Natural Language Processing

Assignment 3 – Report

Tommy Yu

May. 25, 2018

1. Language Modeling

Dataset:

<s> what drink would you like, coffee or tea </s>

<s> what drink would you like, coffee or Coke </s>

<s> what drink would you like, coffee or Sprite </s>

<s> what drink would you like, tea or coffee </s>

<s> what drink would you like, tea or Coke </s>

<s> what drink would you like, tea or Sprite </s>

<s> what drink would you like, Coke or coffee </s>

<s> what drink would you like, Coke or tea </s>

<s> what drink would you like, Coke or Sprite </s>

<s> what drink would you like, Sprite or coffee </s>

<s> what drink would you like, Sprite or tea </s>

<s> what drink would you like, Sprite or Coke </s>

<s> you drink </s>

Model U:

	</s>	what	drink	would	you	like	,	coffee	tea	Coke	Sprite	or
<s>	/	12	0	0	1	0	0	0	0	0	0	0
you	0	0	1	0	0	12	0	0	0	0	0	0
drink	1	0	0	12	0	0	0	0	0	0	0	0

$$probability(< s > \ you\ drink\ \ </s >) = \frac{1}{13} \times \frac{1}{13} \times \frac{1}{13}$$

Model S (after add-1 smoothing):

	</s>	what	drink	would	you	like	,	coffee	tea	Coke	Sprite	or
<s>	/	13	1	1	2	1	1	1	1	1	1	1
you	1	1	2	1	1	13	1	1	1	1	1	1
drink	2	1	1	13	1	1	1	1	1	1	1	1

$$probability(< s > \ you\ drink\ \ </s >) = \frac{1}{12} \times \frac{2}{25} \times \frac{2}{25}$$

The probability under Model S is larger.

2. POS Tagging

All models are trained with a maximum of 50 epochs and default batch size of 32, and early stopping is based validation set (dev) accuracy.

Number in brackets refer to the number of epochs when early stopping occurs.

Baseline

1 hidden layer with width 128, w=1, “tanh”: Accuracy: 83.59% (26)

Varying w

w	0	2
Accuracy	80.69% (20)	83.24% (25)

As expected, merely considering the center word (w=0) without its context yields a worse result than the baseline, however, increasing the window size to 2 didn’t improve performance either.

Change non-linearity functions

f	identity	ReLU	sigmoid
Accuracy	83.86% (29)	83.59% (28)	74.55% (39)

Identity and ReLU yield similar results to the baseline, while the sigmoid function performs terribly in this task.

Change hidden layers

Number of layers	Accuracy	
0	79.95% (50)	
1	Small: 128 - baseline	Large: 256
		83.22% (22)
2	Small: 256, 128	Large: 512, 256
	84.14% (16)	84.51% (16)

Increasing the number of layers lead to better performance of the model. So does increasing the width of layers, as we can see from comparing the 2-layer models. The slight decrease in accuracy after doubling the layer width in the 1-layer models may largely be due to the difference in the number of epochs trained.