SAPIENZA
UNIVERSITÀ DI ROMA

Word-in-Context Disambiguation

NLP Homework 1

July 2021

Olga Sorokoletova, 1937430

Overview

- Given: **Word-in-Context Disambiguation task**
  - **Word-level approach**
  - **Sequence encoding approach**
- Goal: **Obtain the best-performing model (in terms of accuracy)**
  - A priori
  - Exploit potential power of the sequence encoding approach

- Models:
  - **Baseline**
  - **Baseline 2**
  - **2a**
  - **2b**

Baseline

Pre-processing

Pre-processing: Single sentence processing

Given: `'Over 5,000 now hold legal immigrant documents, which, after five years of annual renewal, entitles the` **holder** `to apply for permanent residence.'`

1. Numbers Removal	`Over , now hold legal immigrant documents, which, after five years of annual renewal, entitles the holder to apply for permanent residence.`
2. Punctuation Removal	`Over  now hold legal immigrant documents which after five years of annual renewal entitles the holder to apply for permanent residence`
3. Lower Casing	`over  now hold legal immigrant documents which after five years of annual renewal entitles the holder to apply for permanent residence`
4. Tokenization	['over', 'now', 'hold', 'legal', 'immigrant', 'documents', 'which', 'after', 'five', 'years', 'of', 'annual', 'renewal', 'entitles', 'the', 'holder', 'to', 'apply', 'for', 'permanent', 'residence']
5. Stop words removal	['hold', 'legal', 'immigrant', 'documents', 'five', 'years', 'annual', 'renewal', 'entitles', 'holder', 'apply', 'permanent', 'residence']

Stop words removal problem

Before:

```
'At the police station he did not make any such claims, but had alleged torture
```
only `at the district court trials.'`

After:

```
['police', 'station', 'make', 'claims', 'alleged', 'torture', 'district', 'court', 'trials']
```

Needs to be manually corrected!

Embeddings

GloVe 50d

['hold', 'legal', 'immigrant', 'documents', 'five', 'years', 'annual', 'renewal', 'entitles', 'holder', 'apply', 'permanent', 'residence']

[**embedding**('hold)', **embedding**('legal'), **embedding**('immigrant'),...]

Compute mean

```
[ 0.3866, -0.2908, -0.1011,
0.1910,  0.1187,  0.1503,  0.1034,
0.3662, 0.2403, -0.2460, -0.2139,
-0.4824, -0.3774, -0.4342,
0.5688, -0.1132,...]
```

50d tensor of numbers

Pre-processing: Join

'sentence1': 'This growth is the direct result of the increased number of baccalaureate holders, who form the potential market for higher education.'

'.'

'sentence2': 'Over 5,000 now hold legal immigrant documents, which, after five years of annual renewal, entitles the holder to apply for permanent residence.'

[0.1160, 0.3436, 0.2790,...]

50d tensor of numbers

[1.5164e-01, 3.0177e-01, -1.6763e-01,...]

50d tensor of numbers

[-0.0470, 0.5142, -0.0584,...]

50d tensor of numbers

[0.1160, 0.3436, 0.2790,..., 1.5164e-01, 3.0177e-01, -1.6763e-01,...,-0.0470, 0.5142, -0.0584,...]

