A New Evaluation Approach for Sign Language Machine Translation

Abdulaziz ALMOHIMEED, Mike WALD and Robert DAMPER
University of Southampton
{aia07r, mw, rid}@ecs.soton.ac.uk

Abstract. This paper proposes a new evaluation approach for sign language machine translation (SLMT). It aims to show a better correlation between its automatically generated scores and human judgements of translation accuracy. To show the correlation, an Arabic Sign Language (ArSL) corpus has been used for the evaluation experiments and the results obtained by various methods.

Keywords. Sign Language, Evaluation Metric, Machine Translation, Accessibility

Introduction

One of the major challenges for building any sign language (SL) translation system is evaluating it. In general, there are two ways to evaluate SL translation output. First, one can use human judgement to assess the translation quality. This method is considered the most reliable way to evaluate any translation system. However, it is expensive and time-consuming. Second, one can use existing automatic evaluation techniques. The problem with these techniques, however, is that they are designed to evaluate natural language (NL), which has different representation. NL representation is linear, while SL is multi-linear. Therefore, we present a new technique that is an extension of the Word Error Rate (WER) technique, which is one of the most widespread evaluation techniques.

1. Background

SL is the native language of the deaf community and, as such, is used as a linguistic medium of communication among deaf persons for expression in the same manner as spoken language among hearing people. SL uses movements of the hands, called manual features (MFs), in conjunction with other parts of the body (such as facial expressions, shoulder movements, head tilts) as a parallel representation (multi-linear representation). Other parts of the body in signs, including facial expressions, are called non-manual features (NMF). MFs are considered essential for use in signs, and NMFs play an important role in expressing signs in conjunction with MFs[1]. NMFs can be classified into three types in terms of their roles. The first is essential: if the NMF is not expressed as part of the sign, the sign will have a completely different meaning. An example of an essential NMF is the sign sentence, “Theft is forbidden,” where closed eyes in the sign for “theft”

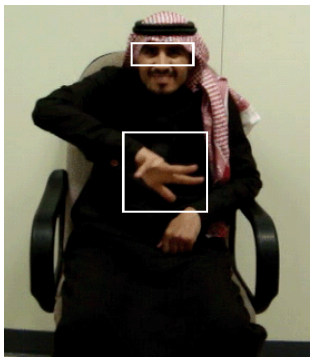

Figure 1. The sign for “theft.” The signer uses the right hand while closing his eyes.

are essential (figure 1). If the signer does not close his or her eyes, the sign will give a completely different meaning; it will mean “lemon.”

The second type of NMF is quality or emotion. In spoken language, inflections, or changes in pitch, can express emotions such as happiness and sadness; likewise in SL, NMFs are used to express emotion. The third type of NMF actually plays no role in the sign. In some cases, NMFs remain from a previous sign and are meaningless. A native viewer naturally discards any meaningless NMFs based on his or her knowledge of SL.

2. Problem definition

Widely used automatic techniques have been designed to evaluate representations in linear sequence. These techniques fail when attempting to measure multi-linear sequences of SL; however, most research in the field of SLMT has evaluated using these techniques and by discarding NMFs and combining MFs as one linear output or by considering the entire sign as one block and combining MFs and NMFs in linear representation[2]. In some cases, when NMFs are discarded, a completely unrealistic evaluation score results. For example, the sign for “theft” would be seen as the sign for “lemon” because “lemon” shares all the manual features of “theft.” In addition, when the sign is treated as one block, the metric score is usually unrealistic, specifically in cases where NMFs are deleted or non-existent NMFs are inserted in signs. Measurements, in these cases, are equivalent to the score of signs with an extra MF or to a sign that is completely different from the original sign and that shares no NMFs or MFs with the original sign. Therefore, an evaluation metric for SL that agrees with human judgements while automatically generating scores is badly needed.

