International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

A Translation System That Converts English Text to American Sign Language Enhanced with Deep Learning Modules

Lalitha Natraj, Sujala D. Shetty

Abstract*: A recent surge in interest to create translation systems inclusive of sign languages is engendered by not only the rapid development of various approaches in the field of machine translation, but also the increased awareness of the struggles of the deaf community to comprehend written English. This paper describes the working of SILANT (SIgn LANguage Translator), a machine translation system that converts English to American Sign Language (ASL) using the principles of Natural Language Processing (NLP) and Deep Learning. The translation of English text is based on transformational rules which generates an intermediate representation which in turn spawns appropriate ASL animations. Although this kind of rule-based translation is notorious for being an accurate yet narrow approach, in this system, we broaden the scope of the translation using a synonym network and paraphrasing module which implements deep learning algorithms. In doing so, we are able to achieve both the accuracy of a rule-based approach and the scale of a deep learning one.

Keywords: Artificial Intelligence, Knowledge Based Systems, Natural Language Processing, Neural Networks, Sign Language.

I. INTRODUCTION

American Sign Language or ASL is a semantically and syntactically rich language, that is used by the hearing impaired primarily in North America. In fact, most deaf people prefer ASL over English. This could be because although they can communicate effectively through ASL, research has shown that the English literacy of deaf people in North America does not exceed a fourth-grade level [1]. Therefore, a translation system from English to ASL is essential to minimize the existing language barrier which excludes the deaf community from much of the written and spoken information available today. Consequently, the differences between these two languages must be addressed by the system developed.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Lalitha Natraj*, Birla Institute of Technology and Science - Pilani, Dubai, U.A.E. Email: lalithanatraj99@gmail.com
Sujala D. Shetty, Birla Institute of Technology and Science - Pilani, Dubai, U.A.E. Email: sujala@dubai.bits-pilani.ac.in

ASL uses hand movement and facial expression as its modality of communication, much in contrast to most other languages such as English which uses both written word and speech for the same. In fact, although there are many notations such as the 'glosses' notation and the 'HamNoSys' notation [2], there does not exist a uniformly accepted written script for ASL. This incongruity of modality is one of the major impediments in the translation from a written language to a signed language. This also leads to a shortage of classical translation datasets in which the source and the target languages can be found in similar formats. Although ASL-English datasets exist, highly precise video processing modules would have to be developed to effectively extract accurate data from the ASL side.

This is the reason that although machine translation of written languages has seen a rise in statistical and deep learning approaches, the translation of signed languages stagnates at a rule-based mapping. Be that as it may, within the confines of the limited rules, a rule-based approach provides a highly accurate translation.

In light of this, the system developed in this paper adopts a rule-based approach, enhanced with deep learning to expand the scope of the translation. The translation comprises of two major phases. The first phase is the translation of the English text to an intermediate notation (similar to the glosses notation) using transformational rules followed by the second phase in which the intermediate terms issue a sequence of appropriate ASL animations.

The major bottleneck in this approach so far is the limited transformational rule base that identifies the grammatical structure of the English text and maps it to the appropriate structure in ASL. However, by introducing a deep learning paraphrasing module, the once narrow rule base expands indefinitely, thereby doing away with the bottleneck altogether.

In addition to this, the database of ASL animations are not only effectively obtained using a deep learning posing software but also expanded using a synonym network. In this way, a basic rule-based framework acts as a sturdy backbone and deep learning components are introduced to enhance the scale of the system.

The rest of the paper is organized as follows. Section II discusses the previous research in the field of machine translation of sign languages. Section III gives an overview of the architecture of the system developed.

Section IV and V describe in detail the processes involved in the two phases of the architecture discussed. Section VI explains the role and working of the deep learning paraphrasing module. Section VII discusses the wordnet enhancement module. The results are shown in Section VIII and upon analysis of the same, future work is discussed.

