Journal of Ambient Intelligence and Humanized Computing (2023) 14:9807–9821
https://doi.org/10.1007/s12652-021-03640-9

ORIGINAL RESEARCH

Sentence2SignGesture: a hybrid neural machine translation network for sign language video generation

B. Natarajan[1] · **R. Elakkiya**[1] · **Moturi Leela Prasad**[1]

Received: 13 July 2021 / Accepted: 1 December 2021 / Published online: 26 January 2022

Abstract
The development of Neural Machine Translation (NMT) systems has attained prominent position in language translation tasks. However, it faces huge challenges in translating the new words and out-of-vocabularies. This problem is identified as a major drawback of conventional NMT systems in language translation results more copied outputs. In addition to that, it places the risks in understanding multilingual language structures and word relationships. In this paper, we propose novel deep stacked GRU algorithm based NMT System to address the aforementioned challenges and handles multilingual sentences based translation tasks efficiently. We aimed to develop the proposed model for translating the spoken sentences into sign words. The generated sign words (glosses) are mapped with sign gesture images to automate the sign gesture video generation process using deep generative models. The proposed hybrid NMT model has been evaluated qualitatively and quantitatively using different benchmark sign language datasets. The improved BLEU Score shows the outperformance of our model compared with earlier approaches. We also evaluated the proposed model using our self created Indian sign language corpus (ISL-CSLTR). The final result shows the achievement of greater translation results with minimal processing cost.

Keywords Sign language · Recognition · Indian sign language · Machine translation · Gated recurrent unit · Deep learning · Video generation

1 Introduction

In machine translation, translation of one language sentences into another produces impressive results. This benefits the common public to understand the different language inputs. Sign languages are highly differ from traditional languages due to visual and gesticulated nature. In order to translate the spoken sentences into sign gestures, the model requires intelligent mechanism to understand the different articulation styles. Sign languages are expressive and compose manual and non-manual cues in a single pose. The translation of continuous spoken sentences into sign gestures demands higher attention and places huge risks in development. The model development should consider the various features like hand poses, hand orientation, hand velocity, eye expressions, body poses and mouth movements to produce plausible sign gesture output. From which it summarizes the sign language communication uses various upper body components of humans (Ong et al. 2005). Earlier approaches only focus the native languages based translation tasks becomes degenerative to our goal.

To address all these aforementioned challenges we propose the novel approaches in this paper. Understanding and communicating through the sign gestures pose huge difficulties in real life creates gap between normal people and deaf-mute. The translation of traditional language sentences into sign gestures requires a lot of supervision and attention. The research studies on development of sign language based software assistive systems produces plausible results at word level. The sentence level continuous recognition of spoken sentences and translating them into sign gesture based videos adds higher complexities to the researchers in the area of computer vision and deep learning. The recent advancements in recurrent neural networks (RNN) and its variants produce intelligible results in sequential text processing

✉ R. Elakkiya
elakkiyaceg@gmail.com

B. Natarajan
rec.natarajan@gmail.com

Moturi Leela Prasad
121003172@sastra.ac.in

[1] School of Computing, SASTRA Deemed to be University, Thanjavur, Tamilnadu 613401, India

tasks. The continuous sign language based communications incurs greater efforts in learning sign gestures. Due to this intrinsic complexity nature, real-time development of such systems still growing stage and mostly handles word level isolated sign gestures.

The emergence of RNN networks and its variants has achieved extraordinary results in handling sequential data and time series analysis. The two major RNN versions are Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). The recent advancements in RNN networks have transformed the effectiveness of RNN models into new dimensions. These models have produced fruitful results in Text Generation, Speech to Text Translation, Natural Language Processing applications, Weather forecasting and Sentiment Analysis. However, the vanilla RNN struggles to handle long term dependencies. The advent of LSTM and GRU models address this challenge and explore the solutions to vanishing grading problems. Both of these networks perform equally in sequential applications and performance slightly differs in other domain applications. The language translation tasks uses sequence of words in order to remember the sequences and translate it into matching target sequences involves sequential text processing tasks. This can be achieved with the help of RNN networks. The RNN model has been used powerfully in NMT applications to translate different benchmark datasets. Most of the research studies cover the bilingual corpus like English to French, Tamil to English, and Malayalam to Hindi datasets. We consider English and German languages sentences taken from benchmark datasets to identify its equivalent sign words and sign gesture images.

The research studies on conventional NMT systems mainly focus the translation performance on native language to other language conversions. Although a lot of advances made in earlier research works, still demands better model to produce sustainable improvements. The conventional NMT system learns the language structures of source and target language sentences and yields the translation results. The introduction of earlier models like Statistical Machine Translation Models(Chiang et al. 2005; Koehn et al. 2005; Carpuat et al. 2007) phrase-based models(Koehn et al. 2003; Simard et al. 2007; Utiyama et al. 2007), and syntax oriented models (Wang et al.2007; Pust et al.2015), produces reasonable results in translation but has limited performance issues. The advent of sequence-based translation models (Sutskever et al. 2014; Neubig et al. 2017) attained acceptable results still performance bottlenecks in accurate translation. To address this challenge, the author (Bahdanauet al. 2014), introduced attention mechanism that improves the model translation performance to newer heights. Similar to these approaches, the researchers introduced attention based deep neural models (Bengio et al. 2003; Wołk et al. 2015; Tu et al.2016),Google NMT (GNMT) systems proposed by the author Wu et al. (2016). Although these models produces quantitative and qualitative results in language translation tasks, but still requires improvements in handling unknown words. Because of these shortcomings, the copying penalty rate (Liu et al. 2021) will be high and lacks the intelligence of model development. Based on the research studies it is evident that the common issues persist in earlier works are learning the language structures of source and target, ordering the tokens, processing the new words and identifying the sequence relationship between the lexicons. In practice, translation of language sentences from one grammatical structure to another encounters huge challenges, it equally needs a lot of supervision over the translation tasks by considering the different components of language sentences as depicted in Fig. 1. Each token of the sentences needs to be investigated and a matching target element should be chosen.

The deep neural networks handle such huge computationally intensive, sequences based tasks using various techniques. The RNN networks and their variants LSTM and GRU handle sequence learning-related tasks efficiently and achieve impressive results. In addition to that, the extraction of sign glosses from spoken sentences requires high expertise and powerful models. The GRU-based models evident the solution for these kinds of expert tasks. Experiments on the benchmark dataset show that the improved performance of the proposed hybrid NMT model compared with baseline models.

In this paper, we aimed to develop hybrid NMT systems for the translation of spoken text into sign glosses which helps to automate the sign gesture video generation process much easier. Sign glosses denote the class identification of sign gestures that aids the sign gesture video production for the spoken sentences. The proposed model assists deaf-mute

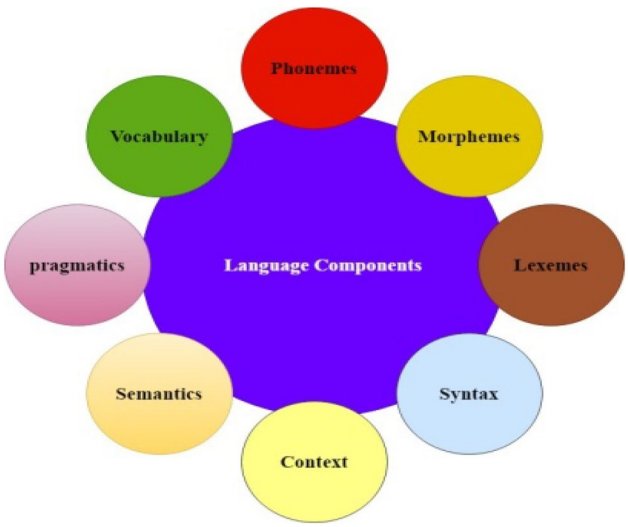

Fig. 1 The various components of languages

people to convey their needs into sign language actions. The development of such models poses potential research challenges in mapping language sentences with sign action sequences. We proposed the deep-stacked GRU algorithm for translating the spoken sentences into sign glosses. The proposed model produces comparatively sustainable amount of improvements in translation results compared to earlier approaches. It has already been proven that GRU produces better results for various sequence learning and prediction problems. However, there is a great demand for improving the performance of GRU for sequential learning tasks.

