r06922115
鄭皓謙

HW1 report

Task 2

task3與task2作法相似，差異在如何利用時間的資訊。

首先產生fake link，我猜測在testing set完全沒有link的情況下，產生一些fake link可以幫助node2vec。

接著根據title abstract產生document embedding。

最後將兩者當作classifier的輸入，做binary classification。

不同於doc2vec的做法，這邊產生doc embedding的方式不是pre-train，而是直接跟link prediction訓練，所以embedding可能不具有generalization的特性

task2我嘗試了node2vec作為graph embedding，再用NN處理word embedding，同時利用兩個feature用nn做classifier。

Classifier

每個node會有word embedding跟graph embedding作為feature，將兩種feature接起來後，link prediction則用兩個node的feature接起來，配合BCE loss，用NN做binary classification。

Doc Embedding

為了利用title與abstract的資訊，我參考Google的論文，Attention is all you need，裡面transformer的架構，嘗試將title+abstract轉成一個向量，為每個node得到doc embedding。會選擇如此原因是我曾實作過這部分，減少實作時間。

原本transformer model的目標是用來訓練seq2seq model做機器翻譯，翻譯的原文序列為K, key，而翻譯後的文本序列為Q, query，Q K分別會做self-attention，然後再用Q對K做attention，象徵對每個input sequence的attention weight。每個node，title與abstract會接在一起，將Q作為citation link的destination node，K作為source node。

由於原先transformer是seq2seq model，所以輸出是sequence，我增加了一個trainable weight再對輸出的sequence做一次attention，當作是summary，如此一來便對每個link pair有了text embedding。使用glove當作pre-trained word embedding。

Graph Embedding

首先根據Fake link產生的graph，使用node2vec的feature當作graph embedding。

Negative sampling

根據task1同學的分享，藉由計算adjacent matrix, A^2 = A * A，特別挑選那些兩步之間有連線的node，而一步卻沒有link的node pair。0.8機率sample上述情況，剩下0.2機率則uniform sample training node set。

Fake link

由於testing set完全沒有資料，這次要做test node的fake link是計算training set的out-degree的mean與std，然後為每個test node從normal distribution sample一個out-degree，然後對training set做uniform sample，產生fake link。

嘗試與心得

由於graph embedding的理解不夠深，task2嘗試在word embedding上找出比較好的feature，所以這次用了之前實作的code，配合其他作squad的論文，使用attention mechanism，對title及abstract得出一個doc embedding。

原先negative sampling只有使用uniform sampling的方式時，public Score是0.50292，而增加了A * A的資訊後，有上升到0.50492，而後再增加attention的複雜度則提升到0.50669。但是我覺得問題在於graph embedding的失敗，如果有好的方式可以產生graph embedding，應該可以再提升，這是這次作業我沒有深入的地方。

另外在task1很有效的cosine similarity，這次由於沒有testing link，成效相當的低，所以利用text的額外資訊相當重要，而我認為缺少testing link，對於產生graph embedding應該有更好的做法，更進一步實驗的話，想嘗試單使用text資訊做logistic regression，然後為每個testing node產生相應的weight，再由此產生graph embedding，單是如何挑選哪些fake pair也很困難。

實驗圖 accuracy

藍色部分是training set 以及 negative sampling的link

而紅色為validation set ，不帶有negative sampling，全部都是positive sample

分別為accuracy 與 loss的變化，我解釋為training 跟validation的曲線相似，沒有相異的趨勢，所以沒有overfit，而訓練中止則是因為時限的緣故。