GRADUATE CERTIFICATE IN PRACTICAL LANGUAGE PROCESSING MODULE REPORT

Machine-Translation-Hindi-to-English , FAQ & Translation Chatbot

Institute of Systems Science, National University of Singapore, Singapore 119615

ABSTRACT

This investigation is designed to shed light on identification and translation of idiomatic expression from Hindi to English language. The main problem encountered in idiomatic translation was investigated. Identification of idiom is utmost important resource for machine translation system and research. Machine translation system has been developed for many languages such as we have google translator, Bing translator etc. But they are failed to provide the correct translation of sentences containing idioms. The idiom parallel corpus was manually created to test the generated resource. The sentences containing idioms are translated with google translate system and noticed that it does not provide figurative meaning which makes the translation of idioms rather difficult than any text translation. They are the real challenge for machine translation from the preliminary stage of machine translation development. A lot of research has been done for extraction and translation of text in many languages, but no significant research has been captured in Hindi to English idiom translation. In this paper we have given a rule based approach for the identification and translation of idiom using machine translation. The aim of a proper idiom translation is achieving equivalent sense and provide figurative meaning, strategies, cultural aspects and effects. The output is evaluated manually for intelligibility and accuracy. Further This Hindi to English idiom translation system can be expanded for other language pairs to improve their translation by encapsulating correct idiom translation with their ordinary translation.

1. INTRODUCTION

Machine Translation automatically converts one natural language to another through computer. It converts the text of source language (SL) preserving its meaning to the fluent text in target language (TL). Machine translation is a computer application for converting the text from one language to other with or without human assistance as it may require a pre-editing and a post-editing phase. An Idiom is phase or expression that has a figurative meaning. An idiom's figurative meaning is different from the literal meaning. Meaning of an idiom is not predictable from usual meaning of its constituent elements. Every language has its own set of idioms, English and Hindi are abundant in idioms. Idioms are important part of conversation and are frequently used in wide variety of situations from friendly conversations to business and more formal and written context. An idiomatic expression may convey a different meaning, that what is evident from its words thats why they are hard to translate into other language. Therefore, it is important to identify the idioms and then replace them with the suitable and idiom carry appropriate meaning in any other language while translation. This is still an important topic for research and serious drawback of many machine translators like google.

The Good

translation is necessary because it can significantly and positively impact its benefactors

Obtaining Information By Translating Other Sources

Social media translation can be used as inputs to aid the building of a brand and ensure relevant decision-making.

Benefits of social media translation for businesses

On the other hand, social media translation can also be used as outputs to improve the popularity and accessibility of the business.

Reach More People

Translating the original social media posts into ones of different languages is incredibly beneficial to an enterprise. This is because the translated work can now be read by many more people who may not be fluent in the original language.

Cater to Specific Audiences

Instead of broadening the market holistically, you can also choose specific audiences to engage in your posts. Translating and speaking using their everyday vernacular is one way to spark their interest.

What happens to brands not utilizing social media translation?

However, as much as it can give benefits to those that do it, it can also be a downfall for those that choose not to translate their social media posts. This is especially true for businesses that are just starting out and need exposure to fuel them.

Machine-Translation-English-to-Hindi

The model translates English text to Hindi text with the help of LSTM. The project was implemented in Keras Framework on TensorFlow. An encoder was used to convert the English phrases to feature vectors that can be trained upon and a decoder converts the output vector back to normal Hindi text.

Encoder - Model

Encoder takes the English data as input and converts it

into vectors that is passed to an LSTM model for training. We discard the encoder output and only keep the states.

Decoder - Model

The decoder takes in Input the states of the encoder and the Hindi data points corresponding to the English input of Encoder. It trains an LSTM to produce the translated phrase in output. The decoder used SoftMax layer.

2. PROBLEM STATEMENT

translation is necessary because it can significantly and positively impact its benefactors On the other hand, social media translation can also be used as outputs to improve the popularity and accessibility of the business. Instead of broadening the market holistically, you can also choose specific audiences to engage in your posts. Translating and speaking using their everyday vernacular is one way to spark their interest.