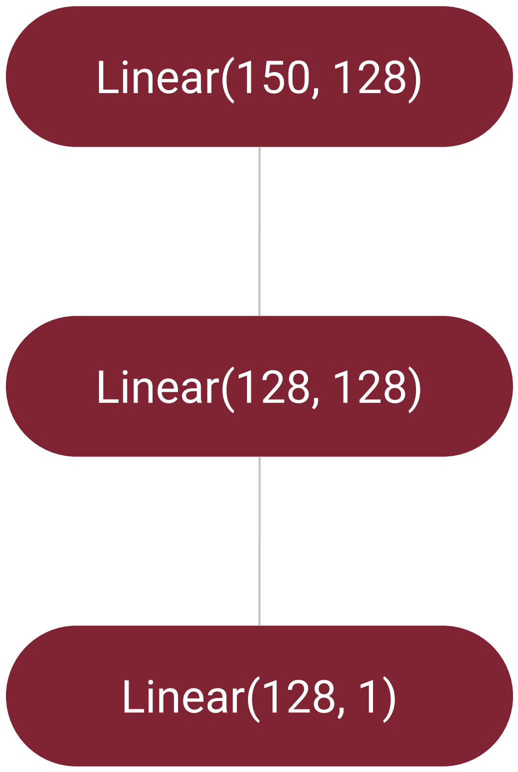
150d tensor of numbers

Baseline

Model

Model

Architecture



Hyper-parameters

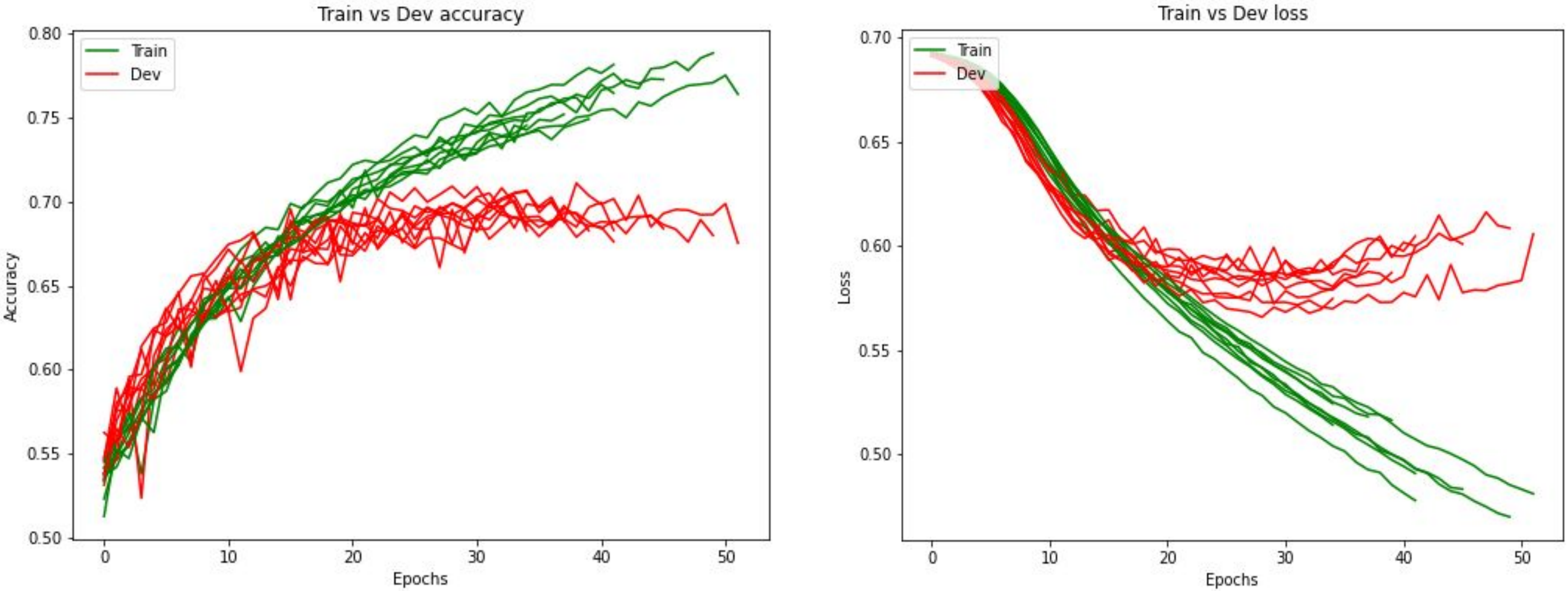
Epochs	50
ES patience	7
ES threshold	0.009
Batch size	64
Embedding dim	50
N features	150
N hidden units	128
N hidden layers	2
Activation	ReLU
Optimizer	Adam
Learning Rate	0.0001

Baseline

Performance

Performance: Training vs Validation Accuracy and Loss

- Best accuracy: **0.7236**
- Problem: **Overfitting about 30-40 epoch (regularization does not help!)**



Over 10 independent runs

Baseline 2

Pre-processing

Pre-processing: Single sentence processing

Given: `'It will place as many demands on our material resources as on our intellectual capabilities.'`

1. Numbers Removal
2. Punctuation Removal
3. Lower Casing
4. Tokenization
5. Stop words removal

['place', 'many', **'demands'**, 'material', **'resources'**, 'intellectual', **'capabilities'**]

6. Lemmatization

['place', 'many', **'demand'**, 'material', **'resource'**, 'intellectual', **'capability'**]

Pre-processing: Join

'sentence1': 'This growth is the direct result of the increased number of baccalaureate holders, who form the potential market for higher education.'

'.'

'sentence2': 'Over 5,000 now hold legal immigrant documents, which, after five years of annual renewal, entitles the holder to apply for permanent residence.'

['growth', 'direct', 'result', 'increased', 'number', 'baccalaureate', 'holder', 'form', 'potential', 'market', 'higher', 'education']

['.']

['hold', 'legal', 'immigrant', 'document', 'five', 'year', 'annual', 'renewal', 'entitles', 'holder', 'apply', 'permanent', 'residence']

['growth', 'direct', 'result', 'increased', 'number', 'baccalaureate', 'holder', 'form', 'potential', 'market', 'higher', 'education',

'.',

'hold', 'legal', 'immigrant', 'document', 'five', 'year', 'annual', 'renewal', 'entitles', 'holder', 'apply', 'permanent', 'residence']

Pre-processing: Prepare input for RNN

['growth', 'direct', 'result', 'increased', 'number', 'baccalaureate', 'holder', 'form', 'potential', 'market', 'higher', 'education', '.', 'hold', 'legal', 'immigrant', 'document', 'five', 'year', 'annual', 'renewal', 'entitles', 'holder', 'apply', 'permanent', 'residence']

Indexed vocabulary with 2 special indices:

- **0 - for padding token**
- **1 - for Out-of-Vocabulary**

```
[554, 1496, 714, 1043, 225, 31822, 6101, 685, 1158, 213, 611, 633, 4, 804, 832, 5660, 2883,
176, 64, 942, 9239, 53808, 6101, 3517, 2275, 3700, 0, 0, …, 0]
```

Padding of a sequence with ML - L zeros:

- **ML - max length of a sequence in the batch**
- **L - length of the current sequence**

