Intent Retrieval from Online News

NTU_R07922104_
R07922104 林傳祐 R07922108 陳鎰龍
B06502149 張琦琛 B06902127 鄭人愷

1. problem study:
   - BERT:

     Masked LM:

     從輸入的句子中預測出重要的字詞

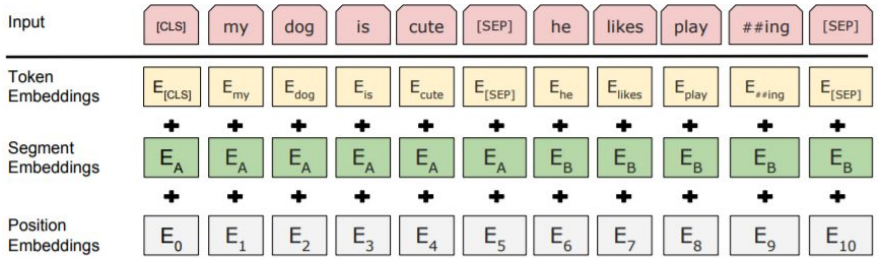
     ```
     Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
     Labels: [MASK1] = store; [MASK2] = gallon
     ```

     Next Sentence Prediction:

     判斷同一文章中兩個句子是否相鄰

     Embedding:

     最後的word vector 由以下三種向量組成

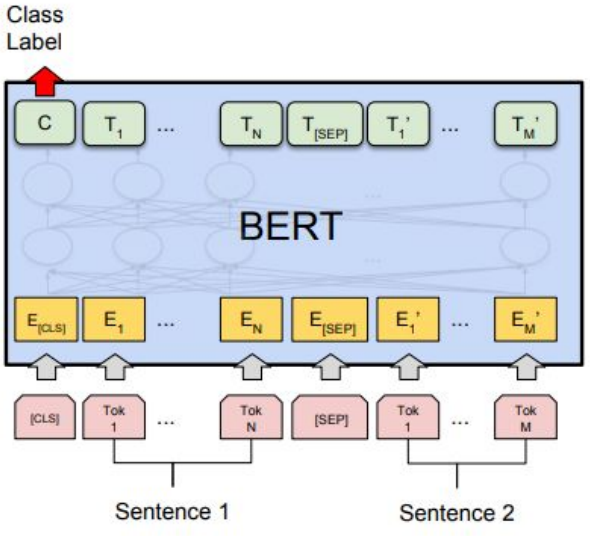
     

     Token Embeddings:詞向量

     Segment Embeddings:用來區別兩種不同句子，做分類用

     Position Embeddings:代表詞出現相對位置的向量

     Model架構:

     利用上圖架構，並且使用類似QQP(判斷兩個句子語意是否相同)的方式，來判斷文章與query之間的關係

- Universal Sentence Encoder:

  Universal Sentence Encoder為將encoding sentences轉為embedding vectors的模型，而透過sentence embedding的transfer learning所訓練出來的model會比word embedding的transfer learning有更好的表現。
- Doc2Vec:

  Doc2Vec在原有word2vec的training model新增了paragraph vector，預測下一個字詞時也對paragraph的vector進行修改，訓練完後便能夠利用該vector表示整個paragraph。

2. proposed method:
   - TF-IDF:計算query、document的TF-IDF，再利用TF-IDF為每個query與document計算相似度分數，取得每個query對應的前300名document，即為輸出

     計算相似度方法:

     $\log (1 + \log (1 + \text{query\_tf})) * \text{idf} * \log (1 + \log (1 + \text{document\_tf}))$
   - 預計導入BERT、Universal Sentence Encoder、Doc2Vec等較新的機器學習方式，進行自然語言處理，希望可以改進準確度。
3. reference:
   - BERT:https://arxiv.org/pdf/1810.04805.pdf
   - Universal Sentence Encoder:https://arxiv.org/pdf/1803.11175.pdf
   - Doc2Vec:https://cs.stanford.edu/~quocle/paragraph_vector.pdf