CS 224n: Assignment #4

1. Neural Machine Translation with RNNs
(g) (3 points) (written) The generate sent masks() function in nmt model.py produces a tensor called enc masks. It has shape (batch size, max source sentence length) and contains 1s in positions corresponding to 'pad' tokens in the input, and 0s for non-pad tokens. Look at how the masks are used during the attention computation in the step() function (lines 295-296).

First explain (in around three sentences) what effect the masks have on the entire attention computation.

Answer:
When applying enc masks, it sets the attention score for the pad token with most minimum value(-inf), so when we calculate attention distribution by applying softmax, the attention probability of padding token will be zero.

Then explain (in one or two sentences) why it is necessary to use the masks in this way.

Answer:
Attention score/attention distributions reflects the attention possibility for a target word to each source words. 'pad' token we added is only for Tensor Operation convenient doesn't have language meaning at all so it has no context with either source and target words, so we must filter them out for model accuracy.

(i) BLEU = 22.69

```
(local_nmt) zheng@zheng-ubuntu-t420:~/git/CS224N/homework4/a4/a4$ sh run.sh test
load test source sentences from [./en_es_data/test.es]
load test target sentences from [./en_es_data/test.en]
load model from model.bin
Decoding:   0%|                                                  | 0/8064 [00:00<?, ?it/s]
/home/zheng/anaconda3/envs/local_nmt/lib/python3.5/site-packages/torch/nn/functi
onal.py:1320: UserWarning: nn.functional.tanh is deprecated. Use torch.tanh inst
ead.
  warnings.warn("nn.functional.tanh is deprecated. Use torch.tanh instead.")
Decoding: 100%|                                                  | 8064/8064 [07:12<00:00, 18.63it/s]
Corpus BLEU: 22.69125595415839
(local_nmt) zheng@zheng-ubuntu-t420:~/git/CS224N/homework4/a4/a4$
```

(j) (3 points) In class, we learned about dot product attention, multiplicative attention, and additive attention. Please provide one possible advantage and disadvantage of each attention mechanism, with respect to either of the other two attention mechanisms. As a reminder, dot product attention is e t,i = s Tt h i , multiplicative attention is e t,i = s Tt Wh i , and additive attention is e t,i = v T (W 1 h i + W 2 s t).

Answer:

	Advantage	Disadvantage
Dot Product attention	Simple, no extra linear layer required.	1. s_t and h_i have to be same dimension. 2. cost is high for large dimension s_t and h_t
Multiplicative Attention	s_t and h_i do not need to have same dimension.	1. Cost is high for high dimension 2. added one more training parameter(W) to model.
Additive attention	Performs better for high dimension	Added two more training parameter(W1 and W2), and more hyperparameter(d3 in lecture (the attention dimensionility) to tune.

(a)

1. Identify the error in the NMT translation.
2. Provide a reason why the model may have made the error (either due to a specific linguistic construct or specific model limitations).
3. Describe one possible way we might alter the NMT system to fix the observed error.

i. (2 points) Source Sentence: Aquí otro de mis favoritos, “La noche estrellada”.
Reference Translation: So another one of my favorites, “The Starry Night”.
NMT Translation: Here’s another favorite of my favorites, “The Starry Night”.

Answer:
Error: favorite in “favorite of my favorites”
Reason: Model Limitations, low-resource language pairs.
Possible Fix: Try add more training data on this kind of language pairs.

ii. (2 points) Source Sentence: Ustedes saben que lo que yo hago es escribir para los niños, y, de hecho, probablemente soy el autor para niños, ms ledo en los EEUU.
Reference Translation: You know, what I do is write for children, and I’m probably America’s most widely read children’s author, in fact.
NMT Translation: You know what I do is write for children, and in fact, I’m probably the author for children, more reading in the U.S.

Answer:
Error: “more reading in the U.S.” has ambiguity

Reason: model limitation,error Maintaining context over longer text model limitation. Sentence is long, since encoder encodes a variable length sentence into a fixed-size vector representation, the neural network may fail to encode all the important details.

Possible Fix: Try increase hidden size of LSTM cell, Add more training data on long sentences

iii. (2 points) Source Sentence: Un amigo me hizo eso – Richard Bolingbroke.
Reference Translation: A friend of mine did that – Richard Bolingbroke.
NMT Translation: A friend of mine did that – Richard <unk >

Answer:
Error: Richard <unk >

Reason: model limitations, Bolingbroke is Out of vocabulary words.