The proceeding sections of the paper are organized as follows. The Sect. 2 deals with related works which are carried out in previous research studies. The Sect. 3 elaborates the development of proposed systems and its technical details. The Sect. 4 compares the performance of our model using different datasets and quality metrics. Finally the conclusion part summarizes the merits of the entire work and points the future research directions.

2 Related works

In NMT Systems, the translation capability of the model is encountered as a major component to produce intellectual results. Further investigations of NMT systems directed the researchers to achieve reasonable results in image captioning, Healthcare, E-commerce, Government Services, tourisms and services as a tool for communication. The development of NMT models encompasses variety of techniques to produce plausible outputs. The translation process considers the various modalities of a language structures which includes grammatical rules, articles, verb, pronouns, idioms, adjectives and context meaning of the word or sentences. Understanding the source and target language structures and mapping the matching words accommodates high level expert skills and training process. Such systems can aid deaf and dumb people to get education by automating the spoken sentences into sign gestured videos, online ticket booking, and food order services. From the model development point of view, one observable benefit of dealing with the hybrid NMT system is the ability to automate the communication process among various language-based people to understand the other languages in a better way. These systems also handle large vocabularies and relevant field of domain knowledge can be expertise to the next level.

The research studies about the earlier developments of NMT systems helps to analyze the drawbacks of the past models. The variants of NMT systems are shown in Fig. 2. The primary version of NMT systems are introduced by the authors (Sutskever et al. 2014; Bahdanau et al. 2014) handles the sequential text inputs and generates the better results. However, these versions lose the capability to handle longer text sequences and discard the parallelization ideas. The Statistical Machine translation models are introduced by the researchers (Koehn et al. 2005; Carpuat et al. 2007; Utiyama et al. 2007; Koehn et al. 2007; Bengio et al. 2003;Chiang et al. 2005), Phrase-Based Machine Translation (PBMT) (Wang et al. 2007; Pust et al. 2015), and Syntax-Based Machine Translation (SBMT) techniques (Xiao et al. 2012;Pust et al. 2015) are developed and produce concrete results in language translation tasks. The performances of these versions are downgraded due to the model capability to handle large scale datasets and lack of recognizing power in case of handling the new words.

In addition to that, the author Camgoz et al. (2018) proposed the idea reverse of our work by fetching the text information form sign gesture images using spatial details and word embedding techniques. The author has developed a neural sign translation model for generating spoken language sentences from sign videos. This model analyses the spatial representation and employs embedding techniques. The results of the model inherit word ordering problems for generated sign glosses. Graves et al. (2006) discussed Connectionist Temporal Classification (CTC) Loss for calculating errors by considering source and target sequences alignment. CTC achieved greater performance in computer vision applications and speech recognition (Graves et al. 2013; Amodei et al. 2016; Camgoz et al. 2018) tasks. Due to the source and target sentences are similar order constraints and inherent complexity, while training a neural network, CTC failed to perform well in machine translation tasks which led to the development NMT system Graves et al. (2006). Initially, the traditional encoder-decoder system using RNN is proposed by Sutskever et al. (2014)later revised by Bahdanauet al.

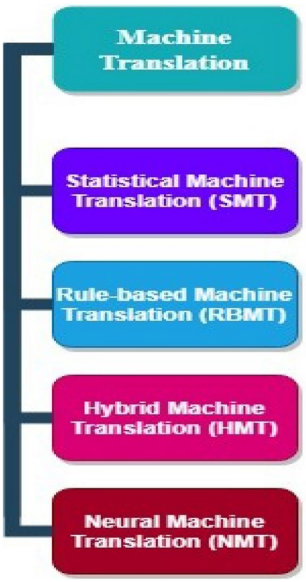

Fig. 2 The variants of NMT system

(2014)using separate RNN for encoders and decoders. Sennrich et al. (2015) introduced the new approaches for attention mechanism and handling inputs and GNMT model (Wu et al. 2016) produces significant improvements in translation performance over benchmark datasets.

The neural network-based Machine Translation models step away from statistical phrase-based approaches due the lot of drawbacks imposed by such models (Guo et al. 2018). The conventional NMT models are predicting the target sequences based on the constraints imposed by the attention mechanisms and probability values yield by context vectors. The RNN based encoder-decoder architecture primarily done this kind of transliteration task. The attention mechanism compensates the problem of encoding longer sequence inputs into a single vector (Ko et al. 2019) directs the decoder to consider the various parts of source inputs for making perfect translation results. The NMT model (Sennrich et al. 2015) addresses the challenges of the open-vocabulary problem by translating the new words as subunits. Subunit components capture all the modalities of tokens and find the equivalent target components. The encoder encodes these rare and unknown words as subunits. The sub-word level segmentation techniques and byte pair encoding techniques are applied to enhance the model performance in English to German language translation. The Coverage Models (Tuet al. 2016) are employed to address the under translation and over translation issues and fully focus all the words present in a sentence (Guo et al. 2018). From a modelling perspective, one vibrant advantage of dealing with the NMT framework is that the translation performance needs to be accurate and there is an associated domain knowledge that can be exploited. More specifically, Conventional approaches lack robustness. More formally, the task here is to improve the translation quality given (i) the large sentences corpus, (ii) the new unseen words, and (iii) the source and target work pairs. Current approaches (Stoll et al. 2020) typically use RNN (Carpuat et al. 2007) for machine translation tasks.

The encodings of the source sentences generate context vectors and fed them to the decoder to generate the target sentences. However, these methods do not exploit the exactness of target sentences and human evaluation results. Such systemsare improved the performance using attention mechanism (Konstantinidis et al.2018), as demonstrated by recent works on syntax-aware neural machine translation(Wang et al.2018; Stoll et al.2018, 2020). Handling rare words and copying words or word pieces models are developed by Carpuat et al. (2007) improves the performance to the next level which is used with our proposed NNMT Systems. Vaswani et al (2018) discussed stacked self-attention layers equipped the Transformer model to improve the translation performance and training speed, but this model incurs a high cost for memory usage. Sennrich et al. (2015) discussed Byte Pair Encoding (BPE) techniques for NMT using rare words and sub-word unit approaches which deal out of vocabulary words problem. The NMT architecture converts the source language sentences into fixed-size vectors in an encoder phase of translation, the decoder based on the values of vector it generates the target sentences, in which both sentences are mapped earlier. The authors (Sutskever et al. 2014; Bahdanauet al. 2014) introduced RNN based sequence-to-sequence models that achieved greater results in machine translation, later it was refined by LSTM Techniques (Ko et al. 2019). Although, the models (Koller et al. 2016) and (Wang et al. 2017) produce good translation results, fail to perform well in longer sequence sentences. The emergence of attention mechanisms has revolutionized the performance of NMT systems. To address the difficulties in handling long-range sequences, the first attention mechanism proposed by Bahdanau et al. (2014) provides the solution. The attention mechanism applied on neural networks concentrates specific portions of input words and ignores the other things.

