CS224 Winter 2019 Assignment 4

Name: Dat Nguyen
Date: 2/26/2019

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

1. Neural Machine Translation with RNNs (45 points)

(g) The mask set $\mathbf{e}_t$ to $-\infty$ where the corresponding position in the source sentence is 'pad' token. Since $\mathbf{e}_t$ is passed to softmax to produce $\mathbf{a}_t$, this has the effect of ignoring the attention to the 'pad' token in the source sentence. It is necessary to use the mask in this way since we padded all the source sentences to the same length for easy processing and at the same time only want to focus our attention to the actual words of the source sentences.

(i) The model's corpus BLEU score is 22.539.

(j) Dot product attention does not require the weight matrix $\mathbf{W}$ as in the multiplicative attention so the computing time and memory is lower, however it does not has as much representative power for the interaction between $\mathbf{s}_t$ and $\mathbf{h}_i$ as the multiplicative attention does.
Multiplicative attention can represent the multiplicative interaction between a dimension of $\mathbf{s}_t$ and every dimension of $\mathbf{h}_i$ because of $\mathbf{W}$ in between; whereas the additive attention can only represent elementwise interaction between $\mathbf{s}_t$ and $\mathbf{h}_i$. However multiplication attention does not allow transformation in $\mathbf{s}_t$ while additive attention does by $\mathbf{W}_1$, so it may gain more representation power there.
Additive attention allows transformation in both $\mathbf{s}_t$ and $\mathbf{h}_i$ so it has greater representation power than the dot product attention. However it requires more parameters for $\mathbf{W}_1$ and $\mathbf{W}_2$.

2. Analyzing NMT Systems (30 points)

(a) i
- Error: repetition of the word "favorites"
- Reason: the second word "favorites" is correct but since the first "favorite" is generated before that, there is no way to know the second word to alter it to "one".
- Proposed solution: Use bidirectional decoder and decode the whole sentence at once instead of one word at a time.

ii
- Error: can not capture the meaning of "America's most widely read children's author".

- Reason: it seems that the nmt system is trying to preserve the sentence’s structure as in the Spanish source sentence.
- Proposed solution: use bidirectional decoder, and collect more training examples where the source and the reference translation having significant different sentence structure.

iii
- Error: output <unk>instead of name belonging to unknown words.
- Reason: the system does not have a way to handle unknown word both in source sentence and the generated sentence.
- Proposed solution: add in vocabulary k words reserved for unknown words (for example if k = 5 we might have <UNK1>, <UNK2>, ..., <UNK5>). When processed source sentence, assign the first unknown word to <UNK1>and so on. We then apply the softmax for the extended vocabulary to choose the correct unknowns and output the corresponding words from the source sentence to the generated sentence.

iv
- Error: the meaning of ”go around the block” is translated incorrectly as ”go back to the apple”.
- Reason: even though the source sentence contains the word ”apple” but it should not be literally translated as it is.
- Proposed solution: collect more training example containing idioms, or unusual, not literal combination of phrases and train the system more on those examples.

v
- Error: ”the teachers’lounge” is translated incorrectly into ”the women’s bathroom”.
- Reason: our training set may contains more instance of the word ”women” than ”teacher” so the system may output higher score for ”women” in this case.
- Proposed solution: collect more examples with the relevant, mistranslated words and train the system on those examples.

vi
- Error: ”hectares” is translated incorrectly to ”acres”.
- Reason: ”acres” may appears much more frequently than ”hectares” in the dataset.
- Proposed solution: collect more training examples which have the source and reference having ”hectares” and train the nmt system on those examples.

(b)

(c) i Computation of the BLUE scores for $\mathbf{c}_1$

$$\begin{aligned}
&p_1 = \frac{0+1+1+1+0}{5} = 0.6\\
&p_2 = \frac{0+1+1+0}{4} = 0.5\\
&r^* = 4\\
&BP = 1\\
&BLEU = 1 \times \exp\big(0.5 \times \log(0.6) + 0.5 \times \log(0.5)\big) = 0.548
\end{aligned}$$

Computation of the BLUE scores for $\mathbf{c}_2$

$$\begin{aligned}
&p_1 = \frac{1+1+0+1+1}{5} = 0.8\\
&p_2 = \frac{1+0+0+1}{4} = 0.5\\
&r^* = 4\\
&BP = 1\\
&BLEU = 1 \times \exp\big(0.5 \times \log(0.8) + 0.5 \times \log(0.5)\big) = 0.632
\end{aligned}$$

Since the BLUE score for $\mathbf{c}_2$ is larger than that of $\mathbf{c}_1$ it is considered better translation according to the BLEU score. I agree that it is a better translation.

ii Computation of the BLUE scores for $\mathbf{c}_1$

$$\begin{aligned}
&p_1 = \frac{0+1+1+1+0}{5} = 0.6\\
&p_2 = \frac{0+1+1+0}{4} = 0.5\\
&r^* = 6\\
&BP = e^{1-\frac{6}{5}} = 0.819\\
&BLEU = 0.819 \times \exp\big(0.5 \times \log(0.6) + 0.5 \times \log(0.5)\big) = 0.449
\end{aligned}$$

Computation of the BLUE scores for $\mathbf{c}_2$

$$
\begin{aligned}
&p_1 = \frac{1+1+0+0+0}{5} = 0.4 \\
&p_2 = \frac{1+0+0+0}{4} = 0.25 \\
&r^* = 6 \\
&BP = e^{1-\frac{6}{5}} = 0.819 \\
&BLEU = 1 \times \exp\big(0.5 \times \log(0.4) + 0.5 \times \log(0.25)\big) = 0.259
\end{aligned}
$$

The $\mathbf{c}_1$ translation now receives higher BLEU score. I do not agree that it is a better translation.

iii Because candidate translation with higher match of n-grams generally receives higher BLUE scores, if we only have only one reference translation it may result in bias agaisnt this particular reference translation. In particular, good candidate translation but the word ordering or use of word are not consistent with the reference translation will still receive lower score.

iv Advantages of BLUE compared to human evaluation

- Can evaluate over huge translations automatically.
- Can be used as a more objective benchmark to compare different system since human evaluation tends to vary from person to person.

Disadvantages of BLUE compared to human evaluation

- Often not as accurate as human evaluation.
- May require multiple reference translations to give good score which are not always available.