Sign Language Gloss Translation using Deep Learning Models

Mohamed Amin, Hesahm Hefny, Ammar Mohammed
Department of Computer Science
FGSSR, Cairo University, Egypt

***Abstract*—Converting sign language to a form of natural language is one of the recent areas of the machine learning domain. Many research efforts have focused on categorizing sign language into gesture or facial recognition. However, these efforts ignore the linguistic structure and the context of natural sentences. Traditional translation methods have low translation quality, poor scalability of their underlying models, and are time-consuming. The contribution of this paper is twofold. First, it proposes a deep learning approach for bidirectional translation using GRU and LSTM. In each of the proposed models, Bahdanau and Luong's attention mechanisms are used. Second, the paper experiments proposed models on two sign languages corpora: namely, ASLG-PC12 and Phoenix-2014T. The experiment conducted on 16 models reveals that the proposed model outperforms the other previous work on the same corpus. The results on the ASLG-12 corpus, when translating from text to gloss, reveal that the GRU model with Bahdanau attention gives the best result with ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score 94.37% and BLEU (Bilingual Evaluation Understudy)-4 score 83.98%. When translating from gloss to text, the results also show that the GRU model with Bahdanau attention achieves the best result with ROUGE score 87.31% and BLEU-4 66.59%. On Phoenix-2014T corpus, the results of text to gloss translation show that the GRU model with Bahdanau attention gives the best result in ROUGE with a score of 42.96%, while the GRU model with Luong attention gives the best result in BLEU-4 with 10.53%. When translating from gloss to text, the results report that the GRU model with Luong attention achieves the best result in ROUGE with a score of 45.69% and BLEU-4 with a score of 19.56%.**

***Keywords*—Sequence to sequence model; neural machine translation; sign language; deep learning; LSTM; GRU**

I. Introduction

Sign languages is a visual-gesture based language considered to be the standard language for the deaf. This language operates through gestures and visual channels [1].In sign languages, hand gestures, facial expressions, and body movements are used for communication. According to the World Health Organization[1], around 466 million people worldwide have hearing impairments, out of which 34 million are children. It is estimated that by 2050 over 900 million people will have hearing impairments or difficulties in communication [2].

Also, it is estimated that there are almost 121 types of sign language used worldwide today [3] with less than sufficient number of sign language interpreters to deal with the diversity of sign languages. Hence, there is a need for developing translation systems that make the translation process faster and more accurate. The first step toward automating the translation is to formalize the sign language in standard form. There are existing several forms of sing languages including Stokoe [4], HamNoSys [5], SignWriting [6], and Gloss Notation [7]. Stokoe notation does not include facial expressions and body movements. Thus, this sign language is limited and is not suitable for translation to the deaf. Furthermore, the HamNoSys form is designed to formalize any sign language using 3D animated avatar. However, it does not provide any easy way for describing facial expressions and body movements. The SignWriting notation uses highly iconic symbols,but is difficult to analyze with a computer. Gloss notation [7] on the other hand is a formal sign language that is similar to Braille, finger-spelling, and Morse code. It is used to annotate, represent, and describe sequences of visual-gestural language sequences based on labels on natural language words. This form is a straightforward way that conveys the idea expressed in a natural language, in sign languages. For its simplicity, expressiveness, and formal representation of sign language, glossing has attracted considerable research attention in sign language translation[8], [9], [10], [3].

Several studies have been proposed to translate sign languages to natural languages. Those efforts can be categorized into rule-based [11], [12], example-based [13], [14], [15] and statistical-based approach[8], [9], [10], [3] However, those previous forms are limited in terms of the translation quality and need extra human efforts. For example, the rule-based approach needs domain knowledge of linguistic experts that will be responsible for analyzing the sign language, performing natural language processing tasks, and generating translation rules. Also, natural language processing adds extra complexity as it has many exceptional cases needed to cover using rules. Hence, the number of generated rules is increased. In contrast, example-based machine translation relies on large parallel aligned corpora. It tries to match input sentences with relevant retrieved sentences in a specific corpus. The shortcomings of this translation approach is that it needs massive use-cases to match the input with similar retrieved cases. Also, retrieving similar cases is inefficient and time-consuming [16]. In the statistical approach, translations are generated based on a statistical-based model whose parameters are derived from the analysis of bilingual text corpora. However, this approach needs a large parallel aligned corpus. Moreover, building a corpus with preprocessing tasks is expensive and time-consuming, and it requires collaboration with computer scientists, translators, and linguists. The full process consumes much time. Additionally, the statistical-based approach is tedious to fix mistakes of the translation system, and the precision of translation might become superficial.[17].

