Projects Presentation
Natural Language Processing

Master's Degree in Engineering in Computer Science

SAPIENZA
UNIVERSITÀ DI ROMA

Florin Cuconasu

Named Entity Recognition – NER
Homework 1

Named Entity Recognition – NER

NER is the task of **locating** and **classifying** named entities in a text.

Six Categories (Plus **O** tag for no named entity):

PERson **CORP**oration **LOC**ation **PROD**uct **GR**ou**P** **C**reative**W**ork

Ex:
Sundar Pichai is CEO of **Google** having headquarter in **Mountain View**.

Dataset

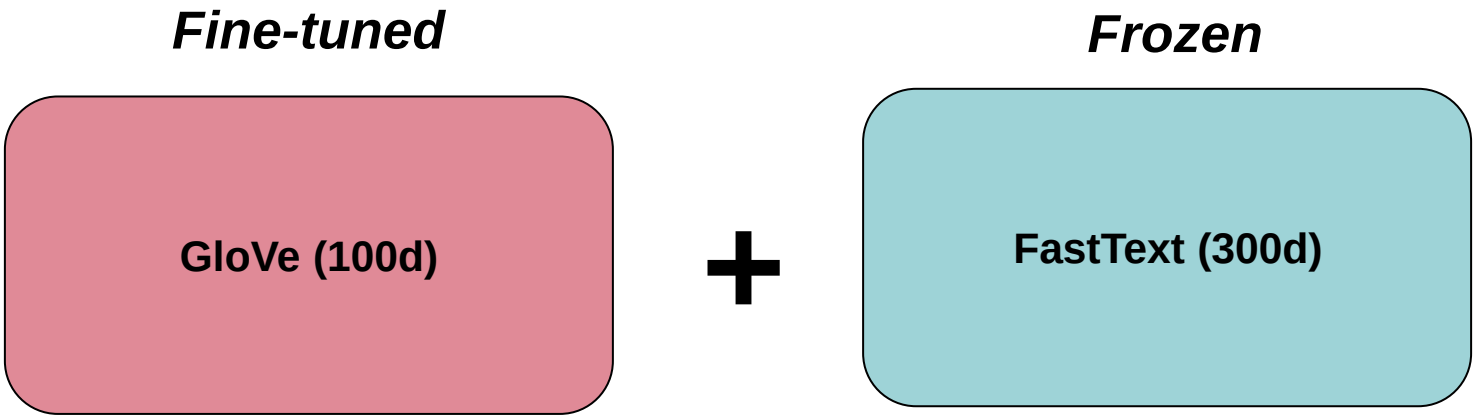
- English
- **14535** training examples
- **765** validation examples
- Already tokenized
- Lowercase tokens only
- Contains non-Latin words

Entity	Train	Dev
O	192841	10240
PER	10895	629
GRP	9459	567
CW	9267	431
LOC	7154	396
CORP	5962	252
PROD	4480	236

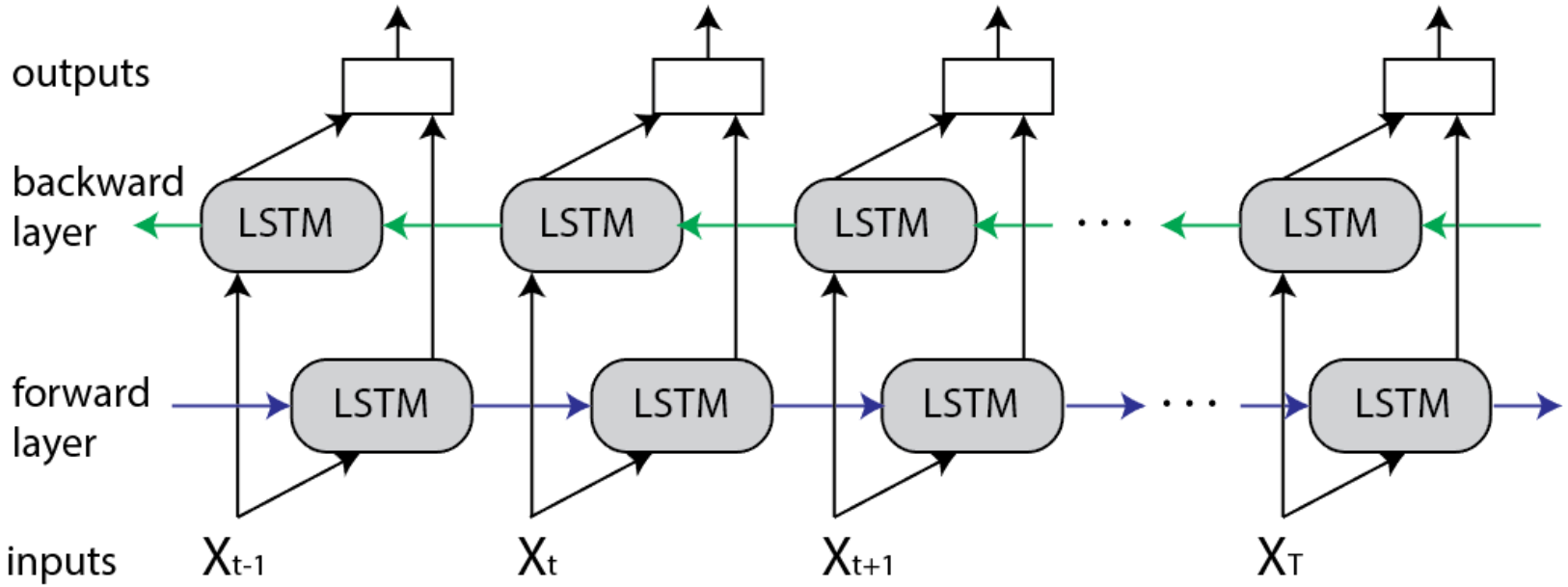
Entity	Train	Dev
PER	23%	25%
GRP	20%	22.6%
CW	19.6%	17%
LOC	15.2%	16%
CORP	12.7%	10%
PROD	9.5%	9.4%

