Towards Automatic Sign Language Annotation for the ELAN Tool

Philippe Dreuw and Hermann Ney

Human Language Technology and Pattern Recognition Group,
RWTH Aachen University, Aachen, Germany
{dreuw,ney}@cs.rwth-aachen.de

Abstract

A new interface to the ELAN annotation software that can handle automatically generated annotations by a sign language recognition and translation framework is described. For evaluation and benchmarking of automatic sign language recognition, large corpora with rich annotation are needed. Such databases have generally only small vocabularies and are created for linguistic purposes, because the annotation process of sign language videos is time consuming and requires expert knowledge of bilingual speakers (signers). The proposed framework provides easy access to the output of an automatic sign language recognition and translation framework. Furthermore, new annotations and metadata information can be added and imported into the ELAN annotation software. Preliminary results show that the performance of a statistical machine translation improves using automatically generated annotations.

1. Introduction

Currently available sign language video databases were created for linguistic purposes (Crasborn et al., 2004; Neidle, 2002 and 2007) or gesture recognition using small vocabularies (Martinez et al., 2002; Bowden et al., 2004). An overview of available language resources for sign language processing is presented in (Zahedi et al., 2006). Recently, an Irish Sign Language (ISL) database (Stein et al., 2007) and an American Sign Language (ASL) database (Dreuw et al., 2008) have been published.

Most available sign language corpora contain simple stories performed by a single signer. Additionally, they have too few observations for a relatively large vocabulary which is inappropriate for data driven and statistically based learning methods. Here we focus on the automatic annotation and metadata information for benchmark databases that can be used for analysis and evaluation of:

- linguistic problems
- automatic sign language recognition systems
- statistical machine translation systems

For storing and processing sign language, a textual representation of the signs is needed. While there are several notation systems covering different linguistic aspects, we focus on the so called gloss notation, being widely used for transcribing sign language video sequences.

Linguistic research in sign language is usually carried out to obtain the necessary understanding regarding the used signing (e.g. sentence boundaries, discourse entities, phonetic analysis of epenthetic movements, coarticulations, or role changes), whereas *computer scientists* usually focus on features for sign language recognition (e.g. body part tracking of head and hands, facial expressions, body posture), or on post-processing and additional monolingual data for statistical machine translation to cope with encountered sign language related statistical machine translation errors.

Therefore some common important features and search goals for these different research areas are e.g.

- body part models and poses, hand poses, facial expressions, eye gaze, ...
- word spotting and sentence boundary detection
- pronunciation detection and speaker identification

In particular, statistical recognition or translation systems rely on adequately sized corpora with a rich annotation of the video data. However, video annotation is very time consuming: in comparison to the annotation of e.g. parliamentary speech, where the annotation real-time-factor (RTF) is about 30 (i.e. 1 hour of speech takes 30 hours of annotation), the annotation of sign language video can have a annotation RTF of up to 100 for a full annotation of all manual and non-manual components.

2. Baseline System Overview & Features

Figure 1 illustrates the components of our proposed recognition and annotation system.

The recognition framework and the features used to achieve the experimental results have been presented in (Dreuw et al., 2007a). The baseline automatic sign language recognition (ASLR) system uses appearance-based image features, i.e. thumbnails of video sequence frames. They give a global description of all (manual and non-manual) features that have been shown to be linguistically important. The system is Viterbi trained and uses a trigram language model (Section 2.4.) which is trained on the groundtruth annotations of main glosses.

The ASLR system is based on the Bayes' decision rule: for a given sign language video input sequence, first features x_1^T are extracted to be used in the global search of the model which best describes the current observation:

$$\begin{aligned} &\arg\max_{w_1^N} \left\{ Pr(w_1^N|x_1^T) \right\} \\ &= \arg\max_{w_1^N} \left\{ Pr(w_1^N) \cdot Pr(x_1^T|w_1^N) \right\} \end{aligned} \tag{1}$$

The word sequence w_1^N (i.e. a gloss sequence) which maximizes the language model (LM) probability $Pr(w_1^N)$ and the visual model probability $Pr(x_1^T|w_1^N)$ will be the recognition result.