Kalchbrenner et al. (2013) introduced an attention mechanism to resolve this problem of long-term dependencies. The conventional NMT systems neglect the underlying rich grammatical and linguistic structures of the languages and perform the translation directly. The relationship between sign glosses and spoken language sentences are mapped clearly as word pairs in the dataset. Luong et al. (Luong et al. 2015) developed an attention mechanism for the NMT system. The proposed attention model achieves significant improvement in BLEU Metric over the earlier re-rank approaches. This model employs local attention and global attention schemes at various levels of source sentence encoding and achieve improved results in the WMT'15 English-German dataset. This model addresses the limitations posed by (Bahdanau et al. 2014; Luong et al. 2015; Xu et al. 2015) attention mechanism and selective attention mechanism. The NMT model has been revised into many versions for providing good translation results. The attention mechanism embedded with NMT systems produces comparable growth. The development of such models and adding the additional components for handling new words and preserving target language word order relationships enhance the model performance to the next level. Developing such models creates greater attention in research studies.

The attention mechanism computes the attention vector values by considering context vector values and softmax functions. The attention vector values represent the weighted average values of the encoder summary. The attention mechanism has the capability to go forward and backward to trace the input values. The activation function relu gives the final output. Instead of one hot vector encoding and unique integer encodings, we have used word embedding techniques to encode the input sentences. The word embedding techniques

Table 1 Existing NMT models—comparison

Author	Modelling	Dataset	BLEU Score	Beam size	No of Hidden layers
Sutskever et al. (2014)	LSTM	WMT'14 English-French dataset	36.5	2	NA
Cho et al. (2014a, b)	grConv		32.77/dev and 35.63/test	10	2000
Cho et al. (2014a, b)	RNNEncoder–decoder		30.64/dev and 33.3/test set	NA	2000
Bahdanau et al. (2014)	BiRNN		33.30	NA	1000
Luong et al. (2015)	LSTM	WMT'14 English-German dataset	25.9	NA	NA
Tu et al. (2016)	Coverage based NMT	NIST	30.14	NA	1000
Wu et al. (2016)	GNMT	WMT'14 English-French dataset	38.95	3	1024
Camgoz et al. (2018)	CNN + RNN	RWTH PHOENIX-Weather 2014 T	19.26	3	1000

are employed to convert the source sentence words into real-valued vectors. These vectors relate the relationship between the words and avoid the sparse nature. The word embedding technique uses dense representation assigns floating-point values to the words and creates a fine-grained relationship between words. These techniques assist the enhancement of output translation quality without losing the word-order relationships.

The new advancements introduced in development of NMT models have been investigated in this work and valuable techniques present in such system are incorporated with the development of proposed hybrid NMT model. The following Table 1 depicts the comparison of various existing techniques involved in the development of NMT systems. This greatly influences the new advancements need to be embedded with the proposed model. Based on the studies about NMT model implementation at various levels, it is a keen observation that the need for the development of NMT models to handle the benchmark dataset consists of language sentences. The proposed model needs to address the various challenges present in the existing approaches. The proposed end-to-end framework handles these challenges and produces greater results on the benchmark datasets. We have evaluated the model performance using three sign language datasets. The proposed model produces significant improvement and faster execution results in translation of spoken sentences to sign glosses which aids the identification of similar sign poses for generating the sign videos. Transformer-based NMT models introduced by Vaswani et al. (2018) are developed mainly for implementing multi-head attention mechanisms using multi-layers. Although, these models eliminate the problem that arises in handling fixed size vectors in sequence models (RNN, LSTM), scaling factors increases as the size of input sentences increases. The multi-head attention layers are stacked in encoder-decoder phases enriches the model's performance to the next level. The encoder-decoder part of the model comprises a totally 6 layers includes a multi-head attention layer, fully connected layers, and normalization layers. Since these models produce impressive results, training the model requires high-end GPU devices as well as large memory requirements. GNMT systems (Wu et al. 2016) are modelled to address the challenges that persist in conventional NMT Models. GNMT model comprises an encoder, decoder, and attention networks. These models introduce residual and attention connections between the decoder to the encoder to speed up the training process.

This model addresses the challenges in handling rare words by incorporating the word piece model (sub word unit) (Sennrich et al. 2015) techniques. The coverage penalty method greatly reduces the error rate by covering all parts of source. The attention-based NMT models proposed by Luong et al. (2015), uses input feeding approaches to reduce coverage effects present in general attention models. These models analyses the source sentences using global and local attention mechanisms. The global attention mechanism handles all the portions of source sentences directly and local attention handles the only subset of words. The word alignment techniques (monotonic and predictive) are effectively implemented by these models. This approach incorporated with our work for identifying the monotonic relationship between words using attention-based NMT Techniques. The Table 1 compares the various parameters of existing NMT implementations to explore the detailed analysis in various dimensions.

Based on detailed studies, we observed some pitfalls in earlier NMT developments. We summarize our main contributions for improving the NMT performance are listed as follows:

- The propose NMT systems handle multilingual datasets using deep neural networks and produces accurate translation results without using any supervision tasks.
- The Deep Stacked GRU algorithm with attention mechanism improves the quality of the machine translation per-

formance and achieves noticeable growth during training and validation.

- The proposed model will be evaluated qualitatively and quantitatively using various benchmark datasets and quality metrics, shows the quality of our work and reduces the translation errors drastically.

The proposed model is aimed to explore the advancements in conventional NMT architecture for achieving improved performance. The model translates the spoken sentences into sign glosses. The proposed model is developed using deep neural architectures and process the large-scale sentence-level multi-lingual sign language datasets.

3 The proposed system

The main challenges for developing NMT models are handling longer sentences, recognizing the new words and producing good translation quality. To address these requirements, we propose novel Deep Stacked GRU Algorithm based NMT model. Here, we discuss the underlying frameworks of various NMT systems (Sutskever et al. 2014; Bahdanauet al. 2014; Sennrich et al. 2015; Vaswani et al. 2018; Luong et al. 2015) upon which we build our novel proposed model to achieve greater translation results.

3.1 RNNencdecModel

The primary version of NMT model is RNN based encoder-decoder model introduced by Sutskever et al. (2014) uses the reverse order of the source sentences and Multi-layered LSTM Technique. This model has secured improved BLEU scores when compared with Phrase Based NMT systems. The main objectives of these models are translation of source to target sentences for bilingual corpora. The basic steps are summarized as follows.

The set of input vectors denoted as (x_1,x_2,x_3..x_n),hidden states are represented as (h_1,h_2,h_3…h_t). The source sentences are converted into vectors and processed by hidden layers. At each time step, the hidden layers remembers the past hidden layer information. The final hidden layer of encoder unit summarizes the source sentences. The decoder unit perform the translation based on encoder summary and context vectors. The probability based LSTM function and softmax computation are described in following Eq. (1), where h_t represents hidden state, h_{t-1} denotes the previous hidden state values. The LSTM network is used for encoder and decoder translation tasks.

$$p\left(\left(y_1, y_2, \ldots y_{K'} | x_1, x_2 \ldots x_K\right)\right) = \prod_{k=1}^{K'} p\left(\left(y_k | v, y_1, y_2, \ldots y_{k-1}\right)\right) \tag{1}$$

The prediction of output sequences follows layered LSTM approaches with conditional probability functions. The final target sequences are represented as (y_1,y_2,y_3…y_k). We incorporate this basic approach in our proposed model.