Entity Frequency without O

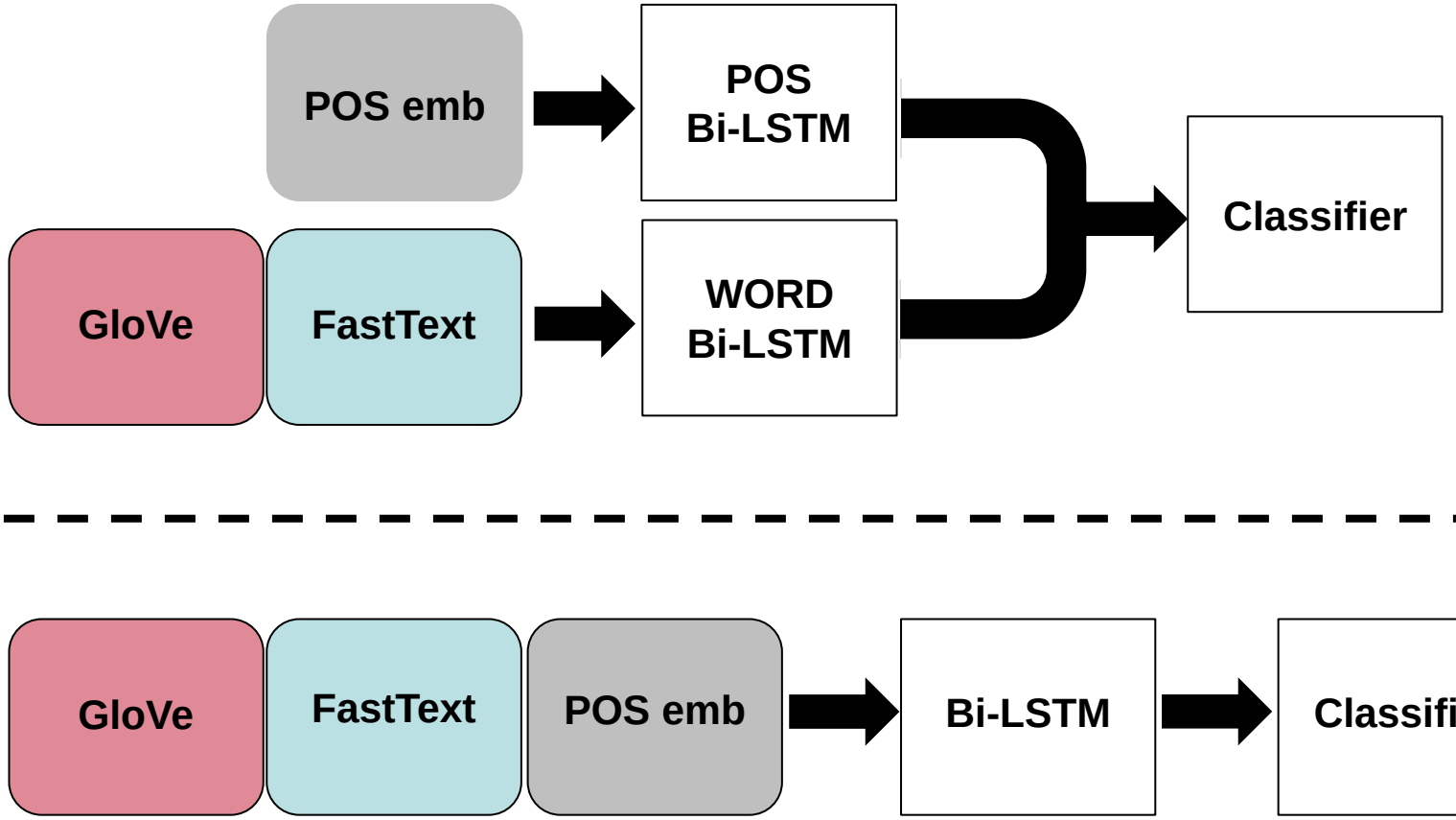
Word Embeddings



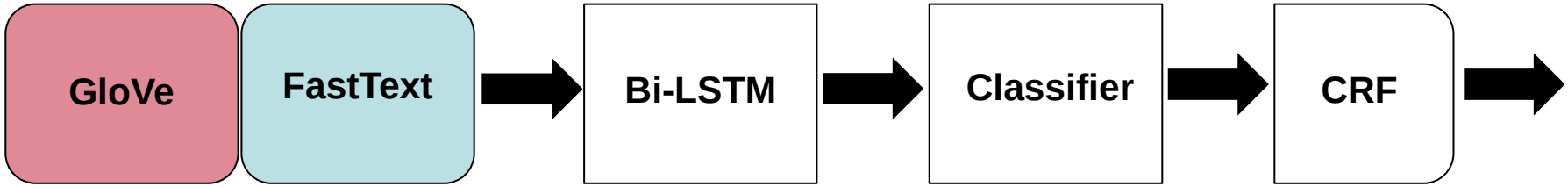
Bi-LSTM


outputs
backward layer
forward layer
inputs
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
X_{t-1}
X_t
X_{t+1}
X_T