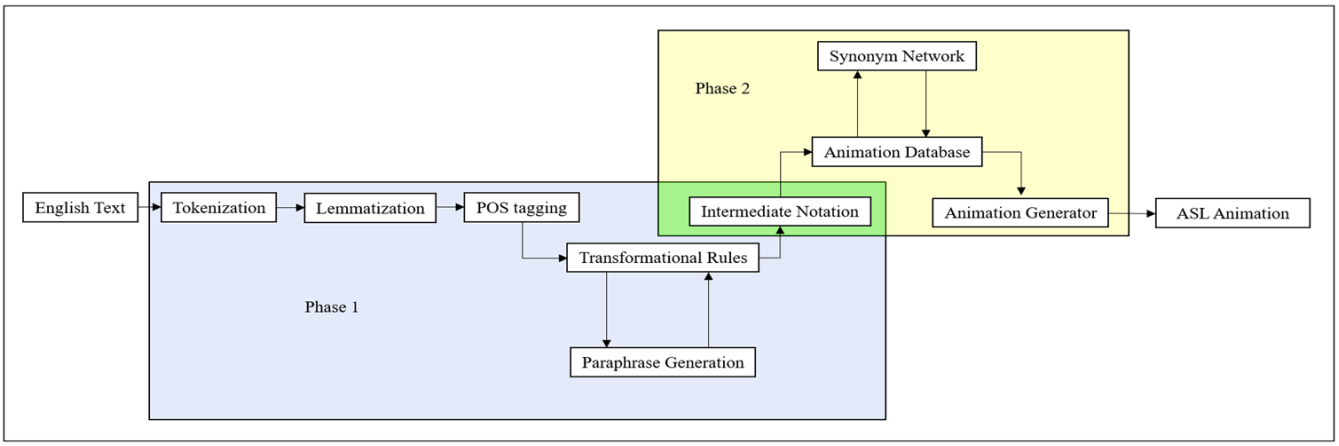

Fig. 1. Architecture Overview of the System

II. LITERATURE REVIEW

The translation of sign languages is an active area of research encouraged by the rapid progress in machine translation technologies. Below is some of the novel developments in this area, the salient features of which, inspired the system developed in this paper.

A. TEAM

In their paper [3] on the machine translation of sign languages, researchers at the University of Pennsylvania, Philadelphia, describe the working of their translation system, which uses the concept of an intermediate language (the glosses notation) with embedded parameters to instruct a 3D avatar to demonstrate the translated signs. They not only took into consideration the grammatical structure of ASL but also included the spatial characteristics of the signs in the form of a vectorized input to the animation avatar.

B. ZARDOS

There have also been systems such as the ZARDOS translator developed by the Hitachi Dublin Laboratory [4], that uses blackboard architecture, an artificial intelligence approach in which common knowledge is loaded onto a common unit called the blackboard and updated by external agents who have access to specialist information. Once the English phrase is analyzed for semantic nature and structure, the sign synthesizer produces an animation of the signed translation using the DCL (Doll Control Language) animation software.

C. Hindi to Indian Sign Language

[5] develops a translation system from Hindi to Indian sign language (ISL). The text processing is done using a dependency parser and part of speech tagger. Using a base of grammar rules and a 'HamNoSys' dictionary of ISL, it generates a written notation of the translation. The limited database of rules is expanded using a wordnet module. The wordnet synonym set also takes into consideration the context of the word to be replaced thereby effectively overcoming the limitation of a limited ISL dictionary.

D. Arabic to Arabic Sign Language

In [6] an Arabic speech to Arabic Sign language (ArSL) translation system is developed. The system is trained on a bilingual corpus and uses a statistical analysis, a robust and upcoming method in the field of natural language processing. The translation is based on transformational rules bolstered by a language model of linguistic background. A sign descriptor module generates an animation of a 3D avatar as the final output. The proposed technique had a sign detection rate of 85% ascertained from experimental results

