Natural Language Processing - INF210
Exercise 2 of part 4

Gketsopoulos Petros | $F3321804$
Gkikas Athanasios | $F3321808$
Euagelou Iordanis | $F3321801$
Kougia Vasiliki | $F3321805$

May 12, 2019

1 Introduction

In this exercise we implemented a part-of-speech (POS) tagger as in the previous exercise, but this time using a bi-directional stacked RNN. We used the EWT corpus for English from the Universal Dependencies treebanks and the same train, validation, development and test sets as in the previous exercise.[1] We employed the ELMo language model in order to obtain pre-trained word embeddings to feed to our model. The hyperparameter values of the RNN resulted from tuning, performed on the validation set. The architecture and the methods we used along with the baselines are described in detail in the following section. In section 3, we present the evaluation results, examples of frequent mistakes and our conclusions.

2 Methods

2.1 RNN

Our input token sequences are of length 25. We choose the prior parameter by applying a simple frequency statistic on our train set and we observed that roughly 81% of our sentences are at most 25 tokens longs. In order to adapt sentences longer than 25 tokens, we used an ngram extractor of length 25 and sentences which were lower than our capacity were padded accordingly with the dummy token *PAD*. Furthermore, our tagger consists of two bi-directional GRU cells. We created an embedding layer initialized with the ELMo pre-trained language model,[2] which we finetune during training. For each word we extract its embedding from the ElmoLayer, which is context-aware and can help improve the POS tagging task. We apply batch normalization after both GRUs to normalize their outputs and ensure regularization along with the use of variational dropout. In each GRU we applied two types of dropout: dropout on the input connections within the GRU nodes and dropout applied on the recurrent state signal of the GRU nodes. We also added a residual connection concatenating the embedding from the ElmoLayer with the normalized output of the first GRU cell. The last dense layer applies a softmax activation and outputs a probability for each of the 18 POS tags. Finally, we decided not to use an extra RNN layer for character-based embeddings since this task is implicitly handled by ELMo weighted average layer.

We used the Hyperas[3] API to tune our model for 10 evaluation trials, on the validation subset from which we sliced a test subject prior to the initialization in order to monitor the accuracy during

[1] `https://github.com/UniversalDependencies/UD_English-EWT`
[2] `https://tfhub.dev/google/elmo/2`
[3] `https://github.com/maxpumperla/hyperas`

tuning. We tuned the number of units of each GRU (25, 50 or 100), the probabilities of drop out and recurrent dropout (0.0, 0.2, 0.5) and the number of the stacked layers. As shown in Figure 1, we chose two GRU layers, while the values of the hyper-parameters as they resulted from tuning for both of them are : units = 100, dropout = 0.2, recurrent_dropout = 0.5.

