1. Machine Translation

(g) Set the pad-corresponding value in e_t to $-\infty$. Then in the softmax stage, the pad would not be contained in the distribution. The encoder hiden state of 'pad' would be weighted 0.

Necessity: Involving pad would make a wrong decision.

(h) 4 Epochs, BLEU = 19.8

(i) i. $e_{t,i} = s_t^T h_i$ is simpler (computational faster) than $e_{t,i} = s_t^T W h_i$
Multiplicative attention could extract more feature.

ii. $e_{t,i} = v^T \tanh(W_1 h_i + W_2 s_t)$ could contain more non-linear feature.
Additive attention is computationally slower than multiplicative one.

2. Analyzing NMT Systems

(a) The embeding layer could compute the multi-characters word's word vector.
If 1-D CNN first, the information would lose and making ambiguity.

(b) i. Grammer error — singular and plural noun
reason: the source chinese sentence don't contain the semantic infomation of number of people
fix: It's ok, no need to fix.

ii Semantic error. — resource is not a person.
reason & fix: word embeding problem, we should use higher dimension word vector or a better algorithm

iii. Vocabulary error — no corresponding word as "国努"
reson & fix: Limitation of lexical resource. → use a larger corpus.

iv. semantic error
reason: the non-semantic meaning of "吗". In Chinese, this word is only about a sound.

(c) source sentence s. k reference translations $r_1 \dots r_k$, candidate translation c

for $n = 1:4$:

$$P_n = \frac{\sum_{ngram \in c} \min\left(\max_{i=1\dots k} Count_{r_i}(ngram),\ Count_c(ngram)\right)}{\sum_{ngram \in c} Count_c(ngram)}$$

i. we will represent the apperance by tuple (in c | in r_1 | in r_2)

ii) c_1: 1-gram: there is a need for adequate and predictable resources
(1,0,0) (1,0,0) (1,0,0) (1,0,0) (1,0,0) (1,0,1) (1,1,1) (1,1,1) (1,1,1)

2-gram: there is | is a | a need | need for | for adequate | adequate and | and predictable | predictable resources
(1,0,0) (1,0,0) (1,0,0) (1,0,0) (1,0,0) (1,0,1) (1,0,1) (1,0,1)

c_2: 1-gram: resources be sufficient and predictable to
(1,1,1) (1,2,1) (1,1,0) (1,1,1) (1,1,1) (1,0,1)

2-gram: resources be | be sufficient | sufficient and | and predictable | predictable to
(1,0,0) (1,1,0) (1,1,0) (1,0,1) (1,0,0)

Also, we have: len(r_1) = 11 len(r_2) = 6 len(c_1) = 8 len(c_2) = 6
(len(r))

$BP_{c_1} = \exp(1 - \frac{3}{4}) = \exp(\frac{1}{4})$ $\quad BP_{c_2} = 1$

$P_1(c_1) = \frac{1}{9}(3) = \frac{1}{3}$ $\quad P_1(c_2) = \frac{1}{6} \cdot 4 = \frac{2}{3}$

$P_2(c_1) = \frac{1}{8} \cdot 0 = 0$ $\quad P_2(c_2) = \frac{1}{5} \cdot 0 = 0$

$\Rightarrow$ $BLEU(c_1) = e^{0.25} \cdot \exp(0.5 \log \frac{1}{3}) = 0.74$

$BLEU(c_2) = 1 \cdot \exp(0.5 \cdot \log \frac{2}{3}) = 0.81$

According to BLEU Score, c_2 would be a better translation. But obviously it's not a better translation.

ii. $P_1(c_1) = \frac{4}{9}$ $\quad P_1(c_2) = \frac{5}{6}$ $\quad BLEU(c_1) = e^{0.25} \cdot \exp(0.5 \log \frac{4}{9} + 0.5 \log \frac{5}{6}) = 0.78$

$P_2(c_1) = \frac{3}{5}$ $\quad P_2(c_2) = \frac{1}{5}$ $\quad BLEU(c_2) = 1 \cdot \exp(0.5 \log \frac{3}{5} + 0.5 \log \frac{1}{5}) = 0.34$

According to BLEU Score, c_1 is a better translation, I agree.

iii. Sometimes single one would be better. But overall, more sample more confident.

iv. advantages: (1) fast / efficient (2) automatic

disadvantages: (1) source-dependent (2) structure-dependent