ANLP Coursework 2

s2173036 s2185658

1. Introduction

In this coursework, we built several taggers to predict part-of-speech (POS) tags from several words embedding methods, which were the most common tagging method, static word embedding (Global Vectors for word representation (GloVe) (Pennington et al., 2014)) and contextualised word embedding (Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al., 2018)). Word embedding is a basic procedure in Natural Language Processing (NLP) that represents words with vectors that can be easily processed by computer (Li & Yang, 2018). To compare different word embedding methods, POS taggers were implemented to probe. We focused on two questions in this coursework. Firstly, we studied the similarity and differences between the most common tagging method baseline, GloVe and BERT. Secondly, we studied the performance of model in each tag and the relationship between tag performance and the frequency of tag appearance. Finally, we concluded that the BERT had the best performance than the most common tagging method baseline and GloVe, and the frequency of tag occurrence had some moderate positive linear relationship with model performance in each tag.

2. Method

2.1. Most Common Tagging Baseline

It was the baseline method that determines the tag of a word with the most frequent tag in the training data. In this coursework, it was implemented by a dictionary whose keys are tokens in training data and values are dictionaries storing all possible tags-occurrence pairs.

2.2. GloVe

The GloVe (Pennington et al., 2014) is one of the static word embedding methods. GloVe combines global statistics and local statistics of a corpus to map a word with a vector that is consistent in all the contexts. In this coursework, each vector had a dimension of 300.

2.3. BERT

The BERT (Devlin et al., 2018) uses contextualised word embedding that maps a word with a vector according to its context. It provides subword representations for each word. For a known word, BERT will generate a vector with 768-dimensional space. If the input word is unseen, BERT utilises a subword tokenisation strategy to handle this situation. In principle, the word will be split into pieces until all sub-token is known which generates a sequence of subword token vectors. There are several methods to combine these vectors back to a single vector for each word. In this coursework, five methods were implemented (First token, Last token, Maxpool, Average and Sum). Finally, each word would have a representative vector.

3. Result

3.1. Question 1: Baseline VS GloVe VS BERT

As shown in Table 2 and Table 4, the accuracy of the most common tag baseline and GloVe were 85.8% and 86.7% respectively. We concluded that the static word embeddings had higher accuracy than the most common tag baseline. However, in the test set, the difference was insignificant, which was less than 1%. We also chose F1-scores as one of the comparing criteria since it takes both false positives and false negatives into account. We found that the two models had a similar macro average and weighted average F1-score. However, there was a particular data that should be focused on, which was the precision of NOUN. The GloVe had 82.5% precision in NOUN. However, the baseline had 67.4% precision in NOUN. Through the confusion matrix of these two models(Figure 1 & Figure 2), we found that there were more false positive samples in the most common tag NOUN. All the unseen words were predicted as NOUN leading to the low precision of NOUN in the most common tagging baseline. On the contrary, GloVe computed the distance between each vector that it would choose the top training vector when facing the unseen words. Therefore, the GloVe should have higher accuracy in unknown words than the most common tag baseline in theory. The reason why Table 2 and Table 4 had simi-

lar average F1-score was that the unseen words were not enough in the test set, so it could be misleading.

The difference between BERT and GloVe are obviously. From the Table 4 and Table 5, we could find that both the accuracy and F1-score were significantly imporved, expecially in the tag ADV, INTJ, SCONJ, NUM and X. The contexts has undoubtedly gains a significant performance enhencement compared with the GloVe. There is an interesting example could show the contextization is useful. From the tree bank in Universal Dependencies (Dependencies, 2014), we found that in the most frequent SCONJ lemma, some of the lemma such as "for" and "of", the frequncy of ADP is at least 20 times greater than SCONJ. The BERT could largely avoid classifying SCONJ to ADP, which was unqualified in the GloVe (reduced from 30.2% to 3.4% in Figure 3 & Figure 4).

3.2. Question 2: Performance On Each Tag

The accuracy of each class in the most common tag baseline, GloVe and BERT were shown above. The 'CCONJ, DET, PART and PUNCT' were identified more often correctly than other tags in the three methods. The frequencies of these four classes in the test set were 2.9%, 7.6%, 2.5%, 12%. In order to find the relationship between accuracy and occurrence frequency, the Pearson correlation coefficient is used. The correlation coefficient is a measure of linear correlation between two sets of data. -1 and 1 means that the two variables are a very strong linear relationship, and 0 means the two are not linear relationships. The formula is as follows:

$$r_{xy} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i^n (y_i - \overline{y})^2}}$$

Where X is the class frequency of occurrence and Y is the accuracy or F1 scores in each class. The result was in Table 3. We used BERT(Reduce_Mean) to represent the whole BERT model because we found that the first token, last token, mean, sum, max had similar accuracy and F1-scores. The three models had a similar correlation coefficient between accuracy and frequency and the similar correlation coefficient between F1-scores and frequency, which were 0.4. It shows that the frequency and the accuracy or F1-scores had some positive linear relationship, but not strong. In order to find more information, we plotted the three models scatter chart(Figure 5& Figure 6), where X and Y are the same in the correlation coefficient. We found that as the frequency increased, the accuracy or F1-scores is similar. Therefore, we concluded that the frequency and the accuracy or F1-scores had some moderate positive linear relationship by correlation coefficient.