3.2 RNNSearchModel

RNNSearch model (Bahdanauet al. 2014) is designed to tune the alignment process and maximizes the quality of translation results. This model considers the important word instead of keeping entire sequence information that results improved performance than earlier approaches (Sutskever et al. 2014; Cho et al. 2014a, b). The basic RNN equations represented as Eq. (2)

$$h_m = f\left(x_m, h_{m-1}\right) \tag{2}$$

where h_t denotes hidden states and x_t refer the input sequences. The context vector (cv) values equivalent to words of sentences in Eq. (3) is computed as

$$cv = g\left(\left\{h_1, h_2, h_3 \ldots h_{Mx}\right\}\right) \tag{3}$$

where hidden state represents $h_m \in \mathbb{R}^n$ at time m. The non-linear functions f and g are applied to compute embedding vector values at each step of the hidden state. The decoder predicts the next words by applying the joint probability function on previously generated output words in Eq. (4).

$$p(y) = \prod_{m=1}^{M} P\left(y_m | \{y_1, y_2, y_3, \ldots . y_{m-1}\}, cv\right) \tag{4}$$

The RNN constructs the conditional probability model in Eq. (5) as follows

$$P\left(y_m | \{y_1, y_2, y_3, \ldots . y_{m-1}\}, cv\right) = g\left(y_{m-1}, hs_m, cv\right) \tag{5}$$

where the non-linear function g uses deep neural networks and produces the output y_m and hs denotes the hidden states.

3.3 Bahdanau's attention mechanism

The conventional sequence-to-sequence models based on encoder-decoder architecture perform well on short sentences but fails in case of long sequences. To address this challenge, attention mechanism approach is proposed by Bahdanau et al. (2014). This attention mechanism considers

only the important portion of the source sentence sequences using which predicts the possible target sequences. It employs the automated soft search techniques and produces comparable performance over earlier approaches (Sutskever et al. 2014). This attention mechanism was developed in bidirectional RNN encoder-decoder architecture, which performs searching while the decoding process takes place. The conditional probability function of the model is defined in Eq. (6)

$$P(y_i|y_1, y_2, \ldots, y_{i-1}, X) = g(y_{i-1}, s_i, c_i) \tag{6}$$

where s_i represents RNN hidden state at time i. S_i is computed using the Eq. (7) represented as follows.

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \tag{7}$$

The distinct context vector c_i values are conditioned using probabilistic functions to map with the target word $y_{i..}$ This approach uses Bi-directional RNN techniques to estimate the forward and backward hidden unit sequences. The context vector uses the annotations $(h_1, h_2, \ldots, h_{T_x})$ to facilitate the mapping of encoder with input sequences. This mapping helps to articulate the position of the important word in a sentence. The annotation of sequences is performed by concatenating the forward pass ($\vec{h}_j^{\,T}$) and backward pass ($\overleftarrow{h}_j^{\,T}$) as $\left[\vec{h}_j^{T}; \overleftarrow{h}_j^{T}\right]$. The alignment model and annotation sequences are further used for computing context vectors. We use Bahdanau's Attention Mechanism in our proposed work for improving accuracy of the proposed hybrid NMT system.

3.4 Handling rare words with subword units

The researcher Sennrich et al. (2015) proposed Byet-pair encoding compression algorithm based NMT systems for handling out-of-vocabulary (OOV) words and rare words. This approaches uses sub-word units idea for segmenting the character pairs. This model highly focuses the morphological changes and transliteration to improve the translation performance. However, it creates spurious ambiguity while handling known words. Other works Unigram Language Model (Kudo 2018), Sub-word Sampling (Provilkov et al. 2019) and BPE-dropout (Kudo et al. 2018) also handles the rare words in different perspectives. We incorporate the sentence piece approach (Kudo et al. 2018) in our model design to provide better translation results for oov words and rare words.

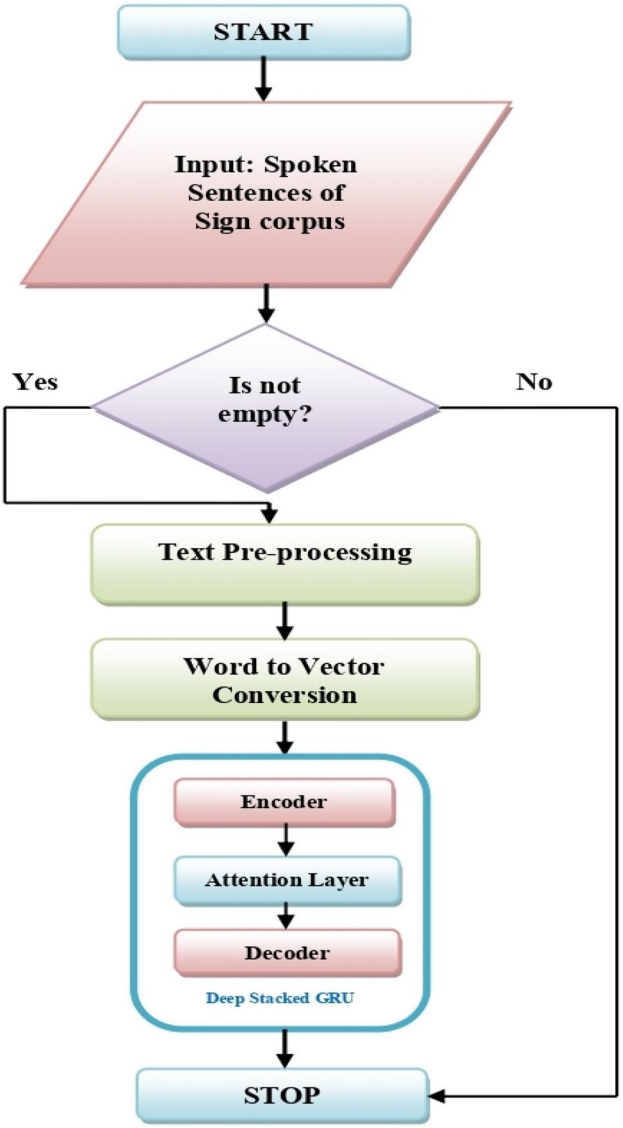

Fig. 3 Flow of execution of the proposed model

3.5 The proposed model

The entire flow of the proposed hybrid NMT model is illustrated in Fig. 3 which demonstrates the step-by-step implementation. Unlike, the conventional sequence to sequence based NMT models, the proposed model challenges human evaluation by employing an attention mechanism (Bahdanauet al. 2014) and sentence piece approach (Kudo et al. 2018). Analysing the various components of source

language sentences and processing them in a large neural network incurs huge challenges, creates high computational complexities. This development requires high power GPU devices to train the model. The proposed hybrid NMT model manages all these challenges using the novel deep stacked GRU algorithm. The proposed hybrid NMT model consists five parts as shown in Fig. 4: Text pre-processing, Encoder Network, Attention Mechanism, Context Vector and Decoder Network. To be specific, the objectives of each module are explained as follows: Text pre-processing, Encoder Network, Attention Mechanism, Context Vector and Decoder Network. To be specific, the objectives of each module are explained as follows:

Text pre-processing. We first apply text pre-processing steps to remove punctuation marks, symbols and special characters from the sentences. Further, we add < start > and < end > tokens to each sentences. The words are converted into word vectors using word embedding techniques. These steps fine-tune the input sentences to be easily processed by the GRU networks using encoder-decoder components. Encoder Network: As shown in Fig. 5, the Deep Stacked GRU based encoder network reads each word vectors of source sentences with time step t and produces summary at the final hidden layer unit. Each hidden layers keep track of previous hidden layer information associated with vector relationships. The forget gets decides which information to be removed and update gate decides the important key word in a sentence.