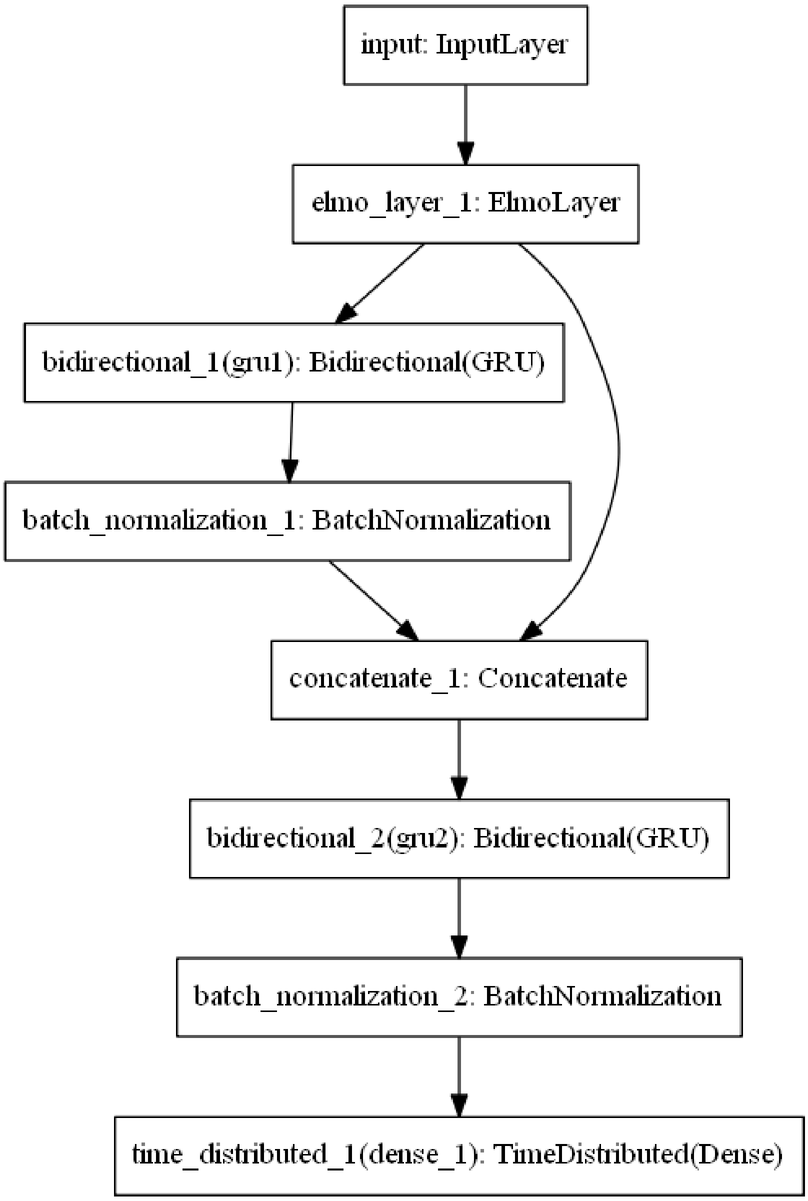

Figure 1: RNN structure

2.2 Baseline

As a baseline we used the same *most-frequent* classifier we used in the previous exercise, which classifies each word of the test set to the most frequent POS tag it had in the train set. If a word is not present in the train set, it is assigned with the "UNK" tag.

3 Evaluation

In the following sections, we demonstrate the results of our experiments. We first analyze the performance of our BiGRU model both during training and at test time. We then, compare our results against the most frequent baseline classifier and the Dense neural model we proposed in the former exercise.

3.1 Evaluation

The main model was trained with the Adam optimizer initialized with the default parameters and categorical cross entropy loss. We also performed early stopping by monitoring the validation loss during training with a patience of 5 epochs. Finally, we set the batch size equal to 128. We illustrate the progression of our model during training in Fig. 2

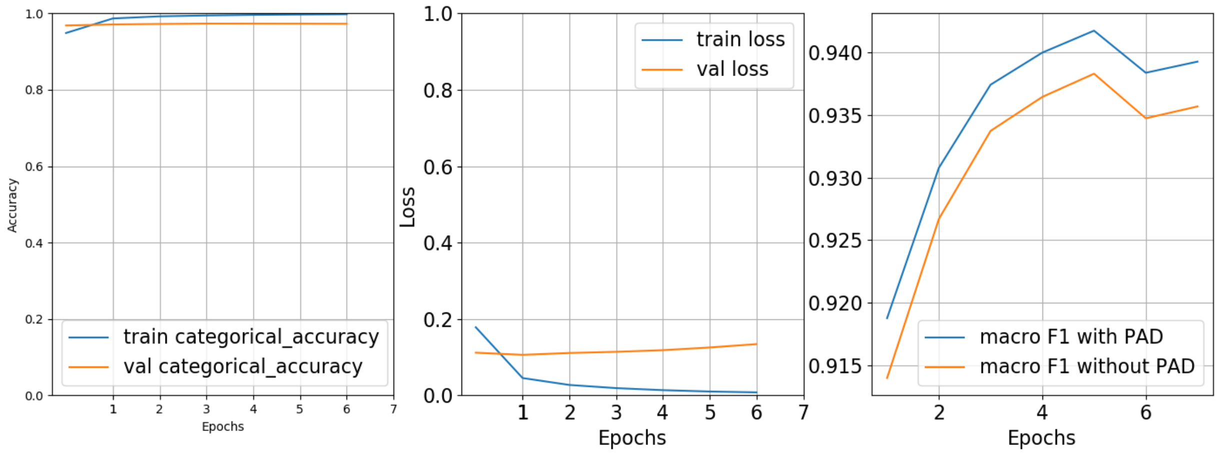

Figure 2: Per epoch evaluation of validation and train subsets. Left, accuracy is moni-tored, while in the right side, the loss of the network.

Additionally, in order to be consistent with the former exercise, we also demonstrate the learning curves for each class of our dataset in Fig. 3.