GloVe 50d

(random for 0 and 1)

```
[embedding(554),
embedding(1496),…,
embedding(0)]
```

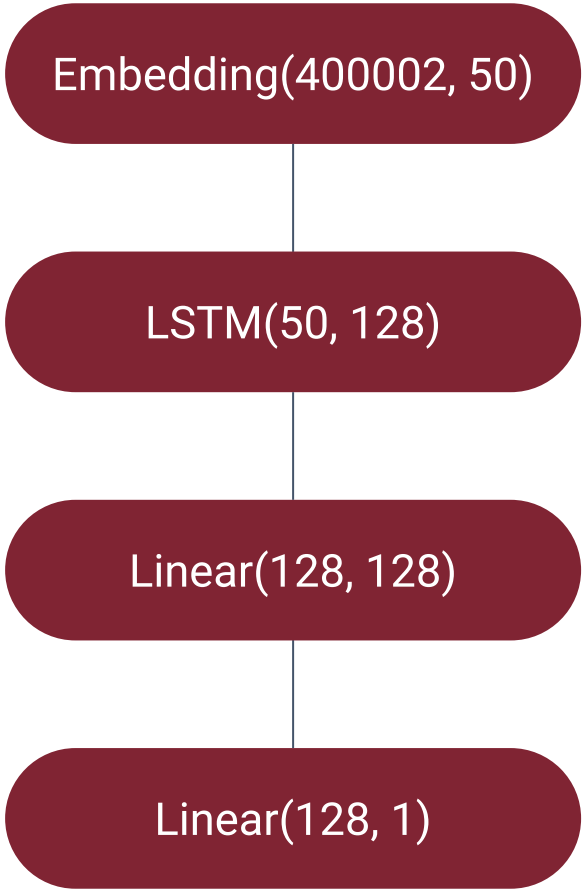
50 x ML tensor of numbers

Baseline 2

Model

Model

Architecture



Hyper-parameters

Epochs	70
ES patience	7
ES threshold	0.01
Batch size	256
Embedding dim	50
N features	50
N hidden units	128
N LSTM cells	1
Activation	ReLU
Optimizer	Adam
Learning Rate	0.0001
Decay Rate	0.000001
Dropout Rate	0.0

