Enhancing a Sign Language Translation System with Vision-Based Features

Philippe Dreuw, Daniel Stein, and Hermann Ney

Lehrstuhl für Informatik 6– Computer Science Department,
RWTH Aachen University – D-52056 Aachen, Germany
{surname}@cs.rwth-aachen.de

Introduction. For automatic sign language translation, one of the main problems is the usage of spatial information and its proper representation and translation, e.g. the handling of spatial reference points in the signing space. Such locations are encoded at static points in signing space as spatial references for motion events.

We present a new approach starting from a large vocabulary speech recognition system which is able to recognize sentences of continuous sign language speaker independently [4]. The manual features obtained from the tracking, are passed to the statistical machine translation system to improve its accuracy. On a publicly available benchmark database, we achieve a competitive recognition performance and can similarly improve the translation performance by integrating the tracking features.

Tracking System. Relevant body parts such as the head and the hands have to be found for feature extraction, but most systems can only produce candidate regions. To extract features which describe manual components of a sign, the dominant hand is tracked in each image sequence. Therefore, a robust tracking algorithm is required as the signing hand frequently moves in front of the face, may temporarily disappear, or cross the other hand. Our head and hand tracking framework is based on the work of [1]. These hand features, which are usually used within the recognition framework, can also be used within the translation framework, in order to improve the translation error rate, too.

Translation System. We use a statistical machine translation system to automatically transfer the meaning of a source language sentence into a target language sentence [2].

Following the notation convention, we denote the source language with J words as $f_1^J = f_1 \dots f_J$, a target language sentence as $e_1^I = e_1 \dots e_I$ and their correspondence as the a-posteriori probability $\Pr(e_1^I|f_1^J)$. Our baseline system maximizes the translation probability directly using a log-linear model:

$$p(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)}, \tag{1}$$

with a set of different features h_m, scaling factors λ_m and the denominator a normalization factor that can be ignored in the maximization process. We choose the λ_m by optimizing an MT performance measure on a development corpus using the downhill simplex algorithm. For a complete overview of the sign language

Fig. 1. Sample frames for pointing near and far used in the translation.

translation system, see [3]. The tracking positions of the dominant-hand were clustered and their mean calculated. Then, for deictic signs, the nearest cluster according to the Euclidean distance was added as additional word information for the translation model.

Experimental Results. The RWTH-Boston-Hands database[1] for the evaluation of hand tracking methods in sign language recognition systems has been prepared. It consists of a subset of the RWTH-Boston-104 videos [4]. The positions of both hands have been annotated manually for 1119 frames in 15 videos. We achieve a 2.30% tracking error rate for a 20×20 search window [1].

In the translation, the incorporation of the tracking data for the deixis words helped the translation system to discriminate between deixis as distinctive article, locative or discourse entity reference function. For example, the sentence `JOHN GIVE WOMAN IX COAT` might be translated into *John gives the woman the coat* or *John gives the woman over there the coat* depending on the nature of the pointing gesture `IX`. Using the tracking data, the translation error rate in preliminary experiments (40 test sentences) improves from 44.2% Levenshtein word error rate to 42.5 %, and from 42.5% position independent word error rate to 40.3%.

Summary & Conclusions. We presented a vision-based approach to continuous ALSR and a statistical machine translation approach for ASLT. The results suggest that hand tracking information is an important feature for sign language translation, especially for grammatically complex sentences where discourse entities and deixis occur a lot in signing space.

References

1. P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In *7th Intl. Conference on Automatic Face and Gesture Recognition*, IEEE, Southampton, pages 293–298, April 2006.
2. A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In *IWSLT*, Kyoto, Japan, pages 103–110, November 2006. Best paper award.
3. D. Stein, J. Bungeroth, and H. Ney. Morpho-Syntax Based Statistical Methods for Sign Language Translation. In *11th Annual conference of the European Association for Machine Translation*, Oslo, Norway, pages 169–177, June 2006.
4. M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, and H. Ney. Using Geometric Features to Improve Continuous Appearance-based Sign Language Recognition. In *17th BMVC*, Edinburgh, UK, pages 1019–1028, September 2006.

[1] `http://www-i6.informatik.rwth-aachen.de/~dreuw/database.html`