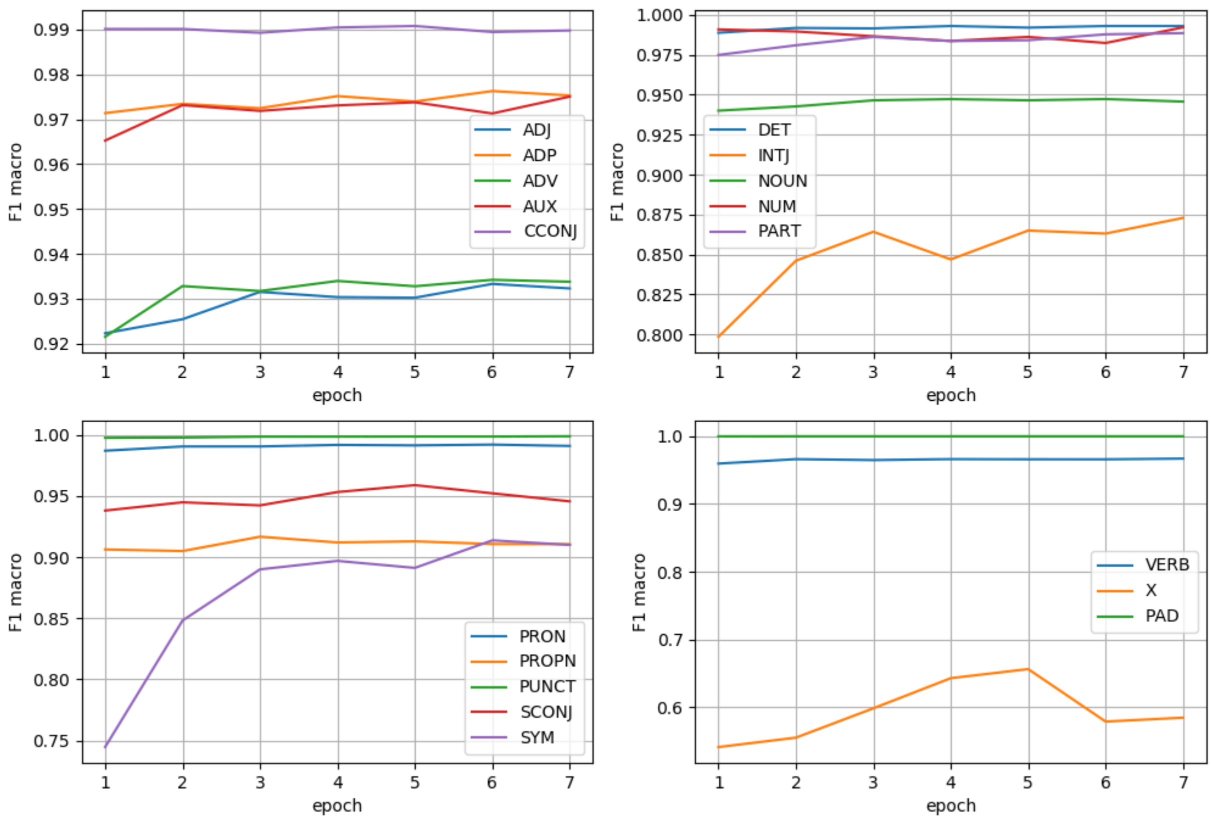

Figure 3: Per *F1* score for each class during training.

3.1.1 Comparative results

	Most frequent classifier			**MLP win5 - Word2Vec**			**BiGRU - ELMo**		
	Precision	**Recall**	**f1-score**	**Precision**	**Recall**	**f1-score**	**Precision**	**Recall**	**f1-score**
ADJ	0.90	0.81	0.85	0.87	0.90	0.89	**0.92**	**0.92**	**0.92**
ADP	0.87	0.88	0.87	0.91	0.96	0.93	**0.96**	**0.99**	**0.97**
ADV	0.93	0.77	0.84	0.91	0.83	0.87	**0.95**	**0.92**	**0.94**
AUX	0.90	0.89	0.90	0.94	0.99	0.96	**0.97**	**0.99**	**0.98**
CCONJ	0.99	0.99	0.99	0.93	0.96	0.95	**1.00**	**0.99**	**0.99**
DET	0.96	0.97	0.96	0.98	0.98	0.98	**1.00**	**0.99**	**1.00**
INTJ	**0.99**	0.69	0.81	0.91	0.82	**0.86**	0.85	**0.86**	0.85
NOUN	0.92	0.76	0.83	0.90	0.91	0.91	**0.93**	**0.95**	**0.94**
NUM	0.90	0.59	0.71	0.84	0.88	0.86	**0.96**	**0.89**	**0.92**
PART	0.66	0.99	0.80	0.92	0.94	0.93	**0.98**	**0.99**	**0.99**
PRON	0.97	0.93	0.95	0.98	0.98	0.98	**0.99**	**0.99**	**0.99**
PROPN	0.89	0.53	0.67	**0.91**	0.83	0.87	0.90	**0.88**	**0.89**
PUNCT	0.99	0.99	0.99	0.93	0.97	0.95	**0.99**	**0.99**	**0.99**
SCONJ	0.62	0.60	0.61	0.89	0.82	0.85	**0.98**	**0.91**	**0.94**
SYM	**0.99**	0.72	0.83	0.93	**0.90**	**0.92**	0.80	0.57	0.67
VERB	0.88	0.81	0.84	0.92	0.93	0.93	**0.97**	**0.97**	**0.97**
X	0.56	0.04	0.07	0.22	0.09	0.13	**0.77**	**0.51**	**0.61**
UNK / PAD	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
micro avg	0.83	0.83	0.83	0.94	0.94	0.94	**0.97**	**0.97**	**0.97**
macro avg	0.83	0.72	0.75	0.88	0.87	0.88	**0.94**	**0.91**	**0.92**
weighted avg	0.91	0.83	0.86	0.94	0.94	0.94	**0.97**	**0.97**	**0.97**
accuracy	0.82			0.93			**0.96**		

