Deep Learning Based Text Translation and Summarization tool for hearing impaired using Indian Sign Language

Anurag Kumar Jha[1], Kabita udhary[1] and Dr. Sujala Shetty[1]
[1]*Birla Institute of Technology and Science, Pilani, Dubai Campus, DIAC Dubai 34055, UAE*

Keywords: Natural Language Processing (NLP), Natural Language Generation (NLG), Bidirectional Auto-Regressive Transformers (BART), Multilingual Bidirectional Auto-Regressive Transformers (mBART), Signing Gesture Markup Language (SiGML), HamNoSys, Indian Sign Language.

Abstract: There have been multiple text conversions emerging with time but there has hardly been any work in the field of sign language. Even in the field of sign language multiple methods have been proposed to convert it into text via image detection, but due to the rarity of sign language corpus not much work has been put into text or speech to sign language. The proposed project intends to create a translation model to convert text or audio into sign language with its designated grammar. The process includes translation of any language to English followed by summarization of a big article or text, removal of stopwords, reordering the grammar form and stemming words into their root form. The translation is performed by mBART model, summarization is performed using BART model, conversion into animation is done via mapping words into a dictionary and replacing words by letters for unknown words. The paper uses HamNoSys [1]., SiGML, BART, mBART and NLP to form the translation system. The paper aims to establish better means of communication with the deaf, dumb and people with hearing issues.

1. INTRODUCTION

1.1. Sign Language generation

Unlike conventional languages sign language utilizes hand gestures, body movements and facial expressions for conveying any information. Translation systems exist between almost all existing languages using machine learning, but sign language stands as an exception. Even in sign language conversion of text into sign language has seen very little development over the years. The objective of the paper is to create a translation system which converts provided text into animated sign language (Indian Sign Language) using animated human figures.

There has not been much work in the field of ISL computerization and those done are mostly in American [4] or British sign language [3]. The underlying architecture for these systems is mostly based on [6]:

- Direct translation of input into target words. The biggest drawback of this system is that output is not grammatically correct and difficult to understand.
- Statistical machine translation which is ruled out in our case because of the lack of a large parallel corpus
- Transfer based: These include proper grammatical rules in place from proper translation from one language to another

The proposed method is to create/collect video animation for the entire pool of ISL words which are around 1500 in total. The input text is manipulated to abide by the grammatical syntax of ISL and then mapped to the dictionary of the video animations. Words not present are broken into letters and shown one by one. These can be for things such as a name or a place.

A major challenge in the system is the conversion of one language to another with a completely different set of grammatical syntax in place.

1.2. Text Summarization

BART is a sequence-to-sequence model trained as a denoising autoencoder It is applicable to many types of tasks like sequence classification (categorizing input text sentences or tokens), summarizing text, machine translation like translation between multiple language, question answering. Its pretraining has mainly two phases. Assign corrupted text with an arbitrary noise and sequence-to-sequence model is learned to rebuild the actual text. It is evaluated with a different noise approach as shown in fig1, like randomly shuffling the order of the original text and using a novel in-filling scheme (in this scheme length of span of text are replaced with mask token). It is an unsupervised language model which can be fine-tuned to a specific application like medical chatbots, generating summary of meeting, natural language to programming language, language translation etc. As it is already pretrained with very large amount of data, a small data set can be used to fine-tune it.

Bart used different noising schemes such as:

- Token masking: Some random tokens are replaced with masks in a sentence and the model is trained to predict the single token based on the rest of the sequence.
- Token deletion: Some random tokens are deleted in a sentence and the model learns to find the deleted token and from where it was deleted.
- Text Infilling: Some contiguous tokens are deleted and replaced with a single mask and the model learns the missing token and the content.
- Sentence permutation: Sentences are permuted, and the model learns the logical implication of the sentence.
- Document Rotation: Here the documents are rearranged randomly. This helps the model to learn how the documents are arranged.

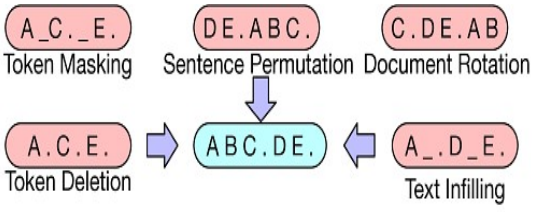

Figure 1. Transformations for noising the input for BART fine-tune

By using the BART transformer model, we can automate the text summarization task. Text summarization can be done in 2 ways.