3. Sign Language Error Rate (SiER) technique

A new technique has been designed to extend Word Error Rate (WER) measurement methods of multi-linear sequence representation. The basic idea is simply to assign a weight to each MF and NMF. This weight should fall between 0 and 1, and the total weight for all MFs and NMFs in one sign is equal to 1. The formula for SiER is

$$SiER = 100 \times \frac{Insertion + \sum_{i=1}^{n}(Substitution_i + Deletion_i) \times Weight_i}{\textit{Total Signs in the reference sentence}} \quad (1)$$

where n is the total number of features. The insertion is only for inserted manual features. Factors considered when the formula was designed are the following: (1) sign has at least one MF; (2) useless NMF features are usually inherited from previous signs and do not affect the meaning of the sign–these are naturally discarded by a native signer; (3) the quality of signs; (4) according to the ArSL corpus, an essential NMF, when it exists, exists only for one NMF; and (5) the number of essential NMFs in the corpus compared to the number of quality NMFs are very limited.

First, in computing SiER, the automated procedure determines whether any NMFs or MFs have been substituted or deleted, or whether any MFs have been inserted between the candidate and reference sign sentences. Second, it distributes the weight of each NMF and MF, assigning 0.375 to the right hand (RH) feature, if the RH feature exists, and 0.375 to the left hand (LH) feature, if the LH feature exists; otherwise, it assigns 0.75 to the only existing MF. It then breaks up the 0.25 weight for all NMFs that exist in the reference sign sentence. If no NMF exists, it will add the 0.25 to the MF weight by adding 0.125 to each MF or 0.25 to the sole MF, if only one exists. Third, it adds the difference for each NMF and MF, then multiplies that sum by the NMF and/or MF weight. Finally, it divides the final score by the total number of signs in the reference sentence. The result is then multiplied by 100 to transform the rate into a percentage[1]. In addition, the Sign Recognition Rate (SiRR) can be calculated from SiER by subtracting 100 from SiER.

Figure 2 is an example of a simple alignment between a reference and candidate sign sentence from the ArSL corpus, showing the types of differences between the two.

	Sign1	Sign2
Reference – RH Gloss	THEFT	FORBIDDEN
LH Gloss		
Eyes Gloss	closed	

Candidate – RH Gloss	THEFT	CRIME
LH Gloss		
Eyes Gloss		

Figure 2. A simple alignment between a reference and candidate sign sentence. The gloss notations for ArSL are translated from Arabic to English.

Applying the traditional WER technique to evaluate the candidate sign sentence,

$$WER = \frac{1+1}{2} \times 100 = 100\% \quad (2)$$

where the WER is equal to 100%. When the SiER is applied,

[1] SiER can be more than 100% when the number of MF insertions are high and the candidate has more words number of reference sentence

$$SiER = \frac{0 + ((1 \times 0.25) + (1 \times 1))}{2} \times 100 = 62.5\% \quad (3)$$

where the SiER is equal to 62.5%, which shows a better correlation.

4. Experiment

To evaluate and test SiER, a set of gloss notations was manually created. The set contained 8 groups, each with 5 gloss notations, with the groups in each set created as follows: (1) an extra MF, (2) an extra NMF, (3) a deleted MF, (4) a deleted essential MF, (5) a deleted quality NMF, (6) a substituted MF, (7) a substituted essential NMF, (8) a substituted quality NMF. The test was conducted by two native signers and one interpreter. The interpreter read the gloss notation and mimed it exactly as written in the notation, then received feedback. After that, he put it in context and received feedback (see Table1).

Table 1. Manual evaluation results and correlations between it and WER1, WER2, and SiER.