Statistical machine translation (SMT) is a data-driven translation method that was initially inspired by the so-called noisy-channel approach: the source language is interpreted

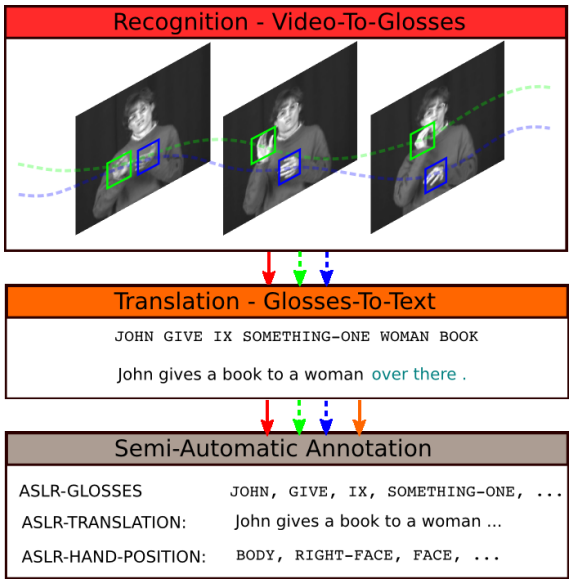

Figure 1: Complete system setup with an example sentence: After automatically recognizing the input sign language video, the translation module has to convert the intermediate text format (glosses) into written text. Both system outputs and features can be used to automatically generate annotations.

as an encryption of the target language, and thus the translation algorithm is typically called a decoder. In practice, statistical machine translation often outperforms rule-based translation significantly on international translation challenges, given a sufficient amount of training data.

A statistical machine translation system presented in (Dreuw et al., 2007b) is used here to automatically transfer the meaning of a source language sentence into a target language sentence. 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 $\Pr(e_1^I|f_1^J)$. The sentence $\hat{e}_1^I$ that maximizes this probability is chosen as the translation sentence as shown in Equation 2. The machine translation system accounts for the different grammar and vocabulary of sign language.

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \Pr(e_1^I|f_1^J) \right\} \quad (2)$$

$$= \arg\max_{e_1^I} \left\{ \Pr(e_1^I) \cdot \Pr(f_1^J|e_1^I) \right\} \quad (3)$$

For a complete overview of the translation system, see (Mauser et al., 2006).

2.1. Body Part Descriptions

The baseline system is extended by hand trajectory features (Dreuw et al., 2007a) being similar to the features presented in (Vogler and Metaxas, 2001). Similar as presented in (Bowden et al., 2004; Yang et al., 2006), features such as the relative position and pose of the body, the hands or the head could be extracted. The proposed system can be easily extended by other feature extraction methods which could

Figure 2: Sample frames for pointing near and far used in the translation.

extract further user specific metadata information for the annotation files.

To enhance translation quality, we propose to use visual features from the recognition process and include them into the translation as an additional knowledge source.

2.2. Pronunciation Detection and Speaker Identification

Given dialectal differences, signs with the same meaning often differ significantly in their visual appearance and in their duration. Each of those variants should have a unique gloss annotation.

Speakers could e.g. be identified using state-of-the-art face detection and identification algorithms (Jonathon Phillips et al., 2007).

2.3. Sentence Boundary Detection and Word Spotting

Temporal segmentation of large sign language video databases is essential for further processing, and is closely related to sentence boundary detection in speech recognition (ASR) and tasks such as video shot boundary detection (Quenot et al., 2003).

In addition to audio and video shot boundary detection, which is usually done just at the *signal level*, we could use the hand tracking information inside the virtual signing space from our sign language recognition framework to search for sentence boundaries in the signed video streams (e.g. usage of neutral signing space). Due to the different grammar in sign language, a word spotting of e.g. question markers (e.g. so called `ONSET`, `OFFSET`, `HOLD` or `PALM-UP` signs (Dreuw et al., 2008)) could deliver good indicators for possible sentence boundaries.

2.4. Language Models

Due to the simultaneous aspects of sign language, language models based on the (main gloss) sign level versus independent language models for each communication channel (e.g. the hands, the face, or the body) can be easily generated using e.g. the SRILM toolkit (Stolcke, 2002) and added as metadata information to the annotation files.

3. Automatically Annotating ELAN Files With Metadata Information

The ELAN annotation software[1] is an annotation tool that allows you to create, edit, visualize, and search annotations for video and audio data, and is in particular designed for the analysis of language, sign language, and gesture. Every ELAN project consists of at least one media file with its corresponding annotation file.

[1] `http://www.lat-mpi.eu/tools/elan/`

Our proposed automatic annotation framework is able to:

- convert and extend existing ELAN XML annotation files with additional metadata information
- automatically annotate new media files with glosses (video-to-glosses), translations (glosses-to-text), and metadata information from the automatic sign language recognition framework

The richness of gloss annotation can be defined by different user needs (e.g. sentence boundaries, word spotting, main glosses, facial expressions, manual features, etc.) (c.f. Section 2.), and can depend on the confidence of the sign language recognition or translation framework: the linguist might search for a specific sign and would need high quality annotations, whereas the computer scientist could only import annotations with low confidences and erroneous recognition or translation for a fast analysis and correction of the automatically generated annotations in order to use them for a supervised retraining of the system.
Currently our proposed framework converts a recognizer output file with its corresponding word confidences generated by the `sclite` tool [2] from the NIST Scoring Toolkit (Fiscus, 2007) into a tab-delimited text file, which can be imported by the recently released ELAN 3.4.0 software. The file contains for each tier the begin, end, and duration times of each annotation value.