[1]https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

In contrast to traditional methods, machine and deep learning have shown great success in several application domains for years [18], [19], [20]. Several researchers have shown interest in the study of machine translation for translating sign languages using a neural network [21], [22], [23], [24], [25]. The recent translation approach based on neural networks is the Neural Machine Translation (NMT) [26], [27] It is an end-to-end learning approach for an automated translation [28]. It consists of two parts: encoder and decoder. To enhance the learning process, an attention mechanism [27] has been lately proposed to allow a neural network to pay attention to only a specific part of an input sentence while generating a translation similar to that of human translations. Although NMT approaches are successful compared to the traditional machine translation approaches, most neural-based studies ignore the sign language's linguistic properties. They assume that there is only a one-to-one mapping of sign-to-spoken words. Additionally, most of the current neural machines focus on the translation from the gloss sign language to the natural language. However, the second direction from natural language to gloss sign language is important to fully automate the translation systems in both directions.

The primary contributions of this paper can be summarized as follows: First, it proposes a sequence-to-sequence deep learning models using LSTM [29] and GRU [30] that translate gloss sign language to natural language text. Second, it introduces a sequence-to-sequence deep learning model that translates natural language text to sign language gloss. In both directions, deep learning models use Bahdanau [27] and Luong [31] attention mechanisms. Third, this paper experiments the proposed models on two different corpora: ASLG-PC12 [32], [33] and Phoenix-2014T [21]. The performance of the results is evaluated using different metrics, e.g., BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores. Also, the best model of the experiments is compared to similar work on the same corpus.

The rest of the paper is organized as follows: Section II presents a brief background on sign languages. Section III discusses several related works. Section IV introduces the proposed approach. Section V discusses the experimental results. Finally, section VI concludes the paper.

II. Background

This section briefly introduces the concept of sign language and machine translation.

A. Sign Language

Sign languages are languages that apply the visual-manual form to convey meaning [34]. The articulators of sign languages are different compared to spoken languages. The primary articulators in spoken languages are the throat, nose, and mouth, whereas the main articulators in sign languages are the fingers, hands, and arms. There are several linguistic features of sign language, and one of those common features is the so-called non-manual feature. The later feature is a parameter of a sign that has meaning. It is not made with hands. but with facial expression, eyebrow movement, movement of the eyes/cheeks, mouth patterns, tilting of the head, movement of the upper body, and shoulder movements. It should be noted that without a non-manual feature, a sign language statement will be meaningless regardless of whether the syntax is in the proper order. Sign language relies on non-manual signals to convey the difference between declarative, imperative, and interrogative sentences.

Furthermore, sign language can be expressed using different ways like Stokoe [4], HamNoSys [5], SignWriting [6], and Gloss Notation [7]. Stokoe, HamNoSys, and SignWriting are iconic representations for a sign language that are hard to read and interpret by deaf people, as translation systems use them to generate 3D animations.

On the contrary, Gloss notation is used to annotate, represent, and describe sequences of signs in a visual-gestural language based on labels-words. It is an interlinear translation used by linguists for transcription. Also learners of sign languages for analysis also use it. The gloss notation is considered an effective way to focus on the grammar and word order, which separates it from the vocabulary. Also gloss notation is written above the natural words using CAPITAL letters. Table I shows pairs of (English, American sign language) sentences.

TABLE I. English and American Sign Language Pairs

English Sentences	ASL Gloss
What is your name?	NAME YOU WHAT^{WH}
He doesn't like pizza.	PIZZA IX-boy DOESN'T-LIKE
Help me.	HELP-ME (one sign)
See you later.	SEE-YOU-LATER (one sign)
Don't know.	DON'T-KNOW (one sign)
Today is Friday, October 28th.	NOW+DAY FRIDAY fs-OCT 28

B. Machine Translation

Early work on machine translation used traditional approaches like rule-based, example-based, and statistical-based. However, these approaches are inefficient in terms of the quality of translation, the limitation of their underlying models, and the exerted efforts of human domain experts.Recently, NMT [26], [27] approach has achieved great progress in machine translation. It is an end-to-end learning approach for automated translation[26].