Build an domain adaptive NMT (Neural MT) system when training data (parallel sentences in the concerned source and target language) is available in a domain. However, tested on some other domain data.

3. THEORY

Attention model : Attention is an interface between the encoder and decoder that provides the decoder with information from every encoder hidden state. With this setting, the model is able to selectively focus on useful parts of the input sequence and hence, learn the alignment between them. This helps the model to cope effectively with long input sentences.

Word Embeddings : Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. They are a distributed representation for text that is perhaps one of the key breakthroughs for the impressive performance of deep learning methods on challenging natural language processing problems.

Encoder Vector : This is the final hidden state produced from the encoder part of the model. This vector aims to encapsulate the information for all input elements in order to help the decoder make accurate predictions. It acts as the initial hidden state of the decoder part of the model.

Sequence to Sequence model: The "sequence-to-sequence" neural network models are widely used for NLP. A popular type of these models is an "encoder-decoder". There, one part of the network encoder encodes the input sequence into a fixed-length context vector. This vector is an internal representation of the text. This context vector is then decoded into the output sequence by the decoder.

4. LINGUISTIC BACKGROUND

HINDI

Hindi language is derived from Hindustani which is derived from Sanskrit language. It contains much vocabulary from Sanskrit language and is also written as such. It is the official language of India, where majority of population communicates using this language and it is 4th most spoken language in the world. It used to be written in Brahmi script but now it is written in Devanagari script. Devanagari consists of 11 vowels and 33 consonants and is written from left to right. Unlike Sanskrit, Devanagari is not entirely phonetic for Hindi, especially failing to mark schwa dropping in spoken Standard Hindi.

ENGLISH

English is a West Germanic language, and 3rd most spoken language of the World. It is closely related to Frisian languages but vocabulary is influenced by other Germanic languages. There are noticeable variations among the accents and dialects used in different countries. It is mostly analytic pattern with little inflection, a fairly fixed SVO word order and a complex syntax.

5. LITERATURE REVIEW

In this section the main focus is on the work done in the Indian context instead of discussing idiomatic translation. Idiomatic Machine Translation efforts in Indian context dated to October 2012 with Gaule & Josanal [3] in their research the investigation is designed to shed light on the identification and translation of idiomatic expressions from English to Hindi is analysed. They gave different strategies of idiom in machine translation Using an idiom of similar meaning and form, Using an idiom of similar meaning but dissimilar form, Using an Idiom Translation by paraphrase, Using an Idiom Translation by Omission and Online MT Systems There are following MT systems that have been developed for various natural language pair.

Systran is a rule based Machine Translation System developed by the company named Systran. It was founded by Toma in 1968. It offers translation in about 35 languages. It provides technology for Yahoo! Babel Fish and it was used by Google till 2007. Design overview- identification of idiom and process them and the processed sentence will be used as input by translation system. The result for evaluation were-30% sentences were correctly translated by sentences directly obtained from goggle translation system and 70% sentences were correctly translated by sentences preprocessed by our system and then translated from goggle translation. Further with Rajesh Kumar Chakrawarti, Himani Mishra, Dr. Pratosh Bansal [4] they witnessed several significant advancements in Natural Language Processing , which has let text and speech processing to make huge gateway to world-wide information source [5].

The paper focuses on the techniques and approaches like corpus-based, rule-based, direct and hybrid approach. used for machine translation systems together with their example systems. They identified the problem of Structural Divergences, Approach used and ambiguity, cultural problem and

named entity. There were several limitations identified such as.