Attention Mechanism. We apply attention mechanism (Bahdanau et al. 2014) to achieve good translation results. Attention mechanism uses softmax function and applies more attention to the important vectors with high probability values and low values to other vectors. This yields good translation output results. Context Vectors. The context vectors values are computed using average sum of attention vector values at each time step t. It can be further used by the decoder network to predict the next words in a sequence translation tasks. Decoder Network. The first hidden unit of decoder network takes last hidden state summary of encoder unit and context vectors to predict the first word of the target sentence. Further using previous state values and context vectors it completes the entire translation process.

Algorithm 1: Deep Stacked GRU
Begin
Input: Source language (S_M) and Target Language (T_N) Sentence Pairs Dataset
Output: Generation of Target Language (T_N)

Procedure

1. Let Source sentences be $X = \{x_1,x_2,x_3,....x_M\} \in S_{x_i}$ and Target sentences be $Y = \{y_1,y_2,y_3,....y_N\} \in T_{y_j}$ Where $x_1,x_2,x_3,....x_M$ denotes words of source sentence and size M and $y_1,y_2,y_3,....y_N$ denotes words of target sentence and size N

2. Apply GRU encoder(voc_sz, em_dim, enc_ut, bat_sz) on S_x where voc_sz denotes vocabulary size, em_dim embedding dimensions, enc_ut encoder unit , bat_sz batch size

3. Initialize voc_sz, em_dim, enc_ut, bat_sz
   Apply embedding on S_X(voc_sz, em_dim)
   Apply GRU on S_X ($enc\_ut, r_{seq}, r_{state}, rec_{int}$)
   where r_{seq} denotes return_sequences, r_{state} return_state, rec_{int} recurrent_initializer

4. Calculate Attention vector
$$\boldsymbol{a}_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t])$$

5. Calculate Attention Weights
$$\alpha_{ts} = \frac{\exp\left(\text{score}\left(h_t, \bar{h}_s\right)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}\left(h_t, \bar{h}_{s'}\right)\right)}$$

6. Calculate context vector
$$c_t = \sum_s \alpha_{ts}\, \bar{h}_s$$
   return(batch_size,units)

7. Apply GRU decoder(voc_sz, em_dim, enc_ut, bat_sz) on T_y

8. Initialize bat_sz,dec_ut
   Apply embedding on T_y(voc_sz, em_dim)
   Apply GRU on T_y($dec\_ut, r_{seq}, r_{state}, rec_{int}$)
   Add dense layer(voc_sz)
   Apply Bahdanau Attention
$$\alpha_{ts}, c_t = attention(H, e_{output})$$
   expand c_t and concatenate em_dim + H_{size}

9. return(output)
End

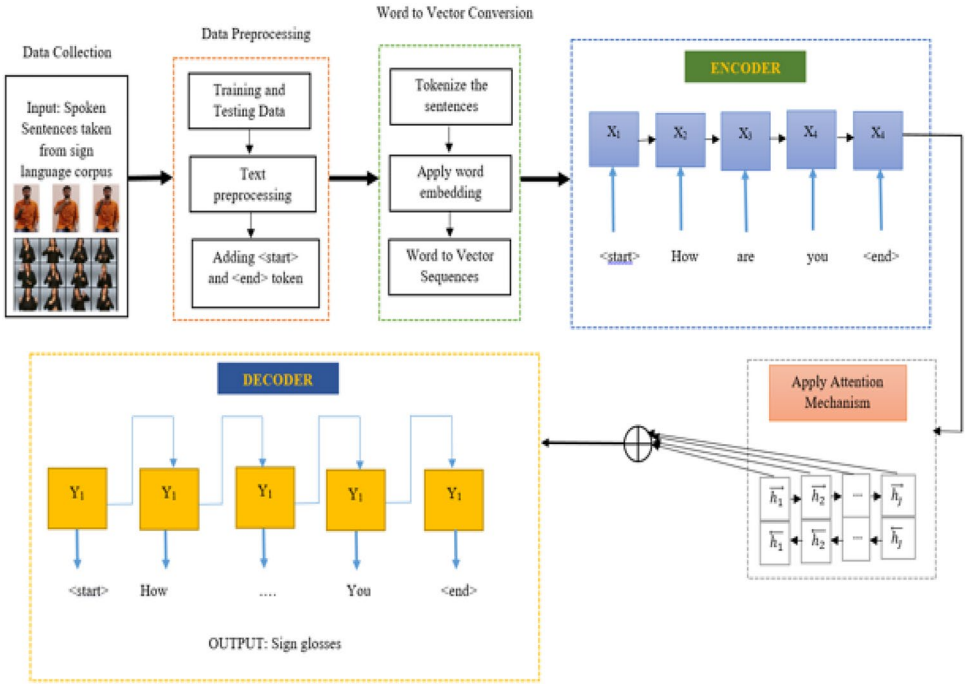

Fig. 4 The proposed hybrid NMT system architecture

The proposed model uses the three benchmark sign corpus datasets—RWTH PHOENIX Weather 2014 T dataset (Koller et al. 2015a, b), How2Sign (Duarte et al. 2021), and ISL-CSLTR Dataset (Elakkiya et al. 2021). These three corpus represents three countries sign languages provide the capability to handle multilingual datasets by our proposed model. The RWTH PHOENIX Weather 2014 T dataset is created for German Sign Language (GSL), How2Sign dataset for American Sign Language (ASL), and ISL-CSLTR represents Indian Sign Language (ISL).

The source sentences are represented as $S = \{x_1, x_2, x_3, \ldots . x_M\}$ where M denotes the varying length and the target language sentences are represented as $T = \{y_1, y_2, y_3, \ldots . y_K\}$ of length K. The proposed model handles the varying length of source sentences and translates them into target languages of different size. The prediction of target sentences by learning source sentence structure is performed by applying conditional probability denoted in Eq. (8) as follows.

$$P\left(T_{target} | S_{source}\right) = P\left(x_1, x_2, x_3, \ldots . x_M\right) = \prod_{i=1}^{K} P\left(y_i | y_0, y_1, y_2, \ldots . y_{i-1}; x_1, x_2, x_3, \ldots . x_M\right) \tag{8}$$

The dense vectors are generated by applying word embedding to the source sentences. The context vector values are generated for source sentences by averaging the weight values of source units. The attention weights are calculated using context vectors to compare the target hidden units with source hidden units depicted in Fig. 4. The attention vector is calculated by adding context vector and target node hidden unit values. The attention vector process the next time step tasks by feeding it as input. The attention weights are computed in Eq. (9).

$$\alpha_{ts} = \frac{\exp\left(\text{score}\left(h_t, \bar{h}_s\right)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}\left(h_t, \bar{h}_{s'}\right)\right)} \tag{9}$$

The context vector equation calculated (Eq. 10) as follows

$$c_t = \sum_s \alpha_{ts} \bar{h}_s \tag{10}$$

The attention vector computed (Eq. 11) as follows

$$a_t = f\left(c_t, h_t\right) = \tanh\left(W_c\left[c_t; h_t\right]\right) \tag{11}$$

The score function compares both target and source hidden units and the results are normalized for yielding attention weights. The attention mechanism called as Bahdanau's additive styles (Eq. 12) is used here for mapping source and target components and passes the attention vectors to the

next nodes and flows the previous decision made in the network.

$$\text{score}\left(\mathrm{h}_t, \overline{h}_s\right) = \upsilon_a^{\mathrm{T}} \tanh(\mathrm{W}_1 \mathrm{h}_t + \mathrm{W}_2 \overline{h}_s) \tag{12}$$

The input sentences are passed through the encoder units, which return the output of the encoder and hidden states. These output values are read back by decoders for making predictions using decoder hidden units and finally computes the loss. This training loops process is performed endlessly still reaching the end of the word in a sentence, stops while encountering < end > token. The gradients are calculated for optimizing the network and back propagation techniques.

