NATURAL LANGUAGE PROCESSING APPLICATIONS

Arushi Singhal (201516178)

B.Tech and MS by research in Building Science and Engineering

Simran Singhal (201516190)

B.Tech and

MS by Research in Building Science and Engineering

International Institute of Information Technology Hyderabad

“ Domain Based Neural Machine Translation”

TABLE OF Contents

CHAPTER	Content	Page No
	ACKNOWLEDGMENTS	3
CHAPTER 1	PROBLEM STATEMENT	4
CHAPTER 2	THEORY	5
CHAPTER 3	LITERATURE REVIEW	6 - 9
CHAPTER 4	MODELS AND RESULTS	10 - 14
CHAPTER 5	CHALLENGES	15
CHAPTER 6	REFERENCES	16

ACKNOWLEDGMENTS

We are really grateful that we managed to complete our NLP Applications semester project within the timeframe.

We would like to express our special thanks of gratitude to **Dr. Manish Srivastava** who gave us the golden opportunity to do this wonderful project, which also helped us in doing a lot of Research and helped us in exploring field of Natural Language Processing. We sincerely thank him for the guidance and encouragement in finishing this project and also for teaching us in this course.

We would like to express our gratitude to the PhD. scholar and Teaching Assistant of the course **Saumitra Yadav** for the support, guidance and friendly advice during the project work as well as for providing necessary information and dataset required for the project.

Problem Statement

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.

THEORY

1) 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.

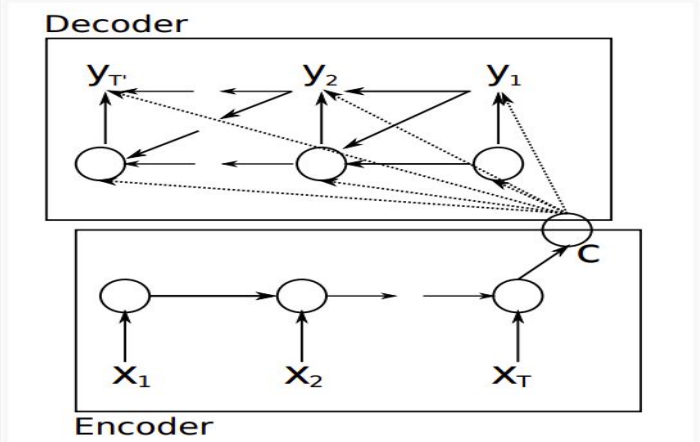
2) Word Embeddings :- Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. They are a distributed representation for text that is perhaps one of the key breakthroughs for the impressive performance of deep learning methods on challenging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

3) 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.

4) 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.

LITERATURE REVIEW

1) Sequence to Sequence Learning with Neural Networks
    a) https://arxiv.org/abs/1409.0473
    b) Summary:-
        i) 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.
        ii) The model consists of 3 parts: encoder, intermediate (encoder) vector and decoder.
        iii) 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.
        iv)



Encoding of Source Sequence to a Context Vector Which is Then Decoded
Taken from "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," 2014.

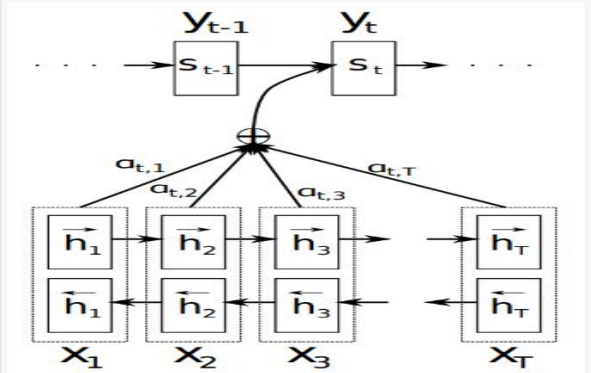
2) Neural Machine Translation by Jointly Learning to Align and Translate
    a) https://arxiv.org/abs/1409.0473
    b) Summary
        i) In any case, the background is that models that use RNN (specifically LSTM) encoders and decoders have done pretty

well on this task, but are limited in their ability to track long-term dependencies and even, tellingly, lose their ability to translate the end of long sentences correctly. The cause of this limitation is that this "basic encoder-decoder" architecture encodes everything about the input sentence in a single fixed-length vector. This is not ideal, since we expect intermediate hidden states to contain useful information.

ii) 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.

iii) 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.

iv)



Example of the Encoder-Decoder model with Attention
Taken from "Neural Machine Translation by Jointly Learning to Align and Translate," 2015.