III. ARCHITECTURE OVERVIEW

The overall architecture of the system developed is discussed below and diagrammatically represented in Fig 1. English text is given as the input. In Phase 1, the English text is converted to an intermediate notation. First, the text is processed by tokenization and lemmatization and then labelled by a part of speech tagger. From the sequence of the part of speech terms in the sentence, the sentence structure of the input can be identified. Given an identified sentence structure, the knowledge base of transformational rules will generate the target sentence structure of the ASL sentence. Once this is ascertained, the intermediate notation is generated, by plugging the input lemmas into the target structure. The enhancement module to this phase is the paraphrase generator which uses deep learning. The function of this module is that, if a particular sentence structure is not identified by the transformational rule base, then the paraphrase module will generate an equivalent sentence with a different structure; hopefully one which is identified by the rule base. The paraphrasing will continue until a match is inevitably found. In phase 2, the intermediate notation must spawn the appropriate ASL animations. An animation database is created with animations corresponding to different lemmas. Therefore, the lemmas present in the intermediate notation will cause the respective ASL animation. The enhancement to this phase is done using a synonym network. By adding a synonym network, if a word is not found in the animation database, the synonyms of the word will be generated.

If the synonym is found in the animation database, the ASL animation of the synonym is performed. Therefore, this increases the scope of the animation database generated, without compromising on the overall meaning of the sentence.

In this way, traversing through the aforementioned phases and modules, the system described in this paper translates English text to ASL animation.

IV. ENGLISH TEXT TO INTERMEDIATE NOTATION

Phase 1 of the translation system converts English text to an intermediate notation. This is done in the following stages.

A. Tokenization and Lemmatization

The English text input is broken up into lexical units called tokens. Tokens can be considered as units of language such as words in a sentence in this case. The stream of tokens in the form of a list of words are then sent to the following stages for further analysis. The stream of tokens is then individually analyzed for morphological context and the root of the word, the lemma, is ascertained. In this way, a list of token and lemma tuples are formed and sent to the next stage.

B. POS Tagging

The The list of token lemma pairs goes through a part of speech (POS) tagger. The POS tagger then assigns a part of speech to each token. Table-I [7] shows the abbreviations used for the different parts of speech identified, formulated. This module also takes into consideration the context of a word in a sentence for ambiguous instances. For example, the word 'lift' in 'They lift weights' and 'They waited for the lift', has two different meanings. In the first sentence lift is identified as a verb, and in the second it is identified as a noun. These tags are then associated with the token lemma tuples.

Table-I: List of Part of Speech Tags

Abbreviation	Part of Speech	Example
CC	Coordinating Conjunction	and, but, or
CD	Cardinal Digit	1, one
DT	Determinant	the, a
IN	Preposition/Subordinating Conjunction	around, although
JJ	Simple Adjective	Smart
JJR	Comparative Adjective	smarter
JJS	Superlative Adjective	smartest
MD	Modal Verb	could, would
NN	Singular Noun	City
NNS	Plural Noun	cities
NNP	Singular Proper Noun	America
NNPS	Plural Proper Noun	Americans
PDT	Predeterminant	all of them
POS	Possessive Ending	student's
PRP	Personal Pronoun	I, he, she

C. Transformational Rule Base and Intermediate Notation Generation

The transformational rule base is a set of rules which when given a sentence structure of the source language, maps to a structure in the target language. The sentence structure is ascertained from the POS tags and searched for in the rule base. Once identified, the target structure file is indexed at the same index the source structure was found at to obtain the target structure. The rule base currently consists rules which covers most of the common structures of English sentences. The transformational rules were formulated upon study and research of the semantics of sign language from [8] [9] and [10]. This is how the target structure is ascertained. Consequently, the corresponding lemmas which correspond to the parts of speech in the target sequence are compiled. This gives the intermediate notation. This is similar to the 'glosses' notation, one of the accepted written scripts for ASL. The rule base works on a sentence by sentence transformation and can therefore be extended to paragraphs as well. Fig 2 shows the step wise conversion of an example sentence in English and the intermediate notation for the same.