The variants of RNN family GRU are implemented using a gating mechanism, which works much similar to LSTM. GRU has update and reset gates. Both these gates monitor the information flow in memory cells. Update gate regulates the flow of the information passing in memory cells and the reset gate limits the information passing out of memory cells. GRU models are designed to avoid the occurrences of vanishing gradient problems. GRU achieved remarkable results in speech signal modelling and character recognition and handwriting letter recognition.

Cho et al. (2014a, b) introduced GRU is an enhanced model of RNN was developed to restrict the vanishing gradient problems that occur in neural networks. The GRU uses update and reset gates to decide what information be passed to the output unit. GRU collects input from current sources and previous hidden layers as vectors and Computes the next value using chain sequences of the next states. The update gate ($\mathrm{z_t}$) is computed at time step t using the following Eq. (13)

$$Z_t = \sigma\left(W^{(z)} x_t + U^r h_{t-1}\right) \tag{13}$$

The current input (x_t) is multiplied with its weight values W and the previous hidden state values (h_{t-1}) are multiplied with its weights (U) and finally added. The activation function sigmoid (σ) applied on added values to brought results between 0 to 1. The update gate decides the amount of information that needs to be carried for the next states. To calculate the amount of information to be forgotten from memory reset gate used, which are computed by the following Eq. (14)

$$r_t = \sigma\left(W^{(r)} x_t + U^r h_{t-1}\right) \tag{14}$$

The reset gate (r_t) is evaluated by multiplying the current input (x_t) and weights (W) and previous hidden node values (h_{t-1}) withweight values (U) and finally sigmoid activation performed on it. The current memory content ($\mathrm{h_{curr}}$) is computed (Eq. 15) as

$$\mathrm{h_{cur}} = \tanh\left(W x_t + r_t \odot U h_{t-1}\right) \tag{15}$$

The inputs are multiplied with their weight values, further the Hadamard product (element-wise multiplication) operation is applied on values of reset gate and previous hidden states values, multiplied with its weight values. At the end, the non-linear activation tanh is applied to the results. The final information content present in memory ($\mathrm{h_f}$) at time step t computed using Eq. (16).

$$h_f = Z_t \odot h_{t-1} + \left(1 - Z_t\right) \odot h_t^{'} \tag{16}$$

In the final step, final memory cell values based on time step t are computed using the update gate. The Hadamard operation is performed on update gate values and $\left(1 - Z_t\right)$ using hidden state vectors.

The proposed model uses a Deep-stacked GRU algorithm which learns the patterns of source and target sequences using deep neural networks and it has 1024 hidden layers and neurons based on the input size, this process is performed repeatedly till < eos > token is encountered. The sequences of source text learned by GRU-encoders convert it into context vectors. The attention layer computes the attention vector for the encoder context vector values. The decoder reads the encoder values and translates it as target sentences. The proposed model also incorporates the techniques for handling rare words (Kudo et al. 2018) and word embedding techniques to find the non-monotonic relationship between words. This proposed model automatically learns the relationship among the words and underlying structure of source and target languages. Algorithm 1 illustrates the execution flow of Deep Stacked GRU implementation.

The proposed Deep stacked GRU algorithm uses stacked layers of GRU to effectively process the sequential inputs and translates into target form. We apply Bahdanau et al. (2014) attention mechanism to compute distinct context vector values and yields good results. The recursive nature of GRU process the entire source sentences and translates them into target sentences. We use beam size 10 and tanh and sigmoid activation function. The proposed model totally process 40 k sentences by combining multilingual sign corpus collected from different sources.

3.6 Time complexity analysis

The proposed hybrid NMT model has been developed using Intel Xeon Silver 4210 CPU 2.20 GHz and 2.20 GHz (2 processors) and Windows 10 (64-bit system) Operating system environment. The time complexity analysis deals the efficiency of our algorithm run time. The spoken sentences are pre-processed to remove the special symbols, and punctuation marks. The proposed model uses 256 hidden nodes

at each stack of GRU layer. The time taken for complete execution takes 41 min to produce plausible results.

4 Experiment settings

We have compared the performance of the proposed hybrid NMT model with three benchmark sign language corpus. The RWTH-PHOENIX-Weather 2014 T, How2Sign and ISL-CSLTR dataset are used. The BLEU score based quality evaluations shows the significant improvements in the model performance. These models are not designed for specific purposes and datasets. These models will work effectively for different language corpora and yields better results. The proposed model translation results indicate the quality of translation over benchmark datasets. We have trained our proposed model for 200 epochs with 864 parameters. The dropout value is set as 0.1.

4.1 RWTH-PHOENIX-Weather 2014 T dataset

The RWTH-PHOENIX-Weather 2014 T dataset (Koller et al. 2015a, b) was developed for German Sign Language based on weather broadcast information. These continuous sentences corpus comprises 1225 signs includes 5356 continuous sentences performed by 9 native signers. This corpus was collected for the years 2009–2011 from public Television Centre, Germany. The contents are related to everyday news information and weather reports. The videos are captured using multiple signers. The dataset was properly labelled and uses dark clothes and uniform background and the sample images are shown in Fig. 6.

4.2 How2Sign dataset

The How2Sign dataset have been developed for conducting research over American Sign Language. The dataset consists of 80 h duration video for continuous sentences with proper annotations. This corpus was created using 11 different signers and provides different modalities of data like pose information, multi-view and speech and consists of 16 k vocabularies and the sample images are shown in Fig. 7.

4.3 ISL-CSLTR: continuous sign language translation and recognition dataset

The Indian Sign Language dataset (sentence level) is developed for exploring research in continuous SLRT model development (Elakkiya et al. 2021). The dataset contains sign language videos for commonly used spoken English sentences. This novel corpus was primarily developed for 100sentences captured at SASTRA Deemed University, Thanjavur, Tamilnadu, India. The ISL-CSLRT corpus has been published and open access for researchers at https://data.mendeley.com/datasets/kcmpdxky7p/1.

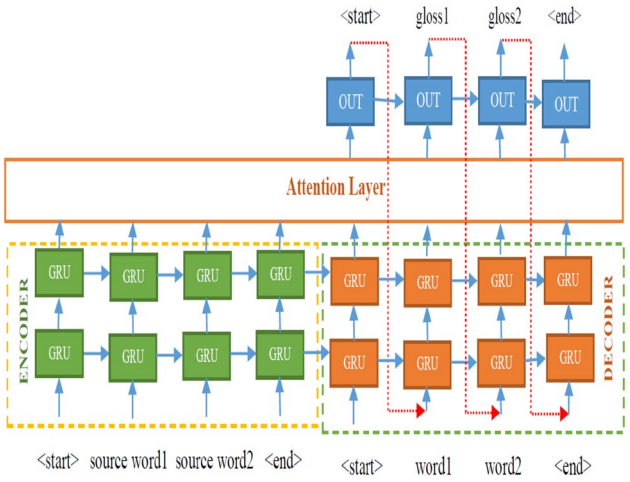

Fig. 5 Proposed Deep Stacked GRU

The ISL-CSLRT corpus comprises totally 700 sign videos. This corpus has been designed for handling large vocabularies of language sentences. Totally 35 k sign frames are present in this novel corpus. The videos are annotated properly for easier processing. This corpus supports the multi-signer environment, captured using 5 male signers and 2 female signers of various age groups. Table 2 shows the statistics of the ISL-CSLRT corpus.

