AI6127 Assignment 2 Report

Zhang Ziyi

G2102704G

Question 2a.

In order to practically completing all the cross-validation tests by using the current available computer, unfortunately, all the datasets are still getting filtered by the condition of MAX_Length = 20. The other important hyperparameters used in the experiment are hidden state size = 256, learning rate = 0.01, training iterations = 75000.

As for the details of implementing the beam search, there are three different settings tested in this experiment, which are with length normalization and without max length limitation (denoted as “default” in the data table); with length normalization and with max length limitation (denoted as “N Max L” in the data table); without length normalization and without max length limitation (denoted as “No N No Max L” in the data table). The detailed test results and data can be seen in the appended tables. The % performance column is all calculated regarding the greedy search performance as the baseline.

Some interesting findings after analysing the data are:

- Beam search generally doesn’t outperform greedy search in this experiment. When applying the setting as without length normalization and without max length limitation, beam search’s performance is even much worse than the greedy search (average -85% across all the four datasets). This result demonstrates the importance of length normalization factor in beam search’s scoring formula.

- The default setting of beam search (with length normalization and without max length limitation) in general slightly improves the BLEU-3 score (average +5% improvement) while has in average -5% performance decrease in terms of BLEU-1 score. Although the difference is not significant, but it is quite consistent across all the datasets. The takeaway here is that the increase in the BLEU-3 score should be more important than the decrease in the BLEU-1 score. The reason is that high BLEU-1 score may not really reflect a good translation result. As what could be seen from the random translation results, many translated sentences consist of repetitive stop words, in order to achieve a higher unigram overlap score. In contrary, increase in BLEU-3 score can better reflect the sentence fluency and similarity to the natural sentences. As a result, when using default setting, beam search should be considered to improves the MT performance when comparing using the greedy search, due to the improvement in BLEU-3 score.
- The beam search setting with length normalization and with max length limitation generally is indifferent from greedy search. The assumption here is that the two parameters' effect kind of offset with each other.
- The last ru-en dataset generally has lower performance compared to the other three datasets, across all the parameters. The reason is assumed to be lack of normalisation methods. Because all the other three languages (cs, de, fr) can be converted to English alphabet-like characters, they can all be applied with the predefined alphabet normalisation methods. However, Russian characters cannot be mapped to alphabets, so they don't get pre-processed in the same way as the other three languages.

Question 2b.

For this part, word embeddings weights are pretrained by the word2vec model offered in the Gensim package. Both the encoder and decoder corresponding word embedding layers are replaced with the pretrained ones, and by default the pre-trained embedding layers' weights are frozen.

Some interesting findings are:

- Generally all the datasets get comparable improvements, especially the ru-en dataset. Pre-training the word embeddings is definitely helpful for increasing the network performance. As for the significant increase in the ru-en dataset (average + 60% increase), one possible reason could be due to the overly poor performance in the past. Another potential reason might be by pretraining the word embedding layer, the network gets to know more about the language, thus mitigating the effect of lacking effective normalisation.
- Another finding is that the de-en dataset doesn't improve much comparing to the randomly initialised embedding case. In fact, most of the parameters, except training loss, decrease. However, when trying to set the pretrained embedding weights also trainable, nearly all the parameters get improvements over both the baseline and frozen embedding case. (Please refer to the Improvement over baseline (%) and Improvement over freeze embedding (%) columns) Nevertheless, not all datasets can get further improvement by allowing the pretrained embedding layer to also get updated. For

example, the cs-en dataset has relatively same performance when comparing to the frozen embedding case.

Question 2c.

i. There are two main differences between the attention decoder at lecture and attention decode used here. The first difference is the attention score's calculation. At lecture, the attention score is calculated by the dot product between the decode hidden state at this time step and all the encoder hidden states. However, in this network, attention score is calculated by concatenating the decoder input embedding and previous time step decoder hidden state, then passing the concatenated result into a liner layer. The second difference is that the attention final output is not concatenated with the decoder hidden state and then passing to the probability function, as what is described during lecture. In this network, the attention output is concatenated with the decoder input embedding and passed to the decoder GRU as a composite input.

ii. According to the table appended, when comparing to the results from question 2b, both the two variants multiplicative attention and additive attention could have positive impact on the implemented network here. In other words, both the two attention variants are better than the original attention mechanism in the network. The reason could be due to the different attention score calculation method. The original attention score calculation, as described in the part i, is by concatenating the decoder input and decoder previous hidden state, which does not relate the decoder hidden state with the encoder hidden state. In contrast, the two variants here do. As a result, the stronger connection established between the decoder and encoder could assist the decoder to focus on a specific part in the encoder, which in the end improves the overall performance.