There are many factors that make NMT performance exceed other traditional approaches [28] First, NMT optimizes all the translation learning parameters simultaneously to automatically decrease network output loss. Second, it has distributed representations with many improvements by sharing statistical strengths among similar words or phrases. Third, it can exploit the context of translations better. The more source and target text, the bigger context that NMT can learn.Thus, NMT is more efficient and has better quality than other approaches.

One of the NMT approaches is a sequence-to-sequence model implemented as a coupled network of encoder and decoder with attention mechanism [27]. In this model, a source sentence $x = \{x_1, x_2, .., x_I\}$ of length I words is given, The model converts this sentence into a target sentence $y = \{y_1, y_2, .., y_J\}$.

The encoder network is responsible for converting source sequences into a list of vectors, one vector per input. whereas the decoder network is responsible for generating one symbol

at a time until the special end-of-sentence symbol. In what follows, we briefly describe the encoder and decoder network.

The encoder network can be encoded as a Recurrent Neural Network (RNN) function. It takes the input x_i and a previous hidden state h_{i-1}, and then generates a current hidden state h_i. Without an attention mechanism, the encoder generates a context vector representing the input sentence. The later context vector is fed to the decoder in the first-time step. However, in the consequent time steps, the decoder forgets the context vector. To remedy the forgotten part, either the context vector is copied to each time step in the decoder or to use an attention mechanism. The later mechanism is better as it focuses on the important part in the input sentence [35].

The decoder network, on the other hand, is represented by a function RNN, The RNN takes an input as the decoder hidden state s_{j-1}, the context vector c_j, and the output of the previous time step y_{j-1}, and then generates the current state s_j. Finally, to generate the output, the hidden states s_j are squashed by a non-linear function g, which is passed to the softmax function to calculate the probabilities.

III. Related Works

Recently, there have been many research efforts to automate sign language translations. Those efforts depend on several types of algorithms and machine translation approaches.

Similar to the work proposed in this paper, several authors used neural machine translation of sign languages. For example, the authors in [21] presented a neural sign Language translation that translates gloss sign language to natural language. In their work, they applied sequence-to-sequence neural model and experimented their results on phoenix-2014T[2] corpus. Their proposed GRU model with Luong attention mechanism achieved BLEU on the range of 1 to 4 grams with scores 44.13%, 31.47%, 23.89%, and 19.26% respectively, and ROUGE score 45.45%.

Another similar work that used sequence-to-sequence model was reported in [23]. The authors proposed to translate gloss sign language into text. They used ASLG-PC12 corpus on several network architectures for their experiments with three different attention functions: dot, general, and concat. The evaluation of BLEU score on the range of 1 to 4 gram achieved are 86.70%, 79.50%, 73.20%, and 65.90% using GRU with dot attention function hidden size 800 units.

Similarly, the authors in [24] proposed a sequence-to-sequence translation model based on human key point estimation. In their work, they build KETI sign language corpus [24], which consists of 14,672 videos of high resolution and quality with the corresponding gloss translation. The corpus was divided into 64% training set, 7% development set, 29% test set. Their model based on a sequence-to-sequence model based on GRU cells achieved an accuracy score of 55.28%, a BLEU score of 52.63%, and a ROUGE score of 63.53 on gloss level.

Furthermore, the authors in [36] proposed sign language transformers: joint end-to-end sign language recognition and translation. They experimented their proposed work on Phoenix-2014T dataset, The evaluation of their proposed model with BLEU scores are 48.9%, 36.88%, 29.45%, 24.54%

Also, the authors in [22] proposed a translation system based on transformers models. They experimented their proposed work on Phoenix-2014T [21] and ASLG-PC12 [32], [33] corpora. The evaluation of their proposed model on Phoenix-2014T achieved BLEU on the range of 1 to 4 grams with scores 48.40%, 36.90%, 29.70% and 24.90% using Transformer on Phoenix-2014T dataset. Moreover, they achieved BLEU scores of 92.88%, 89.22%, 85.95% and 82.87% using Transformer on ASLG-PC12.

Also the author in [37] proposed Sign Language Semantic Translation System using Ontology and Deep Learning. Where CNN trained model used in the recognition process with adding the semantic layer. Collected signs of 10 Arabic gestures and their meanings in English and French sign languages used in training and testing the system.