Dictionary used

- Translation will be greatly affected by the depth and richness of dictionary used.
- As it is a machine, failure of the machine can't be predicted. Like all other system, it may crash down at any instance.
- Idioms are difficult to interpret as they point to some other meaning than the words used. In this survey paper, they studied various MT approaches, techniques, and many machine translation systems together with their benefits and limitations in a longitudinal and latitudinal way. Further, approaches of idiom translation are described by Rajesh Kumar Chakrawarti, Himani Mishra, Dr. Pratosh Bansal
- This research paper proposes a different system architecture for idiom translation from Hindi to English. This architecture resembles a Rule-based approach in which Transfer-based method is used which converts the idioms having similar meaning and either similar or dissimilar form from Hindi to Tokens and then in English and Interlingua based method is also used which converts the typical idioms whose meaning is not given as it is in the used database to their simple meanings and then they are converted in English. This architecture contains two phases Phase I- Comparison phase, in which the input is compared to the database Phase II- Translational phase, in which the translation happens. This work is a modified version of Machine translational system which can be embedded with other machine translational systems to get better results. Further Survey of machine translation system in India is given by Garje and Kharate
- They focused on different approaches used in the development of Machine Translation Systems and also briefly described some of the Machine Translation Systems along with their features, domains and limitations. They gave brief history of machine translator system at International level staring from 1948 to 2010. They listed machine translation systems and various approaches used in developing these systems.
- Direct Machine Translation Systems - Anusaaraka systems among Indian Languages (1995), Punjabi to Hindi MT System (2007, 2008), Web based Hindi-to-Punjabi MT System (2010), Hindi-to-Punjabi MT System (2009, 2011).
- Transfer-Based MT Systems - Mantra MT (1997), MANTRA MT(1999), An English–Hindi Translation System (2002), MAT (2002), Shakti (2003), English-Telugu MT System (2004), Telugu-Tamil MT System (2004), OMTrans (2004), The MaTra System (2004, 2006), English-Kannada machine-aided translation system (2009), Tamil-Hindi Machine-Aided Translation system (2009), Sampark System: Automated Translation among Indian Languages (2009).
- Interlingua Machine Translation Systems - ANGLABHARTI (2001), UNL-based English-Hindi MT System (2001), AnglaHindi (2003).
- Hybrid Machine Translation Systems
- Example Based Machine Translation (EBMT) Systems
- Statistical Machine Translation Systems.

In this paper author described MT techniques in a longitudinal and latitudinal way with an emphasis on the MT development for Indian languages as well as non-Indian languages and concluded that almost all existing Indian language MT systems are based on rule-based, hybrid and statistical approaches. D. Brar & R. Kaur [8] discussed the problems associated with the idiomatic translation. They presented the definition of idioms to see what they are. Then, classified the idioms into different categories and in the end, gives some techniques and procedures to translate them. They grouped idioms into five categories of colloquialisms, proverbs, slang, allusions and phrasal verbs. And concluded that idioms are arguably the most complex and problematic task for translators.

Sequence to Sequence Learning with Neural Networks

Sequence to Sequence (often abbreviated to seq2seq) models is a special class of Recurrent Neural Network architectures that we typically use (but not restricted) to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc.

- a sequence to sequence model aims to map a fixed length input with a fixed length output where the length of the input and output may differ.
- The model consists of 3 parts: encoder, intermediate (encoder) vector and decoder.
- The power of this model lies in the fact that it can map sequences of different lengths to each other. As you can see the inputs and outputs are not correlated and their lengths can differ. This opens a whole new range of problems which can now be solved using such architecture.
- This paper addresses the single node bottleneck problem in two ways: first by using a bidirectional LSTM for input and second by introducing an alignment model, a matrix of weights connecting each input location to each output location. This can be thought of as

an attention mechanism that allows the decoder to pull information from useful parts of the input rather than having to decode a single hidden state.

- A potential issue with this encoder–decoder approach is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector. This may make it difficult for the neural network to cope with long sentences, especially those that are longer than the sentences in the training corpus.

- An attentional mechanism has lately been used to improve neural machine translation (NMT) by selectively focusing on parts of the source sentence during translation. This paper examines two simple and effective classes of attentional mechanism: a global approach which always attends to all source words and a local one that only looks at a subset of source words at a time. In this paper Minh-Thang Luong, et al. propose an attention mechanism for the encoder-decoder model for machine translation called "global attention."

