Experimental Results for NMT 2M Chn2Eng News

Qiang Li

1. Experimental Results

1.1 Generalization for source & target training corpora

Source : 39,425,342 tokens, 1,858,452 lines, 21.21 tokens/sent
Target : 44,964,569 tokens, 1,858,452 lines, 24.19 tokens/sent

	10k	20k	30k	40k	50k	~	70k
Chn	0.93	0.97	**0.98**	0.99	~		1
Eng	0.97	0.99	~		1.00		

Development/Test sets	Mt06	Mt04	Mt05	Mt08
Length Ratio (Eng/Chn)	1.26	1.32	1.20	1.23

Systems	Settings	Perplexity	Dev	Test		
		Train / Valid	MT06	MT04	MT05	MT08
NiuTrans. SMT	$number, $date, $time					
	baseline	- / -	32.1	36.7	31.3	26.0
	+ online_LM	- / -	33.9 + 1.8	38.2 + 1.5	32.8 + 1.5	27.9 + 1.9
Open-source NMT	1 layer, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune	- / -	29.6	33.9	28.6	23.1
	finetune	- / -	34.3	-	-	-
	+ unk	- / -	34.9	41.1	34.3	27.1
	ensemble (4)	- / -	37.3 + 5.2	43.5 + 6.8	35.9 + 4.6	29.3 + 3.3
NiuTrans. NMT	4 layers, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune (m13)	8.68 / 13.02	27.9	34.9	27.5	20.6
	tune (m15)	9.09 / 12.15	30.4	37.5	29.6	23.5
	finetune (m12)	7.59 / 9.84	36.5	42.2	34.6	29.3
	finetune (m13)	7.39 / 9.75	36.8	42.9	35.2	29.7
	finetune (m14)	7.34 / 9.72	36.9 + 4.8	42.9 + 6.2	35.5 + 4.2	29.6 + 3.6
	finetune (m15)	7.33 / 9.71	36.7	42.8	35.5	29.5
	ensemble (3)	- / -	-	43.1	35.5	-
	ensemble (4)	- / -	-	42.9	-	-
	2 layers, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune (m15)	7.40 / 9.41	38.3	44.2	37.2	30.2
	finetune (m15)	6.70 / 9.42	38.7	44.7	37.2	30.7
	+ unk	**- / -**	**39.1** **+ 7.0**	**45.1** **+ 8.4**	**37.7** **+ 6.4**	**31.3** **+ 5.3**
	1 layer, 30k src & 30k tgt vocab, 1000 lstm, $number, $date, $time					
	tune (m15)	8.01 / 10.58	35.9	40.8	34.5	28.8
	finetune (m15)	7.56 / 10.60	36.2	41.2	34.5	28.7
	+ unk	- / -	36.7 + 4.6	41.4 + 4.7	35.0 + 3.7	29.2 + 3.2

说明：

1. 所有系统与 NiuTrans.SMT 中 baseline 进行比较
2. 使用的训练数据/开发集/测试集均为泛化的数据，NiuTrans.NMT 评价在泛化的数据下进行
3. Settings 下(m12)/(m13)/(m14)/(m15)为使用同一实验设置下不同轮数的模型
4. NiuTrans.NMT 所有 tune/finetune 均为 10 epoch

结论：

1. 2M 泛化新闻数据上，同样设置下，2 层效果好于 4 层、1 层
2. finetune 在 2 层/1 层系统上效果不明显，需要选取更好的 finetune 初值。在 4 层系统上效果明显（tune 时使用 10 epoch 训练 4 层不充分)
3. PPL 越低，BLEU 越高
4. Ensemble 在 NiuTrans.NMT 效果不明显

1.2 Ungeneralization for source & target corpora

Source : 38,764,129 tokens, 1,847,856 lines, 20.98 tokens/sent
Target : 45,144,464 tokens, 1,847,856 lines, 24.43 tokens/sent

Systems	Settings	Perplexity	Dev	Test		
		Train / Valid	MT06	MT04	MT05	MT08
NiuTrans. NMT	4 layers, 30k src & 30k tgt vocab, 1000 lstm					
	tune (m15)	8.28 / 11.71				
	finetune (m15)	7.38 / 11.58	33.8	43.4	32.2	27.8
	+ unk	- / -	34.2	43.8	32.5	28.3
	4 layers, 200k src & 50k tgt vocab, 1000 lstm					
	tune (m15)	7.75 / 11.44	34.6	44.2	32.6	28.9
	+ unk	- / -	34.8	44.4	32.8	29.3
	finetune (m15)	6.86 / 11.18	35.1	45.3	32.8	29.8
	+ unk	- / -	35.4	45.5	33.0	30.1