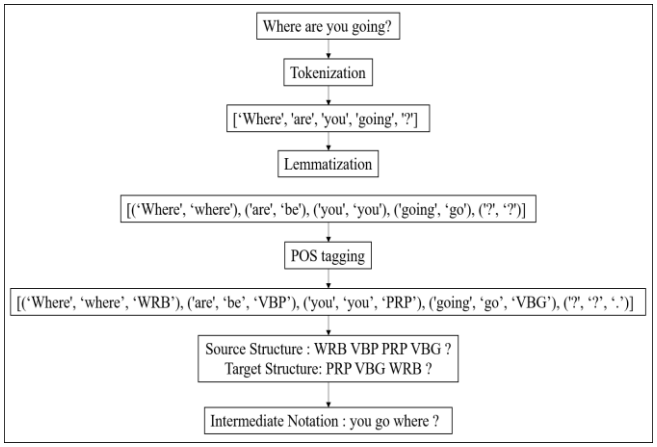

Fig. 2. Diagrammatic representation of the stages of Phase 1 of the system

V. INTERMEDIATE NOTATION TO ASL ANIMATION

Phase 2 of the translation system takes the intermediate notation and spawns appropriate ASL animations. The animation database was formulated such that the lemmas identified in the intermediate notation could call the appropriate animation. A 3D animation avatar was created using the 'Make Human' software, exported with a detailed hand rigging module to 'Blender' to produce the animation. The hands are made slightly larger than the standard size for clarity. Videos of the signs are obtained from various YouTube databases such as [11]. These videos are then sent to the 'Open Pose' an open source pose recognition software developed by researchers Carnegie Mellon University to identify the movements of the bones of the hand and the facial expressions for the sign in the video. Once identified, the bone placement is mimicked onto the rig of the 3D avatar created. Fig 3 depicts the creation of the sign for the letter 'L' in this process. The lemmas in the target sequence calls the appropriate animation set generated by this method and produces the ASL animation for the sentence to be translated.

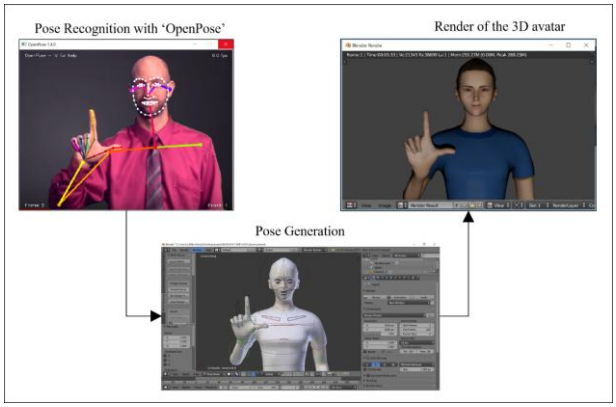

Fig. 3. Pose estimation using deep learning and rendering of the 3D avatar

VI. Deep Learning Paraphrase Module

The structure of an input sentence is ascertained by the sequence of its POS tags. If the structure of the input sentence does not match any of the plausible source structures anticipated by the transformational rule base, a paraphrase of the sentence will be generated. This paraphrase will convey the same meaning but will have a different structure. This new structure may be present in the transformational rule base. Therefore, paraphrases are generated of the input sentence until one of the paraphrases generated has a structure that is identified by the rule base. In this way, the structure of the input sentences is no longer restricted to the rule base defined.

A recent and significant contribution to the topic of paraphrase generation was shown in [12], in which they use stacked residual LSTM (Long Short-Term Memory) Networks to perform Neural Paraphrase Generation. Paraphrase generation is a task that must consider long range dependencies. In so doing, however, RNNs either suffer from the vanishing or exploding gradient problems. The LSTM structure is known to bolster the networks so as to be immune to these problems. So far stacking of LSTMs has been limited to only a few layers due to the difficulty in training. Upon adding residual connections between multiple stacked LSTM networks, we can stack more layers of LSTM successfully without overfitting the data. The paraphrasing module in this paper, along with some common mappings, is developed as in [12], the architecture of which is described in Fig 4 taken from [12].