4. Experimental Results

An independent multi-channel training and recognition will allow automatic annotation of e.g. head and hands. The current whole-word model approach only allows for complete main gloss annotations. However, in another set of experiments presented in (Dreuw et al., 2007b), for incorporation of the tracking data, 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. For a given word boundary, these specific feature informations can be added as an additional tier and imported to the ELAN tool (see `ASLR-HAND` tiers in Figure 3).
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` (see `ASLR-TRANSLATION` tier in Figure 3). This helped the translation system to discriminate between deixis as distinctive article, locative or discourse entity reference function.
Preliminary results for statistical machine translation with sign language recognizer enhanced annotation files have been presented in (Dreuw et al., 2007b; Stein et al., 2007). Using the additional metadata, the translation improved in performance from 28.5% word-error-rate (WER) to 26.5% and from 23.8% position-independent WER to 23.5%, and shows the need for further metadata information in corpora annotation files.

[2] `http://www.nist.gov/speech/tools/`

Preliminary annotation results for word boundaries, sentence boundaries, and head/hand metadata information are shown in Figure 3. Depending on a word confidence threshold of the recognition system, the amount of automatically added glosses can be controlled by the user (see `ASLR-GLOSSES` and `ASLR-CONFIDENCES` tier in Figure 3). This also enables to search for pronunciations (if modeled as e.g. in (Dreuw et al., 2007a)). Furthermore body part and spatial features as proposed in (Stokoe et al., 1965; Bowden et al., 2004) can be added as additional information streams (see `ASLR-HAND` and `ASLR-FACE` tiers in Figure 3).

5. Summary & Conclusion

Here, we presented and proposed an automatic annotation extension for the ELAN tool which can handle automatically generated annotations and metadata information from a continuous sign language recognition and translation framework.
Challenging will be multiple stream processing (i.e. an independent recognition of hands, faces, body, ...), pronunciation detection, and speaker identification, as well as the extraction of better visual features in order to improve the quality of the automatically generated annotation files. It will enable to automatically add rich annotations (e.g. head expression/position/movement, hand shape/position/movement, shoulders, eye brows/gaze/aperture, nose, mouse, or cheeks) as already partly manually annotated in (Neidle, 2002 and 2007).
Interesting will be unsupervised training, which will improve the recognition and translation performance of the proposed systems. The implicitly generated ELAN annotation files will allow for fast analysis and correction.
A helpful extension of the ELAN software would be an integrated video annotation library (e.g. simple box drawing or pixel marking) which would allow to use ELAN as a groundtruth annotation tool for many video processing task, and would furthermore allow for a fast and semi-automatic annotation and correction of sign language videos.

6. References

R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. 2004. A linguistic feature vector for the visual interpretation of sign language. In *European Conf. Computer Vision*, volume 1, pages 390–401.

Onno Crasborn, Els van der Kooij, and Johanna Mesch. 2004. European cultural heritage online (ECHO): publishing sign language data on the internet. In *Theoretical Issues in Sign Language Research*, Barcelona, Spain, October.

P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. 2007a. Speech recognition techniques for a sign language recognition system. In *Interspeech 2007*, pages 2513–2516, Antwerp, Belgium, August. ISCA best student paper award of Interspeech 2007.

P. Dreuw, D. Stein, and H. Ney. 2007b. Enhancing a sign language translation system with vision-based features. In *International Workshop on Gesture in Human-Computer Interaction and Simulation*, pages 18–20, Lisbon, Portugal, May.