Despite the success of the previous neural network translation approaches except this paper, most of these approaches, however, focus on one direction-translation, particularly from gloss sign language to natural language.

IV. Proposed Approach

This section shows the proposed approach that translates from natural language text to gloss sign language and vice versa. The proposed approach is divided into two directions. The first direction translates text to gloss notation, while the second direction translates from gloss notation to text. We describe the details of each direction as follows.

A. Text to Gloss Notation Approach

In the text to gloss notation approach, shown in Fig. 1, the input text is fed to the NMT, which translates the text to gloss notation. The NMT consists of two phases, preprocessing and encoding-decoding phase.

In the preprocessing phase natural language processing occurs as Convert natural language text to lowercase and convert gloss notation to uppercase, Stripe whitespaces, and remove numbers and punctuation. Then text is embedded into continuous vector space. The second phase consists of an encoder-decoder neural network model augmented with an attention mechanism that translates the embedded text into gloss notation language. The neural network of the last phase consists of an encoder and decoder. Generally, the encoder transforms a source sentence into a list of vectors, one vector per input symbol. Given this list of vectors, the decoder produces one symbol at a time until the special end-of-sentence symbol (EOS) symbol is produced. The encoder and decoder are connected through the attention model. The attention model allows a neural network to pay attention to only part of an input sentence while generating a translation, similar to the human translator.

B. Gloss to Text Approach

The second direction of the proposed approach is shown in Fig. 2.

[2]https://www-i6.informatik.rwth-aachen.de/ koller/RWTH-Phoenix-2014-T/

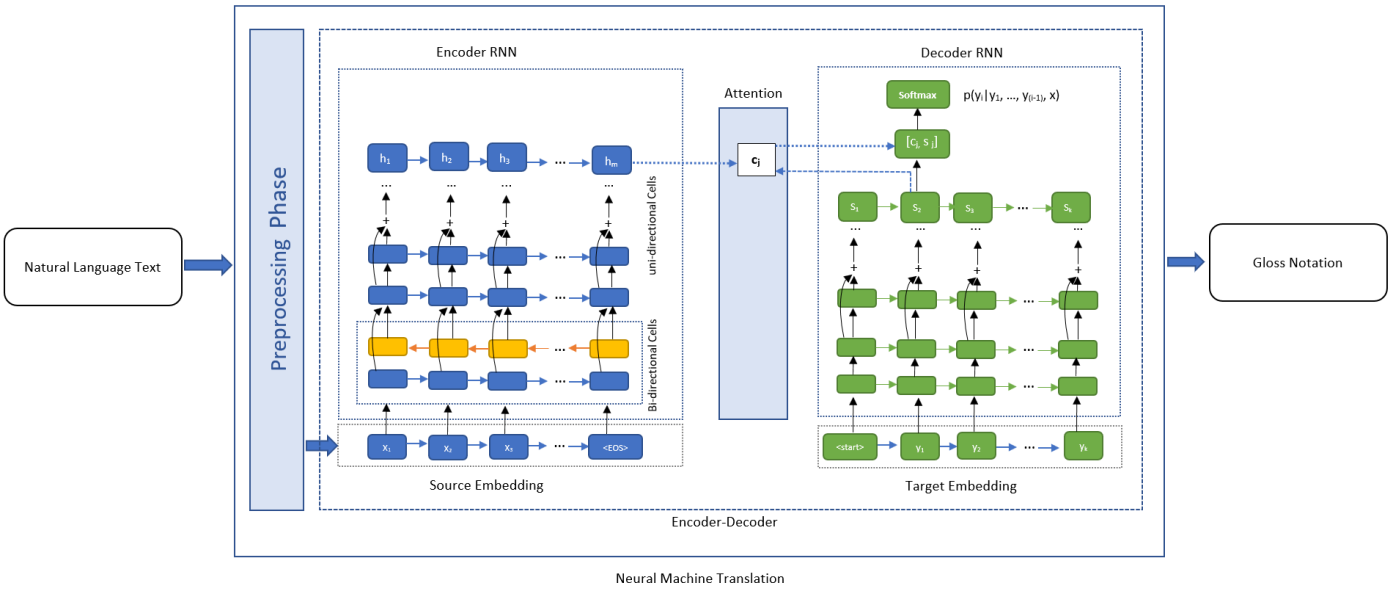

Fig. 1. Natural Language Text to Sign Language Gloss Model.