- Extractive summarization: It provides the important text present in the given input.
- Abstractive summarization: It provides the actual summary of the given input. So it is more challenging as it has to read complete text and understand the meaning of the text and provide us with the summary.

1.3. Text Translation

For translation of text from any language to English we are using mBART[14]. It is a multilingual neural machine translation model. mBART supports up to 50 languages. Initially the mBART model trains on twenty-five different languages and is fine-tuned for different tasks. For translation, it was fine-tuned on bitext (bilingual finetuning). Later for mBART50, the multilingual nature of the pretraining is used for fine-tuning.

2. BACKGROUND AND LITERATURE REVIEW

About 63 million people, which accounts for about 6.3% of the Indian population, suffer from hearing issues(Indian Census, 2011). A vast majority ranging from 76% to 89% of this population have no knowledge of sign language. The low literacy rate can be attributed to lack of work put into this field and absence of proper translation systems.

There have been some research [5] into machine translation used in other sign languages which are:

- Direct Translation: The architecture works on word to word translation. The biggest drawback is the lack of context and meaning. There is no syntactical analysis and grammatical syntax is ignored. There is direct translation without any reordering which has a massive issue in the sense that ordering of sentences is completely different in sign language as compared to English. The format used in the English language is Subject-Verb-Object compared to Subject-Object-Verb in ISL.
- Transfer based (Rule based): In this architecture input is passed through syntactic and semantic transformation to convert it into intermediate text which is

then converted into target language using linguistic rules.

- Interlingua based: It is an alternative to the above architectures and is based on Interlingua which is a language independent semantic structure formed by the semantic analysis of the input. This is then used to generate the target language.

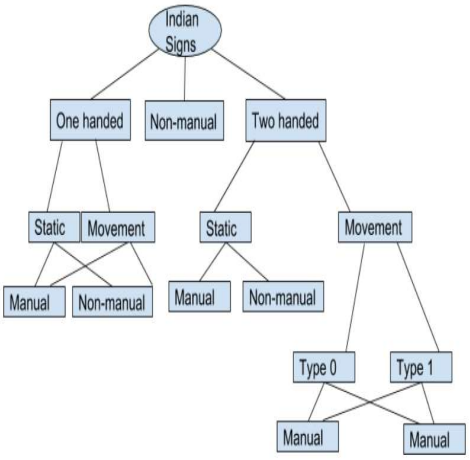

Figure 2: ISL Type Hierarchy(Type 0 refers to use of both hands and in type 1, use of one had is dominant)[1]

In order to formulate an algorithm to translate English text to sign language, the following table of sign language details must be kept in mind:

Table 1: Important details of sign languages.

1	NOT the same all over the world
2	NOT just gestures and pantomime but do have their own grammar.
3	Dictionary is smaller compared to other languages.
4	Finger-spelling for unknown words.
5	Adjectives are placed after the noun for most of the sign language.
6	Never use suffixes
7	Always sign in present tense
8	Do not use articles.
9	Do not uses I but uses me.
10	Have no gerunds
11	Use of eyebrows and non-manual expressions.
12	NOT been invented by hearing people.

As mentioned before, sign language grammar is not similar to conventional languages and has certain distinct features which are:

Table 2: Features of sign languages

1	Number representation are done with hand gestures for each hand.
2	Signs for family relationships are preceded by male or female.
3	In interrogative sentence, all the WH questions are placed in the back of the sentence.
4	It also consists of many non-manual gestures such as mouth pattern, mouth gestures etc.
5	The past, present and future tense is represented by signs for before, then and after.

3. BLOCK DIAGRAM OF THE MODEL

The complete sequence for the conversion of any language to Indian sign language is shown in figures 3.

