Chapter22 BERT

1. 1-of-N Encoding -> Word Class -> Word Embedding
2. Contextualized Word Embedding
   a) A word can have multiple senses
   b) Each word token has its own embedding, even though it has same word type
   c) The embeddings of word tokens also depend on its context
3. ELMO Embeddings from Language Model
   a) RNN-based language model
   b) Each layer in deep LSTM can generate a latent representation (all use)
   c) $output = a_1 emb_1 + a_2 emb_2$ a_1, a_2 are learned with down stream tasks
4. BERT Bidirectional Encoder Representations from Transformer
   a) Learned from a large amount of text without annotation
   b) BERT = Encoder of Transformer
   c) Although "word" used here, "character" may be a better choice in Chinese
5. Training of BERT
   a) Masked LM : Predicting the masked word by a linear multi-class classifier
   b) Next Sentence Prediction
      i) [SEP] : the boundary of two sentences
      ii) [CLS] : the position that outputs classification results
   c) Approaches 1 and 2 are used at the same time
6. How to use BERT
   a) Case 1 : From single sentence to class [Sentiment Analysis / Doc Classification]
      i) [CLS] + Sentence
      ii) Linear Classifier on output of [CLS] from BERT
      iii) BERT -> Finetune
      iv) Linear Classifier : Train from Scratch
   b) Case 2 : From single sentence to class of each word [Slot Filling]
      i) [CLS] + Sentence
      ii) Linear Classifier on output of words from BERT
   c) Case 3 : From two sentence to class [Natural Language Inference]
      i) Given a 'premise', determining whether a 'hypothesis' is T/F/unknown
      ii) [CLS] + Sentence 1 + [SEP] + Sentence 2
      iii) Linear Classifier on output of [CLS] from BERT
   d) Case 4 : Extraction-based Question Answering [Q&A]
      i) Document : $D = \{d_1, \dots, d_N\}$ Query : $Q = \{q_1, \dots, q_M\}$
         D and Q are sent into QA Model and output two integers (s, e)
         Answer : $A = \{d_s, \dots, d_e\}$
      ii) [CLS] + question + [SEP] + document
      iii) Two vector learned from scratch dot product with the output of words in the document from BERT, then use softmax to determine the s and e.
7. ERNIE Enhanced Representation through Knowledge Integration
   a) BERT mask on character for Chinese
   b) ERNIE mask on word for Chinese

8. GPT Generative Pre-Training
    a) GPT-2 : 1542M
    b) GPT = Decoder of Transformer
    c) Zero-shot Learning
        i) Reading Comprehension
        ii) Summarization
        iii) Translation