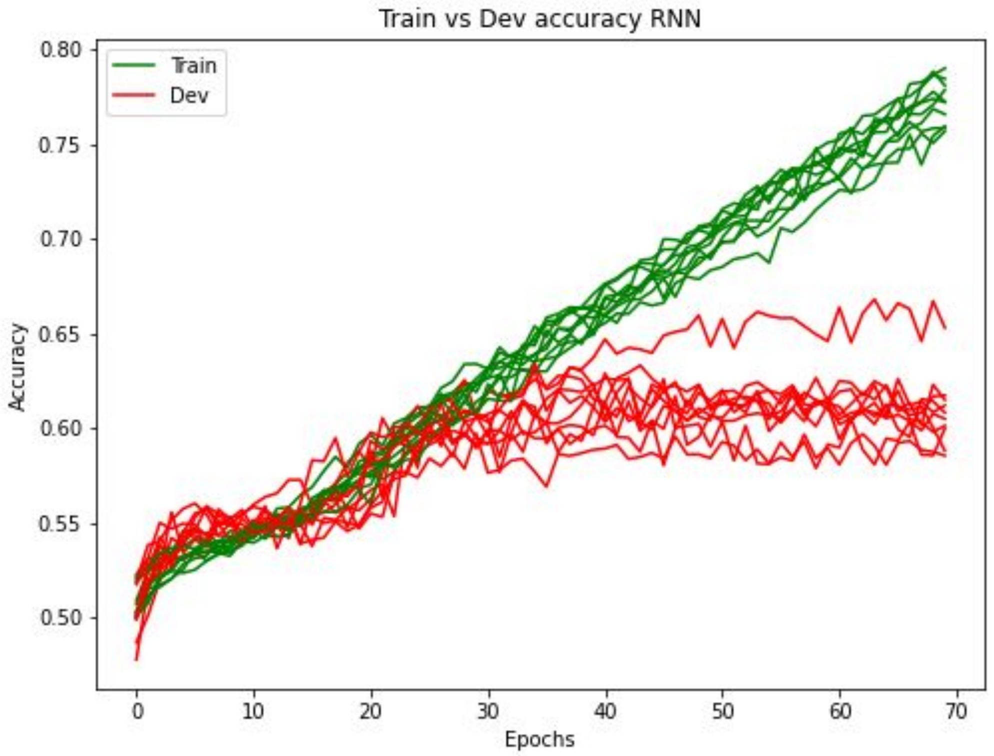
Baseline 2

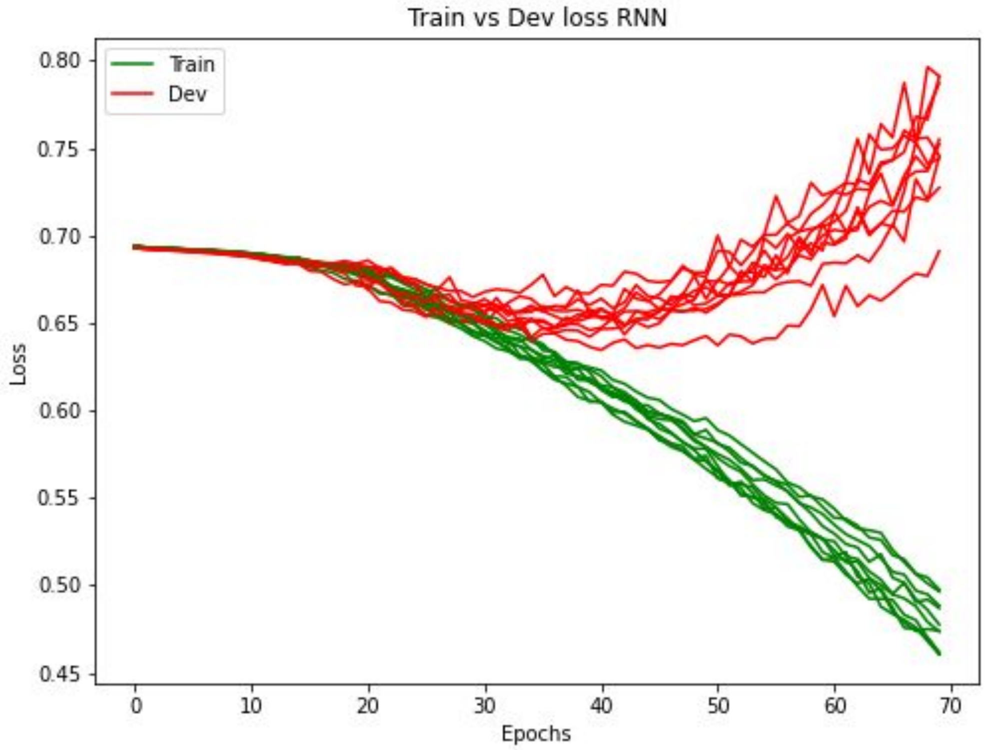
Performance

Performance: Training vs Validation Accuracy and Loss

SAPIENZA
UNIVERSITÀ DI ROMA

- Best accuracy: **0.6681 (vs 0.7236 we had before)**
- Problem: **Overfitting about 50 epoch**