	Human Judgement	WER1	WER2	SiER
Group 1	6	100%	100%	100%
Group 2	1	100%	0%	0%
Group 3	4	100%	100%	37.5%
Group 4	2-3	100%	0%	25%
Group 5	2	100%	0%	25%
Group 6	5	100%	100%	37.5%
Group 7	3	100%	0%	25%
Group 8	1-2	100%	0%	25%

Table1 shows the average manual evaluation results for each group and the correlations between it and WER1, WER2, and SiER. WER2 is the error rate only for MFs, while WER1 considers both MFs and NMFs as an equal. To show the correlation clearly, WER1, WER2, and SiER were calculated based on a reference that has one sign. The manual evaluation was given a scale from 1 to 6 for each feedback, and the average feedback for each group was provided in the table. Scale 1 means the sign is of high fidelity and fluent; scale 2 is given when the sign shows the correct meaning but the evaluators felt a little confused. Scale 3 is given when the sign demonstrates its main meaning but cannot be fluently fitted into the context. Scale 4 is given when a part of the sign is known. Based on the evaluator's knowledge, the missing part could hardly verify its meaning. Scale 5 is given when a small detail of the sign is known. However, evaluators could not determine the meaning of the missing part. Scale 6 is given in the case of a completely unknown sign.

Regarding the result, for group 1, the evaluators had to distinguish between the extra MF, in cases where it is not at all related to SL, such as picking up a pen from the table while signing with the other hand (this case is discarded naturally by the viewer and does not affect the meaning of the sign), and between the actual movement as a part of the sign. Therefore, the first case was omitted from the results. In group 2, in general, adding NMFs to the sign was naturally discarded by the viewer and did not affect the sign. This has been clearly shown in the evaluators' feedback. For groups 4, 5, 7, and 8, the evaluators were able to determine the meaning when the sign added to a context.

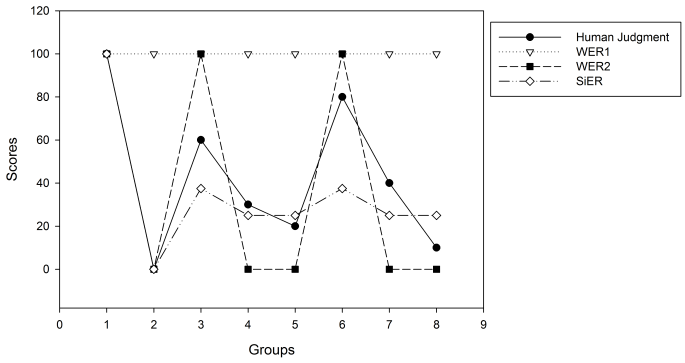

Figure 3. The correlations graph. The human judgement score of 1 is equal to a 0% error rate, 2 is equal to 20%, 3 is equal to 40%, 4 is equal to 60%, 5 is equal to 80%, and 6 is equal to 100%.

Figure3 clearly shows the correlation between WER1, WER2, SiER, and human judgements. In general, SiER scores showed a better correlation with human judgement, especially with respect to NMFs. However, with respect to MFs, WER1 showed a better correlation in some cases. In group 6, WER1 showed (in all five conducted signs) a much better correlation than SiER, which was unrealistic. In group 3, WER1 showed a slightly better correlation than SiER. In some signs in this group, WER1 had a better correlation.

5. Conclusion

A new evaluation approach for SLMT was proposed. It has been designed to extend WER measurement methods for multi-linear representation. It also takes into account that each feature in the representation has a different impact on the evaluation score. The idea behind it is simply to assign a weight to each MF and NMF, considering some facts about SL, such as the insertion of NMF, which has no impact to the meaning of the sign. The experiments show the new approach is promising and, most of the time, more realistic than WER evaluation scores, especially for NMFs. This approach opens the door for further investigations in regards to multi-linear evaluation techniques.

Acknowledgements

This work would not have been done without the hard work of the evaluation team: Mr. Ahmed Alzaharani, Mr. Kalwfah Alshehri and Mr. Abdulhadi Alharbi.

References

[1] Johnston, T. & Schembri, A. (2007). *Australian Sign Language (Auslan): An Introduction to Sign Language Linguistics*. Cambridge University Press.

[2] San-Segundo, R., Pérez, A., Ortiz, D., D'haro, L., Torres, M. and Casacuberta, F. (2007), 'Evaluation of Alternatives on Speech to Sign Language Translation', *Proceedings of the Interspeech07*, pp. 2529–2532.