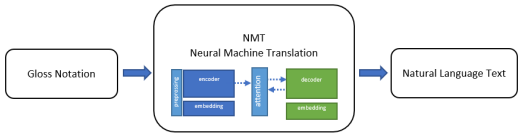

Fig. 2. Sign Language Gloss to Natural Language Text Model.

Here The main task is to translate gloss notation into text. First, the machine translation component receives a gloss notation and performs natural language preprocessing tasks on the gloss notation where the gloss is embedded on a continuous vector space. Second, the embedded gloss is then passed through an encoder-decoder neural network model augmented with an attention mechanism that translates the embedded gloss into text. The architecture of the encoder and decoder is like the one in Fig. 1.

V. Experimental Results

This section shows the experimental results of the proposed approach on two corpora: namely, ASL-PC12 and Phoenix-2014T. We begin by describing the details of each corpus before showing the results. In each corpus, we describe data splitting criteria that are used in the experiments. We described the criteria of each corpus using the following terms: sentence, Running words, vocabulary size, Singletons, and Out of Vocabulary (OOV). Sentences represents number of examples that exist in the corpus. The Running words stands for the number of words in the corpus. Vocabulary size is several tokens that measure how many words a particular model knows. Singletons represents the number of those words that occur only once in the training set. OOV expresses the number of words that occur in test data, but not in training data.

The first corpus, ASLG-PC12, was proposed in [32], [33] as a big parallel corpus between English written texts and American Sign Language Gloss. The ASLG-PC12 is a bilingual corpus of 87,710 sentences. The total number of "running words" is $1,027,100$ for English words and $906,477$ for gloss words in addition to $4,662$ singletons for English words and $6,561$ singletons for gloss words. The vocabulary of both sign gloss annotation and spoken language are $16,788$, and $12,344$, respectively. In the experiments, we split the corpus into $52,626$ sentences for training in the experiments, $17,542$ sentences for validation, and $17,542$ sentences for testing. Table II describes the statistics of the corpus.

TABLE II. Key Statistics of ASLG-PC12

	English			Gloss		
	Train	Dev	Test	Train	Dev	Test
Sentences	52,626	17,542	17,542	52,626	17,542	17,542
Running Words	610,129	207,760	209,211	538,681	183,242	184,554
Vocab Size	16,788	10,121	10,264	12,344	7,470	7,571
Singletons	4,662	-	-	6,561	-	-
OOV	-	2,671	3,027	-	1,949	2,330

The second corpus, Phoenix-2014T, is the German sign language of weather-forecast news. Phoenix-2014T [21] is an extended version of the continuous sign language recognition benchmark dataset found in [38]. It is a gloss annotation, video segments, and spoken language translations matching the sign language. It contains 8257 sequences with 9 different signers. The total running words is $113,717$ for German words and $75,786$ for gloss words. Additionally, it contains 1077 singletons for German words and 337 singletons for gloss words. The vocabulary of both sign gloss annotation and spoken language are 1236 and 2892 respectively. In the experiments, we split the corpus into $7,096$ sentences for training in the experiments, 519 sentences for validation, and 642 sentences for testing. Table III describes the statistics of the corpus.

TABLE III. KEY STATISTICS OF PHOENIX-2014T

	German			Gloss		
	Train	Dev	Test	Train	Dev	Test
Sentences	7,096	519	642	7,096	519	642
Running Words	99.081	6,820	7,816	67,781	3,748	4,257
Vocab. Size	2,892	956	1006	1,236	397	415
Singletons	1077	-	-	337	-	-
OOV	-	57	60	-	19	22

A. Results

The experimental results are reported based on the two previous corpora on 4 types of encoder-decoder architectures with an attention mechanism. For this purpose, we applied two encoder-decoder architectures using GRU and LSTM. Also, we augmented each type of architecture with either Bahdanau or Luong's attention mechanism. Two ways of training from text to gloss and from gloss to text for each combination of the attention mechanism with encoder-decoder architecture were applied. Thus, for both corpora, we totally perform 16 different models in the experiments. The hyper-parameters of the trained models are shown in Table IV.

