



1. a-f) code

g) pad tokens do not contribute to att comp since  
set to  $-\infty$ , which when passed to softmax get  
turned to 0, ensures model doesn't learn anything  
from pad tokens. 


h) BLEU (test) = 19.35 

i) i) dot prod > mult att, b/c cheaper

" < "


Sensitive to magnitudes


ii) add att. > "

( $\rightarrow$  softmax  $\rightarrow$  att: saturation) 

can learn more complex relations


" < "

more expensive: 

2. a) 1D conv. layer can help capture local dependencies,  
since chars can have completely diff meanings when  
combined: 


b) i) 1) no plurality

2) missing "I" char, usually after subject  $\rightarrow$  plural

3) more examples of implicit plurality 


ii) 1) wording

2) repetition implies long context issues

3) more advanced pos, etc. 


iii) 1) Chinese-specific phrase missed

2)  $\square$  = day, but translated too literally

3) more data with these e.g.  $\rightarrow$  bigger models 

iv) 1) idiom missed

2) not big enough  model to memorize

3) bigger hidden layer size(s) + more data 





$$ii) \frac{c_1}{p_1} = \sum \left[ \overset{\text{redundant}}{\min(\max(0, 1))}, \# \text{ there} \right. \\ \left. \min(\max(1, 1)), \# \text{ adequate} \right] \\ 9$$

$$= 1, \\ p_2 = \sum \left[ \dots \right. \\ \left. \max(1, 1), \# \text{ adequate and} \right] \\ 9$$

$$= 1, \\ B.P = 1 \quad (9 > 6), \\ \boxed{BLEU = 1}, \\ \frac{c_2}{p_1} = \sum \left[ \max(1, 1), \# \text{ resources} \right] \\ 6$$


$$= 1, \\ p_2 = 1, \\ B.P = 1 \quad (6 = 6), \\ \boxed{BLEU = 1}$$


$\Rightarrow c_1 = c_2$  when  $c_1$  is better.  $\square$

iii) more references evaluate text better (different wordings, sentence structures, etc.), whereas single missed it (ii). BLEU w/ multiple ref. = avg # n-grams appearing in both  $r_k$  &  $c_i$  vs: single ref = # n-grams that are exact match.  $\square$

iv) BLEU pros: cheap to compute, quantifiable; but cons: not as qualitatively good, only 1.  $\square$

d) i) 200: - i - would like to make a number of issues on the basis,

3000: - i - would also like to clarify a number of issues in the conference,  on the conference,

17200: - i - have also  clarified a clarification on a number of matters raised by the conference,

slightly better b/c more precise

ii) 2nd: - I have also clarified clarification on a number  
- of matters raised by the meeting.

\* 3rd: - I would also like to clarify a number of

4th: - matters raised by the conference.  
- I have also clarified clarification on a  
- number of matters made by the conference.

→ 3rd is best b/c lack of repetition.