

NLP HW1

1.1 Describe how you build your model ?

這次的作業是按照以下的步驟完成:

(a). 使用公開的Pretrained weight來完成Word Embedding

--> glove_6B

(b). 利用Attention得知各個word vector在句子中的意思

--> 考慮上下文的關係

(c). 利用softmax算出來的權重, 將所有attention產生出來的結合在一起來代表句子的意思。

--> 權重較重的代表句子中較重要的單字

(d). 通過兩層Linear Layer來預測我們要的結果

1.2 How did you do to preprocess your data from dataset ?

我做了以下的資料前處理:

(a). 將所有大寫的字母轉成小寫

--> 因為Pretrained weight只有小寫單字

(b). 將所有不認識的單字轉化成空字串

--> 大部分不認識的單字皆為表情符號或帳號密碼(不重要資訊)

(c). 將所有句子padding到一個固定的常數

--> 為了Train加速

1.3 The distribution of the concern is imbalanced, what did you do to improve the macro F1 score on those concern which are in small scale ?

我統計了我在validation dataset上的結果, 以下圖表為我統計結果:

	ineffective	unnecessary	pharma	rushe d	side-effect	mandatory	count ry	ingredients	political	none	conspiracy	religious
F1 score	0.42	0.28	0.579	0.4667	0.6478	0.7633	0.4879	0.5161	0.4527	0.4592	0.4160	0.5445
Precision	0.2852	0.2558	0.4917	0.3324	0.5	0.7238	0.4763	0.4898	0.4191	0.3385	0.4041	0.59776
Recall	0.8621	0.3197	0.7007	0.7889	0.9234	0.8072	0.5	0.5453	0.4921	0.7139	0.4287	0.5
Proportion	0.14	0.06	0.10	0.12	0.31	0.06	0.01	0.03	0.05	0.05	0.04	0.01

從以上統計數據來看，我的模型受到class imbalance的影響很小，因此我只有根據類別的佔比改變loss的權重。以上作法只有稍微改善效能，但模型沒有顯著的進步。

2.1 Have you tried pretrained word embedding ? (e.g. Glove or Word2vec) What was their influence on the result after using them ?

我有嘗試使用pretrained word embedding.

使用之後有以下的改變：

- (a). loss 變得比較穩定
- (b). model的泛化性(Generalization)更好

3.1 Have you tried attention on your model ? What was its influence on the result?

我有使用attention在我的模型裡面，它造成了以下的影響：

- (a). 模型變得更快
 - > 比起LSTM Attention更適和做平行化運算
- (b). 模型的效能更加
 - > LSTM會有位置越遠影響越小的特性, 而這在句子中是不合理的, 因為可能比較遠的單字比起近的單字更能描述該單字的意思。

3.1 When your model predict the concern, what was it focusing on ? Do some case studies.

黃色部份是GT標記，有產生concerns的句子。而藍色部份是attention較高的地方，我挑了三句句來做 case studies:

- (a).

Original Text:

@OANN @realDonaldTrump You mean @realDonaldTrump, the Johnson & Johnson that participated in the simulated Coronavirus event 201 on October, 18 2019? You have got to be kidding me....people are dying so some damn Pharma company can make a vaccine? It was planned. <https://t.co/0seSnKeirV>

Tokens:

[illegible]

○

- (b).

Original Text:

@TilersinLeeds @glenn_coin @VonSplashback @BorisJohnson The vaccine doesn't stop transmission

Tokens:

[illegible]

Original Text:

@HHSgov @NIH @NIAIDNews @KizzyPhD @NIHDirector Oh ah oh, miraculous \"modern\" vaccine? So at the end is rather a NIH vaccine right?

Tokens:

[illegible]

可以看出模型可以大致抓出重要的地方