- It is proposed as a simplification of the attention mechanism proposed by Bahdanau, et al. in their paper "Neural Machine Translation by Jointly Learning to Align and Translate." In Bahdanau attention, the attention calculation requires the output of the decoder from the prior time step. Global attention, on the other hand, makes use of the output from the encoder and decoder for the current time step only.

- The model evaluated in the Luong et al. paper is different from the one presented by Bahdanau, et al. (e.g. reversed input sequence instead of bidirectional inputs, LSTM instead of GRU elements and the use of dropout), nevertheless, the results of the model with global attention achieve better results on a standard machine translation task.

6. DATASET

Context

The dataset consist of 1000000 English phrases along with their Hindi translations. The data is given in utf-8 format.

Pre-Processing

- Data was cleaned manually, as the english sentence and their corresponding hindi translation was not parallel.

- Lots of spelling mistakes, around 9000 words were found which were not present in the Glove vocabulary and most of these words are due to spelling mistakes while others are due to name of city, place etc.

- we build our vocabulary of unique words (and count the occurrences while we're at it)

- we replace words with low frequency with ¡UNK¿

- create a copy of conversations with the words replaced by their IDs

- we can choose to add the ¡SOS¿ and ¡EOS¿ word ids to the target dataset now, or do it at training time

- ¡PAD¿: During training, we'll need to feed our examples to the network in batches. The inputs in these batches all need to be the same width for the network to do its calculation. Our examples, however, are not of the same length. That's why we'll need to pad shorter inputs to bring them to the same width of the batch

- ¡UNK¿: For training the model on real data, the resource efficiency of model can be vastly improved by ignoring words that don't show up often enough in the vocabulary to warrant consideration. By replacing those words with ¡UNK¿.

- Teacher forcing: Models work a lot better if we feed the decoder our target sequence regardless of what it's timesteps actually output in the training run.

Each record in the data-set consists of the following attributes

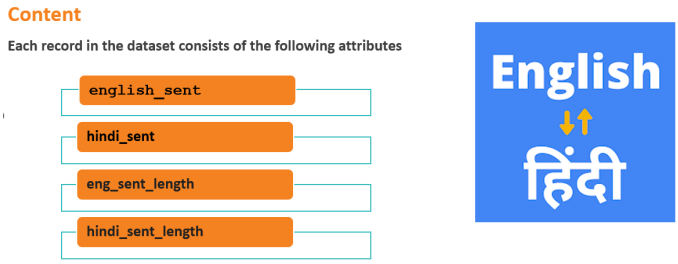

Fig. 1. Machine-Translation-Hindi-to-English

7. METHODOLOGY

For the better translation of idioms from Hindi to English, a custom algorithm is used. This algorithm requires the input file of list of all idioms and their corresponding English idioms rather than their word to word conversion unlike google translate. It then checks the sentence given as input by the user for a matching idiom in the file and replaces with its English counterpart, then translates the sentence, resulting in appropriate replacement of the idiom. This algorithm works in steps, as follow:

Collection of parallel idiom corpus

Parallel Database of 1000 Hindi and English idiom is created manually from different source and a text file is created. A text file is provided to the program as a collection of idioms in Hindi and English. Each line of this file stores Hindi idiom and its English counterpart, separated by a hyphen. A single line can only have one hyphen present. The program will read the file, line by line making two lists on idioms. First list will store Hindi idioms and the second list contains matching English idioms at the same indices.

Idiom identification

Input is taken as a Hindi sentence containing an idiom. This sentence is then split into word stored in a list. This list is matched to all the idioms in the list, word by word. At the end of the loop, a final idiom is identified.

Idiom replacement

Once the idiom is identified, Using the index of the final idiom in first list, the idiom in the second list is extracted, and the idiom in the sentence is replaced by it.

