Natural Language Processing
Assignment №4

Pooya Kabiri

Department of Computer Science

Iran University of Science and Technology

May 2021

Contents

List of Figures

1 Neural Machine Translation with RNNs

g.

Because not every sentence has the same length, we use padding in the beginning. But when an attention vector i.e. a weighted sum of all encoder hidden states is to be created, we don't want any `PAD` tokens to have a impact on the attention vector, so this mask operation changes the places corresponding to pad tokens to have a. negative infinity value, and in this way the probability of a pad token in the attention vector becomes zero.

h.

Model's BLEU Score: 12.4

i.

i.

Dot product attention is the most simple attention, and it is computationally very fast, because it can be implemented using high speed matrix multiplication, but compared to Multiplicative attention, it has less strength in generalization, because of the W weights used in Multiplicative attention.

ii.

In Additive attention, number of parameters to be trained is more (we have three matrices: v, W_1 and W_2), so it is computationally more expensive and slower in high dimensional data, but it can handle more complex situations better, due to the this computation complexity it has more computational strength when dimensionality is high.

2 Analyzing NMT Systems

a.

In a polysynthetic language, words are created from many morphemes, which are independent or dependent parts of a word (they are subwords). So when Cherokee is the source language, and it's polysynthetic language it only makes sense that we use subword embeddings so that we can extract the meaning of words more accurately, because words are created heavily depending on the used subwords, hencce the word's meaning is also highly tangled with the meaning of it's subwords.

b.

When training a NLP model. usually the model is required to understand the semantic meaning of the text and not just the arrangement of specific words. In order to infer this semantic meaning from words, we can use the creating units of the words, which are characters and subwords. In this way the model can have a better semantic understanding.

c.

When using Massive NMT, English is usually used as an either source or target language between all translation language pairs.
in MNMT, the training is done for more than two languages, and the concepts and semantic information of sum high resource languages like German, French and English can be used for low resource languages. In this way, with a small size of corpus, translation to and from these languages can have a much better result.

This is achieved using a procedure for transfer learning called **Positive Transfer**.

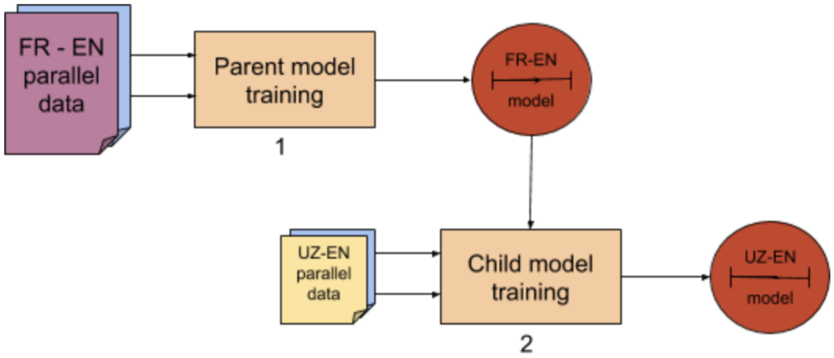

Figure 1: Transfer Learning in NMT

d.

i.

Reason: Limitation in data, Model performs poorly.
Fix: Produce more training data.

ii.

Reason: Model pool performance in capturing the relationship and get that joyful refers to she.

Fix: More complex model, with more data.

iii.

Reason: Didn't get that Littlefish doesn't mean a small fish here, and it is some sort of a slang maybe. it is possible that this word wasn't remembered by the model, if used before in the corpus.
Fix: Increase embedding size to include Littlefish.

e.

i.

"Yes, I'm going to stand with you," he said.

This sentence is generated by my NMT model, but it is not present in the training data. This shows that an NMT system can learn the principals and rules of a language, and hence can generate sentences with understandable semantics, and it is a lot beyond word by word translation.

ii.

And they also took Sodom and Gomorrah, and they that were nigh unto him, according to their fornication, and in the mouth of the flesh.

I think this shows that the NMT model here is having difficulty applying the learned principals or words when the sentence is going to be big. Probably the attention mechanism or the ability of encoder to efficiently encode all the important information in the context vector is not enough for sentences with this big size, and a lot of information are wasted. Due to this, the decoder cannot decode the big sentences within a context, and after some words it diverges to another unrelated things. If the attention mechanism was better, it could have a lot of impact in these cases.

f.

i.

For C1:

Unigram	Max
the	0
love	1
can	1
always	1
do	0

Bigram	Max
the love	0
love can	1
can always	1
always do	0

$$BLUE = e^{(0.5*\log 0.65+0.5*\log 0.5)} = 0.54$$

For C2:

Unigram	Max
love	1
can	1
make	0
anything	1
possible	1

Bigram	Max
love can	1
can make	0
make anything	0
anything possible	1

$$BLUE = e^{(0.5*\log 0.8+0.5*\log 0.5)} = 0.63$$

BLEU score for C2 is better, and it is a better translation.

ii.

For C1:

Unigram	Max
the	0
love	1
can	1
always	1
do	0

Bigram	Max
the love	0
love can	1
can always	1
always do	0

$$BLUE = e^{-0.2} * e^{(0.5*\log 0.65+0.5*\log 0.5)} = 0.44$$

For C2:

Unigram	Max
love	1
can	1
make	0
anything	0
possible	0

love can	1
can make	0
make anything	0
anything possible	0

$$BLUE = e^{-0.2} * e^{(0.5 * \log 0.4 + 0.5 * \log 0.25)} = 0.25$$

BLEU score for C1 is better, and it is not a better translation.

iii.

Language translation is not something that can be done in 1 way. If the reference is only 1 sentence, the model might overfit and lose the ability to learn the general translation.

iv.

Pros

1. Faster than human evaluation.
2. No need to understand both source and target languages.

Cons

1. Doesn't account for semantics or grammar.
2. No quality checking (this can be done by human evaluation).