CS224n: hw5 written part Yukun Jiang

I. Character-based convolutional encoder for NMT

(a) • embedding size for character-level is typically lower than that used for word embedding, because by feeding into a CNN for example, many characters compose one word and such different way of combinations gives diversity and makes the character-level embedding powerful enough.

• And one reality is that, mostly the character space is smaller than word space, so we may afford smaller character-embedding. □

(b)

character-based embedding:

$$(V_{char} \times e_{char}) + ((f \times e_{char} \times k) + f) + ((e_{word} \times e_{word}) + e_{word})$$

↑ embedding ↑ convolutional ↑ Highway

$$= (96 \times 50) + ((256 \times 50 \times 5) + 50) + (256 \times 256 + 256)$$

$$= 134848$$

word-based embedding:

$$V_{word} \times e_{word} = 128000000$$

so roughly the parameter numbers of word-based embedding is close to a hundred times more than that of character-based embedding. □

(c) using a convNet could better catch "local meaning" of a word, for example some prefix, surfix may carry particular meaning. This may be hard to be caught by LSTM since it operates on the "longer" whole sequence.

A convNet allows us to naturally do character-level n-gram by setting the kernel-size equal n. While RNN can only visit every character left to right, so that building up word embeddings from character sequences using RNN may not embed the morpheme-like representations we'd enjoy from 1D convolution Net.

(d)

max-pooling is to catch the highest indicator in the selected filter size. This has the effect to like "an activation function" to catch local peak information.

Avg-pool is taking the average of the selected filter size This gives the local average idea of information.

Ads of max-pooling: in most cases, provide the best summary of its region and locally invariant to small changes.
In contrast, avg-pooling may be "watered" down considerably by noises.

Ads of avg-pooling: it can take regions with a majority of large features signals and output a smooth summary of the neighborhood so as to maintain some information provided by a minority of weak signals.

(h.i) we borrow somebody else's sanity check code for both the CNN and Highway embedding. We mostly focus on the correctness of shape, with batch included.

3. Analyze NMT system

(a) only "traducir" and "traduce"

In word-based NMT, the other four will be translated to <unknown> which is not informative. And actually by using character-based NMT model can recognize the important part "tradu" has already appeared, so that could mostly likely to translate out the meaning of "translate".

(b)

i.

financial	· economic
neuron	· nerve
Francisco	· San
naturally	· occurring
expectation	· norms

ii.

financial	· vertical
nearon	· neurons
Francisco	· Francisco
naturally	· practically
expcotation	· expectations

iii.

- word2Vec models similarity by context proximity. Closer words tend to have similiar semantic meaning.
- characterNN models similarity by close 2 words "spell". So closer words tend to spell alike, enjoying spelling proximity.

(c).

(a) acceptable example.

"Tenemos la idea de que ser mujer es tener identidad femenina"

Reference: So we have the concept that what it means to be a woman is to have a female identity.

Word-based: We have the idea that being a women is to have <UNK> identity.

Character-based: We have the idea that being woman is to have female identity.

The character-based can recognize the Spanish "femenina"

(b) unacceptable example.

"Una mujer autista llamada Zosia Zaks dijo una vez".

Reference: An autistic [man] name Zosia Zaks once said.

Word-based: A autistic woman named <unk> once said.

character-based: A autistic woman named Mosor Zamir once said,

This is unacceptable because it mis-translate the name of that person, and this create confusion. ■