POS Embeddings with spaCy



Best Model

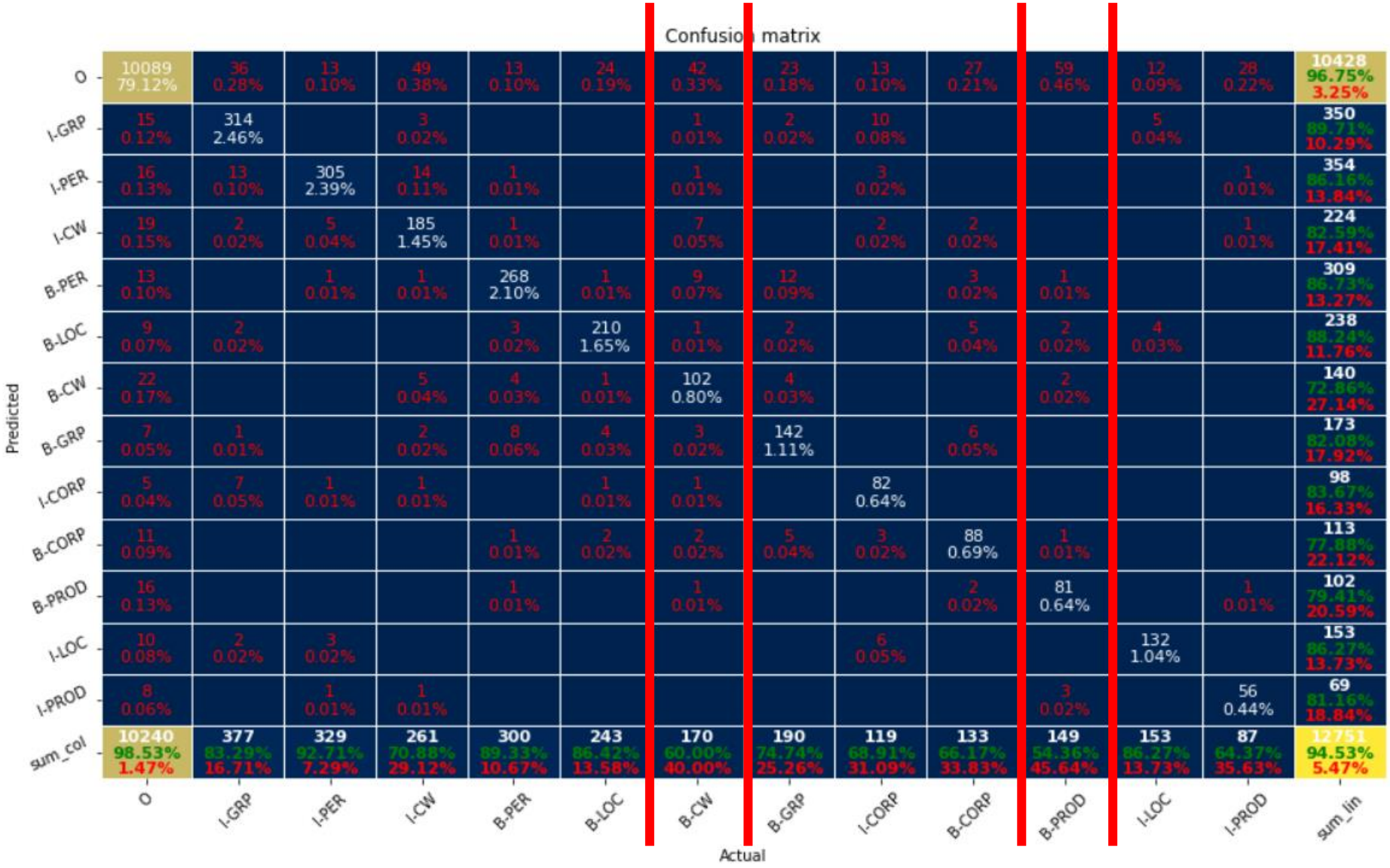


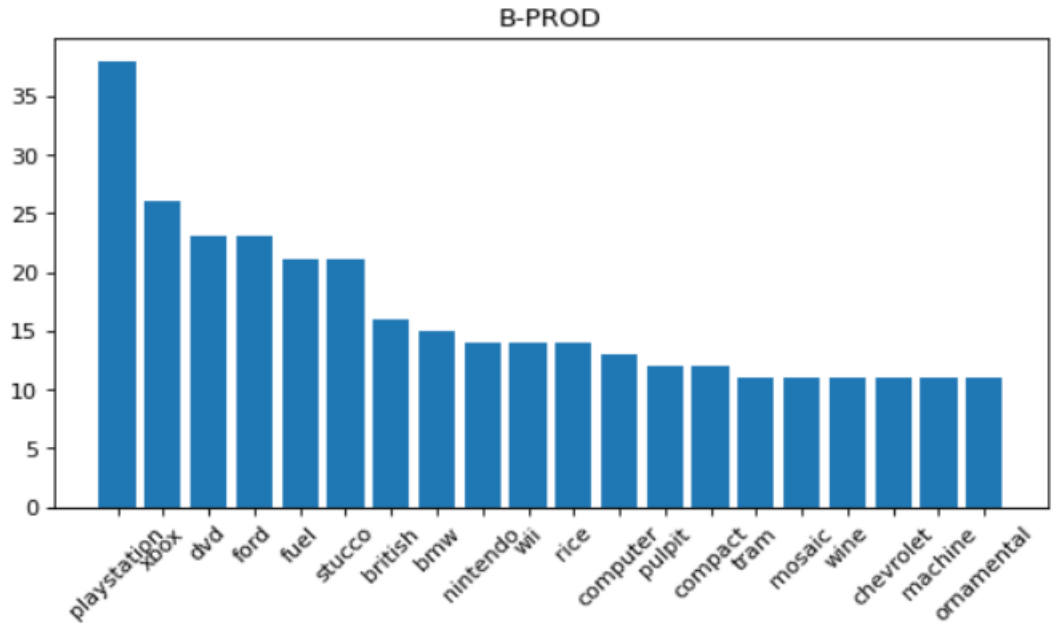
Hyper-parameter	Value
Epochs	12
Batch Size	256
Loss Function	CRF
Optimizer	Adam
Learning Rate	0.003
Gradient Clipping	0.7
StepLR Scheduler	step size=5, $\gamma = 0.2$
Vocabulary Size	20 000
GloVe Embed Dim	100
FastText Embed Dim	300
Bi-LSTM Hidden Dim	512
Bi-LSTM Stacked Layers	3
Dropout	0.4
Activation Function	LeakyReLU