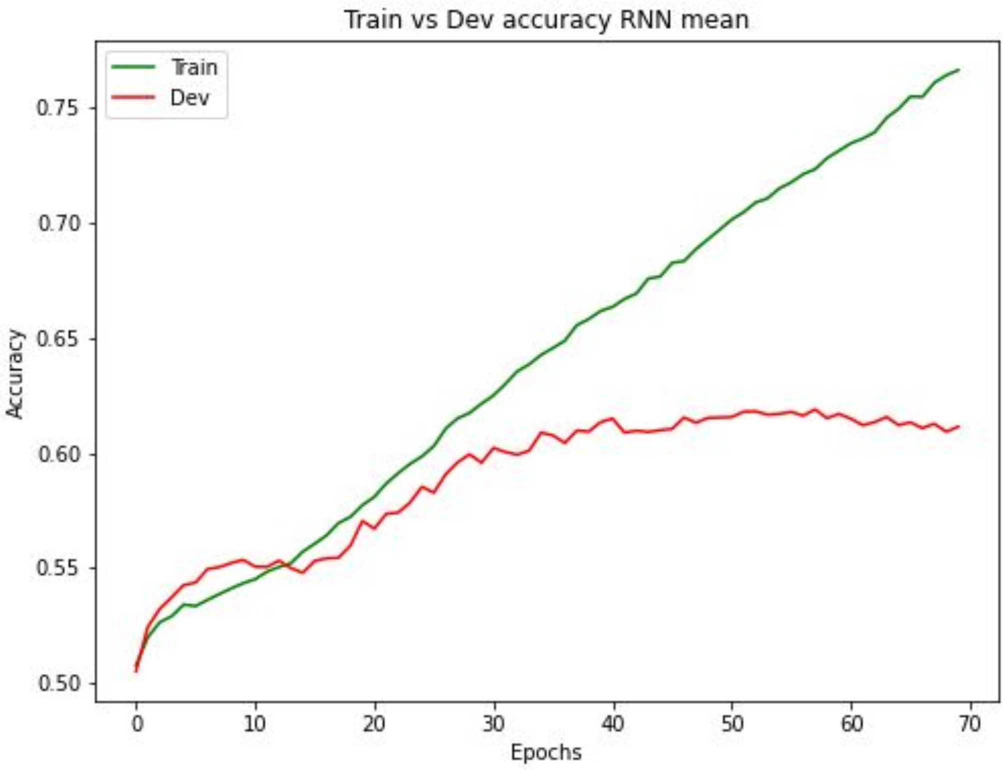


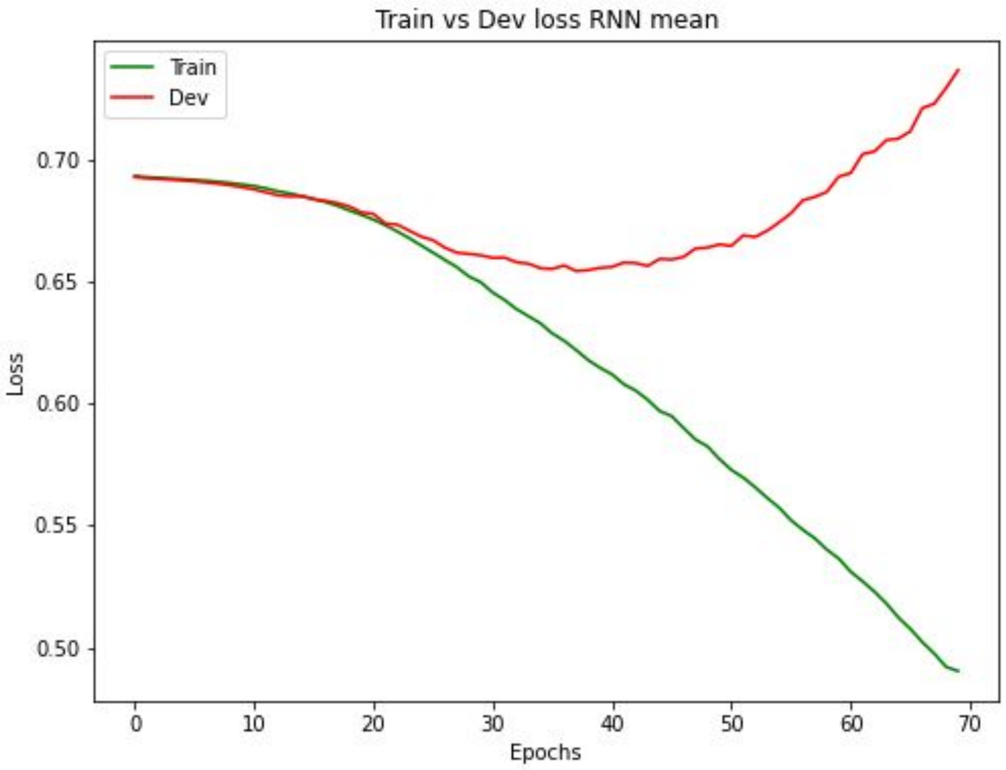


Over 10 independent runs

Performance: Averaged graphs

- Accuracy: got stabilized about **0.60 (vs 0.66 the best accuracy)**
- Problem: **overfitting, the best performance achieved by lucky initialization**