Translation of idiomatic sentence

Once the idiom is replaced in the sentence, the sentence is then translated to English using mtranslate module for python, which works as Google Translator. Since, the final sentence contains Hindi words and the correct English translated idiom, the output sentence by the module considers the idiom and put it as it is, giving a properly translated idiomatic sentence.

8. MODEL RESULT

- There were approximately 25000 pairs of sentences of each domain.
- Approximately 9000 words were misspelled in the data
- Approximately 2000 sentences were mismatched. (corresponding hindi translation was not matching) And other than these there were many wrong translations. Like presence of “Complete it”, “repeated it”, etc.
- After cleaning, the final dataset has 49896 pairs of sentences (with 8000 misspelled words).

Sequence to Sequence model:

- The “sequence-to-sequence” neural network models are widely used for NLP.
- A popular type of these models is an “encoder-decoder”.
- There, one part of the network encoder encodes the input sequence into a fixed-length context vector.
- This vector is an internal representation of the text. This context vector is then decoded into the output sequence by the decoder.

Sequence to Sequence (multi-layer)

- Stacked 4 LSTM encoder layer and 2 LSTM decoder layer
- The input sequence is passed into the embedding layer.
- The output from the embedding layers is fed into the encoder. ...
- The cell state and hidden state are then fed into the decoder, which is also a LSTM based network, along with the decoder input.

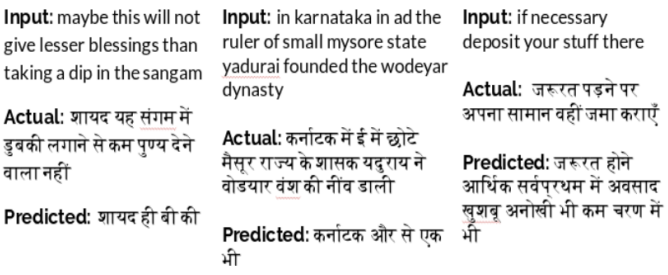

Fig. 2. Stacked 4 LSTM encoder

Sequence to Sequence Bi-directional model

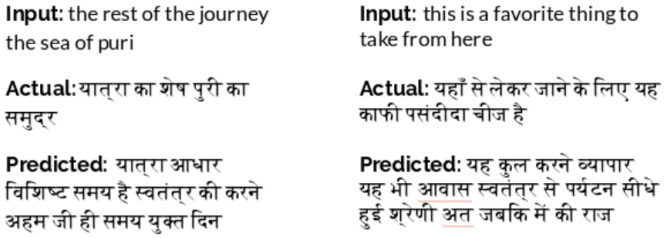

Fig. 3. Sequence to Sequence Bi-directional model

Attention Model

The attention mechanism in NLP is one of the most valuable breakthroughs in Deep Learning research in the last decade. It has spawned the rise of so many recent breakthroughs in natural language processing (NLP), including the Transformer architecture and Google’s BERT

Attention mechanisms enhance deep learning models by selectively focusing on important input elements, improving prediction accuracy and computational efficiency. They prioritize and emphasize relevant information, acting as a spotlight to enhance overall model performance.

Challenges

- Despite of this challenge other challenge faced is data was not clean.
- Time taken for training is long (more than 32 hrs). So making multiple models was relatively difficult.