Possible Fix: Add unknown word to training set

iv. (2 points) Source Sentence: Solo tienes que dar vuelta a la manzana para verlo como una epifanía.
Reference Translation: You’ve just got to go around the block to see it as an epiphany.
NMT Translation: You just have to go back to the apple to see it as a epiphany.

Answer:
Error: apple

Reason: kind of domain mismatch , "manzana" has multiple meanings. One of them is apple and one of them is block. "block" has more way to represent in Spanish than apple in Spanish. However , in training sets, "manzana" are more represent to "apple" than "block"

Possible Fix: add more training data on "manzana" represents "block"

v. (2 points) Source Sentence: Ella salvó mi vida al permitirme entrar al baño de la sala de profesores.
Reference Translation: She saved my life by letting me go to the bathroom in the teachers' lounge.
NMT Translation: She saved my life by letting me go to the bathroom in the women's room.

Answer:
Error: women

Reason: model limitations, translation has bias. Similar to model lack of data may have high bias, the occurrence for women in training set is way more than the occurrence of profesores(teacher) in the training set. So the "signal" for women is stronger. When do prediction with attention, it is possible get wrong translation.

Possible Fix: add more training examples on profesores.

vi. (2 points) Source Sentence: Eso es más de 100,000 hectáreas.
Reference Translation: That's more than 250 thousand acres.
NMT Translation: That's over 100,000 acres.

Answer:
Error: acres

Reason: model limitations, Common sense error, acres and hectares are different unit. Checked the training data , there are 19 occurrence for hectares in the training set , and most of them are modified by "millions" "billions" "thousands of" words directly(next to it). However, for acres, has more occurrence in the training set, 41. and most of them are modified by numbers(next to it).

Possible Fix: Add more training examples with numbers modifying hectares(hectáreas**) directly**

(b)(4 points) Now it is time to explore the outputs of the model that you have trained! The test-set translations your model produced in question 1-i should be located in outputs/test outputs.txt. Please identify 2 examples of errors that your model produced. 2 The two examples you find should be different error types from one another and different error types than the examples provided in the previous question. For each example you should:
1. Write the source sentence in Spanish. The source sentences are in the en es data/test.es.
2. Write the reference English translation. The reference translations are in the en es data/test.en.
3. Write your NMT model's English translation.
outputs/test outputs.txt.
The model-translated sentences are in the
4. Identify the error in the NMT translation.
5. Provide a reason why the model may have made the error (either due to a specific linguistic construct or specific model limitations).
6. Describe one possible way we might alter the NMT system to fix the observed error.

Answer:

example 1:

Source Sentence: El 5 de noviembre de 1990
Reference Translation: On November 5th, 1990
NMT Translation: On five of November 1990
Error: five
Reason: Model lack of lcontext for date in this format.
Fix: add more data on date translate between English and Spanish.

example 2:

Source Sentence: Y mis amigos hondureos me pidieron que dijera: "Gracias TED".
Reference Translation: And my friends from Honduras asked me to say thank you, TED.
NMT Translation: My friends were asked to say, "Thank you."
Error : Translation missing informatin , like "from Honduras", "TED"
Reason: Model limitation. Could be due to output word miss alignment with input word.
Fix: Try different attention model, e.g. Additive attention.

(c)
i. (5 points) Please consider this example:
Source Sentence s: El amor todo lo puede
Reference Translation r 1 : Love can always find a way
Reference Translation r 2 : Love makes anything possible
NMT Translation c 1 : The love can always do
NMT Translation c 2 : Love can make anything possible
Please compute the BLEU scores for c 1 and c 2 . Let $\lambda_i = 0.5$ for $i \in \{1, 2\}$ and $\lambda_i = 0$ for $i \in \{3, 4\}$ (this means we ignore 3-grams and 4-grams, i.e., don't compute p_3 or p_4).
When computing BLEU scores, show your working (i.e., show your computed values for p_1 , p_2 , c, $r*$ and BP).
Which of the two NMT translations is considered the better translation according to the BLEU Score? Do you agree that it is the better translation?