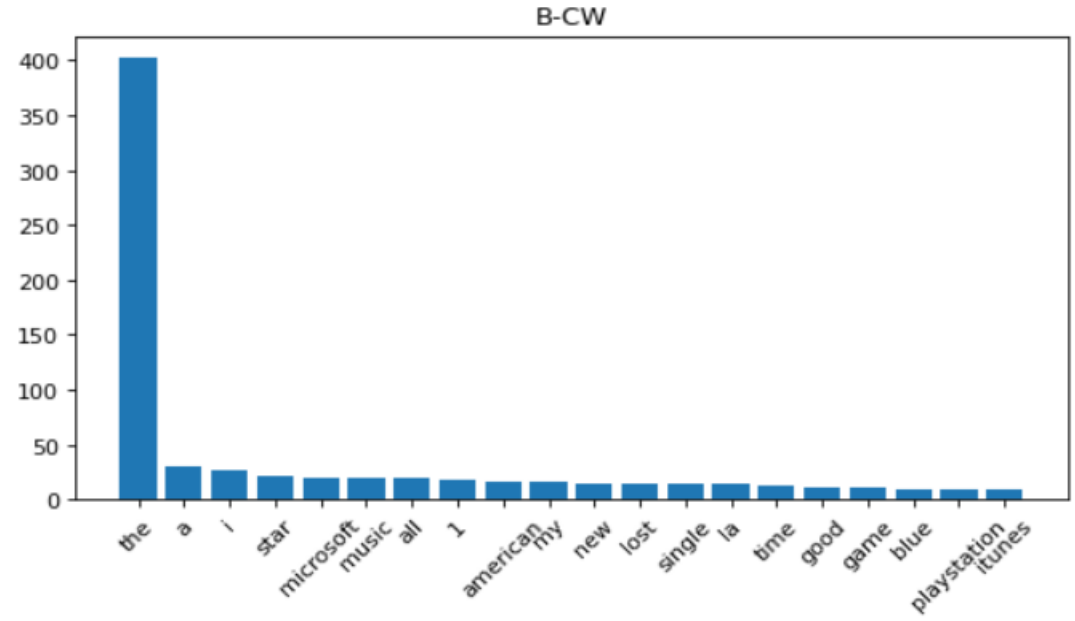
Results

Model	F1 DEV	Epochs
Bi-LSTM + CCE	62.0	15
Bi-LSTM	66.5	14
Bi-LSTM + (F | G* | DEP)	70.3	10
Bi-LSTM + (F | G* | POS)	71.3	7
Bi-LSTM + (F | G)	72.0	9
Bi-LSTM + (F | G*)	**72.5**	12

Table 1: F_1 scores. CCE, F and G stands respectively for Categorical Cross-Entropy, FastText embeddings and GloVe embeddings. G* indicates that the weights are learned. If not specified the loss is CRF.


Confusion matrix
Predicted
Actual
O
I-GRP
I-PER
I-CW
B-PER
B-LOC
B-CW
B-GRP
I-CORP
B-CORP
B-PROD
I-LOC
I-PROD
sum_col
sum_lin
10089 79.12%
36 0.28%
13 0.10%
49 0.38%
13 0.10%
24 0.19%
42 0.33%
23 0.18%
13 0.10%
27 0.21%
59 0.46%
12 0.09%
28 0.22%
10428 96.75% 3.25%
15 0.12%
314 2.46%
3 0.02%
1 0.01%
2 0.02%
10 0.08%
5 0.04%
350 89.71% 10.29%
16 0.13%
13 0.10%
305 2.39%
14 0.11%
1 0.01%
1 0.01%
3 0.02%
1 0.01%
354 86.16% 13.84%
19 0.15%
2 0.02%
5 0.04%
185 1.45%
1 0.01%
7 0.05%
2 0.02%
2 0.02%
1 0.01%
224 82.59% 17.41%
13 0.10%
1 0.01%
1 0.01%
268 2.10%
1 0.01%
9 0.07%
12 0.09%
3 0.02%
1 0.01%
309 86.73% 13.27%
9 0.07%
2 0.02%
3 0.02%
210 1.65%
1 0.01%
2 0.02%
5 0.04%
2 0.02%
4 0.03%
238 88.24% 11.76%
22 0.17%
5 0.04%
4 0.03%
1 0.01%
102 0.80%
4 0.03%
2 0.02%
140 72.86% 27.14%
7 0.05%
1 0.01%
2 0.02%
8 0.06%
4 0.03%
3 0.02%
142 1.11%
6 0.05%
173 82.08% 17.92%
5 0.04%
7 0.05%
1 0.01%
1 0.01%
1 0.01%
1 0.01%
82 0.64%
98 83.67% 16.33%
11 0.09%
1 0.01%
2 0.02%
2 0.02%
5 0.04%
3 0.02%
88 0.69%
1 0.01%
113 77.88% 22.12%
16 0.13%
1 0.01%
1 0.01%
2 0.02%
81 0.64%
1 0.01%
102 79.41% 20.59%
10 0.08%
2 0.02%
3 0.02%
6 0.05%
132 1.04%
153 86.27% 13.73%
8 0.06%
1 0.01%
1 0.01%
3 0.02%
56 0.44%
69 81.16% 18.84%
10240 98.53% 1.47%
377 83.29% 16.71%
329 92.71% 7.29%
261 70.88% 29.12%
300 89.33% 10.67%
243 86.42% 13.58%
170 60.00% 40.00%
190 74.74% 25.26%
119 68.91% 31.09%
133 66.17% 33.83%
149 54.36% 45.64%
153 86.27% 13.73%
87 64.37% 35.63%
12751 94.53% 5.47%


