Hand in Hand: Automatic Sign Language to English Translation

Daniel Stein, Philippe Dreuw, Hermann Ney
Computer Science Department,
RWTH Aachen University, Germany
{stein, dreuw, ney}
@cs.rwth-aachen.de

Sara Morrissey, Andy Way
School of Computing
Dublin City University, Ireland
{smorri,away}
@computing.dcu.ie

Abstract

In this paper, we describe the first data-driven automatic sign-language-to-speech translation system. While both sign language (SL) recognition and translation techniques exist, both use an intermediate notation system not directly intelligible for untrained users. We combine a SL recognizing framework with a state-of-the-art phrase-based machine translation (MT) system, using corpora of both American Sign Language and Irish Sign Language data. In a set of experiments we show the overall results and also illustrate the importance of including a vision-based knowledge source in the development of a complete SL translation system.

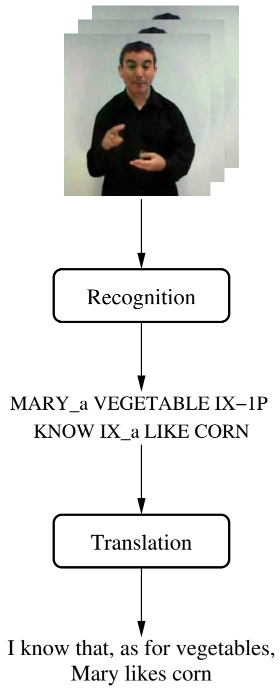

Figure 1: System setup with sample sentence

1 Introduction

The communication between deaf and hearing persons poses a much stronger problem than the communication between blind and seeing people. While the latter can talk freely by means of a common spoken language in which both are equally proficient, the deaf have their own, manual-visual language.

In this paper, we present an approach to automatically recognize sign language and translate it into a spoken language by means of data-driven methods. While the recognizer output is not easily intelligible because of different grammar and annotation format, we show that translation into the spoken language using standardized statistical machine translation (SMT) methods gives reasonable results, even for extremely small corpora. In preliminary experiments, we also give an outlook of how to incorporate vision-based features used in the recognizer to improve the overall translation result. Our work focuses on translating American Sign Language (ASL) and Irish Sign Language (ISL) into English (see Figure 1).

The remainder of the paper is constructed as follows. Section 2 introduces sign languages and gives an overview of the transcription methodology employed for capturing descriptions of sign languages. The area of gesture recognition is presented in section 3. Section 4 details data-driven MT approaches for SLs and describes the MT system we have employed. The experiments carried out are described in section 5 and the results are discussed in section 6. Finally, we conclude the paper in section 7 and outline the future avenues for our work.

2 Sign Languages

In spite of common misconceptions, SLs are natural, indigenous and independent means of communication for deaf and hard-of-hearing communities worldwide. Since the languages have not been created artificially but rather evolved naturally, it is no surprise that most countries have their own particular SL as well as local dialects. SLs are grammatically distinct from spoken languages and the grammar makes extensive use of the possibilities of a visual/gestural modality: locations, verb inflections, pronouns and many other linguistic devices are conveyed by spatial information in front of the signer. Apart from the obvious employment of the hands as information carriers, SLs also use affected facial expressions, tilts of the head and shoulder as well as the velocity of the sign to incorporate information such as comparative degree or subclauses.

For example, ISL, one of the SLs used in this paper, is the primary language of the Irish Deaf community. Despite this, the language is not recognised as an official language in Ireland, however, the 5000 strong community is joined by the Irish Deaf Society[1] and the Centre for Deaf Studies[2] in promoting ISL awareness and research across the country.

2.1 Sign Language Transcription

One of the striking differences between signed and spoken languages is the lack of a formally adopted writing system for SLs. There have been some attempts to develop writing systems for SLs, many of which are based on the seminal work of (Stokoe, 1960) and describe the handshape, location and articulated movement of a sign. These include the Hamburg Notation System (HamNoSys) (Hanke, 2004) and SignWriting (Sutton, 1995). Developed as handwriting systems, they use simple line drawings that are intuitively and visually connected to the signs themselves.

Despite the development of these approaches, they currently fall short of being either computationally useful or comprehensive enough for use in SL MT. For this reason we have chosen to use an approach referred to as *annotation* (Pizzuto and Pietrandrea, 2001). This involves the manual transcription of sign language taken from video data that is reproduced in a *gloss* format. The gloss is a semantic representation of sign language where, conventionally, the semantic meaning of the sign is transcribed in the upper case stem form of the local spoken language. The annotation "IX" signifies a deictic reference signed by a pointing gesture with the index finger. Additional spatial and non-manual information may also be added. An example of annotated glosses taken from our data is shown in Table 1. The first sentence is written in ASL glosses. The narrator (indicated by IX-1P) knows that Mary, at the spatial position referenced as "_a" and in the subordinate clause, likes corn. Here, the deixis "IX_a" serves as a pronoun to pick up the object of the subordinate clause again. A second sentence closer to the English grammar is written in ISL glosses. Note that, although both ISL and ASL are glossed in English, the grammar and vocabularies of the two sign languages are completely different.