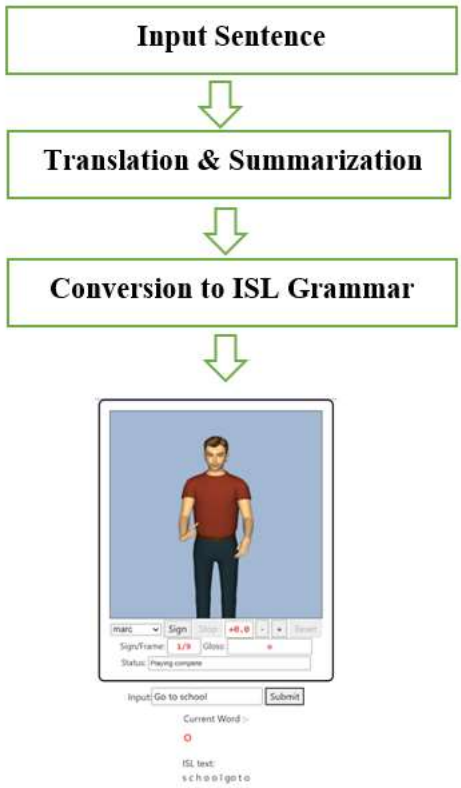

Figure 3: Block diagram for text Translation and Summarization in ISL

4. HYPOTHESIS

For transfer of text to sign language we will be using a dictionary having English words and their equivalent sign. The sign can be in the format of video, images or code signs. All these are compared in the following table:

Table 3: Comparison between different media representing sign languages.

Kind of media	Pros	Cons
Video Signs	• Realistic • Easy to create	• Time consuming to • create • High memory consumption • Not supported by translation system
Pictures	• Very less memory consumption	• Time consuming to create • Not realistic as compared to videos • Not supported by translation system
Code Sign Language Text	• Minimal Memory Consumption • Supported by translation system as it is the written form and can be processed very easily	• Very difficult to read and understand • Required to be learnt

An analysis of the table gives the estimate that although videos are more time consuming to create and require a higher amount of memory, they are best suitable for easy understanding.

5. METHODOLOGY

We have used http://www.indiansignlanguage.org/ to download video clips to map to our English word dictionary. These videos are then manually labelled.

Input is taken in the form of English text. Text parsing is done with the help of Stanford parser[7] which creates its grammatical phase structure. This is reordered in accordance with ISL. English Language grammar follows Subject-Verb-Object structure which is Subject-Object-Verb in the case of ISL. The irrelevant stop words are removed.

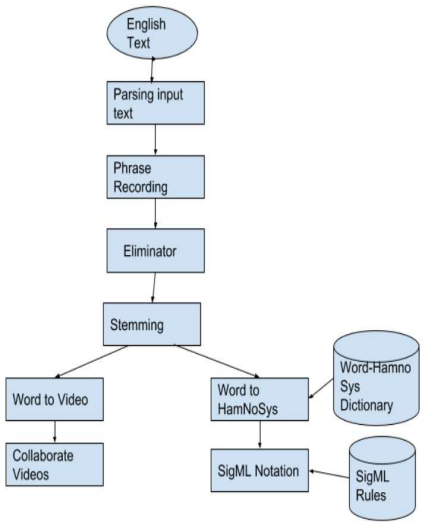

Figure 4: Algorithm for text to ISL

5.1. Solution Structure

- Parsing the input: To carry the translation of one language to another, both their grammatical structure must be known. Parsing is used to obtain this grammatical structure. For parsing the input Stanford parser[7] is used which breaks input into part of speech tagged text, CFG and type dependency representation.
- Grammar rules for conversion from English to ISL: The grammatical pattern of ISL varies from that of English language. ISL requires the verb patterns to be shifted after nouns as shown in the following table 4.
- Eliminating stop words: The English language includes words which don't have any meaning in ISl which include modals, foreign words, possessive ending, coordinating conjunction, determiners, adjectives, comparative and superlative, nouns plural, proper plural, particles, symbols, interjections and non-root verbs.
- Lemmatization: Stemming is used to convert words into their root form using Port stemmer rules. Each of the words is checked in the dictionary, if it doesn't exist, it is tagged to its synonym.