3.2 Conclusions

Our bi-directional GRU POS tagger outperformed the previous MLP POS tagger, achieving an accuracy of 96%. In the previous exercise we have pointed out that the MLP classifier with the Word2Vec embeddings could not classify words with ambiguous meaning correctly, e.g., it could not distinguish the cases where the word "love" was used as a verb from the ones it was used as a noun. This time, using a more complex model with context-aware embeddings we expect to successfully deal with these cases.

In the following tables we present words used in two sentences with different semantic meanings at each and hence, assigned a different POS tag in each sentence. We extracted these example sentences from the test subset and examined that similar cases with these words exist in the train subset as well. The predicted tags generated from the previous MLP classifier and the current *most-frequent* baseline and BiGRU classifier are reported in order to conduct a meaningful comparison between the three systems.

First, the *most-frequent* classifier simply depends on which tag was assigned to the word in question the most times in the train set. For that reason, even though it may assign the correct tag to a word this is caused because of the given train set and it does not actually have the ability to distinguish the correct meaning.

he MLP classifier and the BiGRU classifier, both achieve good performance but the latter has a slightly better ability to classify ambiguous words, e.g. the word "love" in "Let's make love". However, in some cases both models fail to correctly predict the tags. For example, in the sentence "Bland and over cooked.", none of the models classifies the word "cooked" as an ADJ.

	Feels	like	you	are	in	Brooklyn	,	but	people	wathcing	is	entertaining	.
Gold	VERB	SCONJ	PRON	AUX	ADP	PROPN	PUNCT	CCONJ	NOUN	NOUN	AUX	ADJ	PUNCT
Most frequent	UNK	ADP	PRON	AUX	ADP	PROPN	PUNCT	CCONJ	NOUN	VERB	AUX	ADJ	PUNCT
MLP win5 - Word2Vec	VERB	SCONJ	PRON	AUX	ADP	PROPN	UNK	CCONJ	NOUN	VERB	AUX	ADJ	UNK
BiGRU - ELMo	VERB	SCONJ	PRON	AUX	ADP	PROPN	PUNCT	CCONJ	NOUN	VERB	AUX	VERB	PUNCT

	We	honestly	can	not	think	of	even	1	thing	we	did	n't	like	!
Gold	PRON	ADV	AUX	PART	VERB	ADP	ADV	NUM	NOUN	PRON	AUX	PART	VERB	PUNCT
Most frequent	PRON	ADV	AUX	PART	VERB	ADP	ADV	NUM	NOUN	PRON	AUX	PART	ADP	PUNCT
MLP win5 - Word2Vec	PRON	ADV	AUX	PART	VERB	UNK	ADV	NUM	NOUN	PRON	AUX	PART	VERB	UNK
BiGRU - ELMo	PRON	ADJ	AUX	PART	VERB	ADP	ADV	NUM	NOUN	PRON	AUX	PART	VERB	PUNCT

	Bland	and	over	cooked	.
Gold	PROPN	CCONJ	ADV	VERB	PUNCT
Most frequent	ADJ	CCONJ	ADP	VERB	PUNCT
MLP win5 - Word2Vec	PROPN	UNK	ADV	VERB	UNK
BiGRU - ELMo	PROPN	CCONJ	ADV	VERB	PUNCT