说明：
1. 使用的训练数据/开发集/测试集均为不泛化的数据
3. Settings 下(m15)为 tune/finetune 最后一轮输出的模型
4. NiuTrans.NMT 所有 tune/finetune 均为 10 epoch

结论：
1. 2M 不泛化新闻数据上，扩大源语言/目标语言可有效减系统的困惑度，继而 BLEU 提高
2. finetune 在两种设置下，开发集/测试集没有得到较大的性能提高（提高幅度在 1 个 BLEU 点左右）
3. 不泛化与泛化实验结果相比，mt06,mt05,mt08 的 BLEU 下降 2-3 个 BLEU,mt04 增加 0.5

2. Runtime

GeForce GTX 1080

Tune : 4100 tokens/s (4 hidden layers, 1000 lstm size, 30k src & tgt word types)
FineTune : 4100 tokens/s (4 hidden layers, 1000 lstm size, 30k src & tgt word types)
Decoding : 2.7 sents/s (beam_size=20)

Tune : 3362.4 minutes (10 Epoch, src=39,425,342 tokens, tgt=44,964,569 tokens)
FineTune : 3495.0 minutes (10 Epoch, src=39,425,342 tokens, tgt=44,964,569 tokens)
Decoding : 2 sents/s (beam_size=20)

Training corpora:
Source data: 39,425,342 tokens, 1,858,452 lines, generalization
Target data: 44,964,569 tokens, 1,858,452 lines, generalization

Settings:
src_vocab=30k, tgt_vocab=30k, lstm_size=1000, minibatch=64, dropout=0.2, learning_rate=0.7, epoch=10, beam_size=20, finetune=true

	Exp	**Mt06**	**Mt04**	**Mt05**	**Mt08**
Generalization	SMT-baseline	32.1	36.7	31.3	26.0
	SMT-online-lm	33.9	38.2	32.8	27.9
	NMT-o-finetune-1layer-unk	34.9	41.1	34.3	27.1
	NMT-o-ensemble (4)	37.3	43.5	35.9	29.3
	NMT-i-finetune-4layers	36.7	42.8	35.5	29.5
	NMT-i-finetuen-2layers	38.7	44.7	37.2	30.7
	NMT-i-finetune-2layers-unk	**39.1 (+7.0)**	**45.1 (+8.4)**	**37.7 (+6.4)**	**31.3 (+5.3)**
Ungeneralization	**NMT-i-tune-4layers-200k-50k-unk**	34.8	44.4	32.8	29.3
	NMT-i-finetune-4layers-200k-50k-unk	35.4	45.5	33.0	30.1

Experimental Results for NMT 10M Chn2Eng Oral

1. Experimental Results

1.1 Generalization for source & target training corpora

Source : 141,663,980 tokens, 9,977,755 lines, 14.20 tokens/sent
Target : 152,920,696 tokens, 9,977,755 lines, 15.33 tokens/sent

Tune/Finetune:
Src_vocab_size : 30k
Tgt_vocab_size : 30k
LSTM_size : 1000
Minibatch : 64
Dropout : 0.2
Learning_rate : 0.7
Epoch : 10

Decoding:
Beam_size : 20

	10k	**20k**	**30k**	**40k**	**~**	**80k**
Chn	0.91	0.95	**0.97**	0.98	~	0.99
Eng	0.93	0.96	**0.97**	0.98	~	0.98

Exp	**Mt06**	**Mt04**	**Mt05**	**Mt08**
NiuTrans_baseline				
NiuTrans_online_LM				
NMT-opensource				
tune				
finetune				
unk				
ensemble (4)				
NiuTrans.NMT				
tune (m12)				
tune (m13)				
tune (m14)				
tune (m15)				
ensemble (4)				
avg (4)				
finetune (m12)				
finetune (m13)				
finetune (m14)				

finetune (m15)				
ensemble (4)				
avg (4)				

2. Runtime