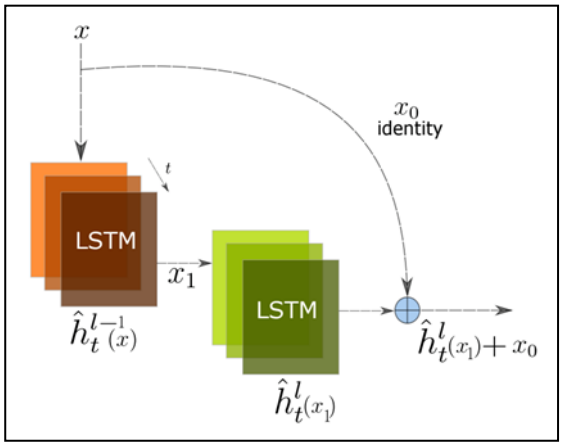

Fig. 4. Diagrammatic representation of the stacked residual LSTM network for paraphrase generation

VII. Synonym Network

If the sign for a particular lemma is not found in the animation database created, then the synonym network will search for the synonyms of the lemma and check whether any of the synonyms have signs in the animation database. The synonym set is retrieved from the wordnet module [13], in addition to a set of restricted rules to expand the same. In this way, the words that can appear in the input sentence is not restricted to the signs in the animation database but can be extended to a network of their synonyms as well.

VIII. Results

A. Detailed Results of One Example Sentence

The following describes the working of an example sentence in the system SILANT. The sentence is 'Dubai is a smart city.'. It is tokenized into words, lemmatized and tagged with parts of speech as shown in Fig 5. The source structure and target structure are ascertained, and the intermediate notation is generated. The intermediate notation spawns the ASL animations as shown in Fig 6. The word 'Dubai' is recognized as a proper noun and is therefore spelled letter by letter while the rest of the words produce their appropriate signs. The working of the paraphrasing module is shown by the sentence 'Dubai is a city that is smart' which produces the same animation for the sentence. The working of the synonym network is shown by the sentence 'Dubai is a smart metropolis' which produces the same animation. These two enhancements are shown diagrammatically in Fig 7 and 8 respectively.

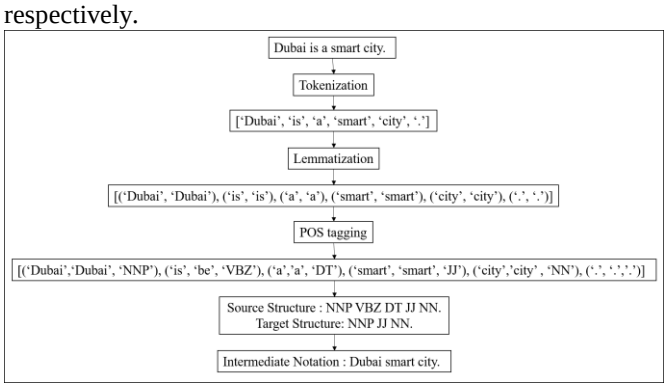

Fig. 5. Phase 1 conversion of the sentence "Dubai is a smart city"

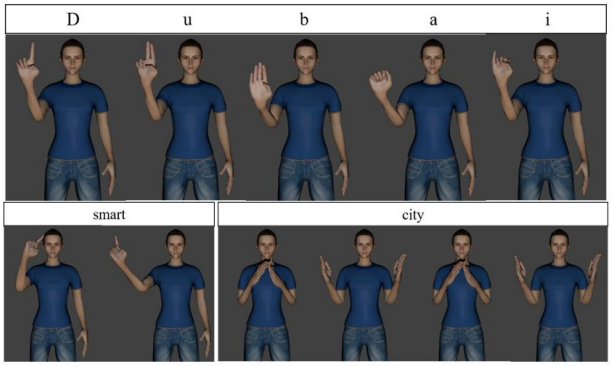

Fig. 6. ASL animation output for the sentence "Dubai is a smart city"