Table 4: Grammatical reordering from English to ISL

Verb Pattern	Rule	Input Sentence	Parsed Sentence	Output Sentence
verb+object	VP TO NP	Go to school	(VP (VB Go)(TO to) (NP (NN school)))	School to go
Subject + verb	NP V	Birds fly	(NP (NNS birds)) (VP (VBP fly))	Birds fly
subject + verb + subject complement	NP V NP	His brother became a soldier	(NP (PRP$ his) (NN brother)) (VP (VBD became) (NP (DT a) (NN soldier)))	His brother a soldier became
subject + verb + direct object + preposition +prepositional object	NP V NP PP	She made coffee for all of us	(NP (PRP She)) (VP (VBD made) (NP (NN coffee)) (PP (IN for) (NP (NP (DT all)) (PP (IN of) (NP (PRP us))))))	she coffee for all of us made

- Output generation: Upon the execution of the above steps we receive the ISL equivalent of the input. It is then checked to corresponding keys in our text-animation dictionary. If a word is found it is displayed as video by passing it through a HamNoSys[2] generator, otherwise the word is broken and finger-spelling used to express the word.

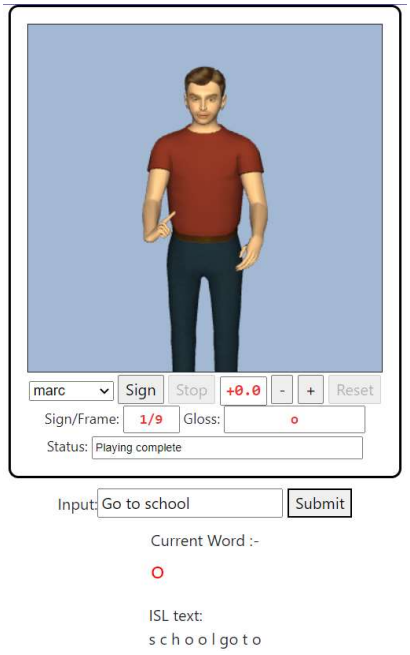

Figure 5: Conversion of English to ISL Grammar

6. BART ARCHITECTURE

The purpose of a NLP model is to not only understand the whole text given to read but also to understand the sequence of the text, like what comes before and after a token. The sequence of input tokens plays a very important role. For example let's say the statement is, "We are going to the theater to watch a movie." So if we mask "theater " by adding some noise and pass it to the model like:" We are going to the [mask]to watch a movie" The BART model should read thoroughly the provided text and also understands the sequence of words to Predict the masked word.

For a NLP model it is imperative to completely read the sentence and understand each and every token in context of their sequence. Such a case, the input sequence can be interpreted by using a bi-directional approach.

Bart uses the bi-directional approach as shown in, figure 5 to find the masked token. Hence the first part of the BART model is to use bi-directional encoder of BERT to find the best representation of its input sequence. In the second part It uses an autoregressive model which uses only past input sequences to predict the next word.

.

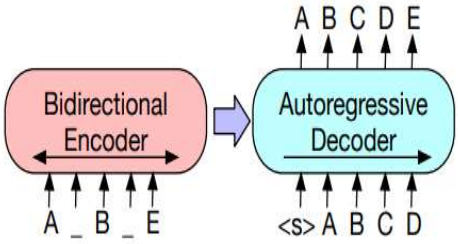

Figure 6: Semantic representation of Bart

HuggingFace provides us the platform to use Bart model for both pretrained and fine-tuned version. We can also use the API for models to summarize text and translate in some other language as well. For our model we are using the "facebook-bart-large-cnn" model for text summarization and the "facebook/mbart-large-50-many-to-many-mmt" model for text translation.

7. DATASET SUMMARY

7.1. Bart

Bart is fine-tuned with CNN, Daily mail dataset which has 300k news articles and all are unique sets, those are written by journalists at CNN and Daily Mail. It supports both type of summarization that are abstractive and extractive summarization.

7.2. mBart

Initially the mBART.cc25 checkpoint [8] available in the fairseq library [12] is used to continue the pretraining process. The monolingual data from XLMR [13] is used to extend the pretraining to a set of 25 languages in addition to the 25 languages mBART model. To be consistent mBART, 250K sentence piece (Kudo and Richardson, 2018) model which was trained using monolingual data for 100 languages from XLMR is used, and thus already supports languages beyond the original 25 mBART was trained on. For pre-training, mBART50 was trained for an additional 300K updates with a batch size of 1700 tokens.