B-PROD
35
30
25
20
15
10
5
0
playstation
xbox
dvd
ford
fuel
stucco
british
bmw
nintendo
wii
rice
computer
pulpit
compact
tram
mosaic
wine
chevrolet
machine
ornamental


B-CW
400
350
300
250
200
150
100
50
0
the
a
i
star
microsoft
music
all
1
american
my
new
lost
single
la
time
good
game
blue
playstation
itunes

Coreference Resolution
Homework 3

Coreference Resolution

Coreference Resolution is the task of finding all **expressions** that **refer** to the **same entity** in a text.

John went home because **he** was tired.

Dataset

- Gender Ambiguous Pronounouns
- English
- **2999** training examples
- **454** validation examples

Pronoun	Train	Dev
his	904 (30.1%)	108 (23.9%)
her	773 (25.8%)	140 (30.8%)
he	610 (20.4%)	93 (20.5%)
she	555 (18.5%)	87 (19.1%)
him	157 (5.2%)	26 (5.7%)

Sample

Pam and Tara clean up Elijah 's body; Jessica , after learning about Russell 's attempt to kill Sookie , decides to call Jason , but Bill asks her why, and refuses.

Sample

Pam and Tara clean up Elijah 's body; Jessica , after learning about Russell 's attempt to kill Sookie , decides to call Jason , but Bill asks her why, and refuses.

her 69

Sample

Pam and Tara clean up Elijah 's body; Jessica , after learning about Russell 's attempt to kill Sookie , decides to call Jason , but Bill asks her why, and refuses.

