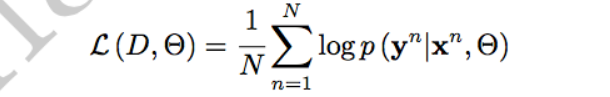
1.

(a)

Maximize Likelihood Estimation



Stochastic gradient descent

(d)

Use synthetic parallel data based on the news texts.

paraphrasing

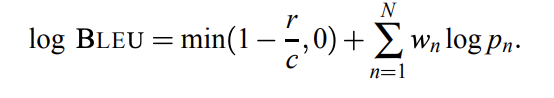
(e)

Apply transfer learning using parallel data, initialize tu-en MT system with de-en MT system parameters

(f)

Human translation references

BLEU score calculate the weighted sum (usually average sum) of 1-4 gram clipped precision (n-gram overlap between system output and reference), penalty term to calculate the length difference between system output and the closest length reference, if system output is too short, the penalty will reflect on BLEU.



(g)

First of all, the dataset’s size is too small to train an effective tu-en model in legal domain. So we use pretrained model such as BERT and apply finetuning on the final output layer, let BERT model suits the task of tu-en translation in legal domain.

2.

(a)

Input words firstly are converted into one-hot vector, then go through a word embedding to become a vector that represent the original word.

Output word was also represented as a vector in the neural network model, through softmax convert the vector into probability distribution.

(b)

Model needs more time to generate the output because softmax need to use all the possible translation result at each step, if increase the vocabulary size, it means every iteration needs more time to finish, as well as the complexity of the model will increase.

(c)

repeatedly replace most frequent symbol pair (’A’,’B’) with ’AB’, until satisfy the stop criteria

c-a -> ca

a-t -> at

r-a -> ra

a-r -> ar

(d)

Hard to map 1-to-many relationship, i.e., cannot replace UNK into correct rare word under different context.

hard to predict inflection with back-off dictionary

For names, if alphabets differ, we need transliteration

(f)

Yes, it could. Because machine translation system usually is encoder-decoder structure and we could use the hidden states of encoder to apply CBOW algorithm, such that we could get the vector representation of the homograph word. After that, we could calculate the similarity between the vector and word vector in our vocabulary, list the top-n similar word, analyse the majority meaning of top-n words, then we could disambiguate the meaning of the homograph word.

(g)

automatic evaluation such as BLEU just calculate the overlap of n-gram between output and reference, if we use synonyms to replace some words in the output, the count of n-gram will decrease, thus the BLEU score will become worse.

3.

(a)

Encoder is taking text as input and it vector representation as output

Decoder is taking encoder’s hidden states as input and it produce the output one by one also depends on decoder output history.

(b)

Compute a function that calculate the similarity between key and query (attention socre), through softmax convert it into probability distribution. By computing the weighted sum of values, we then can get output with fixed length.

(f)

Avoid model learn copying during training.

After removing the masking, the model has the ability to look at the following context, which could be told as information leaking, for seq2seq model the following text is exactly what model trying to predict. The attention of the following words will be assigned with high weight (maybe equals to 1), and the decoder will direct copy the word without updating the parameters inside.