1.

(a)

Simplify model, reduce parameters

Larger dataset

Regularization

(add noise)

(b) ~~weight initialization (all zero),~~ increase batch size, smaller learning rate

(c)

~~RNN. Document is sequence input, which needs model to capture its temporal features.~~

CNN. Could capture the feature of context for example the key word that

RNN. is sequence input, which needs model to capture its temporal features.

CNN. Input data is image, where CNN is more suitable for encoding by using convolution.

CNN. CNN is better at classification task rather than RNN.

(d)

Conditional language model needs also concern about the type of input x when data-to-text generation, because x may be image, signal, databases and log files, which means that special encoder should be used to encode the information from those input.

For summarization, conditional language model should also concern about the problem of fluency and factual inaccuracy. For example, in abstractive summarization, the model also needs to make sure the output summarization is correct in grammar and readable.

Limit data resource and manual validation.

(看piazza)

 Ensuring that facts are true (avoiding hallucinations)

 How to automatically evaluate how truthful the output is to the input

 Understanding/generating out of vocabulary words, including names and numbers

 Avoiding repetition in decoding

 Inference, combining multiple facts, columns or records into a single fact

2.

(a)

The state at timestep t not only depends on x\_t, but also s\_{t-1}, which means that when at timestep t\* which is far away from the initial timestep t\_0, the state value of t\* is taken up little from x\_0, because W is [0,1] after product W for t\* times, the value of x\_0 is close to 0, which means that RNN cannot remember the timestep information if it is too far away from the current timestep, i.e. RNN cannot remember the entire information in the history.

Because RNN use back propagation through time (BPTT), there may be problem of gradient vanishing.

Should only use the content before the utterance takes place. If later content is leaking, the model would probably learn ‘copy’.

Dependency Bias + Gradient vanishing

(b)

LSTM is more complex than RNN, which means that LSTM need more time and memory to train.

LSTM is more sensitive to weight initialization and easy to overfit.

(c)

it does not attempt to encode the entire input sentence into a single fixed-length vector. Instead, it encodes the input sentence as a sequence of vectors and adaptively selects a subset of these vectors when decoding the translation. It means that the model needs to store all the vectors of previous sequence and need more time to traverse.

Computation O(n^2)

(d)

The relation between the last sentence’s mentioned subject with the next sentence’s pronoun in first position. And the indicative relation between subject and object.

In encoder, the model flags the last noun’s part of speech to justify whether it is singular or plural and make this flag works in the next sentence, which increases the probability of the noun’s correct pronoun expression.

(co-reference)

(e)

(piazza上有)

(f)

Precision = correct pronouns/all pronouns

4.

(a)

Lexical feature: by counting the co-occurred words or word-pairs that already have entailment relations.

n-gram feature: similar to the lexical feature, but at a n-gram level, where the statistical property of combination of n words, like co-occurred words counts in both premise and hypothesis sentence.

syntactic feature: the syntax of sentence such as subject-verb-object and subject-linking verb-predicate.

Global features: length of sentence

(b)

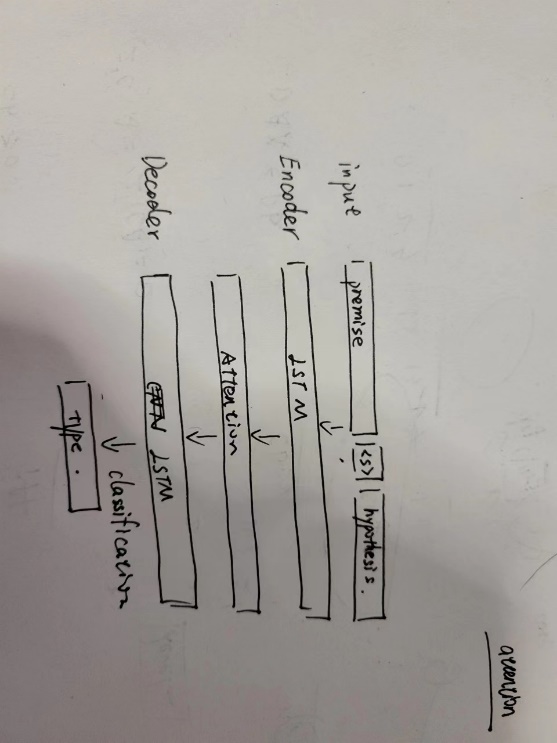
Through word embedding, we can map the words (e.g. premise and hypothesis) into vector space, where the distance between two embedded vector shows the similarity of two sentence. For example, we can use word2vec to embed every single word based on the other words in the sentence.

For classification task, CNN could be a good choice here. The input of the classifier is the concatenation of premise and hypothesis embedding, which is divided by a symbol <CLS> telling the model which part is premise and which part is hypothesis.

(add a logistic regression/ MLP)

Concat hypothesis and premise

(c)



(attention is wrong)

(d)

~~Between the encoder and decoder adds an attention layer. Also, we could use self-attention during encoding.~~

Attention mechanism could use encoded hidden state to get Query, Key and Value, then calculate the attention score of each key, which means that attention could be more likely capture the relations between global and local information than LSTMs, where this ability decreases as the sequence grows longer. Attention supports parallel computing, which could save a lot of time of model training.

The ability of capture long sequence dependency is necessary, because the inpuzt of the model is both premise and hypothesis, which are supposed to be long sequences.

Attention to capture the entailment from the embedding