The ISL-CSLRT corpus videos are captured at SASTRA Deemed University, Tamilnadu, India by students and Research Scholars and 2 native signers from Navajeevan, Residential School for the Deaf, College of Spl. D.Ed&B. Ed, Vocational Centre, and Child Care & Learning Centre, Ayyalurimetta, Nandyal, Andhra Pradesh, India. The recordings of video are done using a Canon Digital SLR Camera. This proposed novel corpus comprises 700 videos for continuous sentence level recognition and 35 k images for isolated recognition of sign languages encourages research in the field of automated Sign Language Recognition and Translation. The ISL-CSLRT corpus videos are captures by considering various angles, different backgrounds, and multiple lighting conditions. This corpus was aligned properly to facilitate easy access to resources present in it. Figure 8 illustrates the sample image frames taken from videos present in the ISL-CSLRT dataset.

5 Results and discussion

The proposed model performance is evaluated using BLEU Score metrics and Table 3 shows the comparative analysis of performance with existing models and clearly shown in Fig. 9. The Bilingual Evaluation Understudy Score (BLEU)

Fig. 6 Sample sign gesture images of RWTH-PHOENIX-Weather 2014 T dataset

Fig. 7 Sample sign gesture images of How2Sign dataset

Fig. 8 Sample sign gesture images of ISL—CSLRT corpus

estimates the performance of our proposed model by comparing generated translation output with ground truth sentences and it is shown in Eq. (17). The proposed model has been developed to compete the human evaluation performance. Our proposed hybrid NMT model has produced improved performance than existing approaches.

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{17}$$

Our model produces the BLEU Score 38 for RWTH PHOENIX Weather 2014 T dataset, 38.6 for How2Sign dataset and 39.1 for the ISL-CSLTR dataset which is comparatively higher than earlier RNN based NMT approaches. The proposed model also greatly decreases the translation errors by 70%.The BLEU scores of earlier approaches are compared with our model displayed in the Table 3 shows the improved performance. Figure 9 depicts the BLEU scores of existing frameworks and the proposed model. This exploits the improvements of translation performance of the NMT model using deep neural approaches and challenges the human evaluation system. The relationship between input language sentences and target language sentences is trained by adopting the syntactical and semantics structures of the language showcases the efficiency of our model. Moreover, the systems RNNencdec, RNNSearch, and grConv secures lower scores over the benchmark datasets, our model achieves remarkable growth in translation performance. We

Table 2 Statistics of the ISL-CSLRT corpus

# signers	7
# frames	35,299
#videos	700
# sentences	100
# glosses	186
Left-handed	83
Two-handed	64
Frame rate	25 fps
Resolution	1920×1080

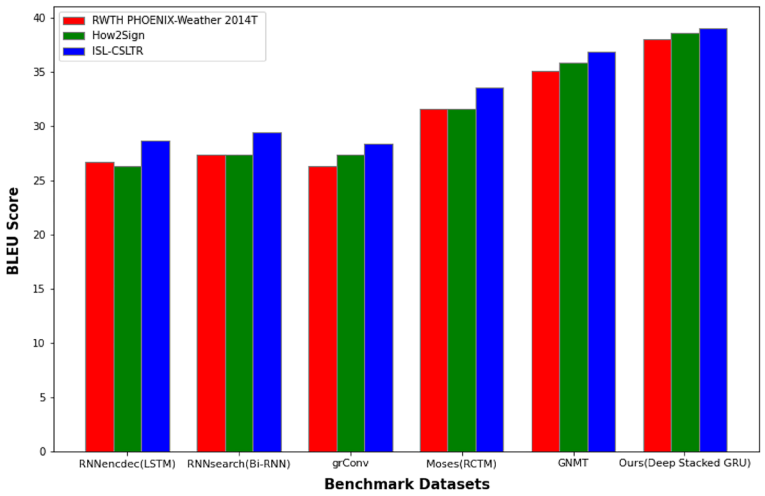

Fig. 9 Comparison of proposed model performance with BLEU Score

Table 3 Results for proposed model performance with BLEU Score

Model	BLEU SCORE		
	RWTH PHOENIX-Weather 2014T	How2Sign	ISL-CSLTR
RNNencdec (Wang et al. 2017)	26.7	26.3	28.7
RNNsearch (Kalchbrenner et al. 2013)	27.4	27.4	29.4
grConv (Amodei et al. 2016)	26.3	27.4	28.4
Moses (Kalchbrenner et al. 2013)	31.6	31.6	33.6
GNMT (Neubig et al. 2017)	35.1	35.9	36.9
Our model	38	38.6	39.1

observed that Fig. 10 processes the language sentences and applies attention on same words in both dimensions x-axis and y-axis, makes the learning process by the model easier and speed up the performance while concentration translation accuracy and other parameters. According to Fig. 10 the word 'repeat' is rarely used in the ISL-CSLTR corpus and requires high attention during translation to find the relationship score with prefixed and following words. The frequency of words, word embedding relationships are taken into account for improving the model performance. The attention plot for sequence learning models is a cognitive process of classifying different segments of sentences as objective or subjective. The alignment of the sentences and glosses are represented. The top side x axis direction denotes the spoken sentences and the target sign glosses are plotted in y axis.

The performance on benchmark datasets RWTH PHOENIX Weather 2014 T dataset, How2Sign, and ISL-CSLTR Dataset during training and testing are illustrated in Fig. 11. In each case, the proposed algorithm produced significant improvements and greatly reduces the translation errors. The proposed hybrid NMT model is evaluated qualitatively and quantitatively using benchmark datasets and produces comparable improvements in translation performance.

Fig. 10 Attention plot for ISL-CSLTR dataset source sentence 'can you repeat that please', target sign gloss 'YOU REPEAT PLEASE'

6 Conclusion

This paper introduces a novel Hybrid NMT system to translate the spoken sentences into sign glosses. The proposed model has been evaluated on multi-lingual sign

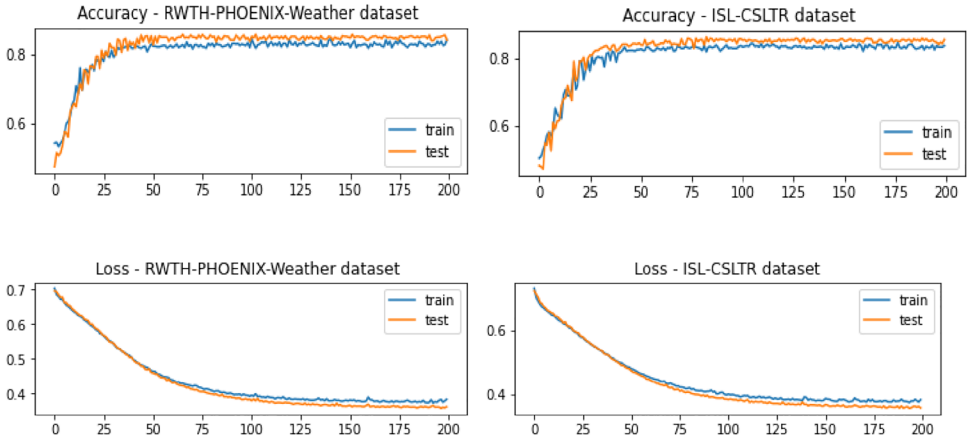

Fig. 11 Analysis of accuracy and loss of the proposed model on benchmark datasets

corpora—RWTH PHOENIX Weather 2014 T dataset, How-2Sign, and ISL-CSLTR Dataset. The experimental results have shown that the proposed model achieves better performance than other state-of-the-art approaches. The proposed hybrid NMT model obtained improved BLEU Scores—38 for RWTH PHOENIX-Weather 2014 T dataset, 38.6 for How2Sign and 39.1 for ISL-CSLTR dataset. The proposed model performance is improved in a greater scale by handling multi-lingual datasets and addressing the issues persists with earlier approaches like handling longer sequences and new words. Henceforth, the proposed Hybrid NMT model addresses the challenges of conventional NMT systems and reaches the new heights in translation performance.