Part 2a Cross-validation Average Results

cs-en dataset			Average values	Improvement (%)
Training loss			4.92414	
Greedy		BLEU-1	0.2377	
		BLEU-2	0.07838	
		BLEU-3	0.02824	
Default	Beam	BLEU-1	0.22464	-5.494320572
		BLEU-2	0.07852	0.178616994
		BLEU-3	0.02992	5.949008499
N Max L	Beam	BLEU-1	0.22468	0.017806268
		BLEU-2	0.07858	0.076413653
		BLEU-3	0.02994	0.06684492
No N No Max L	Beam	BLEU-1	0.03334	-85.16111803
		BLEU-2	0.01112	-85.84881649
		BLEU-3	0.00412	-86.23914496

de-en dataset				
Training loss			4.8609	
Greedy		BLEU-1	0.26282	
		BLEU-2	0.09578	
		BLEU-3	0.03788	
Default	Beam	BLEU-1	0.24906	-5.235522411
		BLEU-2	0.09466	-1.169346419
		BLEU-3	0.03888	2.639915523
N Max L	Beam	BLEU-1	0.24884	-0.088332129
		BLEU-2	0.09462	-0.042256497
		BLEU-3	0.0388	-0.205761317
No N No Max L	Beam	BLEU-1	0.03892	-84.35942774
		BLEU-2	0.0147	-84.46417248
		BLEU-3	0.00606	-84.3814433

fr-en dataset				
Training loss			4.59174	
Greedy		BLEU-1	0.2558	
		BLEU-2	0.09422	
		BLEU-3	0.03666	
Default	Beam	BLEU-1	0.2454	-4.06567631
		BLEU-2	0.09452	0.318403736
		BLEU-3	0.0386	5.291871249
N Max L	Beam	BLEU-1	0.24444	-0.391198044
		BLEU-2	0.09334	-1.248413034
		BLEU-3	0.03778	-2.124352332
No N No Max L	Beam	BLEU-1	0.04378	-82.08967436
		BLEU-2	0.01622	-82.62266981
		BLEU-3	0.0064	-83.05982001

ru-en dataset				
Training loss			5.15864	
Greedy		BLEU-1	0.17754	
		BLEU-2	0.0473	
		BLEU-3	0.01446	
Default	Beam	BLEU-1	0.1678	-5.486087642
		BLEU-2	0.04752	0.465116279
		BLEU-3	0.01568	8.437067773
N Max L	Beam	BLEU-1	0.16788	0.047675805
		BLEU-2	0.04762	0.21043771
		BLEU-3	0.0158	0.765306122
No N No Max L	Beam	BLEU-1	0.0176	-89.51632118
		BLEU-2	0.00494	-89.62620748
		BLEU-3	0.00162	-89.74683544

With pretrained word embedding layer (frozen weights)					
cs-en dataset			Average values	w/o pretrain average	Improvement (%)
Training loss			4.7682	4.92414	-3.166847409
Greedy		BLEU-1	0.25295	0.2377	6.415649979
		BLEU-2	0.08615	0.07838	9.913243174
		BLEU-3	0.032025	0.02824	13.4029745
Default	Beam	BLEU-1	0.2415	0.22464	7.50534188
		BLEU-2	0.086275	0.07852	9.876464595
		BLEU-3	0.033575	0.02992	12.21590909
de-en dataset					
Training loss			4.7383	4.8609	-2.522166677
Greedy		BLEU-1	0.25475	0.26282	-3.070542577
		BLEU-2	0.093575	0.09578	-2.302150762
		BLEU-3	0.03615	0.03788	-4.567053854
Default	Beam	BLEU-1	0.246675	0.24906	-0.957600578
		BLEU-2	0.093375	0.09466	-1.357489964
		BLEU-3	0.0375	0.03888	-3.549382716
fr-en dataset					
Training loss			4.542	4.59174	-1.083249487
Greedy		BLEU-1	0.257666667	0.2558	0.729736774
		BLEU-2	0.092066667	0.09422	-2.28543126
		BLEU-3	0.034266667	0.03666	-6.52845972
Default	Beam	BLEU-1	0.250333333	0.2454	2.010323282
		BLEU-2	0.092366667	0.09452	-2.278177458
		BLEU-3	0.035066667	0.0386	-9.153713299
ru-en dataset					
Training loss			4.950966667	5.15864	-4.02573805
Greedy		BLEU-1	0.238866667	0.17754	34.54245053
		BLEU-2	0.077666667	0.0473	64.20014094
		BLEU-3	0.028766667	0.01446	98.9396035
Default	Beam	BLEU-1	0.230966667	0.1678	37.64402066
		BLEU-2	0.078266667	0.04752	64.70258137
		BLEU-3	0.0297	0.01568	89.41326531
With pretrained word embedding layer (trainable weights)					
cs-en dataset			Average	over baseline (%)	over frozen (%)
Training loss			4.7201	-4.143667727	-1.008766411
Greedy		BLEU-1	0.2474	4.080774085	-2.194109508
		BLEU-2	0.0862	9.977034958	0.058038305
		BLEU-3	0.0333	17.91784703	3.981264637
Default	Beam	BLEU-1	0.2407	7.149216524	-0.33126294
		BLEU-2	0.0871	10.92715232	0.956244567
		BLEU-3	0.0351	17.31283422	4.542069993