	Great	meets	that	are	already	cooked	,	easy	to	take	home	for	dinner	.
Gold	ADJ	NOUN	PRON	AUX	ADV	VERB	PUNCT	ADJ	PART	VERB	ADV	ADP	NOUN	PUNCT
Most frequent	ADJ	VERB	SCONJ	AUX	ADV	VERB	PUNCT	ADJ	PART	VERB	NOUN	ADP	NOUN	PUNCT
MLP win5 - Word2Vec	ADJ	VERB	PRON	AUX	ADV	VERB	UNK	ADJ	UNK	VERB	NOUN	ADP	NOUN	UNK
BiGRU - ELMo	ADJ	NOUN	PRON	AUX	ADV	VERB	PUNCT	ADJ	PART	VERB	ADV	ADP	NOUN	PUNCT

	The	talk	of	the	day	besides	a	more	level	playing	field	with	China	was	North	Korea	.
Gold	DET	NOUN	ADP	DET	NOUN	ADP	DET	ADV	NOUN	NOUN	NOUN	ADP	PROPN	AUX	PROPN	PROPN	PUNCT
Most frequent	DET	VERB	ADP	DET	NOUN	ADP	DET	ADJ	NOUN	VERB	NOUN	ADP	PROPN	AUX	PROPN	PROPN	PUNCT
MLP win5 - Word2Vec	DET	NOUN	UNK	DET	NOUN	ADV	UNK	ADJ	NOUN	VERB	NOUN	ADP	PROPN	AUX	PROPN	PROPN	UNK
BiGRU - ELMo	DET	NOUN	ADP	DET	NOUN	ADP	DET	ADV	NOUN	NOUN	NOUN	ADP	PROPN	AUX	PROPN	PROPN	PUNCT

	The	workers	sped	up	and	down	the	street	with	no	mind	to	the	small	children	playing	.
Gold	DET	NOUN	VERB	ADP	CCONJ	ADP	DET	NOUN	ADP	DET	NOUN	ADP	DET	ADJ	NOUN	VERB	PUNCT
Most frequent	DET	NOUN	UNK	ADP	CCONJ	ADP	DET	NOUN	ADP	DET	NOUN	PART	DET	ADJ	NOUN	VERB	PUNCT
MLP win5 - Word2Vec	DET	NOUN	VERB	ADV	UNK	ADP	DET	NOUN	ADP	DET	NOUN	UNK	DET	ADJ	NOUN	VERB	UNK
BiGRU - ELMo	DET	NOUN	VERB	ADP	CCONJ	ADP	DET	NOUN	ADP	DET	NOUN	ADP	DET	ADJ	NOUN	NOUN	PUNCT

	They	re	probably	just	drawn	for	the	show	anyways	.
Gold	PRON	AUX	ADV	ADV	VERB	ADP	DET	NOUN	ADV	PUNCT
Most frequent	PRON	AUX	ADV	ADV	VERB	ADP	DET	VERB	ADV	PUNCT
MLP win5 - Word2Vec	PRON	AUX	ADV	ADV	VERB	ADP	DET	NOUN	INTJ	UNK
BiGRU - ELMo	PRON	AUX	ADV	ADV	VERB	ADP	DET	NOUN	ADV	PUNCT

	One	of	the	pictures	shows	a	flag	that	was	found	in	Fallujah	.
Gold	NUM	ADP	DET	NOUN	VERB	DET	NOUN	PRON	AUX	VERB	ADP	PROPN	PUNCT
Most frequent	NUM	ADP	DET	NOUN	VERB	DET	NOUN	SCONJ	AUX	VERB	ADP	PROPN	PUNCT
MLP win5 - Word2Vec	NUM	UNK	DET	NOUN	VERB	UNK	NOUN	PRON	AUX	VERB	ADP	PROPN	UNK
BiGRU - ELMo	NUM	ADP	DET	NOUN	VERB	DET	NOUN	PRON	AUX	VERB	ADP	PROPN	PUNCT

	I	love	you	.
Gold	PRON	VERB	PRON	PUNCT
Most frequent	PRON	VERB	PRON	PUNCT
MLP win5 - Word2Vec	PRON	VERB	PRON	UNK
BiGRU - ELMo	PRON	VERB	PRON	PUNCT

	Let	make	love	.
Gold	VERB	VERB	NOUN	PUNCT
Most frequent	VERB	VERB	VERB	PUNCT
MLP win5 - Word2Vec	INTJ	VERB	VERB	UNK
BiGRU - ELMo	VERB	VERB	NOUN	PUNCT