NLA Project

Domain Based Neural Machine Translation

Arushi Singhal (201516178), Simran Singhal (201516190)

Mentors:- Dr. Manish Srivastava, Saumitra Yadav

Problem Statement

Build an NMT (Neural MT) system when training data (parallel sentences in the concerned source and target language) is available in a domain. However, such domain data is of small size. Machine learning is to be used in such a way that the small sized domain data can be combined with the large amount of general data.

Contents

Literature Review

Sequence to Sequence Learning with NMT

1) https://arxiv.org/abs/1409.0473
2) This seq2seq model aims to map a fixed length input with a fixed length output where the length of the input and output may differ.
3) The model consists of 3 parts: encoder, intermediate (encoder) vector and decoder.
4) The power of this model lies in the fact that it can map sequences of different lengths to each other. As we 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.

NMT by Jointly Learning to Align and Translate

1) https://arxiv.org/abs/1409.0473
2) Limitation of previous model is “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.
3) This paper addresses the 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 attention mechanism allows the decoder to pull information from useful parts of the input rather than having to decode a single hidden state.

Effective Approaches to Attention-based NMT

1) https://aclweb.org/anthology/D15-1166
2) It is proposed as a simplification of the attention mechanism proposed by Bahdanau. 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.
3) The model evaluated in the Luong is different from the one presented by Bahdanau (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 MT task.

Nature of Data

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.
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)

Pre-processing of Data

1) Vocabulary of unique words is built (and count the occurrences while we're at it)
2) Words with very low frequency is removed or replaced with <UNK>
3) Word to index and index to word dictionary.
4) Lowercase the whole dataset.
5) All the preprocessing can be seen here (https://bit.ly/2H0m5et)
6) Added the <SOS> and <EOS> word ids to the target dataset.
7) Mixed few percentage of Tourism data to health data (if mixed)

Models And Results

Sequence to Sequence (multi-layer)

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

Sequence to Sequence (multi-layer)

Input: maybe this will not give lesser blessings than taking a dip in the sangam

Actual: शायद यह संगम में डुबकी लगाने से कम पुण्य देने वाला नहीं

Predicted: शायद ही बी की

Input: in karnataka in ad the ruler of small mysore state yadurai founded the wodeyar dynasty

Actual: कर्नाटक में ई में छोटे मैसूर राज्य के शासक यदुराय ने वोडयार वंश की नींव डाली

Predicted: कर्नाटक और से एक भी

Input: if necessary deposit your stuff there

Actual: जरूरत पड़ने पर अपना सामान वहीं जमा कराएँ

Predicted: जरूरत होने आर्थिक सर्वप्रथम में अवसाद खुशबू अनोखी भी कम चरण में भी

Sequence to Sequence Bi-directional model

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

Input: the rest of the journey the sea of puri

Actual: यात्रा का शेष पुरी का समुद्र

Predicted: यात्रा आधार विशिष्ट समय है स्वतंत्र की करने अहम जी ही समय युक्त दिन

Input: this is a favorite thing to take from here

Actual: यहाँ से लेकर जाने के लिए यह काफी पसंदीदा चीज है

Predicted: यह कुल करने व्यापार यह भी आवास स्वतंत्र से पर्यटन सीधे हुई श्रेणी अत जबकि में की राज

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>

Attention Model

1) Code Link (simple attention) :- https://bit.ly/2vz0fsa
2) Code Link (Attention + Bidirectional LSTM) :- https://bit.ly/2WvTHGk
3) 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

Attention Model

Input: king malharav holkar lrb second rrb got made this temple

Actual: महाराजा मल्हाराव होलकर द्वितीय ने यह मंदिर बनवाया था

Predicted: नवरतनगढ़ को इस मंदिर का निर्माण करवाया था

Input: give to the child only the mother milk

Actual: बच्चे को केवल माँ का ही दूध दें

Predicted: बच्चे को माँ का दूध पिलाएँ

Input: cave of ajanta was built in ad

Actual: अजंता की गुफा ई में निर्मित हुई

Predicted: गुफा के किनारे पर करवाया था

Attention Model

Input: delhi is located at an ideal place

Actual: दिल्ली एक आदर्श स्थल पर अवस्थित है

Predicted: दिल्ली एक प्रसिद्ध है

Input: snow falls all around on the mountains

Actual: पर्वतों पर चारों ओर बर्फ़ गिरती है

Predicted: बर्फ पर बैठ कर सैलानी हिमालय की ओर बर्फ पर दिखाई देता है

Input: this is connected with tarmac road

Actual: टकडा पक्की सड़क से जुड़ा है

Predicted: यह मार्ग से जुड़ा है

Attention results with Glove Embeddings

Code :- https://bit.ly/2ZRM8MB https://bit.ly/2vAtGKt

Results with embedding trained during model training

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

Predicted: इसका प्राकृतिक सौंदर्य विशेषज्ञ से ही नाजुक होती है <end>

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

Results with embedding not trained during model training

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

Predicted: इसका सेवन मुख्य रूप है <end>

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

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.

Thank-you