de-en dataset					
Training loss			4.6877	-3.56312617	-1.067893548
Greedy		BLEU-1	0.2586	-1.60566167	1.511285574
		BLEU-2	0.09805	2.370014617	4.782260219
		BLEU-3	0.03995	5.464625132	10.51175657
Default	Beam	BLEU-1	0.25075	0.678551353	1.651971217
		BLEU-2	0.0988	4.373547433	5.809906292
		BLEU-3	0.04115	5.838477366	9.733333333

Attention Variants

				Average values	2b Average values	Improvemets (%)
cs-en dataset						
Multiplicative	Default	Training loss		4.7432	4.7682	-0.524306866
		Greedy	BLEU-1	0.2464	0.25295	-2.589444554
			BLEU-2	0.0893	0.08615	3.656413233
			BLEU-3	0.0339	0.032025	5.854800937
		Beam	BLEU-1	0.2351	0.2415	-2.65010352
			BLEU-2	0.0883	0.086275	2.347145755
			BLEU-3	0.0356	0.033575	6.031273269
Additive	Default	Training loss		4.7432	4.7682	-0.524306866
		Greedy	BLEU-1	0.2453	0.25295	-3.024313105
			BLEU-2	0.0875	0.08615	1.567034243
			BLEU-3	0.0342	0.032025	6.791569087
		Beam	BLEU-1	0.2416	0.2415	0.041407867
			BLEU-2	0.0887	0.086275	2.810779484
			BLEU-3	0.0354	0.033575	5.435591958
de-en dataset						
Multiplicative	Default	Training loss		4.6734	4.7383	-1.369689551
		Greedy	BLEU-1	0.2555	0.25475	0.294406281
			BLEU-2	0.0968	0.093575	3.446433342
			BLEU-3	0.0401	0.03615	10.92669433
		Beam	BLEU-1	0.2486	0.246675	0.780379041
			BLEU-2	0.0983	0.093375	5.274431058
			BLEU-3	0.0422	0.0375	12.53333333
Additive	Default	Training loss		4.672	4.7383	-1.399236013
		Greedy	BLEU-1	0.2444	0.25475	-4.062806673
			BLEU-2	0.0997	0.093575	6.545551697
			BLEU-3	0.0425	0.03615	17.56569848
		Beam	BLEU-1	0.247	0.246675	0.131752306
			BLEU-2	0.1032	0.093375	10.52208835
			BLEU-3	0.0448	0.0375	19.46666667
fr-en dataset						
		Training loss		4.4929	4.542	-1.081021576
			BLEU-1	0.2589	0.257666667	0.478654592

Multiplicative	Default	Greedy	BLEU-2	0.0972	0.092066667	5.575669804
			BLEU-3	0.0386	0.034266667	12.6459144
		Beam	BLEU-1	0.2563	0.250333333	2.383488682
			BLEU-2	0.0993	0.092366667	7.50631541
			BLEU-3	0.041	0.035066667	16.92015209
Additive	Default	Training loss		4.453	4.542	-1.959489212
		Greedy	BLEU-1	0.2461	0.257666667	-4.489003881
			BLEU-2	0.0945	0.092066667	2.64301231
			BLEU-3	0.0384	0.034266667	12.06225681
		Beam	BLEU-1	0.2414	0.250333333	-3.568575233
			BLEU-2	0.096	0.092366667	3.933597979
			BLEU-3	0.0403	0.035066667	14.92395437

ru-en dataset

Multiplicative	Default	Training loss		4.968	4.950966667	0.344040558
		Greedy	BLEU-1	0.2295	0.238866667	-3.921295004
			BLEU-2	0.0804	0.077666667	3.519313305
			BLEU-3	0.0312	0.028766667	8.458864426
		Beam	BLEU-1	0.2174	0.230966667	-5.873863472
			BLEU-2	0.0791	0.078266667	1.064735945
			BLEU-3	0.032	0.0297	7.744107744
Additive	Default	Training loss		4.9606	4.950966667	0.194574797
		Greedy	BLEU-1	0.2471	0.238866667	3.446832263
			BLEU-2	0.0822	0.077666667	5.836909871
			BLEU-3	0.0302	0.028766667	4.982618772
		Beam	BLEU-1	0.2364	0.230966667	2.352431808
			BLEU-2	0.0814	0.078266667	4.003407155
			BLEU-3	0.0311	0.0297	4.713804714