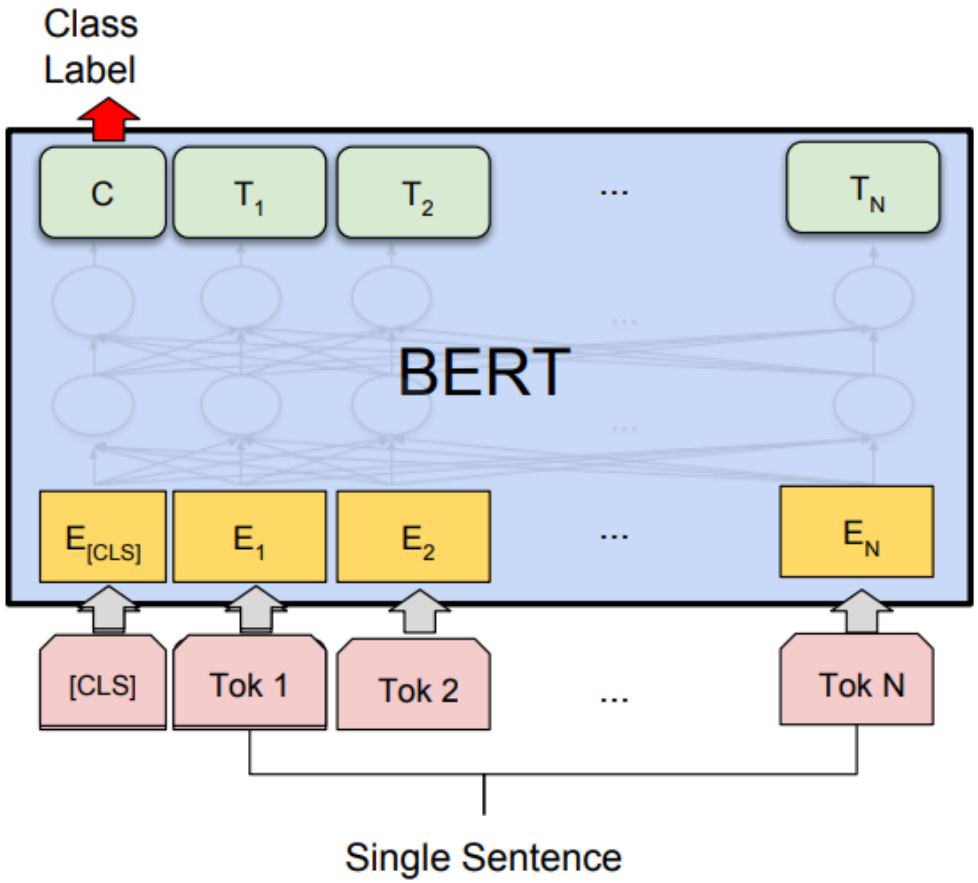
her	69
Jessica	37
Sookie	93

Sample

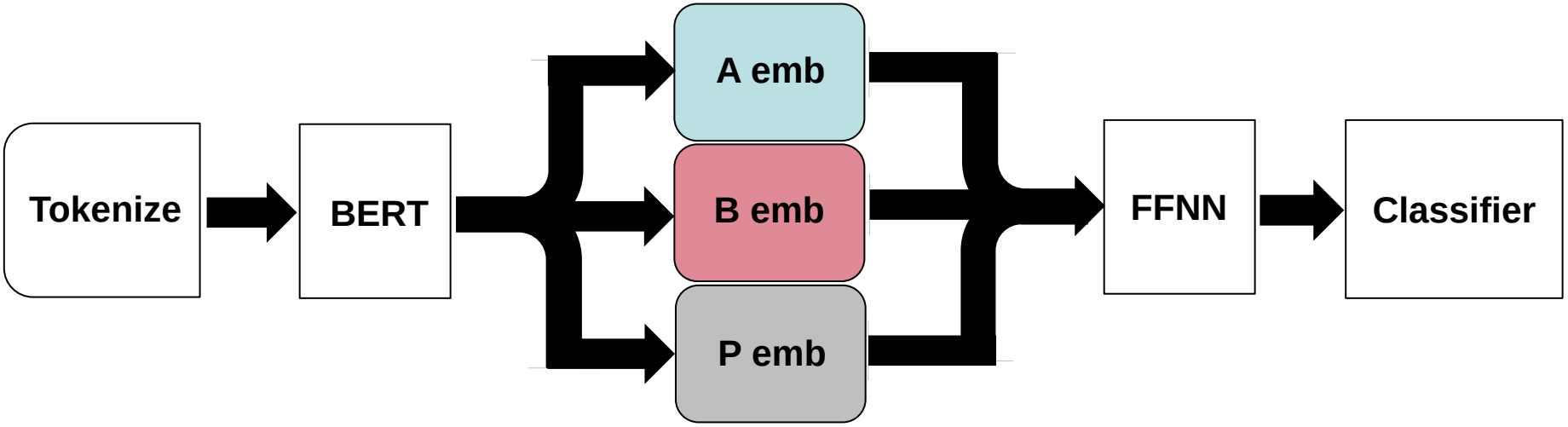
Pam and Tara clean up Elijah 's body; Jessica , after learning about Russell 's attempt to kill Sookie , decides to call Jason , but Bill asks her why, and refuses.

