CS 224n: Assignment #4

1.Neural Machine Translation with RNNs

(g) 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. First explain (in around three sentences) what effect the masks have on the entire attention computation. Then explain (in one or two sentences) why it is necessary to use the masks in this way.

Sol: Essentially, we use pad to make each sentence in a batch share the same length, which indicates that the 'pad' tokens have no information. Thus we give them a very small value, i.e. $-\inf$, in calculating the e_t attention scores. Intuitively, it is said to not to focus on the pad token. Otherwise, padding token would somewhat dampen the attention distribution.

(h) Please report the model's corpus BLEU Score.

Sol: The Corpus BLEU is 23.63.

(i) In class, we learned about dot product attention, multiplicative attention, and additive attention. As a reminder, dot product attention is $e_{t,i} = s_t^T h_i$, multiplicative attention is $e_{t,i} = s_t^T W h_i$, and additive attention is $e_{t,i} = v^T tanh(W_1 h_i + W_2 s_t)$. Explain one advantage and one disadvantage of dot product attention compared to multiplicative attention. Explain one advantage and one disadvantage of additive attention compared to multiplicative attention.

Sol:

Attention	Advantage	Disadvantage
Dot product attention	Easy to compute, memory efficient, and more intuitive, i.e. any source hidden vector that is orthogonal to target hidden vector will be "washed out"	It requires that $s_t, h_i \in \mathbb{R}_m$, i.e. have the same dimension, rigid
Multiplicative attention	Efficient to compute, more flexible than dot product attention	Still not flexible enough than additive attention
additive attention	Both target and source hidden vector share their own project matrix, i.e. more flexibility	Slower to compute than the other two attentions

2. Analyzing NMT Systems

(a) In part 1, we modeled our NMT problem at a subword-level. That is, given a sentence in the source language, we looked up subword components from an embeddings matrix. Alternatively, we could have modeled the NMT problem at the word-level, by looking up whole words from the embeddings matrix. Why might it be important to model our Cherokee-to-English NMT problem at the subword-level vs. the whole word-level?

Sol: Since Cherokee is a polysynthetic language, where a single word may function as a whole sentence, i.e. language characterized by complex words consisting of several morphemes. In this case, it is better to model at the subword level, because if we model it at the word-level, 1) the embedding size would be much bigger as to better characterize the sentence, 2) embedding for a whole sentence would dampen the performance of the model since each single word can function as a sentence, and all words together would lead to much complex linguistic results.

(b) Character-level and subword embeddings are often smaller than whole word embeddings. In 1-2 sentences, explain one reason why this might be the case.

Sol: The information character-level and subword embeddings reflects is fewer than the whole word embedding.

(c) One challenge of training successful NMT models is lack of language data, particularly for resource-scarce languages like Cherokee. One way of addressing this challenge is with multilingual training, where we train our NMT on multiple languages (including Cherokee). You can read more about multilingual training here. How does multilingual training help in improving NMT performance with low-resource languages?

Sol: Many languages share some common features, e.g. grammer structures. In this case, when training a model that is based on multilingual features, the model is capable of capturing the representational similarity across a large body of languages. Thus the model can be generalized to fit the resource-scarce languages like Cherokee.

(d) Here we present three examples of errors we found in the outputs of our NMT model (which is the same as the one you just trained). The errors are underlined in the NMT translation sentence. For each example of a source sentence, reference (i.e., 'gold') English translation, and NMT (i.e., 'model') English translation, please:

1. Provide possible reason(s) why the model may have made the error (either due to a specific linguistic construct or a specific model limitation).
2. Describe one possible way we might alter the NMT system to fix the observed error. There are more than one possible fixes for an error. For example, it could be tweaking the size of the hidden layers or changing the attention mechanism.

i.

- Source Translation: Yona utsesdo ustiyegv anitsilvsgi digvtanv uwoduisdei.
- Reference Translation: Fern had a crown of daisies in her hair.
- NMT Translation: Fern had $\underline{her\ hair}$ with her hair.

Sol: Error: Model limitation. It didn't capture the linguistic construct with slang using. Fix: Increase more similar linguistic data and increase the size of hidden layer.

ii.

- Source Sentence: Ulihelisdi nigalisda.
- Reference Translation: She is very excited.
- NMT Translation: $\underline{It's}$ joy.

Sol: Error: Model limitation. It didn't capture the relationship in certain context. Fix: Increase the sentence size to include more context information. Increase more similar source data.

iii.

