CS224n, Assignment 4

~~i_generate_sent_masks():~~

1 g) for each sentence$_i$ in batch

~~seq of word embeddings~~

generate mask$_i$

where at end of each sentence, thru end of sentence list/buffer (length == max_len), set mask = 1, else zero.

in step():

with ~~attention~~ e_t,

~~fill~~ where ~~mask~~ mask == 1,

~~with~~ set ~~e_t~~ $e_{t,m} = -\infty \Rightarrow \alpha_{t,m}$

& this will ~~set~~ set $\alpha_{t,m}$ ~~= 0~~ $\simeq 0$,

meaning these ~~words in sentence~~ null tokens will get no attention.

j) i) dot-product attention: pro: quick, easy to compute
con: s_t, h_i have to have identical dims. Strong underlying assumption that attention ~ dot product(s_t, h_i)

ii) multiplicative attention: pros: $\dim s_t$ can be diff from $\dim h_i$, still not too hard to compute
cons: still assumes linear dependence, more params to tune

iii) additive attention: attention & s, h
pros: more complex model, more d.f.
cons: harder to compute, ~~2 more~~ 3 more params to tune

(2) a) i)

NMT	Correct
1) a) Here's	→ So
b) favorite	→ one

2) can't generate pronouns, even though understands what antecedent is

3) add more training examples w/ pronouns
-or- ~~update~~ loss function to penalize more heavily pronoun errors
-or- change attention model, do additive

ii) 1)

NMT commas	Reference
<no comma>	→ You know $\boxed{,}$

"most widely read children's author" → "author for children, more reading"

2) problem NMT is doing transliteration, not translation. NMT has trouble w/ translating clauses

3) ~~intro~~ add more residual connections b/t earlier and later nodes in decoder (and encoder?)

② a, cont'd) ~~a~~ iii) 1) Reference NMT

Bolingbroke → <unk>

2) can't do anything w/ word it doesn't recognize

3) when sees word it doesn't recognize, just carry over, character-by-character into translation

iv) 1) Ref NMT

block → apple

2) 'manzana' word ~~e~~ embeddings most similar to apple. multiple

retrain w/ examples of "block" or add word embeddings per word sense

OR for each sense of word?

v) 1) Ref NMT

~~in the~~ teachers' lounge → women's room

2) NMT fills in blank w/ most likely answer, given corpus, but doesn't weight enough ~~words in source language sentence's equivalent slot~~

3) add res. connections b/c encoder & decoder for equivalent slots

vi) 1) Ref NMT

hectares (or equiv) → acres

2) semantically, hectares & acres are very similar, but quantitatively very different in meaning

3) add hectares to target vocab, if not already there, ~~change ref.~~ ~~translation~~ ~~train~~ & add to training examples

2c) i) < see spreadsheet >

B translation (2) has higher BLEU

& ~~better~~ more faithful translation

ii) translation 1 score higher

but to human eye, translation 2

is more faithful, sounds grammatically correct

~~2~~) iii) ~~mult~~ ~~to~~ ~~to~~ many accurate

- translation ~~exist~~ are usually possible, not just one

- ~~single reference translation~~ NMT learns style translation, & learns style, ~~but not underlying structure of language~~ ~~results will~~ some candidate translations that are good will be thrown out, or penalized, for not being worded similarly to ref translation

2c iv)

pros: fast
- easy to calc automatically, no human needed
- objective, ~~~~ whereas multiple humans will have differing judgments

cons:

- doesn't measure grammar correctness, "realism" of diction/style

- ~~doesn't~~ ~~innately measure fact~~ doesn't ~~capture~~ compare semantic meaning of translation vs. original text