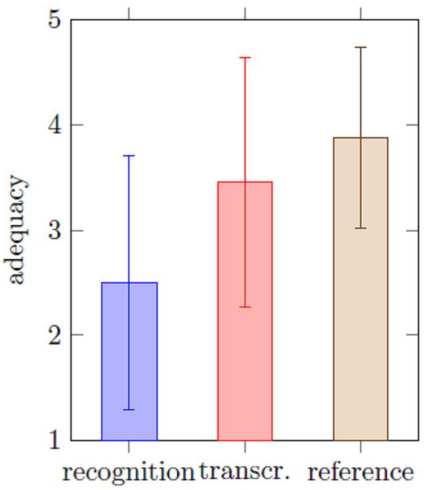

Figure 4. Summary: adequacy

Fluency

Figures 5 and 6 show the fluency scores of the individual videos and the three groups of videos. Interestingly, while the evaluators tended towards a medium score of "3" when evaluating the adequacy of the groups of videos, they applied lower scores for the fluency of the whole group compared to the average of the individual videos. This indicates a rather negative overall judgment in cases of grammatical errors and diffluent sentences.

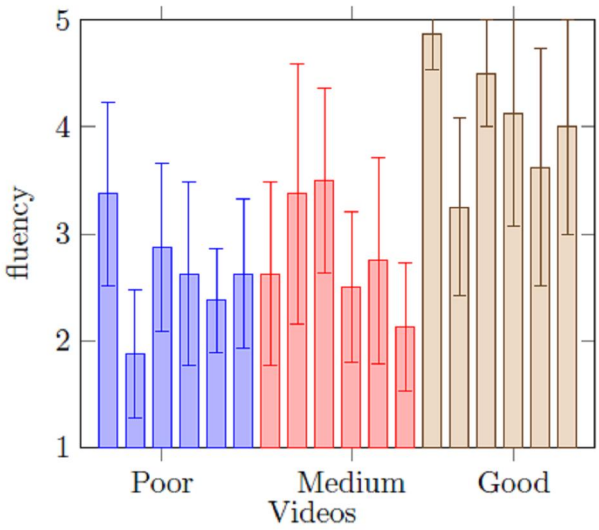

Figure 5. Fluency: Individual videos

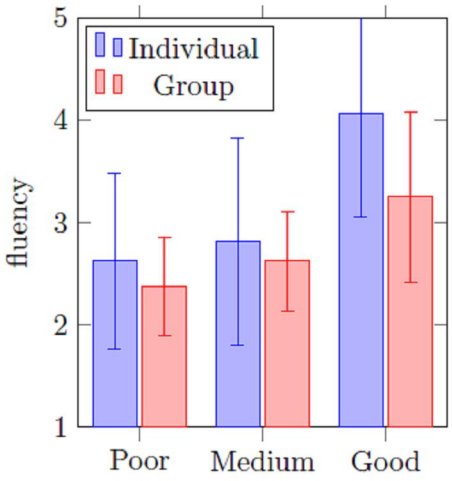

Figure 6. Fluency: Individual videos vs. groups

Figure 7 shows the fluency scores grouped according to the three different setups. The average scores are presented in Figure 8. The fluency score of the whole pipeline is very similar to the score of the translations of the reference glosses. This indicates that the fluency of the translation system does not strongly depend on the quality of the gloss input, but that the fluency has only an average quality even for perfect gloss input. In the discussion which took place after the evaluation, some evaluators expressed a strong rejection of translations which contain any grammatical errors.

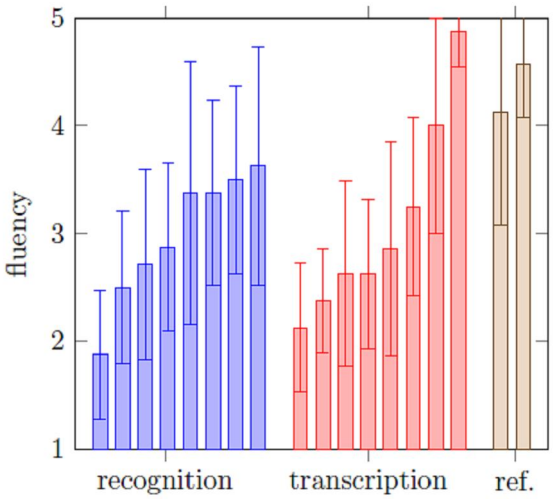

Figure 7. Fluency: Videos using recognition, transcription, reference

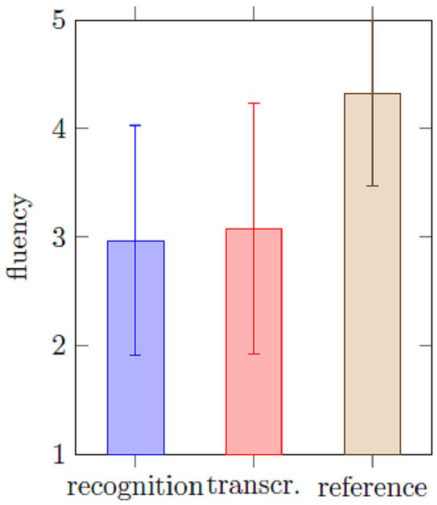

Figure 8. Summary: Fluency

Human evaluation vs. automatic metrices

One issue which arose in the course of the project was the question whether the automatic metrices which compare the system output to the reference text are a valid means to measure the system performance. In this section, we compare the automatic scores to the scores given by the human judges. Since the comparisons have to be performed on the sentence level, the BLEU score cannot be taken into account, because it is only defined on the document level. Two correlation measures are used to compare the adequacy and the fluency scores with the TER:

- The Pearson Correlation measures a linear dependence between two random variables.
- The Spearman's Rank Correlation measures whether the relationship between two random variables is monotonic. If the relationship between two metrices are monotonic, they rank the different hypotheses in the same order.