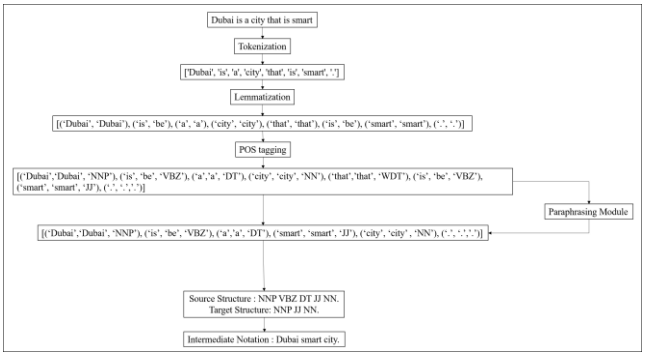

Fig. 7. Working of the paraphrase module

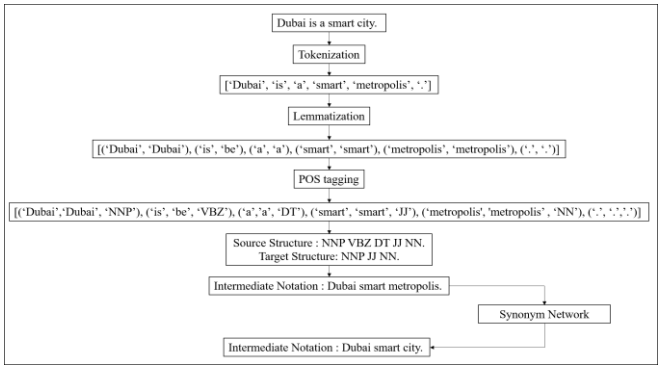

Fig. 8. Working of the synonym network

B. Results for SIGNUM dataset

The working of SILANT was evaluated using the SIGNUM dataset [14], which included 780 common phrases that had to be translated to sign language. Table II illustrates the overall results. Partial translation indicates not finding the word in the dictionary, but fingerspelling the words in the case of potential proper nouns. The accuracy of complete translation by SILANT without the enhancement modules was 54.10% which was increased to 78.72% (for complete translation) and 91.67% (for partial translation) upon inclusion of the enhancement modules. Furthermore, the accuracy of recognizing words increased from 60.12% to 91.82% upon the inclusion of the wordnet module.

Table-II: Performance of SILANT on the SIGNUM dataset

Total Number of Sentences	780
Sentences Translated Completely (without enhancement modules)	422
Sentences Translated Completely (with enhancement modules)	614
Sentences Translated Partially (with enhancement modules)	715
Accuracy of complete translation without enhancement modules	54.10%
Accuracy of complete translation with enhancement modules	78.72%
Accuracy of partial translation with enhancement modules	91.67%

C. Test Case Results

Table III illustrates in detail the results produced by SILANT for 15 test cases. It shows the input sentence and the intermediate notation generated. It also mentions whether the sentence has been paraphrased and whether the synonym network module was used.

Table-III: Performance of SILANT on the SIGNUM dataset

Input Sentence	Intermediate Notation	Paraphrase Module	Synonym Network
The boy kicked the ball.	BOY KICK BALL	No	No
The girl climbed a tree.	GIRL CLIMB TREE	No	No
Mark gave a witty response.	MARK GIVE SMART ANSWER	No	Yes Witty → Smart Response → Answer
The ball was kicked by the boy.	BOY KICK BALL	Yes The ball was kicked by the boy → The boy kicked a ball	No
Where do you live?	YOU LIVE WH?	No	No
Where do you reside?	YOU LIVE WH?	No	Yes Reside → Live
I'm hungry	I HUNGRY	No	No
I'm famished	I HUNGRY	No	Yes Famished→Hungry
What is your name?	YOU NAME WH?	No	No
She is both smart and beautiful.	SHE SMART AND BEAUTIFUL	Yes She is both smart and beautiful → She is smart and beautiful	No
The cake was made by the baker.	BAKER MAKE CAKE	Yes The cake was made by the baker → The baker made the cake	No
I like eating ice-cream.	I LIKE EAT ICE-CREAM	No	No
I can either go by the stairs or by the elevator.	I GO STAIRS OR ELEVATOR.	Yes I can either go by the stairs or by the elevator. → I can go by the stairs or by the elevator	No
Do you comprehend ?	YOU UNDERSTAND?	No	Yes Comprehend → Understand
Has your summer vacation commenced?	YOUR SUMMER VACATION START?	No	Yes Commenced →Started

IX. CONCLUSION

This paper describes the working of SILANT (SIgn LANguage Translator), a machine translation system that converts English text to American Sign Language (ASL). An enhanced rule-based system is used to produce ASL animation as the output with English text as the input,

employing concepts of Natural Language Processing (NLP) enhanced using Deep Learning algorithms.