2.2 The Corpus

Data-driven approaches to MT require a bilingual data set. In comparison to spoken language translation, SL corpora are difficult to acquire. To tune and test our system, we assembled the RWTH-Boston-104 corpus as a subset of a larger database of sign language sentences that were recorded at Boston University for linguistic research (Neidle et al., 1999). The RWTH-Boston-104 corpus consists of 201 video sentences, consisting of 104 unique words. The sentences were signed by 3 speakers and the corpus is split into 161 training and 40 test sequences. An overview of the corpus is given in Table 2: 26% of the training data are singletons, i.e. we only have one attempt to train the models properly. The sentences

[1] http://www.deaf.ie

[2] http://www.centrefordeafstudies.com

Table 1: Gloss annotation examples

ASL gloss	MARY_a VEGETABLE IX-1P KNOW IX_a LIKE CORN
English translation	I know that, as for vegetables, Mary likes corn.
ISL gloss	IX-FLIGHT FLIGHT B A ROUND TRIP IX-FLIGHT palm-up
English translation	Is flight B A a round trip flight?

Table 2: RWTH-Boston-104 corpus statistics

	Training	Test
sentences	161	40
running words	710	178
unique words	103	65
singletons	27	9
OOV	-	1

Table 3: ATIS corpus statistics

	Training	Devel	Test
sentences	482	98	100
running words	3707	593	432
unique words	375	88	128
singletons	144	28	10
OOV	-	30	4

have a rather simple structure and therefore the language model perplexity is low. The test corpus has one out-of-vocabulary (OOV) word. Obviously, this word cannot be recognized correctly using whole-word models.

Apart from this relatively small corpus, few data collections exist that are interesting for data-driven approaches. Much of what is available is in the form of conversation, stories and poetry which is unsuitable for ASLR and MT as illustrated in (Morrissey and Way, 2006). For this reason we chose to create our own corpus. We used the Air Travel Information System (ATIS) corpus of transcriptions from speech containing flight information in English as our base. The corpus consists of 680 sentences. For the purposes of our translation work, we had the data set translated and signed into ISL by native deaf signers. This was then manually annotated with semantic glosses as described in section 2.1.

3 Sign Language Recognition

The automatic sign language recognition (ASLR) system is based on an automatic speech recognition (ASR) system adapted to visual features(Lööf et al., 2006). The word sequence which best fits the current observation to the trained word model inventory (which is related to the acoustic model in ASR) and language model (LM) will be the recognition result.

In our baseline system, we use intensity images scaled to 32×32 pixels as features. To model image variability, various approaches are known and have been applied to gesture recognition similar to the works of (Dreuw et al., 2007). The baseline system is Viterbi trained and uses a trigram LM. In subsequent steps, this baseline system is extended by features that take the hand position and movement into account.

To extract manual features, 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. We use an approach based on dynamic programming which is inspired by the time alignment algorithm in ASR and which is guaranteed to find the optimal path with respect to a given criterion and which prevents taking possibly wrong local decisions. Given the position of the hand, features such as velocity, trajectory, and acceleration can easily be extracted.

4 Data-driven Sign Language MT

SL MT is still a new area of research with work dating back only roughly a decade. Despite the relative novelty of the area in comparison with mainstream MT, it has followed the trend away from ‘second generation’ rule-based approaches towards data-driven methods. An overview of current developments in this area is given in section 4.1 and the translation system used for our experiments is described in section 4.2.

4.1 Related Research

There are currently four groups working on data-driven SL MT. Their approaches are described below:

- (Morrissey and Way, 2005) have explored Example-Based MT approaches for the language pair English–Sign Language of the Netherlands with further developments being made in the area of ISL.
- (Stein et al., 2006) have developed an SMT system for German and German sign language in the domain weather reports. Their work describes the addition of pre- and post-processing steps to improve the translation for this language pairing. However, the methods rely on external knowledge sources such as grammar parsers that cannot be utilized here since our source input are glosses, for which no automatic parser exists.
- (Chiu et al., 2007) present a system for the language pair Chinese and Taiwanese sign language. The optimizing methodologies are shown to outperform IBM model 2.
- (San-Segundo et al., 2006) have undertaken some basic research on Spanish and Spanish sign language with a focus on a speech-to-gesture architecture. They propose a decompensation of the translation process into two steps: first they translate from written text into a semantic representation of the signs. Afterwards a second translation into graphically oriented representation is done. This representation can be understood by the avatar. Note, however, that this is the opposite translation direction as the one proposed here.