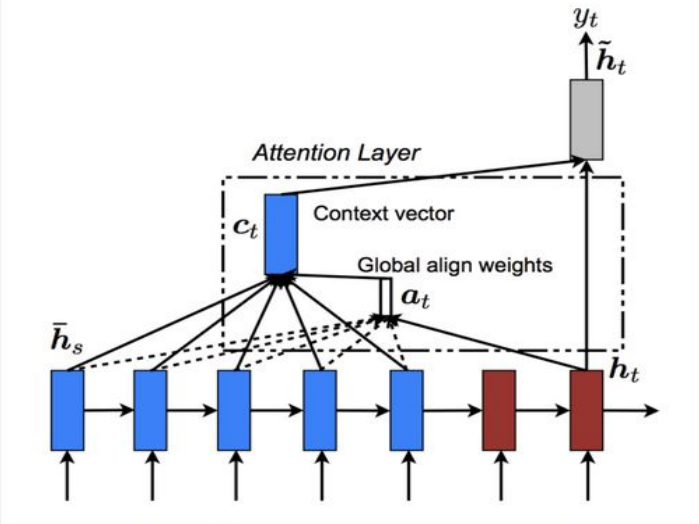
3) Effective Approaches to Attention-based Neural Machine Translation
   a) https://aclweb.org/anthology/D15-1166
   b) Summary
      i) An attentional mechanism has lately been used to improve

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.”

ii) 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.

iii) 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.

iv)



The depiction of Global Attention in an Encoder-Decoder Recurrent Neural Network.
Taken from "Effective Approaches to Attention-based Neural Machine Translation."

CHAPTER 4

MODELS AND RESULTS

4.1 Dataset

1) The dataset we worked with was of two domains (Healthcare + Tourism). (https://bit.ly/2UYNCkj)
2) There were approximately 25000 pairs of sentences of each domain.
3) Approximately 9000 words were misspelled in the data. (https://bit.ly/2WpflMG)
4) Approximately 2000 sentences were mismatched. (corresponding hindi translation was not matching)
5) And other than these there were many wrong translations. Like presence of “Complete it”, “repeated it”, etc.
6) After cleaning, the final dataset has 49896 pairs of sentences (with 8000 misspelled words), 24918 healthcare data :- https://bit.ly/2Jffbn6 (English) https://bit.ly/2DLWDHO (Hindi)

4.2 Pre-Processing

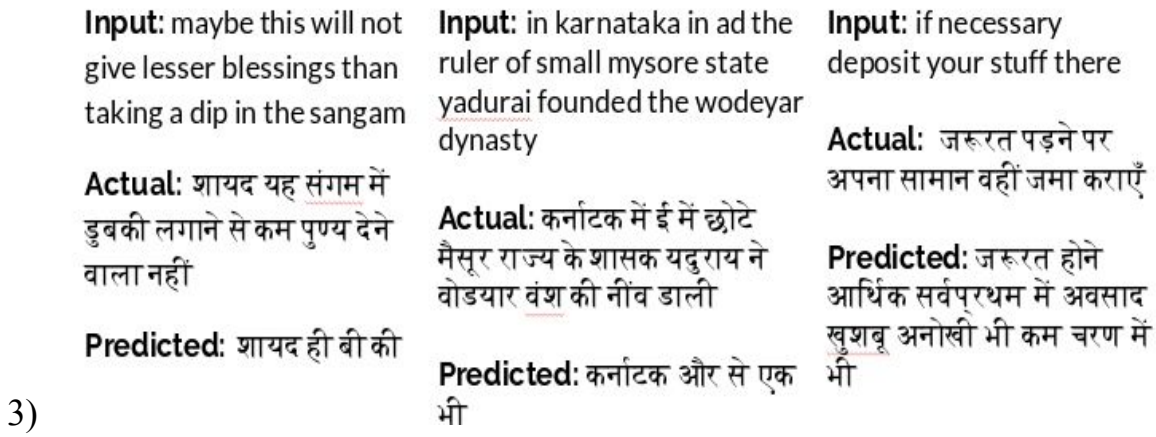
1) Data was cleaned manually, as the english sentence and their corresponding hindi translation was not parallel.
2) Lots of spelling mistakes, around 9000 words were found which were not present in the Glove vocabulary and most of these words are due to spelling mistakes while others are due to name of city, place etc.
3) we build our vocabulary of unique words (and count the occurrences while we’re at it)
4) we replace words with low frequency with <UNK>
5) create a copy of conversations with the words replaced by their IDs