Note that the correlations between the human judgment and TER are negative, because for the human scores a higher number is better, but for TER, a lower number is better. An ideal correlation would therefore be -1. Figure 9 shows the correlation between the accuracy scores given by the human judges and the automatic TER score. The Pearson correlation is -0.61, Spearman's rank correlation is -0.67. The scores indicate that there is only a rough correlation between accuracy and TER. In Figure 10, the correlation between fluency and TER is examined. Here, the correlation is much stronger with a Pearson correlation of -0.80 and a Spearman's rank correlation of -0.77.

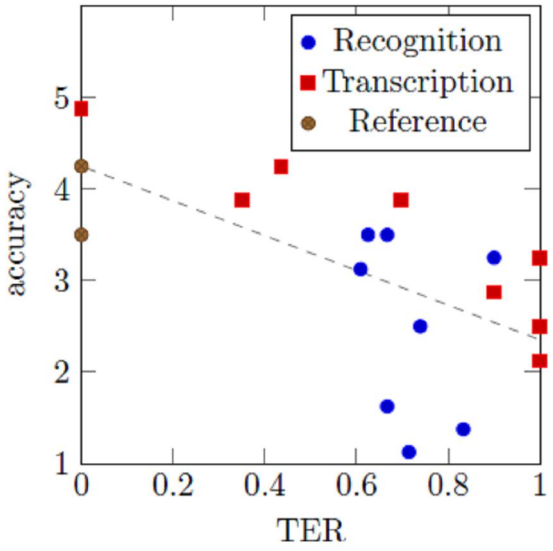

Figure 9. Correlation between adequacy and TER

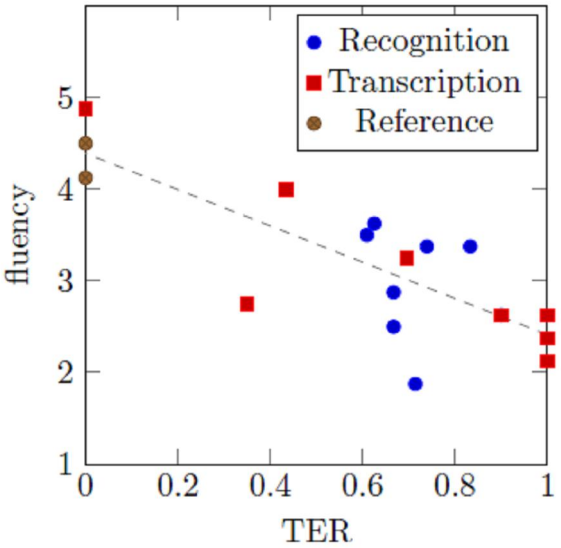

Figure 10. Correlation between fluency and TER

Results of the discussion following the human evaluation

After the formal evaluation of the software, the human judges could express their opinion on general issues concerning the SignSpeak project in the form of an informal discussion. Here are some opinions we received during this discussion:

- Concerning the layout of the evaluation website, one evaluator remarked that he felt the light text on a dark background was strenuous for the eyes, and that he would prefer a dark text on a light background.
- The users complained that the interpreters looked distorted in the video. The reason is that television screens use a different aspect ratio than computer screens. By now, we have adjusted the aspect ratio on the demonstration website.
- Most evaluators said that the domain of weather forecasting was not very relevant to the Deaf community, since usually a weather map or a textual description of the weather forecast is sufficient

Other topics which might be of interest to the Deaf community include:

- Sports
- Learning written German: the Deaf person could sign a sentence, and the system would show how the sentence looks in written German.
- Application forms at the social welfare office or for help at the workplace.
- Subtitles for sign language videos, for example for the website www.vibelle.de produced by the DESIRE team.
- Sign language input for Siri: giving commands to the smart phone or asking questions.
- Simple translations for everyday communication, e.g. when going shopping.
- Signing simple SMS messages, e.g. ”Sorry, I am coming late.”

An important topic in the discussion was the point that many Deaf are not very prefficient in writing texts and would prefer to sign about a topic, and the system would translate their signing into written text.

Advantages of automatic interpretation of sign language:

- Privacy issues, e.g. many Deaf persons do not want to bring along a human interpreter when going to the doctor.

- A computer can interpret when no human interpreter is at hand, e.g. in emergency situations. This would presuppose a mobile application such as a smartphone app.

Disadvantages of automatic interpretation of sign language:

- When the computer-generated translation is of a poor quality, the conversation partner might think that original signing of the Deaf person was incorrect and suppose that the Deaf person is stupid. The Deaf evaluators feared for their reputation.
- Lots of information is conveyed by facial expressions, which the computer might miss.
- Computers only see the video, but do not take into account the context. The human interpreter can take into account the context of the conversation and translate individual utterances accordingly.

What is important for a good translation:

- All information must be conveyed.
- The translation must be grammatically correct.
- Language is used for communication. Aspects such as intention and stressing what is important must be conveyed as well.
- The context of the whole communication must be taken into account for an adequate translation.