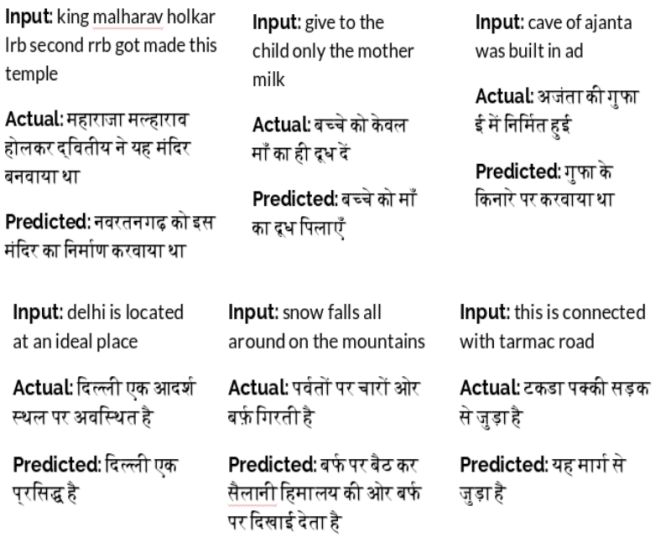

Fig. 4. Attention Model

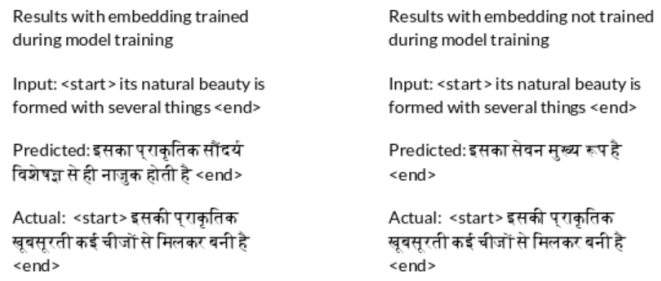

Fig. 5. Attention results with Glove Embedding

9. FAQ CHATBOT

FAQ chatbots are bots designed to answer common questions people usually ask about a company's products or services. Usually, FAQ chatbots are used on websites, ecommerce stores, or customer service apps.

These bots, thanks to natural language processing, operate on the question-answer format which creates a real conversation vibe between the bot and the user, significantly improving the quality of customer service. Also, with their ability to automate, they can help companies save time, money, and effort spent on laborious tasks like responding to repeat questions.

Simply put, an FAQ chatbot is like an FAQ page, but driven by artificial intelligence.

Benefits of FAQ chatbot

An FAQ chatbot will be your great ally while providing great customer service automation. As the FAQ bots are based on artificial intelligence and natural language processing (NLP), it doesn't require any effort from your support team.

BLEU-1:	0.000484
BLEU-2:	0.021989
BLEU-3:	0.080511
BLEU-4:	0.148287
Individual 1-gram:	0.000484
Individual 2-gram:	1.00001
Individual 3-gram: 1.000000	1.00001
Individual 4-gram: 1.000000	1.00001

Fig. 6. Model Result

But here are four main benefits of getting an FAQ chatbot that will help you decide if it's worth it. Let's go through them briefly.

Different kinds of FAQ chatbots

Based on the way FAQ chatbots operate, we can divide them into three main categories:

- Rule-based chatbots)
- Independent (keyword) chatbots
- Natural Language Processing (NLP) chatbots

Let's briefly describe each type, so it will be easier for you to understand the differences between them and choose the right one for your business.

Rule-based chatbots

This is the most basic FAQ chatbot on our list. Also called "button-based" or "menu-based", rule-based chatbots use a series of defined rules to identify and answer users' questions. These bots are also known as decision-tree bots because their conversations are guided by a decision tree. Users are given a set of predefined options which lead them to desired answers.

Remember that these bots have limitations. They won't answer any questions outside of the defined rules. So, if you want your bots to solve complex issues, you should consider more advanced options.

Independent (keyword) chatbots

These response chatbots use keywords to communicate with users. They deeply analyze your customers' queries and prepare the most appropriate answers. Independent bots are also called the "keyword" ones. That's because, when answering queries, these bots use exactly the same keywords as their users did. Thanks to that, response bots are able to find the best possible answers to users' queries.

Natural Language Processing (NLP) chatbots

NLP chatbots, also known as contextual chatbots, are one of the most technically advanced bots. Thanks to NLP and artificial intelligence, these bots can handle multiple requests from one user and simulate a human-like conversation.

Not only do these bots have the ability to learn on their own, but they are also able to understand language nuances and carry on conversations with your clients.