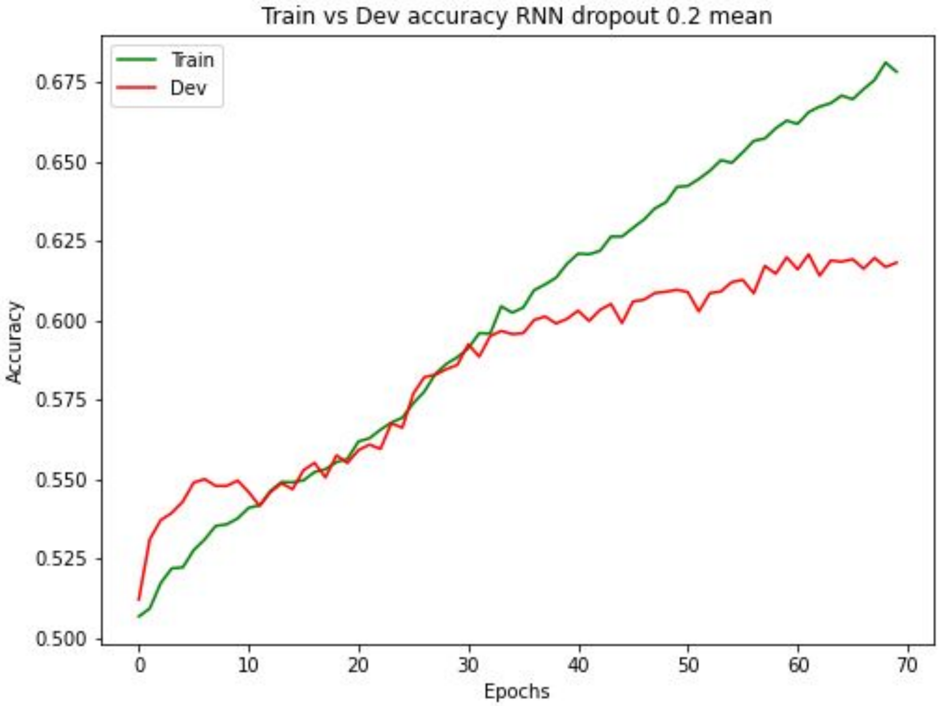


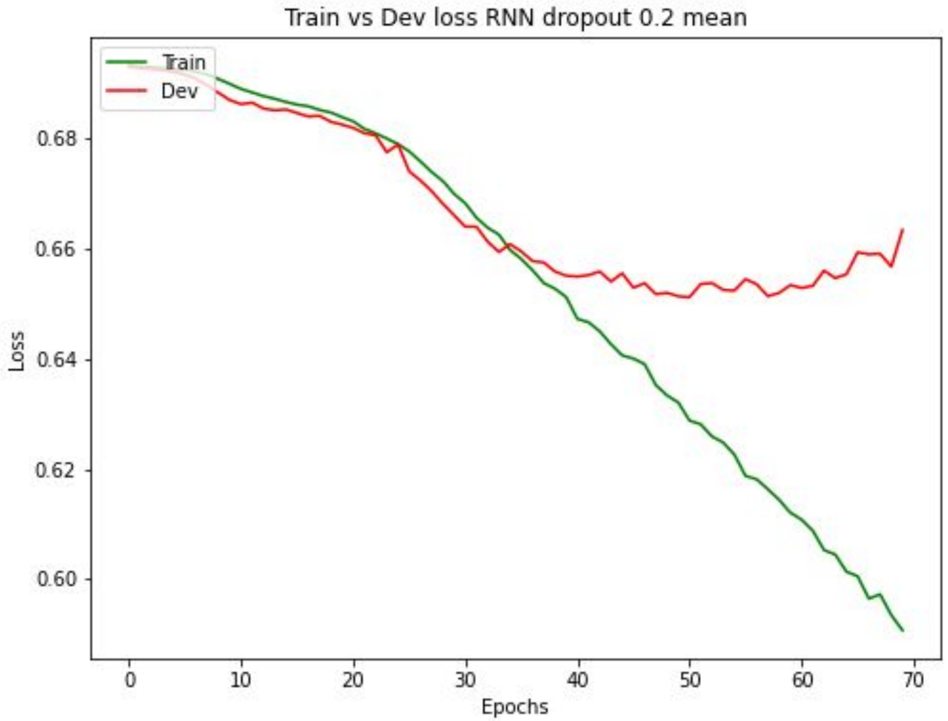


Averaged over 10 independent runs

Performance: Add dropout

- Accuracy: got stabilized about **0.62 (vs 0.60 without dropout)**
- Problem: **overfitting, but less obvious**