TABLE IV. HYPERPARAMETERS

	ASLG-PC12	Phoenix-2014T
Number of Layers	1	4
Initial Learning Rate	10^{-4}	10^{-4}
Batch Size	128	128
Hidden units	1024	1024
Embedding units	1024	1024
Dropout	0.30	0.30
Gradient Clipping	5	5

To apply the proposed approach on ASLG-PC12, we created a deep network model with one layer of the encoder (unidirectional) layer, and one layer of the decoder layer. Also, we used GRU and LSTM cell for each type of network. We used an embedding layer of 1024 units with each recurrent layer containing 1024 hidden units of batch size 128. We also used Adam optimization with a learning rate of 10^{-4} as a default parameter and gradient clipping with a threshold of 5 and dropout connections with a drop probability of 0.3. The model was implemented the model using TensorFlow [39] with eager execution and we use evaluation metrics BLEU and ROUGE score. All our networks are trained in 70 epochs. Tables V and VI illustrate the full results of the proposed approach on ASLG-PC12 in two ways of translation, namely from text to gloss and from gloss to text.

TABLE V. ASLG-PC12 TEXT TO GLOSS MODEL RESULTS

	Test				
	Rouge	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**
LSTM B	91.19	89.47	83.93	79.39	75.38
GRU B	**94.37**	**93.26**	**89.64**	**86.68**	**83.98**
LSTM L	88.88	89.98	81.14	74.82	69.55
GRU L	70.42	71.03	59.58	50.79	43.46

The results of the trained text to gloss models reveal that the encoder-decoder model with GRU of Bahdanau (B) attention achieves the best result with ROUGE score 94.37% and BLEU-4 score 83.98% when compared to other models. Also, the trained gloss-to-text models' results reveal that the encoder-decoder model with GRU of Bahdanau attention achieves the best result with ROUGE score 87.31% and BLEU-4 66.59%.

TABLE VI. ASLG-PC12 GLOSS TO TEXT MODEL RESULTS

	Test				
	Rouge	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**
LSTM B	80.59	81.88	70.99	62.76	55.98
GRU B	**87.31**	**88.65**	**79.68**	**73.23**	**66.59**
LSTM L	79.54	69.69	60.75	60.75	53.57
GRU L	62.78	63.90	51.63	42.66	35.52

To compare the results with other related work, Table VII summarizes our best results [*] against the best models in [23] concerning ASLG-PC12 gloss to text translation.

TABLE VII. COMPARISON TEST SCORE ASLG-PC12 FOR GLOSS TO TEXT WITH OTHER WORK

	Rouge	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**
GRU L [23]	-	86.70	79.50	73.20	65.90
GRU B*	**87.31**	**88.65**	**79.68**	**73.23**	**66.59**

In the experiments for the proposed approach on Phoenix-2014T, we created the deep network model with four stacked layers of the encoder (1 bidirectional [40] and 3 unidirectional layers), and 4 stacked layers of the decoder that support residual connections to avoid exploding and vanishing gradient problems [41], [42]. Also, we used two GRU and LSTM cells for each type of network. Each recurrent layer contains 1024 hidden units and 1024 units of an embedding layer with batch size 128. Furthermore, we used Adam's optimizer [43] with a learning rate of 10^{-4} as a default parameter. We are clipped the gradient with a threshold of 5 and dropout connections with a drop probability of 0.3. Likewise, for the models of ASLG-PC12 corpus, we implemented the models using TensorFlow [39] with eager execution. We equally applied BLEU and ROUGE score as the evaluation metric. All models are trained using 70 epochs. Table VIII and IX summarize the results of the proposed approach on Phoenix2014T for the two ways of the translation, i.e., text to gloss and from gloss to text, respectively.

TABLE VIII. PHOENIX-2014T TEXT TO GLOSS MODEL RESULTS

	Test				
	ROUGE	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**
LSTM B	36.54	38.56	21.16	12.06	07.39
GRU B	**42.96**	**43.90**	**26.33**	**16.16**	10.42
LSTM L	40.21	42.60	24.24	15.34	10.55
GRU L	41.14	42.45	25.27	15.90	**10.53**

The results of trained text to gloss models show that the encoder-decoder model with GRU having Bahdanau (B) attention achieves the best result in ROUGE with a score of 42.96%, whereas GRU with Luong (L) attention achieves the best result in BLEU-4 with 10.53%. Also, the results of trained gloss-to-text models reveal that the GRU encoder-decoder model with Luong (L) achieves the best result in ROUGE and BLEU-4 with a score of 45.69% and 19.56%. respectively.

