Date: a4 - 99521361 تاریخ:

Subject: موضوع:

1) g) et is set to $-\infty$ where enc-mask is 1. so when softmax operation called, it assigns a weight near zero for those and it causes that the model focus on relevant parts. As a result of that, it's necessary to use masks to prevent attending to irrelevant parts. using masks ensures that the attention distribution is concentrated on the non-pad tokens

h) BLEU: 20.3491

i) i) adv: dot product has no learning parameters so its fast to compute.
disadv: because it has no learning parameters, it doesn't know to pay attention to which part.

ii) adv: additive attention compute the similarity in non-linearly transformed between attention and the hidden states
disadv: this computation costs much.

Date:

Subject:

موضوع:

2) a) it helps the NMT system find local dependencies and patterns within the input sequence. this is useful for languages like Mandarin chinese, where the meaning of a character can change depending on its context.
for example like the characters of the hint example, the meaning of the individual characters changes when they are combined

b) i) the model's ability has limitation on handling plural forms; it may be caused by an imbalance in the training data. if part-of-speech tagging added during training, the model will better understand the subject-verb agreement and correctly identify the plurality of nouns.

ii) instead of translating the second part, it repeats the resources.... The model may not effectively capture the semantic meaning of the second part due to limitations in modeling long-rage depandency. Adding transformer-based architectures can help in handeling this problem.

Date: تاريخ:

Subject: موضوع:

iii) The model may lack the contextual knowledge to recognize the phrase as a term associated with "national mourning" and instead interprets it as a simple phrase. One solution is to enrich the training data with more instances of that phrase to enhance the model's familiarity with such specific vocabulary.

iv) The error may be due to the model's limited exposure to idiomatic expressions or proverbs during training. self-attention, can help model better understand and capture the meaning of idiomatic expressions.

c) i) c_1: 1-grams in c_1 = {there, is, a, need, for, adequate, and, predictable, resources}

1-grams in r_1 = {resources, have to, be, sufficient, and, they, predictable}

1-grams in r_2 = {adequate, and, predictable, resources, are, required}

$$P_1 = \frac{0+0+0+0+0+1+1+1+1}{9} = 0,444$$

تاريخ:

Date:

موضوع:

Subject:

2-grams in c_1 = {there is, is a, a need, need for, for adequate, adequate and, and predictable, predictable resources}

2-grams in r_1 = {resources have to, have to be, be sufficient, sufficient and, and they, they have to, be predictable}

2-grams in r_2 = {adequate and, and predictable, predictable resources, resources are, are required}

$$P_2 = \frac{0+0+0+0+0+1+1+1}{8} = 0,375$$

len(c) = 9 , len(r) = 9

BP = 1 $\Leftarrow$ len(c) $\geq$ len(r)

BELU = 1 x exp(0,5 x ln 0,444 + 0,5 ln 0,376) = 0,408

c_2: 1-grams c_2 = {resources, be, sufficient, and, predictable, to}

$$P_1 = \frac{1+1+1+1+1+0}{6} = 0,833$$

2-grams c_2 = {resources be, be sufficient, sufficient and, and predictable, predictable to}

$$P_2 = \frac{0+1+1+1+0}{5} = 0,6$$

Date: تاریخ:

Subject: موضوع:

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

BP = 1 ← len(c) ≥ len(r)

BELU = 1 × exp(.5 × ln 0,833 + .5 ln 0,6) = 0,707

→ C_2 is better than C_1
but I think C_1 is a better Translation.

ii) C_1: $P_1 = \frac{0+0+0+0+0+1+1+1+1}{9} = 0,444$

$P_2 = \frac{0+0+0+0+0+1+1+1}{8} = 0,375$

len(c) = 9 , len(r) 6 → BP = 1

BELU = 1 × exp(.5 × ln(.444) + .5 ln(0,375)) = 0,408

C_2: $P_1 = \frac{1+0+0+1+1+0}{6} = 0,5$

$P_2 = \frac{0+0+0+1+0}{5} = 0,2$

len(c) = 6 , len(r) = 6 → BP = 1

BELU = 1 × exp(0,5 × ln(.5) + .5 × ln(.2)) = 0,316

Date:

Subject: موضوع:

C_1 is better than C_2 and in my opinion it's correct.

iii) when there are multiple reference translations, maximum number of some particular n-grams is taken across all references. the way P_n is not reduced and the higher it is, the higher BLEU is so it may lead to not very accurate BLEU scores.

iv) advantages: - doesn't need human resources, it's reliable and doesn't make errors.
- fast

disadvantages: - it doesn't consider words with the same meanings
- it doesn't consider the grammer of the output for scoring.