4. Conclusion and Future Work

In this coursework, we studied different word embedding methods via analysing the results of POS tag prediction. Two questions were investigated with quantitative and qualitative analysis. We found that BERT gave the best result followed by the GloVE and the most common tag baseline. Also, we found that the 'CCONJ, DET, PART and PUNCT' were identified more often correctly than other tags, and the the frequency of tag occurrence had some moderate positive linear relationship with tag performance, but not strong. However, BERT has many irrelevant words in the sentence, which might mislead or bring heavy computing. Therefore, in the future, we will continue to focus on reducing irrelevant variables to the BERT. Also, we will study the difference and similarity of word embedding methods in different languages.

References

Dependencies, Universal. Universal dependencies. https://universaldependencies.org/, 2014. Accessed November 26, 2021.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li, Yang and Yang, Tao. Word embedding for understanding natural language: a survey. In *Guide to big data applications*, pp. 83–104. Springer, 2018.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

	precision	recall	f1-score	support
NOUN	0.674	0.929	0.781	4136
PRON	0.968	0.933	0.950	2158
PUNCT	0.994	0.985	0.990	3098
VERB	0.884	0.809	0.845	2640
accuracy			0.858	25098
macro avg	0.841	0.780	0.792	25098
weighted avg	0.872	0.858	0.855	25098

Table 1. Baseline Classification Result

	precision	recall	f1-score	support
ADJ	0.912	0.822	0.864	1782
ADP	0.872	0.883	0.878	2030
ADV	0.922	0.743	0.823	1147
AUX	0.909	0.887	0.898	1509
CCONJ	0.991	0.996	0.993	738
DET	0.961	0.968	0.965	1898
INTJ	0.988	0.692	0.814	120
NOUN	0.674	0.929	0.781	4136
NUM	0.907	0.591	0.716	541
PART	0.665	0.992	0.796	630
PRON	0.968	0.933	0.950	2158
PROPN	0.902	0.531	0.668	1985
PUNCT	0.994	0.985	0.990	3098
SCONJ	0.632	0.650	0.641	443
SYM	0.786	0.830	0.807	106
VERB	0.884	0.809	0.845	2640
X	0.333	0.015	0.028	137
accuracy			0.858	25098
macro avg	0.841	0.780	0.792	25098
weighted avg	0.872	0.858	0.855	25098

Table 2. Baseline Classification Test Set Result

	correlation coefficient
Baseline accuracy	0.435
GloVe accuracy	0.429
BERT(Mean) accuracy	0.389
Baseline F1	0.387
GloVe F1	0.435
BERT(Mean) F1	0.412

Table 3. Correlation Coefficient Compare in Three models

	precision	recall	f1-score	support
ADJ	0.855	0.817	0.836	1782
ADP	0.867	0.883	0.875	2030
ADV	0.864	0.709	0.779	1147
AUX	0.900	0.885	0.893	1509
CCONJ	0.991	0.992	0.991	738
DET	0.955	0.975	0.965	1898
INTJ	0.933	0.692	0.794	120
NOUN	0.825	0.869	0.847	4136
NUM	0.887	0.723	0.796	541
PART	0.666	0.995	0.798	630
PRON	0.965	0.926	0.945	2158
PROPN	0.728	0.760	0.744	1985
PUNCT	0.994	0.983	0.988	3098
SCONJ	0.611	0.639	0.625	443
SYM	0.772	0.830	0.800	106
VERB	0.833	0.821	0.827	2640
X	0.231	0.022	0.040	137
accuracy			0.867	25098
macro avg	0.816	0.795	0.797	25098
weighted avg	0.868	0.867	0.866	25098

Table 4. The GloVe Classification Test Set Result

	precision	recall	f1-score	support
precision	recall	f1-score	support	1782
ADJ	0.925	0.917	0.921	1782
ADP	0.966	0.978	0.972	2028
ADV	0.932	0.906	0.919	1147
AUX	0.973	0.986	0.979	1508
CCONJ	0.991	0.992	0.991	737
DET	0.988	0.990	0.989	1898
INTJ	0.907	0.907	0.907	118
NOUN	0.934	0.943	0.938	4135
NUM	0.949	0.959	0.954	540
PART	0.981	0.989	0.985	629
PRON	0.993	0.993	0.993	2156
PROPN	0.912	0.890	0.901	1984
PUNCT	0.994	0.991	0.993	3096
SCONJ	0.955	0.950	0.952	443
SYM	0.823	0.877	0.849	106
VERB	0.967	0.967	0.967	2638
X	0.687	0.672	0.679	137
accuracy			0.958	25082
macro avg	0.934	0.936	0.935	25082
weighted avg	0.958	0.958	0.958	25082