Acknowledgements The research was funded by the Science and Engineering Research Board (SERB), India under Start-up Research Grant (SRG)/2019–2021 (Grant no. SRG/2019/001338). We would like to thank Navajeevan, Residential School for the Deaf, College of Spl. D.Ed & B.Ed, Vocational Centre, and Child Care & Learning Centre, Ayyalurimetta, Nandyal, Andhra Pradesh, India for their support and also, we thank all the students for their contribution in collecting the sign videos and the successful completion of the ISL-CSLTR corpus.

References

Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Zhu Z (2016) Deep speech 2: end-to-end speech recognition in English and Mandarin. In: International conference on machine learning, pp 173–182. PMLR

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: arXiv: 1409.0473

Bantupalli K, Xie Y (2018) American sign language recognition using deep learning and computer vision. In: 2018 IEEE international conference on big data (big data), pp 4896–4899. IEEE

Bengio Y, Ducharme R, Vincent P, Janvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155

Bheda, V, Radpour D (2017) Using deep convolutional networks for gesture recognition in American sign language. arXiv preprint arXiv: arXiv: 1710.06836

Camgoz NC, Hadfield S, Koller O, Ney H, Bowden R (2018) Neural sign language translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7784–7793

Carpuat M, Wu D (2007) Improving statistical machine translation using word sense disambiguation. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 61–72

Chiang D (2005) A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05), pp 263–270

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv: arXiv: 1406.1078

Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv: arXiv: 1409.1259

Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634

Duarte A, Palaskar S, Ventura L, Ghadiyaram D, DeHaan K, Metze F, Giro-i-Nieto X (2021) How2Sign: a large-scale multimodal dataset for continuous American sign language. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 2735–2744

Elakkiya R (2021) Machine learning based sign language recognition: a review and its research frontier. J Ambient Intell Hum Comput 12(7):7205–7224

Elakkiya R, Natarajan B (2021) ISL-CSLTR: Indian sign language dataset for continuous sign language translation and recognition. Mendeley Data. https://doi.org/10.17632/kcmpdxky7p.1

Elakkiya R, Selvamani K (2018) Enhanced dynamic programming approach for subunit modelling to handle segmentation and recognition ambiguities in sign language. J Parallel Distributed Comput 117:246–255

Elakkiya R, Selvamani K (2019) Subunit sign modeling framework for continuous sign language recognition. Comput Electr Eng 74:379–390

Graves A, Fernández S, Gomez F, & Schmidhuber J (2006) Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on machine learning, pp 369–376

Graves A, Mohamed AR, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp 6645–6649. IEEE

Guo D, Zhou W, Li H, Wang M (2018) Hierarchical LSTM for sign language translation. In: Proceedings of the AAAI conference on artificial intelligence, vol 32(1)

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Kalchbrenner N, Blunsom P (2013) Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1700–1709

Ko SK, Kim CJ, Jung H, Cho C (2019) Neural sign language translation based on human keypoint estimation. Appl Sci 9(13):2683

Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. University of Southern California Marina del Rey Information Sciences Institute

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pp 177–180

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: MT summit, vol 5, pp 79–86

Koller O, Forster J, Ney H (2015b) Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comput vis Image Underst 141:108–125

Koller O, Ney H, Bowden R (2015) Deep learning of mouth shapes for sign language. In: Proceedings of the IEEE international conference on computer vision workshops, pp 85–91

Koller O, Zargaran O, Ney H, Bowden R (2016) Deep sign: hybrid CNN-HMM for continuous sign language recognition. In: Proceedings of the British machine vision conference 2016

Konstantinidis D, Dimitropoulos K, Daras P (2018) A deep learning approach for analyzing video and skeletal features in sign language recognition. In: 2018 IEEE international conference on imaging systems and techniques (IST), pp 1–6. IEEE

Kudo T, Richardson J (2018) Sentencepiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv: arXiv: 1808.06226

Kudo T (2018) Subword regularization: Improving neural network translation models with multiple subword candidates. arXiv preprint arXiv: arXiv: 1804.10959

Luong MT, Pham, H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv preprint arXiv: arXiv: 1508.04025

Neubig G (2017) Neural machine translation and sequence-to-sequence models: a tutorial. arXiv preprint arXiv: arXiv: 1703.01619

Ong SC, Ranganath S (2005) Automatic sign language analysis: a survey and the future beyond lexical meaning. IEEE Trans Pattern Anal Mach Intell 27(06):873–891

Provilkov I, Emelianenko D, Voita E (2019) Bpe-dropout: simple and effective subword regularization. arXiv preprint arXiv: arXiv: 1910.13267

Pu J, Zhou W, Li H (2019) Iterative alignment network for continuous sign language recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4165–4174

Pust M, Hermjakob U, Knight K, Marcu D, May J (2015) Parsing English into abstract meaning representation using syntax-based machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1143–1154

Sennrich R, Haddow B, Birch A (2015) Neural machine translation of rare words with subword units. arXiv preprint arXiv: arXiv: 1508.07909

Simard M, Ueffing N, Isabelle P, Kuhn R (2007) Rule-based translation with statistical phrase-based post-editing. In: Proceedings of the second workshop on statistical machine translation, pp 203–206

Stoll S, Camgoz NC, Hadfield S, Bowden R (2020) Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. Int J Comput vis 128(4):891–908

Stoll S, Camgöz NC, Hadfield S, Bowden R (2018) Sign language production using neural machine translation and generative adversarial networks. In: Proceedings of the 29th British machine vision conference (BMVC 2018). University of Surrey

Sutskever I, Vinyal O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp 3104–3112

Tu Z, Lu Z, Liu Y, Liu X, Li H (2016) Modeling coverage for neural machine translation. arXiv preprint arXiv: arXiv: 1601.04811

Utiyama M, Isahara H (2007) A comparison of pivot methods for phrase-based statistical machine translation. In: Human language technologies 2007: the conference of the North American chapter of the association for computational linguistics; proceedings of the main conference, pp 484–491

Vaswani A, Bengio S, Brevdo E, Chollet F, Gomez AN, Gouws S, Uszkoreit J (2018) Tensor2tensor for neural machine translation. arXiv preprint arXiv: 1803.07416

Wang W, Knight K, Marcu D (2007) Binarizing syntax trees to improve syntax-based machine translation accuracy. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 746–754

Wang X, Lu Z, Tu Z, Li H, Xiong D, Zhang M (2017) Neural machine translation advised by statistical machine translation. In: Thirty-first AAAI conference on artificial intelligence

Wang S, Guo D, Zhou WG, Zha ZJ, Wang M (2018) Connectionist temporal fusion for sign language translation. In: Proceedings of the 26th ACM international conference on multimedia, pp 1483–1491

Wołk K, Marasek K (2015) Neural-based machine translation for medical text domain. based on European Medicines Agency leaflet texts. Proc Comput Sci 64:2–9

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Dean J (2016) Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv: 1609.08144