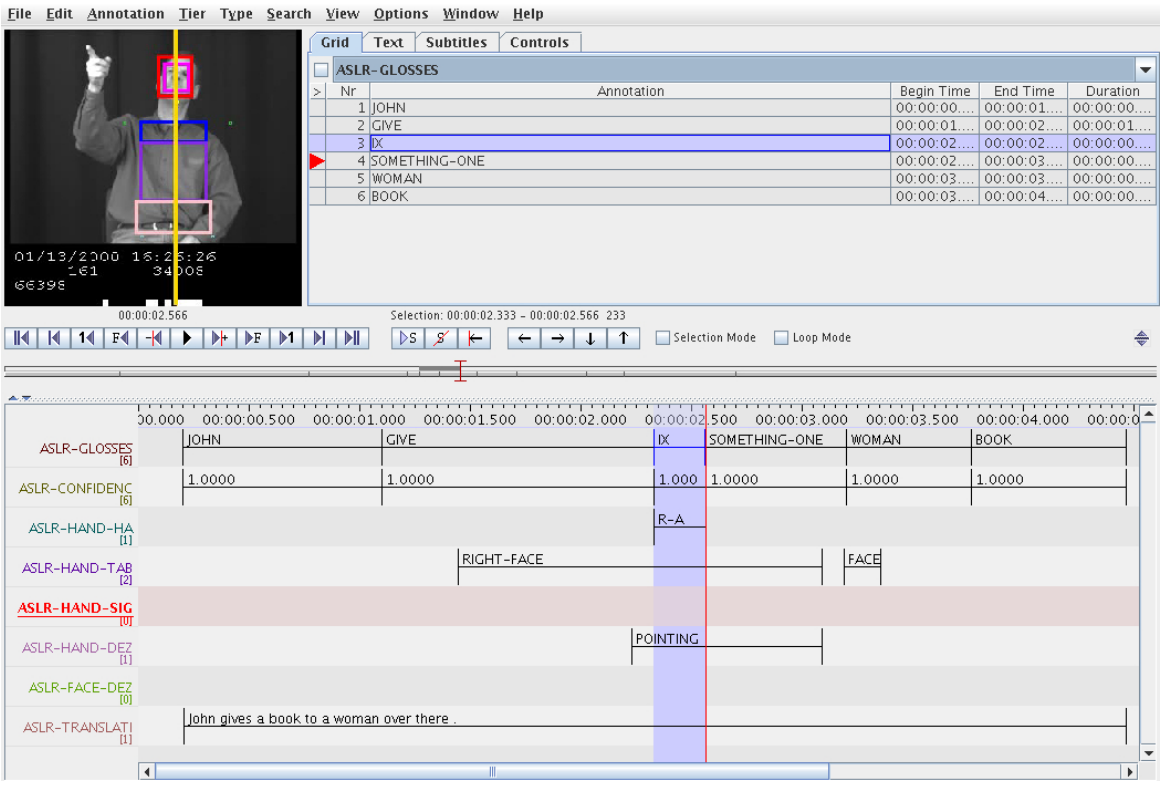

Figure 3: Screenshot of the ELAN tool with automatically generated annotations and video with overlaid features.

P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May. `http://www-i6.informatik.rwth-aachen.de/~dreuw/database.html`.

John Fiscus. 2007. Nist speech recognition scoring toolkit (NIST-SCTK). `http://www.nist.gov/speech/tools/`.

P. Jonathon Phillips, W. Todd Scruggs, Alice J. O'Toole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, and Matthew Sharpe. 2007. Frvt 2006 and ice 2006 large-scale results. Technical Report NISTIR 7408, NIST, March.

A. M. Martinez, R. B. Wilbur, R. Shay, and A. C. Kak. 2002. Purdue RVL-SLLL ASL database for Automatic Recognition of American Sign Language. In *IEEE Int. Conf. on Multimodal Interfaces*, Pittsburg, PA, USA, October.

Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, and Hermann Ney. 2006. The RWTH statistical machine translation system for the IWSLT 2006 evaluation. In *IWSLT*, pages 103–110, Kyoto, Japan, November. Best paper award.

Carol Neidle. 2002 and 2007. Signstream^{TM}Annotation: Conventions used for the American Sign Language Linguistic Research Project and addendum. Technical Report 11 and 13, American Sign Language Linguistic Research Project, Boston University.

G. Quenot, D. Moraru, and L. Besacier. 2003. Clips at trecvid: Shot boundary detection and feature detection. In *TRECVID 2003 Workshop Notebook Papers*, pages 18–21, Gaithersburg, MD, USA.

Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, and Andy Way. 2007. Hand in hand: Automatic sign language to speech translation. In *11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 214–220, Skövde, Sweden, September.

W. Stokoe, D. Casterline, and C. Croneberg. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Gallaudet College Press, Washington D.C., USA.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*, volume 2, pages 901–904, Denver, CO, September.

C. Vogler and D. Metaxas. 2001. A framework for recognizing the simultaneous aspects of American Sign Language. *Computer Vision & Image Understanding*, 81(3):358–384, March.

R.D. Yang, S. Sarkar, B. Loeding, and A. Karshmer. 2006. Efficient generation of large amount of training data for sign language recognition: A semi-automatic tool. In *International Conference on Computers Helping People with Special Needs (ICCHP 06)*.

Morteza Zahedi, Philippe Dreuw, David Rybach, Thomas Deselaers, and Hermann Ney. 2006. Continuous sign language recognition - approaches from speech recognition and available data resources. In *Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pages 21–24, Genoa, Italy, May.