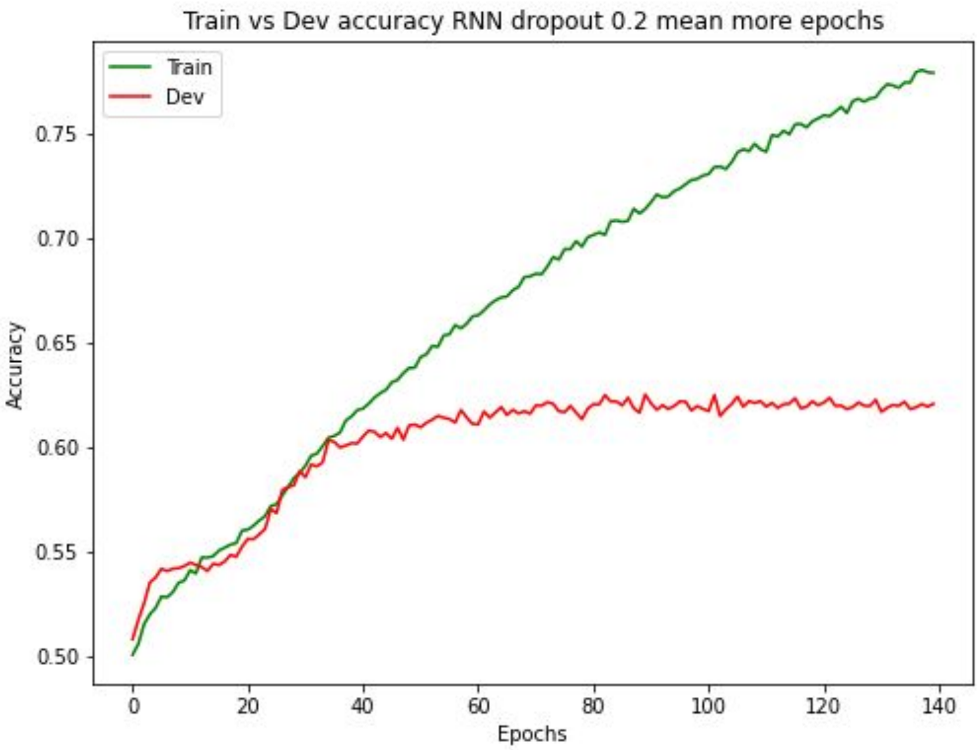


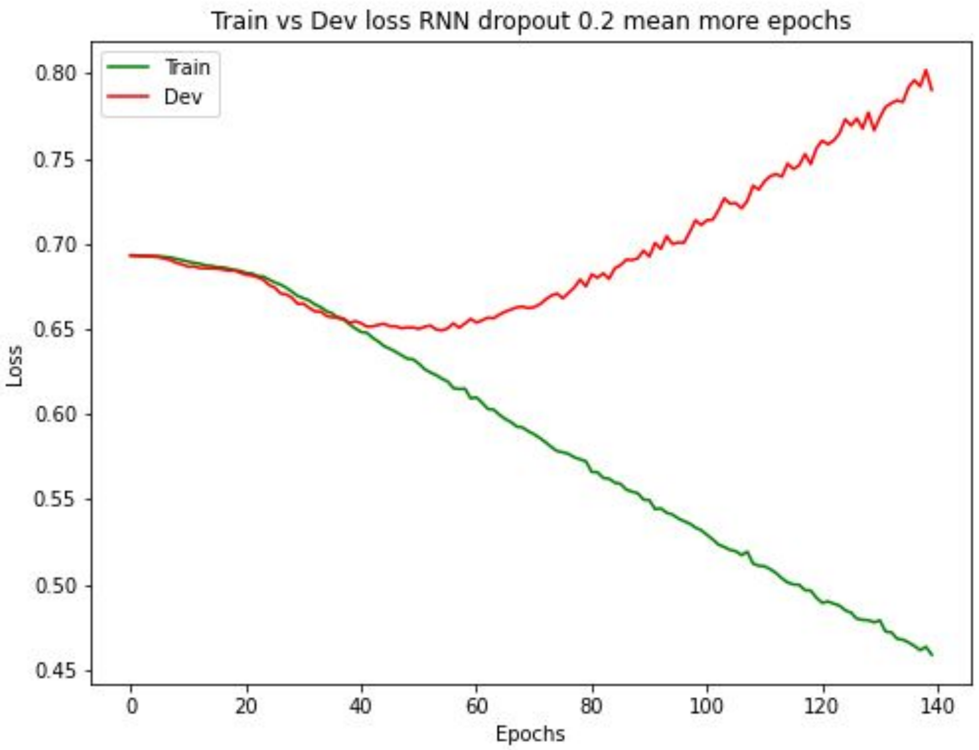


Averaged over 10 independent runs, p = 0.2, 2 LSTM layers, dropout applied after embedding layer, between 2 LSTMs and after them

Performance: More epochs

- Accuracy: got stabilized about **0.62 (vs 0.62)**
- Problem: **overfitting, now it is obvious**

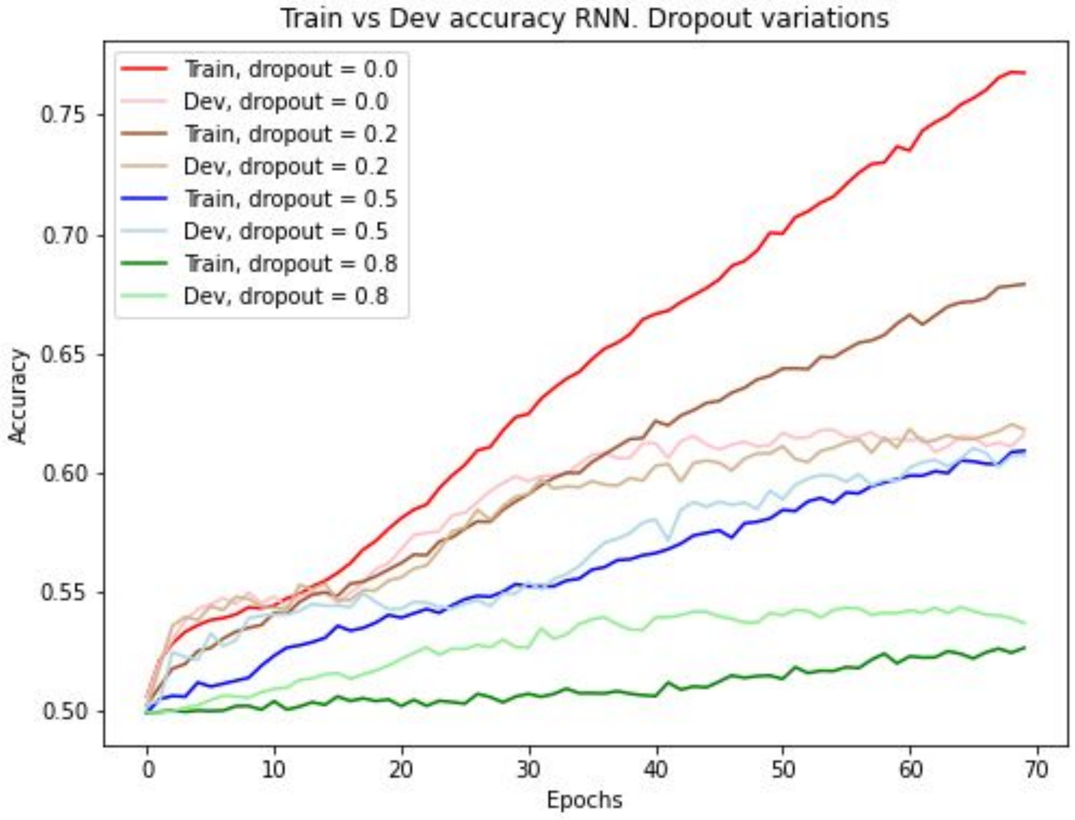




Averaged over 10 independent runs, p = 0.2, 2 LSTM layers, dropout applied after embedding layer, between 2 LSTMs and after them, 140 epochs of training

Performance: Dropout variations

- Accuracy: got stabilized about **0.62 or less**
- Problem: **overfitting, and dropout variation does not help**



Averaged over 10 independent runs

2a

Pre-processing and Model

Pre-processing and Model: differences with Baseline 2

Pre-processing

Model

Need to keep index of the target word in the sequence of indices (additionally to an index of the last not padding token)

Extract two sequence encodings: corresponding to the representation of a whole sentence and corresponding to the target word