Table 5. BERT Classification Test Set Result

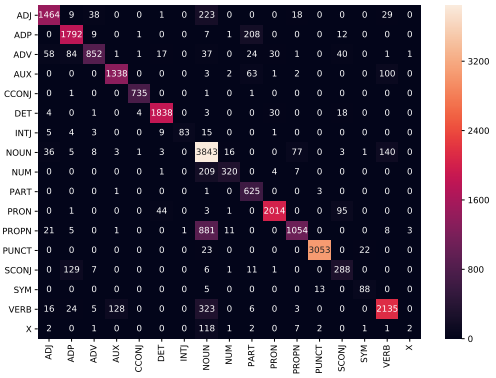

Figure 1. The most common tag baseline confusion matrix

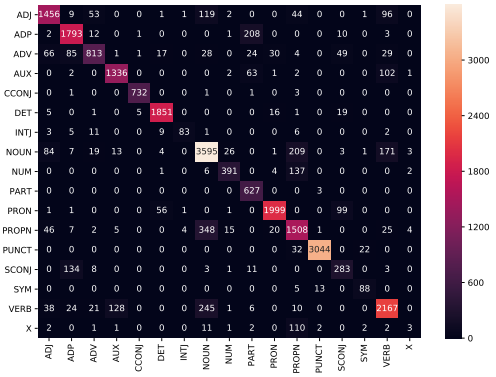

Figure 2. The GloVe confusion matrix

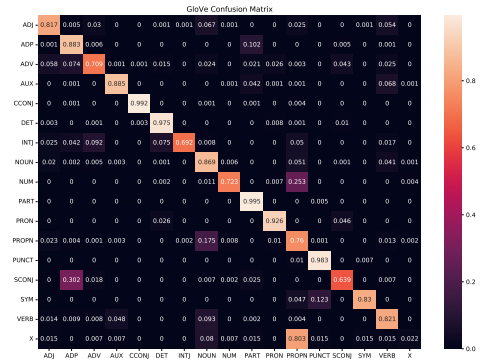

Figure 3. The GloVe confusion matrix (in percentage)

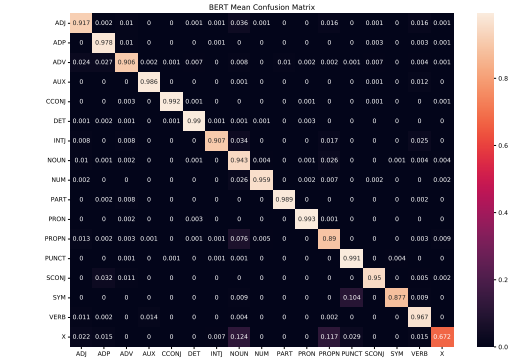

Figure 4. The BERT confusion matrix (in percentage)

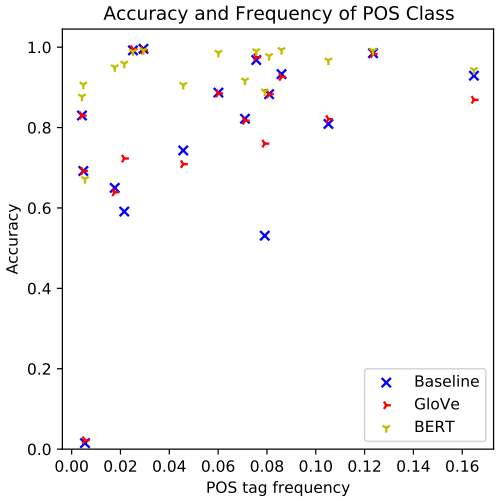

Figure 5. Accuracy / Frequency in each class

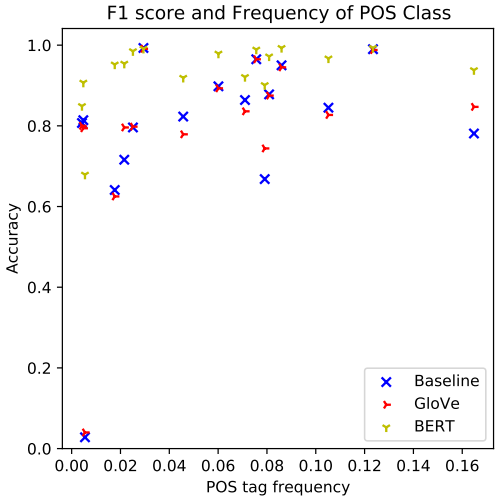

Figure 6. F1-score / Frequency in each class