1. Neural Machine Translation with RNNs

(g) The mask makes padding tokens have very small attention scores, so that decoder's hidden states can pay little attention on these padding tokens, which can improve our model's performance theoretically.

(i) BLEU = 22.62

2. Analyzing NMT Systems

(c) i. for translation c_1: the love can always do 1-gram: the, love, can, always, do $p_1 = \frac{3}{5} = 0.6$ 2-gram: the love, love can, can always, always do $p_2 = \frac{2}{4} = 0.5$

$c = 5, r^* = 4, BP = 1$ $BLEU = exp(0.5 * log0.6 + 0.5 * log0.5) = 0.54772$

for translation c_2: love can make anything possible 1-gram: love, can, make, anything, possible $p_1 = \frac{4}{5} = 0.8$ 2-gram: love can, can make, make anything, anything possible $p_2 = \frac{2}{4} = 0.5$

$c = 5, r^* = 4, BP = 1$ $BLEU = exp(0.5 * log0.8 + 0.5 * log0.5) = 0.63245$

It seems c_2 is better, I agree with that.

---

ii. for translation c_1: $p_1 = \frac{3}{5} = 0.6$ $p_2 = \frac{2}{4} = 0.5$ $c = 5, r^* = 6, BP = exp(-1/5)$
$BLEU = exp(-1/5) * exp(0.5 * log0.6 + 0.5 * log0.5) = 0.44843$

for translation c_2: $p_1 = \frac{2}{5} = 0.4$ $p_2 = \frac{1}{4} = 0.25$ $c = 5, r^* = 6, BP = exp(-1/5)$
$BLEU = exp(-1/5) * exp(0.5 * log0.4 + 0.5 * log0.25) = 0.25890$

It seems c_1 is better now, I don't agree.

---

iii. Naturally, it can be various ways of translations, with only a single reference translation, the BLEU score can sometimes be low with some good translation.

---

iv. Advantages: easy to compute, can be a good metric under certain conditions. Disadvantages: do not consider the synonym or similar transaction, needs multiple reference translation to achieve a good performance.