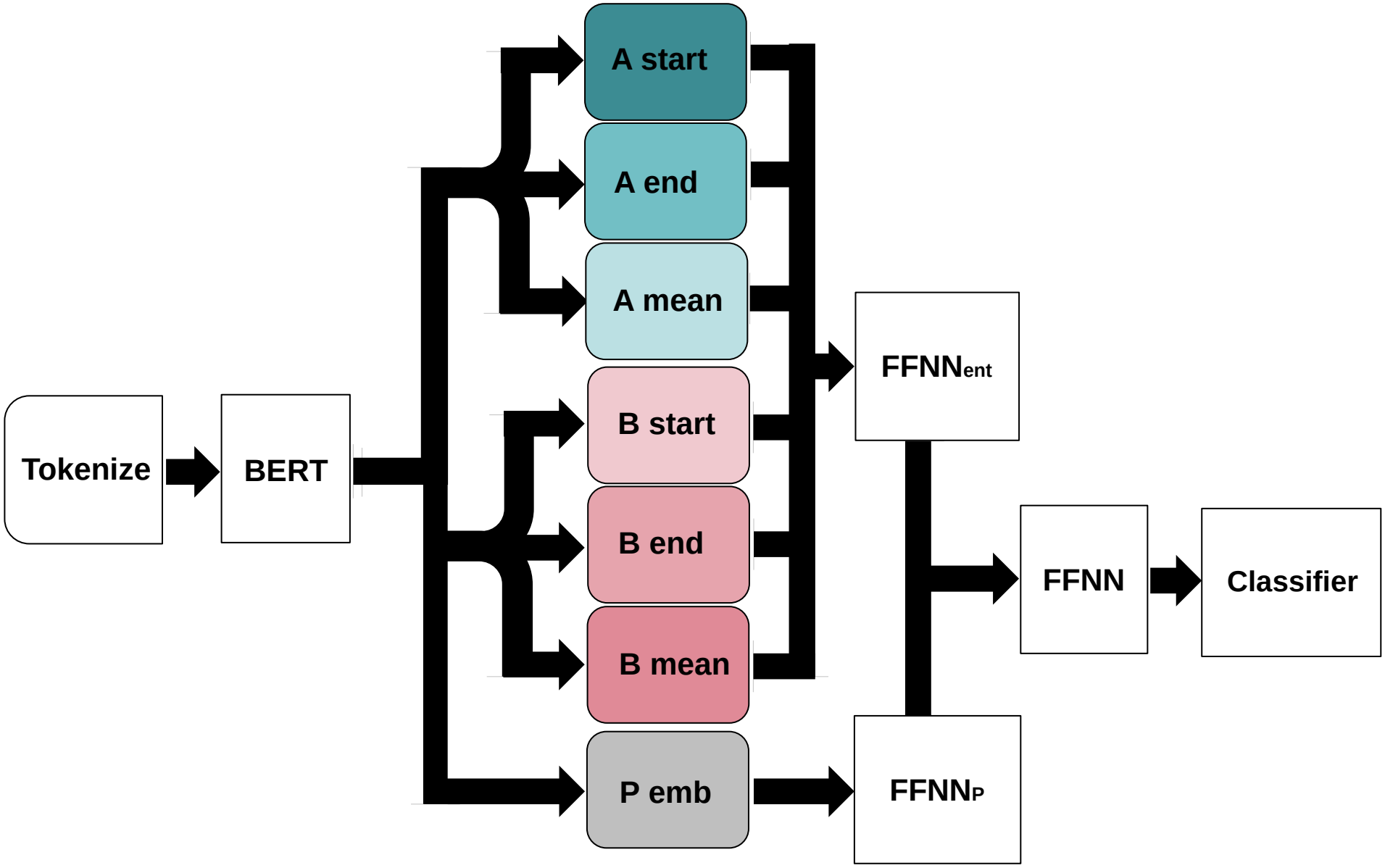
her	69	
Jessica	37	**TRUE**
Sookie	93	**FALSE**

BERT


Class
Label
C
T_1
T_2
...
T_N
BERT
$E_{[CLS]}$
E_1
E_2
...
E_N
[CLS]
Tok 1
Tok 2
...
Tok N
Single Sentence

Baseline




A start
A end
A mean
B start
B end
B mean
P emb
Tokenize
BERT
FFNNent
FFNNP
FFNN
Classifier

Models Hyper-parameters

Hyper-parameter	Value
Epochs	2
Batch Size	4
Loss Function	Cross Entropy
Optimizer	Adam
Learning Rate	8e-6
StepLR Scheduler	step size = 1, $\gamma = 0.8$
BERT Tokenizer	Base Uncased
Head Hidden Dim	512
Linear Hidden Dim	256
Dropout	0.1
Activation Function	GELU

Hyper-parameter	Value
Epochs	6
Batch Size	4
Loss Function	Cross Entropy
Optimizer	Adam
Learning Rate	5e-6
StepLR Scheduler	step size = 3, $\gamma = 0.5$
BERT Tokenizer	Base Cased
Head Hidden Dim	512
Dropout	0.1
Activation Function	LeakyReLU

Results

Model	ACC	Epochs	Time
Multiple Choice (Large)	82.2	3	11m 58s
Baseline Base Uncased	83.0	5	16m 16s
Baseline Base Cased	86.3	6	15m 49s
Baseline Large Uncased	86.8	3	8m 2s
Mention Score Base	87.0	2	5m 31s
Mention Score Large	**87.2**	2	5m 44s

Table 1: Accuracy scores on validation set. The Large models were trained on a Colab Tesla T4 GPU; instead the other model were trained using a GeForce RTX 2060.

Fine fin fund канец kraj край final kraj 終わり konec ende einde lõpp վերջ pää ein son final τέλος vég deireadh beigas 结 galas крајот aħħar slutt დასასრული koniec fim sfârşit 束 конец deireadh крај 종료 konec slutet آخر ахыр кінець oxiri diwedd kawg עק End