4.2 Statistical Machine Translation

We use a state-of-the-art phrase-based statistical machine translation system 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 $f_1^J = f_1 \ldots f_J$, a target language sentence as $e_1^I = e_1 \ldots e_I$ 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:

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

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 description of the system, see (Mauser et al., 2006).

5 Experiments

5.1 RWTH-Boston-104

Baseline. The baseline translation of the annotated gloss data into written English for the RWTH-Boston-104 has a word error rate (WER) of 21.2% and a position-independent word error rate (PER) of 20.1%. Looking at the data, the translation is even more accurate than that – the main problem being the lack of sentence boundary markers like dots and commas in sign language which are then omitted in the translation process.

Recognition. First, we analyze different appearance-based features for our baseline system. The simplest feature is to use intensity images down scaled to 32×32 pixels. As a baseline, we obtained a WER of 33.7%. For reducing the number of features and thus the number of parameters to be learned in the models, we apply linear feature reduction technique to the data, the principal component analysis (PCA). With PCA, a WER of 27.5% can be obtained (see Figure 2).

A log-linear combination of two independently trained models (PCA that include automatic tracking of hand velocity (HV) and hand trajectory (HT), respectively), leads to our best result of 17.9% WER (i.e. 17 del., 3 ins., and 12 subst.), where the model weights have been optimized empirically.

Sign-Language-to-Speech. If we translate these recognized glosses into written English (again, with punctuation mark post-processing), the overall score is 27.6% WER and 23.6% PER.

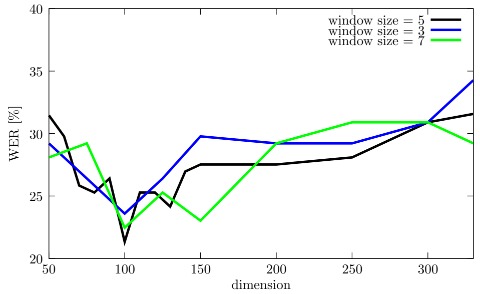

Figure 2: Combination of PCA-frames using PCA windowing

Figure 3: Sample frames for pointing near and far used in the translation.

In another set of experiments, we derive the tracking positions from all of the sentences. 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 (Dreuw et al., 2006). In order to distinguish between locative and descriptive pronouns, 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 (see Figure 3).

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 improves in performance from 28.5% WER to 26.5% and from 23.8% PER to 23.5%.

5.2 ATIS Corpus

The baseline translation of the annotated gloss data into written English has a WER of 45.1% and a PER of 34.7%. While this is a much more challenging result in itself if introduced with an additional error source like recognition, the preliminary recogniton of the ATIS videos had an error rate of 85% WER, with 327 deletions, 5 insertions and 175 substitutions out of 593 words. It is apparent from these result that further translation makes no sense at the moment if we start from the recognized data.

6 Discussion

Although the size of the corpus RWTH-Boston-104 is far too small to make reliable assumptions about the general significance of the results, at the very least we show that statistical machine translation is capable to work as an intermediate step for a complete sign-to-speech system. Even for extremely small training data, the resulting translation quality is reasonable.

We have shown that the recognition output in itself is not directly intelligble, given the different grammar and vocabulary of sign languages and shortages of the existing annotation system, but together with the automatic translation, the overall system can be easily trained on new language pairs and new domains. This set of sentences could without any doubt be translated with a reasonable rule-based system, yet it is not the ultimate goal to translate this corpus but to show that a sign-to-speech system is in principle possible using statistical methods, given reasonable data.

Moreover, adding features from the recognition process like the hand tracking position seems to help the translation quality, as it enables the system to distinguish between certain flexions of common words like the pointing gesture "`IX`". We argue that this can be compared to adding parts-of-speech (POS) information, to discriminate for example between deixis as distinctive article or as locative discourse entity reference.

As no grammar parser exists for sign language annotation, we propose a stemming of the glosses (i.e. leaving out the flexion) during recognition to cope with data sparseness problems. The missing information can be included by adding the relevant features during the translation process, analogous to morpho-syntactic translation enhancement to sparse language pairs with a rich grammatical parser on the source language side.