6) we can choose to add the <SOS> and <EOS> word ids to the target dataset now, or do it at training time
7) <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
8) <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>.
9) 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.

4.3 Models and Results

4.3.1 Sequence to Sequence (multi-layer)

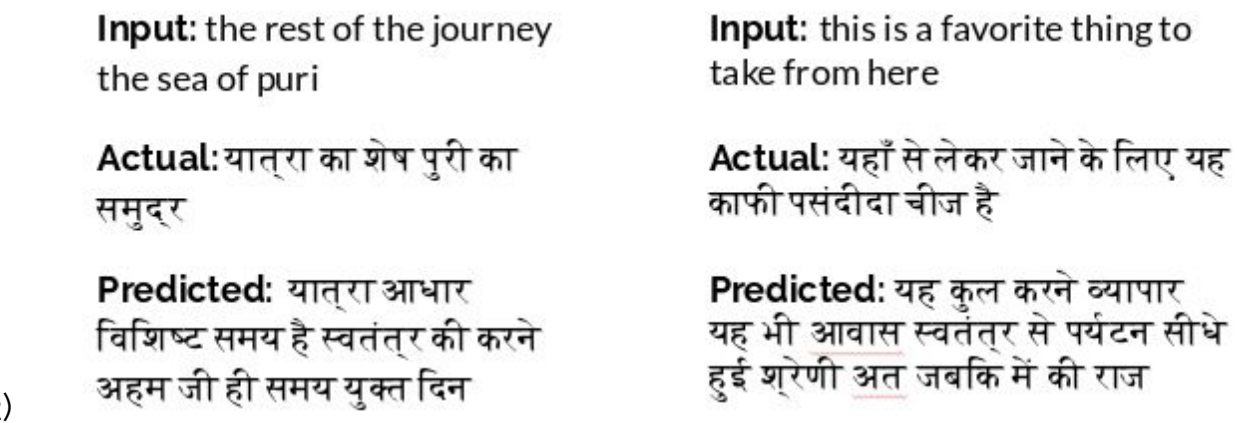
1) Code:- https://bit.ly/2YcNyj9
2) Stacked 4 LSTM encoder layer and 2 LSTM decoder layer
3)



4.3.2 Sequence to Sequence Bi-directional model

1) Code :- https://bit.ly/2GVM0mr

2)



4.3.3 Attention Model

1) Code Link (simple attention) :- https://bit.ly/2vz0fsa
2) Code Link (Attention + Bidirectional LSTM) :- https://bit.ly/2WvTHGk
3) Got good results comparatively
    a) BLEU-1: 0.000484
    b) BLEU-2: 0.021989
    c) BLEU-3: 0.080511
    d) BLEU-4: 0.148287
    e) Individual 1-gram: 0.000484
    f) Individual 2-gram: 1.000000
    g) Individual 3-gram: 1.000000
    h) Individual 4-gram: 1.000000

4)

Input: king malharav holkar lrb second rrb got made this temple	**Input:** give to the child only the mother milk	**Input:** cave of ajanta was built in ad
Actual: महाराजा मल्हाराव होलकर द्वितीय ने यह मंदिर बनवाया था	**Actual:** बच्चे को केवल माँ का ही दूध दें	**Actual:** अजंता की गुफा ई में निर्मित हुई
Predicted: नवरतनगढ़ को इस मंदिर का निर्माण करवाया था	**Predicted:** बच्चे को माँ का दूध पिलाएँ	**Predicted:** गुफा के किनारे पर करवाया था