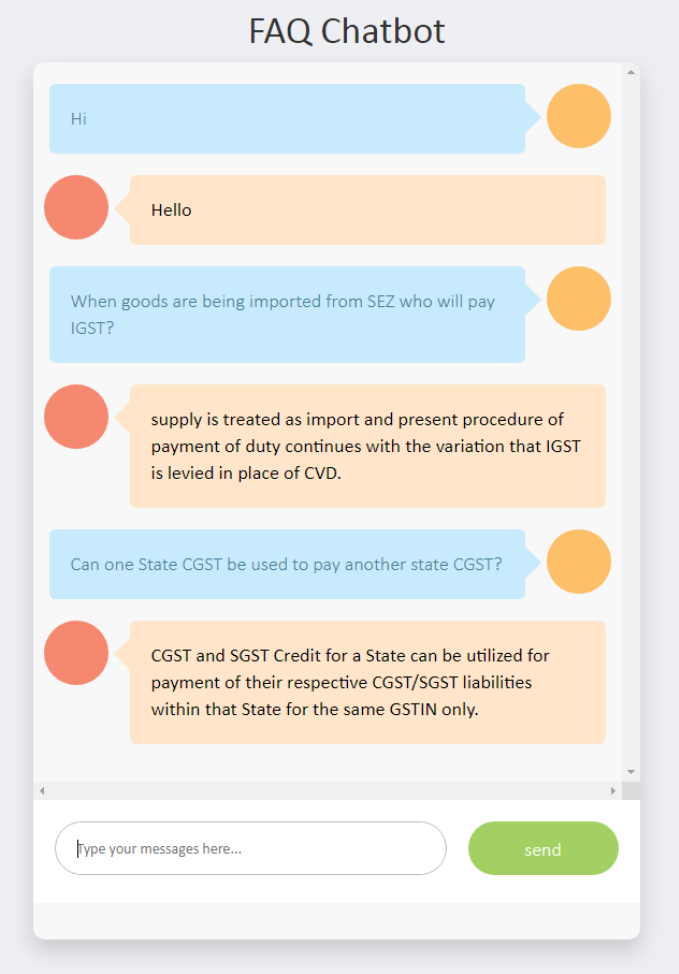

Fig. 7. FAQ Chatbot

Sequence to Sequence Bi-directional model

10. LEARNING OBJECTIVES

- Get familiar with class imbalance through coding.
- Understand various techniques for handling imbalanced data, such as Random under-sampling, Random over-sampling, and Near Miss.
- Apply the relevant models that need to be used for each task
- Apply the major guiding principles when choosing a model for a specific task within NLP
- Decide when to and when not to use neural network based or deep learning methods for a specific task within NLP