To compare the results with other related work, Table X summarizes our best results against the best models in [21] concerning the gloss to text translation. In the evaluation comparison, we did not consider the text to gloss translation, as the authors of [21] focused only on the translation from gloss to text. Our GRU and LSTM models, marked with (*)

TABLE IX. PHOENIX-2014T GLOSS TO TEXT MODEL RESULTS

	Test				
	ROUGE	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**
LSTM B	44.27	45.02	29.92	22.18	17.77
GRU B	45.45	45.38	31.26	23.34	18.64
LSTM L	44.60	44.47	29.55	21.72	17.38
GRU L	**45.69**	**45.38**	**31.81**	**24.17**	**19.56**

outperform the work of [21] in terms of ROUGE and BLEU evaluation metrics.

TABLE X. COMPARISON TEST SCORE PHOENIX-2014T FOR GLOSS TO TEXT WITH OTHER WORK

	Rouge	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**
LSTM L	41.92	41.22	28.03	20.77	16.58
LSTM L*	**44.60**	**44.47**	**29.55**	**21.72**	**17.38**
GRU L	43.73	43.43	30.73	23.36	18.75
GRU L*	**45.69**	**45.38**	**31.81**	**24.17**	**19.56**
GRU B	42.61	42.76	29.55	22.00	17.40
GRU B*	**45.45**	**45.38**	**31.26**	**23.34**	**18.64**

VI. CONCLUSION

In this paper, we proposed an approach that translates sign language to natural language and vice versa. In particular, we proposed a deep learning approach based on sequence to sequence for bidirectional translation, from gloss notation to text and text to gloss for both directions of translation. We used encoder-decoder with attention to Bahndanau and Luong mechanism. In particular, two models of encoder-decoder network with GRU and LSTM were adopted. We have tested the proposed approach on both ASLG-PC12 and Phoenix-2014T corpora. We conducted four models of encoder-decoder with different attention mechanisms per each translation direction for the two corpora. We compared the results of the four models in each direction of translation. The overall experimental results on eight different models applied to the ASLG-PC12 corpus indicated that the GRU model with Bahdanau attention achieved the best performance using the ROUGE metric with an 87.31% score translating from gloss to text. Also, the GRU model with Bahdanau attention achieved the best performance with a ROUGE score of 94.37% when translating from text to gloss. Similarly, the overall experimental results on eight different models applied to the Phoenix-2014T corpus revealed that the GRU model with Luong attention achieved the best performance on ROUGE with a score of 45.69% when translating from gloss to text. In the other direction of translation, the GRU model with Bahdanau achieved the best performance on ROUGE with a score of 42.96%. Moreover, part of the results were compared to similar work on the same corpus in one direction of translation and showed the superiority of the proposed models. We think that one big enhancement of sign language translations is to use the so-called pose estimation[44], [45], [46]. In particular, the translation from text to pose estimation and vice versa is worth investigating as a future research direction.

REFERENCES

[1] A. Othman and M. Jemni, "An xml-gloss annotation system for sign language processing," in *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*. IEEE, 2017, pp. 1–7.

[2] B. O. Olusanya and V. E. Newton, "Global burden of childhood hearing impairment and disease control priorities for developing countries," *The Lancet*, vol. 369, no. 9569, pp. 1314–1317, 2007.

[3] A. Othman and M. Jemni, "Statistical sign language machine translation: from english written text to american sign language gloss," *arXiv preprint arXiv:1112.0168*, 2011.

[4] W. C. Stokoe Jr, "Sign language structure: An outline of the visual communication systems of the american deaf," *Journal of deaf studies and deaf education*, vol. 10, no. 1, pp. 3–37, 2005.

[5] T. Hanke, "Hamnosys-representing sign language data in language resources and language processing contexts," in *LREC*, vol. 4, 2004, pp. 1–6.

[6] M. Stuart, "A grammar of signwriting," Ph.D. dissertation, Thesis in Linguistics, University of North Dakota, 2011.

[7] E. S. Klima and U. Bellugi, *The signs of language*. Harvard University Press, 1979.

