# Progress of Neural Machine Translation with Memory Network

Yang Feng

Jan 3, 2016

## 1 Baseline

**Improve the baseline** by

1. rnn to bidirectional_rnn;
2. SGD optimizer to Adadelta optimization;
3. initializing the initial attention state;
4. seperating hidden_edim and hidden_units (in the baseline, they share a common parameter);
5. changing the frequency of saving models.

**To do:**
change one-best greedy search to **beam search**.

**Cannot get reasonable results of baseline**, shown in Table 1. Investigated in terms of

1. output order chaos;
2. file format.

**Fix this problem** by using greater learning rate of 0.5, and results are shown in Table 2.

## 2 NMT+MN

### 2.1 Adding attended lexical translation to the decoder

### 2.2 Investigate

1. learning rate, shown in Table 2.2

2. the placeholder for the target translation of the source words whose target translation is less than trans-limit, shown in Table 2.2
   v0: all use PAD_ID
   v1: use NULL_ID for words, and PAD_ID for PAD
   v2: all use NULL_ID

3. size of translations (trans-limit), shown in Table 2.2
   For each source word, the number of target translation words

4. training algorithm shown in Table 2.2
   Also tried Adam, but got very bad results.

| System | Speed (per epoch) | Dev (BLEU4 ) | Test (BLEU4) |
|---|---|---|---|
| nmt | 84s | 10.3 | 11.1 (49w epch) |
| nmt$^+$ | 108s | 13.0 | 15.4 (26w epch) |
| nmt$^+$+mn | 180s | 11.1 | 12.5 (30w epch) |

Table 1: configuration: learning_rate=0.001, hidden_edim=310, hidden_units=310, batch_size=80, decay_learning_rate=0.99, SGD

| System | Speed (per epoch) | Dev (BLEU4 ) | Test (BLEU4) |
|---|---|---|---|
| nmt (sgd) | 84s | 29.6 | 31.4 (28800 step) |
| nmt$^+$(*sgd*) | 108s | 33.0 | 38.1 (34200 step) |
| nmt$^+$+mn (ada) _v0 | 180s | 30.6 | 36.0 (71400 step) |

Table 2: configuration: learning_rate=0.5, hidden_edim=310, hidden_units=500, batch_size=80, decay_learning_rate=0.99

| systm | rate=0.001 | | rate=0.005 | |
|---|---|---|---|---|
| | dev | test | dev | test |
| nmt$^+$ | 13.0 | 15.4 (26w epch) | 33.0 | 38.1 (34200 step) |
| nmt$^+$+mn | 11.1 | 12.5 (30w epch) | 30.6 | 36.0 (71400 step) |

Table 3: results of different learning rate

| translimit | 1 | 2 | | 10 |
|---|---|---|---|---|
| trainer | Adadelta | Adadelta | SGD | Adadelta |
| nmt$^+$+mn-v0 | 30.6–36.0 | – | – | – |
| nmt$^+$+mn-v1 | 30.0–35.3 | 29.8-36.6 | 29.9–37.3 | 29.5–36.7 |
| nmt$^+$+mn-v2 | – | 30.0–38.1 | 30.5–37.8 | 30.4–35.8 |

Table 4: comparison of different configurations of **palcehlder**, **trans-limit**, **training algorithm**.

nmt+

BLEU = 38.1, 74.8/47.5/31.6/20.5 (BP=0.978, ratio=0.979, hyp_len=3715, ref_len=3796) (34200)

v0-1-Ada

BLEU = 36.0, 73.3/45.4/29.3/18.9 (BP=0.976, ratio=0.976, hyp_len=3706, ref_len=3797) (71400)

v1-1-Ada

BLEU = 35.3, 72.1/43.5/28.0/18.7 (BP=0.986, ratio=0.986, hyp_len=3786, ref_len=3838) (34600)

v1-2-Ada

BLEU = 36.6, 72.8/45.1/29.5/20.0 (BP=0.982, ratio=0.982, hyp_len=3742, ref_len=3810) (103800)

v1-2-sgd

BLEU = 37.3, 73.8/45.8/30.4/20.3 (BP=0.981, ratio=0.981, hyp_len=3708, ref_len=3781) (48500)

v1-10-Ada

BLEU = 36.7, 72.7/45.8/29.6/19.0 (BP=0.993, ratio=0.993, hyp_len=3814, ref_len=3841) (15000)

v2-2-Ada

BLEU = 38.1, 74.0/46.5/31.0/21.1 (BP=0.984, ratio=0.984, hyp_len=3770, ref_len=3832) 73000)

v2-2-sgd

BLEU = 37.8, 73.2/45.5/30.3/21.3 (BP=0.987, ratio=0.987, hyp_len=3753, ref_len=3802) (50000)

v2-10-Ada

BLEU = 35.8, 73.3/45.0/28.9/18.2 (BP=0.987, ratio=0.987, hyp_len=3723, ref_len=3772) (25000)