5)

Input: delhi is located at an ideal place	**Input:** snow falls all around on the mountains	**Input:** this is connected with tarmac road
Actual: दिल्ली एक आदर्श स्थल पर अवस्थित है	**Actual:** पर्वतों पर चारों ओर बर्फ़ गिरती है	**Actual:** टकडा पक्की सड़क से जुड़ा है
Predicted: दिल्ली एक प्रसिद्ध है	**Predicted:** बर्फ पर बैठ कर सैलानी हिमालय की ओर बर्फ पर दिखाई देता है	**Predicted:** यह मार्ग से जुड़ा है

6)

Input: <start> the construction of these jain temples are unique in itself <end>

Actual: <start> इन जैन मंदिरों की संरचना अपने आप में खास है <end>

Predicted Translation

No Tourism data	5000 Tourism data	1000*2 Tourism data	15000 Tourism data
इन अंशों में शरीर के सबसे पहले एक तरह के रूप में है <end>	इन मंदिरों का उल्लेख अपने जीवन में ही स्थापित दिखाई देता है <end>	इन मंदिरों का निर्माण अपने ही मंदिर परिसर में ही हुआ है <end>	इन जैन मंदिरों का निर्माण कार्य करते हैं <end>

Input: <start> its natural beauty is formed with several things <end>

Actual: <start> इसकी प्राकृतिक खूबसूरती कई चीजों से मिलकर बनी है <end>

Predicted Translation

No Tourism data	5000 Tourism data	1000*2 Tourism data	15000 Tourism data
इसका प्राकृतिक चिकित्सा से छुटकारा मिलता है <end>	इसके प्राकृतिक सौंदर्य से ही प्राकृतिक है <end>	इसका प्राकृतिक सौंदर्य से है <end>	इसका प्राकृतिक चीजें साथ जा रही है <end>

7)

4.3.4 Attention results with Glove Embeddings

1) Codes :- https://bit.ly/2ZRM8MB and https://bit.ly/2vAtGKt

Results with embedding trained during model training	Results with embedding not trained during model training
Input: <start> its natural beauty is formed with several things <end>	Input: <start> its natural beauty is formed with several things <end>
Predicted: इसका प्राकृतिक सौंदर्य विशेषज्ञ से ही नाजुक होती है <end>	Predicted: इसका सेवन मुख्य रूप है <end>
Actual: <start> इसकी प्राकृतिक खूबसूरती कई चीजों से मिलकर बनी है <end>	Actual: <start> इसकी प्राकृतिक खूबसूरती कई चीजों से मिलकर बनी है <end>

2)

CHALLENGES

1. Major challenge was dataset, dataset was very small 25000 for health data for training and 25000 Tourism data for testing.
2. Despite of this challenge other challenge faced is data was not clean.
3. Time taken for training is long (more than 32 hrs). So making multiple models was relatively difficult.

REFERENCES

1) http://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html
2) https://arxiv.org/abs/1409.3215
3) https://machinelearningmastery.com/develop-neural-machine-translation-system-keras/
4) http://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/
5) https://towardsdatascience.com/nlp-sequence-to-sequence-networks-part-2-seq2seq-model-encoderdecoder-model-6c22e29fd7e1
6) https://nlp.stanford.edu/~johnhew/public/14-seq2seq.pdf
7) https://www.analyticsvidhya.com/blog/2018/03/essentials-of-deep-learning-sequence-to-sequence-modelling-with-attention-part-i/
8) https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html
9) https://www.coursera.org/learn/nlp-sequence-models/lecture/ftkzt/recurrent-neural-network-model
10) https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html
11) https://towardsdatascience.com/word-level-english-to-marathi-neural-machine-translation-using-seq2seq-encoder-decoder-lstm-model-1a913f2dc4a7
12) https://towardsdatascience.com/attention-seq2seq-with-pytorch-learning-to-invert-a-sequence-34faf4133e53
13) https://discuss.pytorch.org/t/cuda-changes-expected-lstm-hidden-dimensions/10765/6
14) https://github.com/A-Jacobson/minimal-nmt/blob/master/nmt_tutorial.ipynb


Thank You