- By using huggingface api, we are translating text into English as shown in figure 7.

```
import requests
r = requests.post\
(url="https://kabita-choudhary-translationmodel.\
hf.space/api/predict",
 json={"data": ['''संयुक्त राष्ट्र के प्रमुख का \
  कहना है कि सीरिया \
में कोई सैन्य समाधान नहीं है''',"Hindi"
,"English" ]})
r.json()
```

```
{'data': ['The head of the United Nations says there is no military solution in
Syria'],
 'is_generating': False,
 'duration': 27.792309284210205,
 'average_duration': 38.51862472663691}
```

Figure 7: Sample of text-translation from Hindi to English

- By using huggingface api, we are summarizing text as shown in figure 8.

```
import requests
r = requests.post(url='https://kabita-choudhary-summary.hf.space/run/predict', json={"data": ['''As inflation continues to
increase, so does the probability of a recession, according to several recent economic forecasts. That means more layoffs,
fewer jobs and higher interest rates may soon be on the horizon.A big reason a recession looks imminent is because of
inflation, which is showing few signs of slowing down. Last week's consumer price index (CPI) report revealed year-over-year
inflation reaching 9.1%, the highest rate since 1981.Banks, including Citigroup, Deloitte and PNC Financial Services, previously
predicted a slowdown in 2023, but recent forecasts say a recession could occur in 2022 or earlier in 2023 than formerly expected.
''']})
r.json()
```

```
{'data': ['Inflation has reached 9.1%, the highest rate since 1981. Forecasts say a recession could occur in 2022 or earlier than previously
expected.'],
 'is_generating': False,
 'duration': 30.775837898254395,
 'average_duration': 187.53118216991425}
```

Figure 8: Sample of text-summarization

8. CONCLUSION

We demonstrate that we can translate and summarize any text to english and transform the generated text into Indian sign language. For translation and summarization, the language generation model is used from the huggingface platform. The future scope of this project is gathering information from an audio and video file and converting it into sign language.

9. REFERENCES

Regina Leven, Heiko Zienert, Thomas Hamke, and Jan Henning Siegmund Prillwitz.(1989). *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introductory Guide in Proceedings of International Studies on Sign Language and Communication of the Deaf.* vol. 5, Hamburg, Germany.

Ulrike Zeshan. (2003) . *Indo-Pakistani Sign Language Grammar:A Typographical Outline*. Sign Language Studies, vol. 3, no. 2, pp. 157-212,.

E´va Safar Ian Marshall. (2003). *A Prototype Text to British Sign Language (BSL) Translation. System*Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2 (ACL '03), vol. 2, pp. 113-116

Matthew Huenerfauth. (2003). *A Survey and Critique of American Sign Language Natural Language Generation and Machine Translation Systems*. Computer and Information Sciences, University of Pennsylvania, Philadelphia, September.

Pamela W.Jordan, John W.Benoit, Bonnie J.Dorr. (1999). *A Survey of Current Paradigms in Machine Translation, Advances in Computers.* vol. 49, pp. 1-68

R. San-segundo , J. M. Montero , J. Macías-guarasa , R. Córdoba , J. Ferreiros and J. M. Pardo. (2004). *Generating Gestures from Speech*

Xu, H, Abdel Rahman, S Jiang, M Fan, J.W. and Huang Y, (2011). November. *An initial study of full parsing of clinical text using the Stanford Parser*. In 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW) (pp. 607-614). IEEE.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer, (2020). *Multilingual denoising pre-training for neural machine translation*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel Rahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*,

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Google AI Language.

Colin Raffel,Noam, Shazeer,Adam, Roberts, Katherine Lee, Sharan Narang ,Michael Matena , Yanqi Zhou , Wei Li , Peter J. Liu . (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. (2019) *FAIRSEQ: A fast, extensible toolkit for sequence modeling.* In North American Association for Computational Linguistics (NAACL): System Demonstrations.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettle-´ moyer, and Veselin Stoyanov. (2019). *Unsupervised cross-lingual representation learning at scale.*

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, Angela Fan. (2020). *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.*