Natural Language Processing

Hadi Sheikhi 97521369

1.

g.

In attention computation, we set the values of e_t to –INFINITY for cells where mask is true. By doing so, in calculating alpha, when we use softmax function on very negative values, the probability of attending to that cell is about to 0. And because they are padding cells, it's necessary to have the probability of 0 on attending the decoder output.

h. BELU score on the corpus was: 11.893397615190858

I.

i. Advantages: easier to understand, because it has no learning parameter, it can increase the efficiency of learning process
Disadvantages: Again here because there is no learning parameter, maybe it affects the accuracy of the model. And by adding a learning parameter at this part we can learn how to attend hidden states with each other.

ii. Advantage: Additive attention can find out the nonlinear relations between attention and the hidden states but multiplicative attention can't do so.
Disadvantages: It's not space efficient as multiplicative attention.

2.

a. As the question mentioned, Cherokee is a polysynthetic language, this means words are composed of many morphemes (sub-words that have independent meaning but cannot use them alone) so in these type of languages its better to find these sub-words and for a word, concatenate the embedding vector of sub-words.

b. We don't need extra information which a word might have like word sense and specially in character level we dont even need to store any meaning, so it's obvious we need less information in sub-word level or character level than word level.

c. When we train multilingual model, learning signal from one language should benefit the quality of translation to other languages and that's why we use multilingual training.

d.

i. Seems model have low amount of data about encoder and cut some parts of source translation, so we need to improve hidden layer size and cell information.

ii. Wrong meaning of a word, maybe increasing word embedding size helps.

iii. Really don't know!

e.

i. *"Oh, Charlotte," he said.* In line 34. And yes they are same word by word. The model is trying to find a pairwise relation between words from different languages.

ii. In line 30:

1. Gold: And if thy right hand causeth thee to stumble, cut it off, and cast it from thee: for it is profitable for thee that one of thy members should perish, and not thy whole body go into hell.
2. Model: And if thy hand cause thee to stumble, let it be eaten, and it is: for it is good for thee one of thee, and cast it into thy sight:

Cause of this problem I think is the memory of LSTMs, and as the sequence gets longer irrelevant data is more appearing.

f.

i. Calculating BELU for c_1

$$1-grams\ in\ c_1=\{the, love, can, always, do\}$$

$$1-grams\ in\ r_1=\{love, can, always, find, \text{a}, \text{way}\}$$

$$1-grams\ in\ r_2=\{love, \text{makes}, \text{anything}, \text{possible}\}$$

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

$$2-grams\ in\ c_1=\{\text{the love}, \text{love can}, \text{can always}, \text{always do}\}$$

$$2-grams\ in\ \mathrm{r}_1=\{\text{love can}, \text{can always}, \text{always find}, \text{find a}, \text{a way}\}$$

$$2-grams\ in\ \mathrm{r}_2=\{\text{love makes}, \text{makes anything}, \text{anything possible}\}$$

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

len(c) = 5, len(r) = 4

$$BP=1\ because\ len(c)\geq \mathrm{len(r)}$$

$$\mathrm{BELU}=\mathrm{BP}\times\exp(0.5\times\ln 0.6+0.5\times\ln 0.5)=0.5477$$

Calculating BELU for c_2

$$1-grams\ in\ c_2=\{love, can, make, anything, \text{possible}\}$$

$$1-grams\ in\ r_1=\{love, can, always, find, \text{a}, \text{way}\}$$

$$1-grams\ in\ r_2=\{love, \text{makes}, \text{anything}, \text{possible}\}$$

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

$$2-grams\ in\ c_2=\{\text{love can}, \text{can make}, \text{make anything}, \text{anything possible}\}$$

$$2-grams\ in\ \mathrm{r}_1=\{\text{love can}, \text{can always}, \text{always find}, \text{find a}, \text{a way}\}$$

$$2-grams\ in\ \mathrm{r}_2=\{\text{love makes}, \text{makes anything}, \text{anything possible}\}$$

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

len(c) = 5, len(r) = 4

$$BP=1\ , because\ len(c)\geq \mathrm{len(r)}$$

$$\mathrm{BELU}=\mathrm{BP}\times\exp(0.5\times\ln 0.8+0.5\times\ln 0.5)=0.6324$$

C2 is better than c1.

ii. Calculating BELU for c_1

$$1-grams\ in\ c_1=\{the, love, can, always, do\}$$

$$1-grams\ in\ r_1=\{love, can, always, find, \text{a}, \text{way}\}$$

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

$$2-grams\ in\ c_1=\{\text{the love}, \text{love can}, \text{can always}, \text{always do}\}$$

$$2-grams\ in\ \mathrm{r}_1=\{\text{love can}, \text{can always}, \text{always find}, \text{find a}, \text{a way}\}$$

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

len(c) = 5, len(r) = 6

$$BP=\mathrm{e}^{1-1.2}=0.8187$$

$$\mathrm{BELU}=\mathrm{BP}\times\exp(0.5\times\ln 0.6+0.5\times\ln 0.5)=0.8187\times 0.5477=0.4484$$

Calculating BELU for c_2

$$1 - grams\ in\ c_2 = \{love, can, make, anything, \text{possible}\}$$

$$1 - grams\ in\ r_1 = \{love, can, always, find, \text{a}, \text{way}\}$$

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

$$2 - grams\ in\ c_2 = \{\text{love can, can make, make anything, anything possible}\}$$

$$2 - grams\ in\ \mathrm{r}_1 = \{\text{love can, can always, always find, find a, a way}\}$$

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

len(c) = 5, len(r) = 6

$$BP = \mathrm{e}^{1-1.2} = 0.8187$$

$$\text{BELU} = \text{BP} \times \exp(0.5 \times \ln 0.4 + 0.5 \times \ln 0.25) = 0.8187 \times 0.3162 = 0.2588$$

I DON'T AGREE that c1 is better translation.

iii. Sometimes we can use different word for same meaning and maybe a phrase has many translations with the same meaning, and if we have only 1 reference, maybe the true phrases have bad BELU values just because other translation does not available in our data! And this is bad.

iv. Disadvantages:

1. Cannot count words with same meaning and maybe we get the bad BELU score for a good translation.
2. Doesn't count the grammar of the output.

Advantages:

1. It's just a metric!
2. If in some translations order of words are different BELU still works.