For the more sophisticated ATIS Corpus, translation is possible, at this stage, however, recognition produces far too much noise for a reasonable translation adaption. Given the numbers of singletons alone, these are already quite an obstacle for translation, but if they consist of several frames in a video where the exact starting and end time is not passed on to the recogniser, they are quite challenging for the algorithm. Moreover, sign languages produce quite a large effect known as coarticulation, i.e. the movement between two regular signs, that cannot be as easily trained. To date, we have not carried out experiments on the ATIS data with the addition of several recognition features, so, while time-expensive, there is still ground for improved results. The ratio of the deletions with regard to the number of words also strongly indicate that there is much room for improvement with tuning on the development set.

7 Conclusion

To the best of our knowledge, we present the first approach to combine data-driven methods for recognition output and translation of sign languages. Both these methods alone work on an intermediate notation, that does not provide any support for the target group as it is not used in the deaf community. With our system, we are able to produce a unique sign-language-to-speech system.

Like other poorly resourced languages, sign language research suffers from lack of training material to feed the corpus-based algorithms properly. However, given the data sparseness, a small domain that matches in vocabulary size according to the small sentence number, gives reasonably optimistic results.

We have also shown that the translation improves if it relies on additional recognition data and argue that this can be interpreted as adding external POS information. Other features are likely to improve the error rates as well and should be investigated further, these include: velocity movements, head tracking to measure the tilt of the head (often indicating sub-clauses) or the shift of the upper body (possible indications for direct or indirect speech). Also, a complex entity model can be built up based on the location of the signs. If a new character in the discourse is introduced and stored on the right hand-side of the chest, later deictic pronoun signs pointing to the same position can be interpreted correctly, while pronouns in spoken languages are usually ambiguous.

References

[Chiu et al.2007] Y.-H. Chiu, C.-H. Wu, H.-Y. Su, and C.-J. Cheng. 2007. Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **29**(1):28–39.

[Dreuw et al.2006] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. 2006. Tracking using dynamic programming for appearance-based sign language recognition. In *7th Intl. Conference on Automatic Face and Gesture Recognition*, IEEE, pages 293–298, Southampton, April.

[Dreuw et al.2007] Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. 2007. Speech recognition techniques for a sign language recognition system. In *Interspeech 2007 - Eurospeech*, pages 2513–2516, Antwerp, Belgium, August.

[Hanke2004] T. Hanke. 2004. HamNoSys - Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Workshop on the Representation and Processing of Sign Languages at LREC 04*, pages 1–6, Lisbon, Portugal.

[Lööf et al.2006] J. Lööf, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schluter, and H. Ney. 2006. The 2006 RWTH parliamentary speeches transcription system. In *Ninth ICSLP*, Pittsburgh, Pennsylvania, September.

[Mauser et al.2006] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney. 2006. The RWTH statistical machine translation system for the IWSLT 2006 evaluation. In *IWSLT*, pages 103–110, Kyoto, Japan, November. Best paper award.

[Morrissey and Way2005] S. Morrissey and A. Way. 2005. An Example-based Approach to Translating Sign Language. In *Proceedings of the Workshop in Example-Based Machine Translation (MT Summit X)*, pages 109–116, Phuket, Thailand.

[Morrissey and Way2006] S. Morrissey and A. Way. 2006. Lost in Translation: the Problems of Using

Mainstream MT Evaluation Metrics for Sign Language Translation. In *Proceedings of the 5th SALT-MIL Workshop on Minority Languages at LREC 2006*, pages 91–98, Genoa, Italy.

[Neidle et al.1999] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee. 1999. *The Syntax of American Sign Language*. MIT Press.

[Pizzuto and Pietrandrea2001] E. Pizzuto and P. Pietrandrea. 2001. The notation of signed texts: open questions and indications for further research. *Sign Language and Linguistics (Special Issue - Sign Transcription and Database Storage of Sign Information)*, 4: 1/2:29–43.

[San-Segundo et al.2006] R. San-Segundo, R. Barra, L. F. D'Haro, J. M. Montero, R. Córdoba, and J. Ferreiros. 2006. A Spanish Speech to Sign Language Translation System for assisting deaf-mute people. In *Proceedings of Interspeech 2006*, Pittsburgh, PA.

[Stein et al.2006] D. Stein, J. Bungeroth, and H. Ney. 2006. Morpho-Syntax Based Statistical Methods for Sign Language Translation. In *Proceedings of the 11th Annual conference of the European Association for Machine Translation (EAMT'06)*, pages 169–177, Oslo, Norway.

[Stokoe1960] W. C. Stokoe. 1960. *Sign language structure: an outline of the visual communication system of the American deaf*. Studies in Linguistics, Occasional Paper, 2nd printing 1993: Burtonsville, MD: Linstok Press.

[Sutton1995] V. Sutton. 1995. *Lessons in Sign Writing, Textbook and Workbook (Second Edition)*. The Center for Sutton Movement Writing, Inc.