- Source Sentence: Tsesdi hana yitsadawoesdi usdi atsadi!
- Reference Translation: Don't swim there, Littlefish!
- NMT Translation: Don't know how $\underline{a\ small\ fish}$!

Sol: Error: Model limitation. The word embedding didn't know the other meaning of the slang. Fix: Increase the size of the embedding. Change the attention mechanism to addictive attention as to include more context information.

(f) BLEU

It is usually calculated across the entire test set, but here we will consider BLEU defined for a single example.2 Suppose we have a source sentence s, a set of k reference translations $r_1, \ldots, r_k$, and a candidate translation c. To compute the BLEU score of c, we first compute the modified n-gram precision p_n of c, for each of n = 1, 2, 3, 4, where n is the n in n-gram:

$$p_n = \frac{\sum_{ngram \in c} min(max_{i=1,\ldots,k}(Count_{r_i}(ngram)), Count_c(ngram))}{\sum_{ngram \in c} Count_c(ngram)}$$

Next, we compute the brevity penalty BP. Let len(c) be the length of c and let len(r) be the length of the reference translation that is closest to len(c) (in the case of two equally-close reference translation lengths, choose len(r) as the shorter one).

$$BP = \begin{cases} 1, len(c) \geq len(r) \\ exp(1 - len(r)/len(c)), len(c) < len(r) \end{cases}$$

Lastly, the BLEU score for candidate c with respect to r1, . . . , rk is:

$$BLEU = BP \cdot exp(\sum_{i=1}^{4} \lambda_i log(p_i))$$

where λ1, λ2, λ3, λ4 are weights that sum to 1. The log here is natural log.

i. Please consider this example from Spanish:

- Source Sentence s: el amor todo lo puede
- Reference Translation r1: love can always find a way
- Reference Translation r2: love makes anything possible
- NMT Translation c1: the love can always do
- NMT Translation c2: 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 p3 or p4). When computing BLEU scores, show your working (i.e., show your computed values for p1, p2, len(c), len(r) and BP). Note that the BLEU scores can be expressed between 0 and 1 or between 0 and 100. The code is using the 0 to 100 scale while in this question we are using the 0 to 1 scale. 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?

Sol: c_1: $p_1 = \frac{0+1+1+1+0}{1+1+1+1+1} = 0.6$, $\frac{0+1+1+0}{1+1+1+1} = 0.5$, since len(c)=5≥len(r)=len(r_2)=4, so BP=1. It can be showed that $BLEU_1 = 1 \cdot exp(0.5 \cdot (log(0.6) + log(0.5))) = 0.5477$.

c_2: $p_1 = \frac{1+1+0+1+1}{1+1+1+1+1} = 0.8$, $\frac{1+0+0+1}{1+1+1+1} = 0.5$, since len(c)=5≥len(r)=len(r_2)=4, so BP=1. It can be showed that $BLEU_2 = 1 \cdot exp(0.5 \cdot (log(0.8) + log(0.5))) = 0.6325$.

c_2 is considered to be the better translation with HIGHER BLEU. I agree with the result.

ii. 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?

Sol: c_1: $p_1 = \frac{0+1+1+1+0}{1+1+1+1+1} = 0.6$, $\frac{0+1+1+0}{1+1+1+1} = 0.5$, since len(c)=5<len(r)=len(r_1)=6, so BP=exp(1-6/5). It can be showed that $BLEU_1 = exp(-0.2) \cdot exp(0.5 \cdot (log(0.6) + log(0.5))) = 0.4484$.

c_2: $p_1 = \frac{1+1+0+0+0}{1+1+1+1+1} = 0.4$, $\frac{1+0+0+0}{1+1+1+1} = 0.25$, since len(c)=5<len(r)=len(r_1)=6, so BP=exp(1-6/5). It can be showed that $BLEU_2 = exp(-0.2) \cdot exp(0.5 \cdot (log(0.4) + log(0.25))) = 0.2589$.

c_1 is considered to be the better translation with HIGHER BLEU. I disagree with the result.

iii. 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.

Sol: The translation are somewhat subjective translations, and different representation can have the same meaning. If NMT systems are often evaluated with respect to only a single reference translation, then good translation might receive a low BLEU score due to little n-gram overlaps.

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

Sol:

Advantages	Disadvantages
Fully automated and quantitative evaluation, no subjective influence.	Only measures n-grams overlaps, lacks ability to evaluate structures and positions
Fast to compute as opposed to forcing human to interpret both source and target languages	No measuring sematics, log, corpus