Analyzing the results of the system, the future work that could be done includes adding an English speech to text converter so that SILANT may become an English speech to ASL translation system. SILANT may find its application in fields such as ASL captioning in online videos or even live educational lectures and various other customer service portals. This will enable technology to be accessed by even those who are differently abled thus working towards creating a more inclusive society.

REFERENCES

1. T. E. Allen, "Who are the deaf and hard-of-hearing students," Gallaudet University, Washington, DC, 1994.
2. T. Hanke, "HamNoSys—Representing sign language data in language resources and language processing contexts," in 4th International Conference on Language Resources and Evaluation, LREC, Lisbon, 2004.
3. L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler and M. Palmer, "A Machine Translation System from English to American Sign Language," in 4th Conference of the Association for Machine Translation in the Americas, AMTA, Cuernavaca, 2000.
4. T. Veale and A. Conway, "Cross modal comprehension in ZARDOZ an English to sign-language translation system," INLG '94 Proceedings of the Seventh International Workshop on Natural Language Generation, pp. 249-252, 21-24 June 1994.
5. P. Vij and P. Kumar, "Mapping Hindi Text To Indian sign language with Extension Using Wordnet," in Proceedings of the International Conference on Advances in Information Communication Technology & Computing, Bikaner, 2016.
6. O. H. Al-Barahamtoshya and H. M. Al-Barhamtoshy, "Arabic Text-to-Sign (ArTTS) Model from Automatic SR System," in 3rd International Conference on Arabic Computational Linguistics, Dubai, 2017.
7. Penn Treebank P.O.S. Tags. [Online]. Available: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html. [Accessed: 08-Feb-2019].
8. P. Boudreault and R. I. Mayberry, "Grammatical processing in American Sign Language: Age of first-language acquisition effects in relation to syntactic structure," Language, Cognition and Neuroscience, vol. 21, no. 5, pp. 608-635, 2006.
9. S. Zucchi, "Formal Semantics of Sign Languages," Language and Linguistics Compass, vol. 6, no. 11, pp. 719-734, 2012.
10. C. Valli, L. Ceil, K. J. Mulrooney and M. Villanueva, Linguistics of American Sign Language, 5th Ed.: An Introduction, Washington D.C.: Gallaudet University Press, 2005.
11. "ASL THAT," 10 October 2009. [Online]. Available: https://www.youtube.com/user/chsasl. [Accessed 10 March 2019].
12. A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu and D. Farri, "Neural Paraphrase Generation with Stacked Residual LSTM Networks," in 26th International Conference on Computational Linguistics, Osaka, 2016.
13. Princeton University, "WordNet," Princeton University, 2010. [Online]. Available: https://wordnet.princeton.edu/. [Accessed 9 April 2019].
14. SIGNUM Database. [Online]. Available: https://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/. [Accessed: 09-Apr-2019].

AUTHORS PROFILE

Lalitha Natraj is an undergraduate student from Birla Institute of Technology and Science, Pilani, (BITS Pilani), Dubai Campus. Currently pursuing a degree in Computer Engineering, she is interested in the field of Artificial Intelligence, particularly Natural Language Processing. With a passion for research backed by a strong academic background and proficient coding skills, she aspires to pursue a master's degree in the above field and make a notable contribution to the field of AI one day. She is also a published author of two books and a successful youtuber with a channel dedicated to teaching core concepts of Computer Science.

Dr. Sujala D. Shetty is working as Assistant Professor and Head of Department in the Department of Computer Science at BITS Pilani, Dubai Campus, since 2002. She completed her Ph.D from BITS Pilani, Rajasthan in 2010. She has 22 publications in various journals and international conferences. Prior to joining BITS Pilani, Dubai Campus, she has worked as a Lecturer in Computer Science Department of MIT, Manipal, from 1997 to 2002. Her current areas of research interests include Big Data, Database Applications, Web Services, Artificial Intelligence, Network Security. She has served as Thesis examiner for various universities. She has guided a number of undergraduate projects and Thesis as well as Dissertation of ME and MBA students. She is currently guiding two PhD scholars. She is the faculty in charge of the ACM student chapter at BITS Pilani, Dubai Campus.