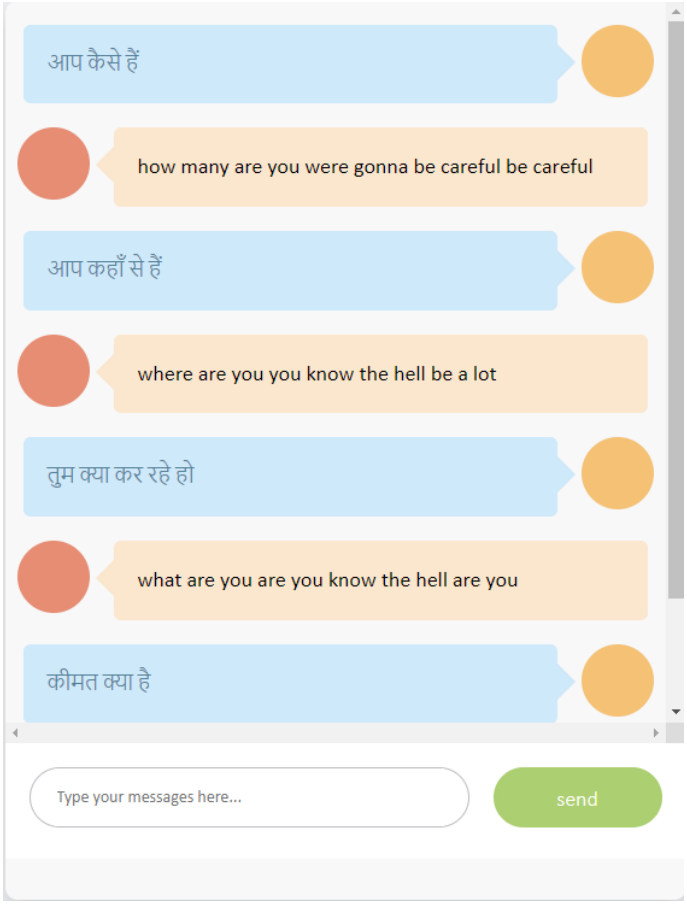

Fig. 8. Translation Chatbot

- Analyze the time complexity involved for a specific NLP algorithm
- Pre-process textual data into suitable representation for text analytics
- Build and evaluate language models using appropriate text processing techniques for tasks like document classification, topic modeling, information extraction, etc.
- Apply deep learning techniques on large amounts of textual data to obtain high quality models

11. REFERENCES

[1] D. Anastasiou, "Idiom Treatment Experiments in Machine Translation", Cambridge Scholars Publishing, 2010.

[2] The Oxford companion to the English language (1992:495f.).

[3] M. Gaule & Josan, G. S., "Machine Translation of Idioms from English to Hindi", International Journal Of Computational Engineering Research, 2(6), 2012.

[4] R. K. Chakrawarti, H. Mishra, P. Bansal, “Review of Machine Translation Techniques for Idea of Hindi to English Idiom Translation”, International Journal of Computational Intelligence Research, 13(5), 1059-1071, 2017.

[5] F. Ciravegna, S. Harabagiu, “Recent Advances in Natural Language Processing”, IEEE magazine, computer.org/intelligent, 2013.

[6] H. Mishra, R. K. Chakrawarti P. Bansal, “A New Approach for Hindi to English Translation”, International Journal on Computer Science and Engineering, Vol. 9 No.07, 0975-3397, Jul 2017.

[7] G. V. Garje G. K. Kharate, “Survey of machine translation systems in India”, International Journal on Natural Language Computing (IJNLC) Vol, 2, 47-67, 2013.

[8] D. Brar R. Kaur, “A Review of Transliteration system from English to Punjabi”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4 (7), 2277 128X, July 2014.

[9] V. Gupta, N. Joshi I. Mathur, “Approach for Multiword Expressions and Recognition Annotation in Urdu Corpora”, Image Information Processing (ICIIP), Fourth International Conference, 2017. IEEE 2017.

[10] V. Gupta, N. Joshi I. Mathur, “Design Development of Rule Based Inflectional and Derivational Urdu Stemmer Usal”, INBUSH-ERA-2015, 7-12, 2015. IEEE 2015.

[11] V. Gupta, N. Joshi I. Mathur, “Rule Based Stemmer in Urdu”, Computer and Communication Technology (ICCCT), Fourth International Conference, 2013. IEEE 2013.

[12] https://arxiv.org/abs/1409.3215 (Research Paper)

[13] http://www.manythings.org/anki/

[14] https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/

[15] https://machinelearningmastery.com/encoder-decoder-long-short-term-memory-networks/

[16] https://machinelearningmastery.com/develop-neural-machine-translation-system-keras/

[17] http://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

[18] https://towardsdatascience.com/nlp-sequence-to-sequence-networks-part-1-processing-text-data-d141a5643b72

[19] https://towardsdatascience.com/nlp-sequence-to-sequence-networks-part-2-seq2seq-model-encoderdecoder-model-6c22e29fd7e1

[20] https://nlp.stanford.edu/ johnhew/public/14-seq2seq.pdf

[21] https://www.analyticsvidhya.com/blog/2018/03/essentials-of-deep-learning-sequence-to-sequence-modelling-with-attention-part-i/

[22] https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html

12. AUTHOR

Thota Siva Krishna , e0943696@u.nus.edu