[8] J. Bungeroth and H. Ney, "Statistical sign language translation," in *Workshop on representation and processing of sign languages, LREC*, vol. 4. Citeseer, 2004, pp. 105–108.

[9] V. López-Ludeña, R. San-Segundo, J. M. Montero, R. Córdoba, J. Ferreiros, and J. M. Pardo, "Automatic categorization for improving spanish into spanish sign language machine translation," *Computer Speech & Language*, vol. 26, no. 3, pp. 149–167, 2012.

[10] R. San-Segundo, R. Barra, R. Córdoba, L. F. D'Haro, F. Fernández, J. Ferreiros, J. M. Lucas, J. Macías-Guarasa, J. M. Montero, and J. M. Pardo, "Speech to sign language translation system for spanish," *Speech Communication*, vol. 50, no. 11-12, pp. 1009–1020, 2008.

[11] R. San-Segundo, R. Barra, L. D'Haro, J. M. Montero, R. Córdoba, and J. Ferreiros, "A spanish speech to sign language translation system for assisting deaf-mute people," in *Ninth International Conference on Spoken Language Processing*, 2006.

[12] I. Marshall and É. Sáfár, "A prototype text to british sign language (bsl) translation system," in *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, 2003, pp. 113–116.

[13] S. Morrissey and A. Way, "Lost in translation: the problems of using mainstream mt evaluation metrics for sign language translation," 2006.

[14] ——, "Joining hands: Developing a sign language machine translation system with and for the deaf community," 2007.

[15] S. Morrissey, "Data-driven machine translation for sign languages," Ph.D. dissertation, Dublin City University, 2008.

[16] P. Antony, "Machine translation approaches and survey for indian languages," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 1, March 2013*, 2013.

[17] C. K. Alexandris, *Issues in the Multilingual Information Processing of Spoken Political and Journalistic Texts*. Cambridge Scholars Publishing, 2020.

[18] A. Mohammed and R. Kora, "Deep learning approaches for arabic sentiment analysis," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 1–12, 2019.

[19] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and trends in signal processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[20] S. A. Abdelaziz Ismael, A. Mohammed, and H. Hefny, "An enhanced deep learning approach for brain cancer mri images classification using residual networks," *Artificial Intelligence in Medicine*, vol. 102, p. 101779, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365719306177

[21] N. Cihan Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7784–7793.

[22] K. Yin and J. Read, "Attention is all you sign: Sign language translation with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshop on Sign Language Recognition, Translation and Production (SLRTP)*, 2020.

[23] N. Arvanitis, C. Constantinopoulos, and D. Kosmopoulos, "Translation of sign language glosses to text using sequence-to-sequence attention models," in *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2019, pp. 296–302.

[24] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho, "Neural sign language translation based on human keypoint estimation," *Applied Sciences*, vol. 9, no. 13, p. 2683, 2019.

[25] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2sign: towards sign language production using neural machine translation and generative adversarial networks," *International Journal of Computer Vision*, pp. 1–18, 2020.

[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[28] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[31] T. Luong, E. Brevdo, and R. Zhao, "Neural machine translation (seq2seq) tutorial. 2017," *URL: https://www. tensorflow. org/tutorials/seq2seq*, 2017.

[32] A. Othman and M. Jemni, "English-asl gloss parallel corpus 2012: Aslg-pc12," in *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC*, 2012.

[33] A. Othman, Z. Tmar, and M. Jemni, "Toward developing a very big sign language parallel corpus," in *International Conference on Computers for Handicapped Persons*. Springer, 2012, pp. 192–199.

[34] T. Supalla, "The classifier system in american sign language," *Noun classes and categorization*, vol. 7, pp. 181–214, 1986.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[36] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 023–10 033.

[37] E. K. Elsayed and D. R. Fathy, "Sign language semantic translation system using ontology and deep learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, 2020.

[38] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney, "Extensions of the sign language recognition and translation corpus rwth-phoenix-weather." in *LREC*, 2014, pp. 1911–1916.

[39] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[41] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[42] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem. corr abs/1211.5063 (2012)," *arXiv preprint arXiv:1211.5063*, 2012.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[45] S. Gattupalli, A. Ghaderi, and V. Athitsos, "Evaluation of deep learning based pose estimation for sign language recognition," in *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2016, pp. 1–7.

[46] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," *arXiv preprint arXiv:1502.06807*, 2015.