Answer:

c1:

1-gram word	Max($Count_r(ngram), Count_c(ngram)$)
the	0
love	1
can	1
always	1
do	0

2-gram word	Max($Count_r(ngram), Count_c(ngram)$)
the love	0
Love can	1
Can always	1
Always do	0

$$p_1 = \frac{0+1+1+1+0}{5} = 0.6$$
$$p_2 = \frac{0+1+1+0}{4} = 0.5$$

c = 5
$r^w = 4$

Since $c \geqslant r^w$ so BP = 1

so $BLEU_{c1} = 1 * \exp(0.5 * \log(0.6) + 0.5 * \log(0.5)) = 0.5477$

c2:

1-gram word	Max($Count_r(ngram), Count_c(ngram)$)
love	1
can	1
make	0
anything	1
possible	1

2-gram word	Max($Count_r(ngram), Count_c(ngram)$)
Love can	1
Can make	0
Make anything	0
Anything possible	1

$$p_1 = \frac{1+1+0+1+1}{5} = 0.8$$
$$p_2 = \frac{1+0+0+1}{4} = 0.5$$

c = 5

$r^w = 4$

Since $c \geqslant r^w$ so BP = 1

so $BLEU_{c2} = 1 * \exp(0.5 * \log(0.8) + 0.5 * \log(0.5)) = 0.632$

According to BLEU score, c2 is “better” translation.

But I do not agree this.

ii. (5 points) Our hard drive was corrupted and we lost Reference Translation r 2 . Please recompute BLEU scores for c 1 and c 2 , this time with respect to r 1 only. Which of the two NMT translations now receives the higher BLEU score? Do you agree that it is the better translation?

Answer:

c1:

1-gram word	Max($Count_r(ngram), Count_c(ngram)$)
the	0
love	1
can	1
always	1
do	0

2-gram word	Max($Count_r(ngram), Count_c(ngram)$)
the love	0
Love can	1
Can always	1
Always do	0

$$p_1 = \frac{0+1+1+1+0}{5} = 0.6$$
$$p_2 = \frac{0+1+1+0}{4} = 0.5$$

c = 5

$r^w = 6$

Since $c < r^w$ so BP = $\exp(1-\frac{6}{5}) = 0.8187$

so $BLEU_{c1} = 0.8187 * \exp(0.5 * \log(0.6) + 0.5 * \log(0.5)) = 0.4484$

c2:

1-gram word	Max($Count_r(ngram), Count_c(ngram)$)
love	1
can	1
make	0
anything	0
possible	0

2-gram word	Max($Count_r(ngram), Count_c(ngram)$)
Love can	1
Can make	0
Make anything	0
Anything possible	0

$$p_1 = \frac{1+1+0+0+0}{5} = 0.4$$
$$p_2 = \frac{1+0+0+0}{4} = 0.25$$

c = 5

$r^w = 6$

Since $c < r^w$ so BP = $\exp(1-\frac{6}{5}) = 0.8187$

so $BLEU_{c2} = 0.8187 * \exp(0.5 * \log(0.4) + 0.5 * \log(0.25)) = 0.2589$

According to BLEU score, c1 is better translation.

Now I agree with it.

iii. (2 points) Due to data availability, NMT systems are often evaluated with respect to only a single reference translation. Please explain (in a few sentences) why this may be problematic.

Answer:

If we use single reference , it increase the possibility for a good translation can get a poor BLEU score because it has low n-gram overlap with the reference translation. If we add more reference, it increase the chance for low n-gram overlap to be higher in a good translation. So that we could possible to give a good translation a relatively high BLEU score.

iv. (2 points) List two advantages and two disadvantages of BLEU, compared to human evaluation, as an evaluation metric for Machine Translation.

Answer:

Advantage:

1. it is Automatic evaluation and it is faster than human evaluation which is manual evaluation, it can evaluate a lot of samples in s short time, which human evaluation can not achieve

2. Algorithm is sample, easy to implement, no need any language knowledge to implement and perform evaluation. Human evaluation requires the person perform evaluation have the knowledge for the target language and source language.

Disadvantage:

1. result could not be stable, a good translation can get a poor BLEU score because it has low n-gram overlap with the reference translation.

2. can not evaluate the grammar of the NMT translated sentence.