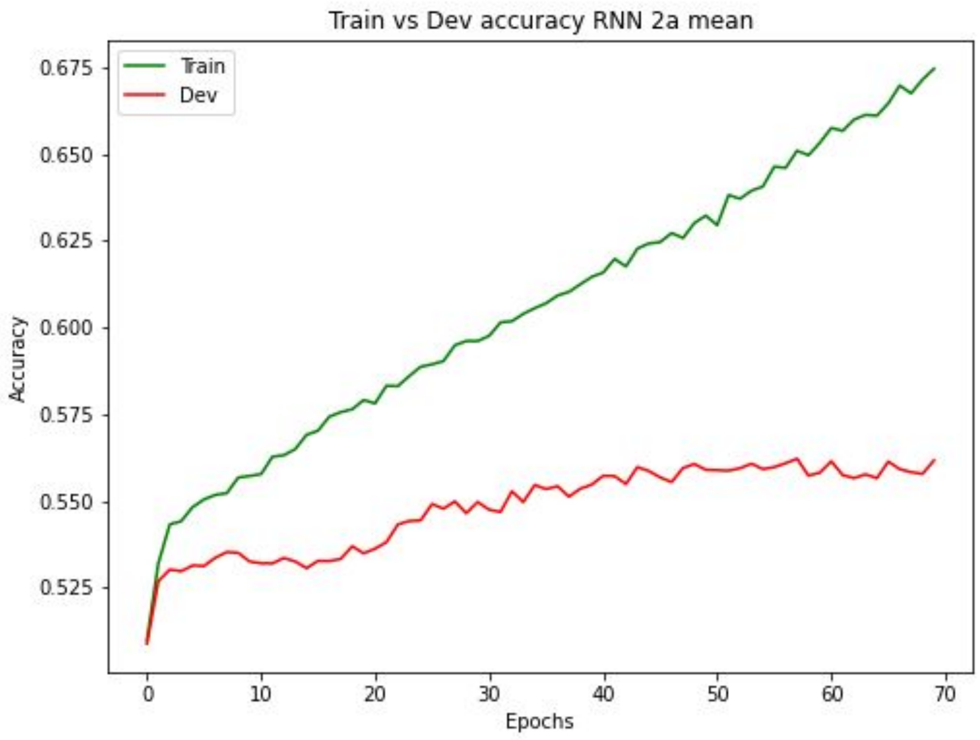
Double-labeling

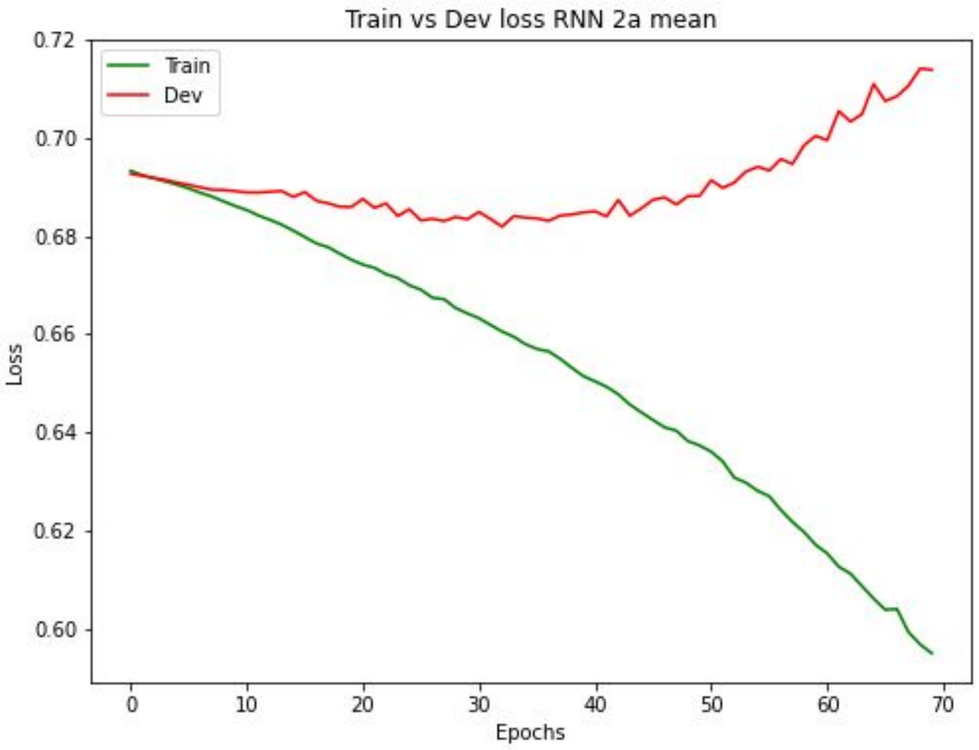
2a

Performance

Performance: Training vs Validation Accuracy and Loss

- Best accuracy: **0.5776 (vs 0.6681 we had for Baseline 2)**
- Problem: **improvement is not achieved**





Averaged over 10 independent runs

2b

Idea

2b

The underlying idea of this model was to perform binary classification over the following representation of the training data:

Feature 1	Feature 2	Feature 3	Feature 4	Label
whole sequence encoding for the sentence 1	sequence encoding corresponding to the target word in the sentence 1	whole sequence encoding for the sentence 2	sequence encoding corresponding to the target word in the sentence 2	gt label

- Best accuracy: **0.5773 - comparable with 2a, but not comparable with Baseline 2**

Conclusion

- Even if sequence encoding approach is potentially more powerful, sometimes **simpler approach can be better** performing;

- Handling **overfitting** is a challenging task, and in our case standard regularization techniques did not help. Therefore, some more sophisticated approaches either to it or to the model architecture design/way of pre-processing are needed;

- However, the best performing model achieved quite **decent performance** by means of common NLP practices for the pre-processing.