Sequence to Sequence Learning with Neural Networks

Ilya Sutskever, Oriol Vinyals, Quoc V. Le *(2014 NIPS)*

NLP Basic Paper Study week4

Presenter : 허환

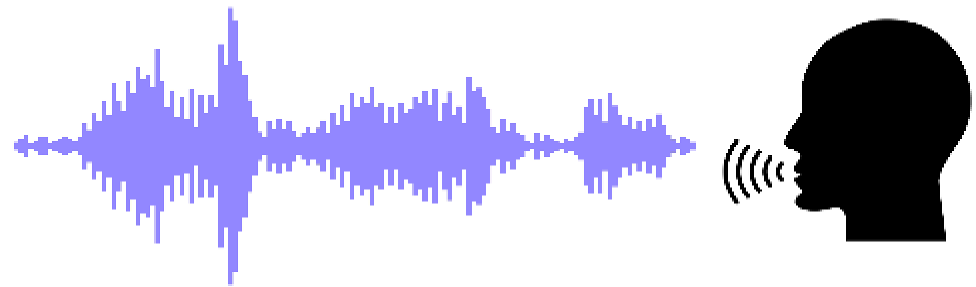
Table of content

1. Introduction

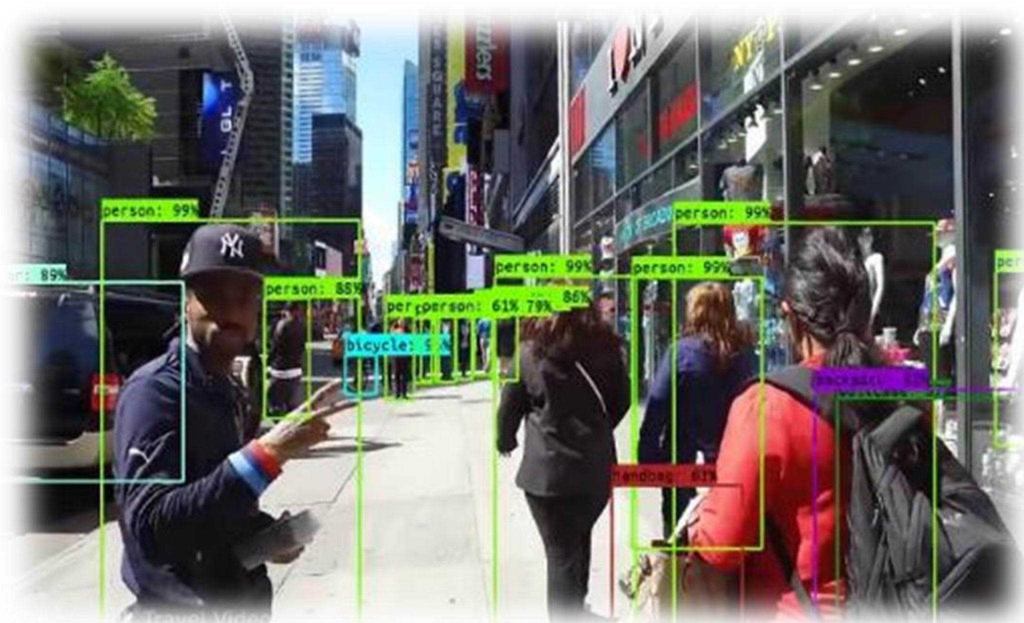
- 1.1 DNN's area
- 1.2 DNN's limitation

1.1 DNN’s Area

- Speech recognition



- Vision recognition



1.2 DNN’s limitation

- Well with large/labeled dataset, with fixed length input

  - Vision recognition, Speech recognition

- How about variable length?

  - Sequence mapping (Machine Translation), Question Answering (Chatbot)

  => BAD ! Need domain independent method

2. Model

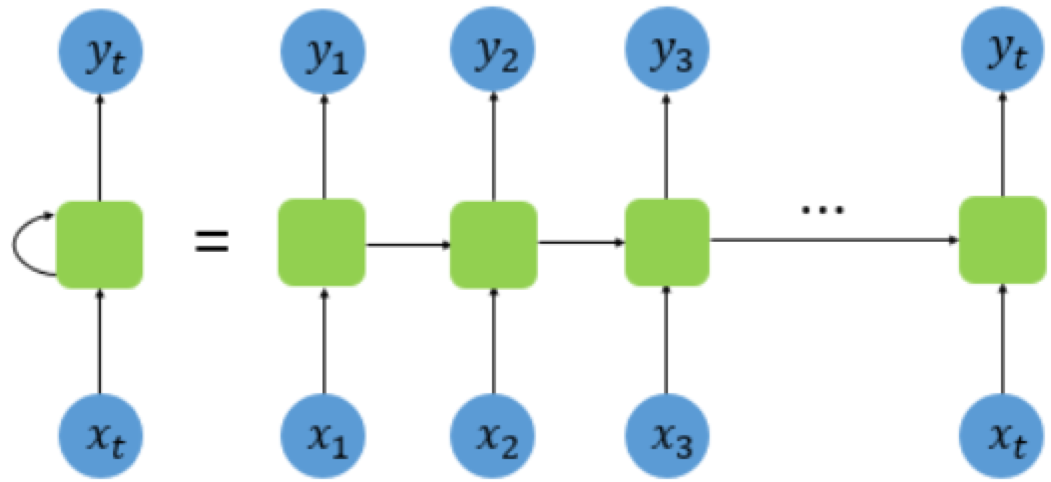
- 2.1 RNN
- 2.2 LSTM
- 2.3 Sequence to Sequence Model

2.1 RNN

- Recurrent Neural Network

<vanilla RNN>

$$
\begin{aligned}
h_t &= \mathrm{sigm}\left(W^{\mathrm{hx}} x_t + W^{\mathrm{hh}} h_{t-1}\right) \\
y_t &= W^{\mathrm{yh}} h_t
\end{aligned}
$$



The result of the hidden layer enters the input of the next calculation!!

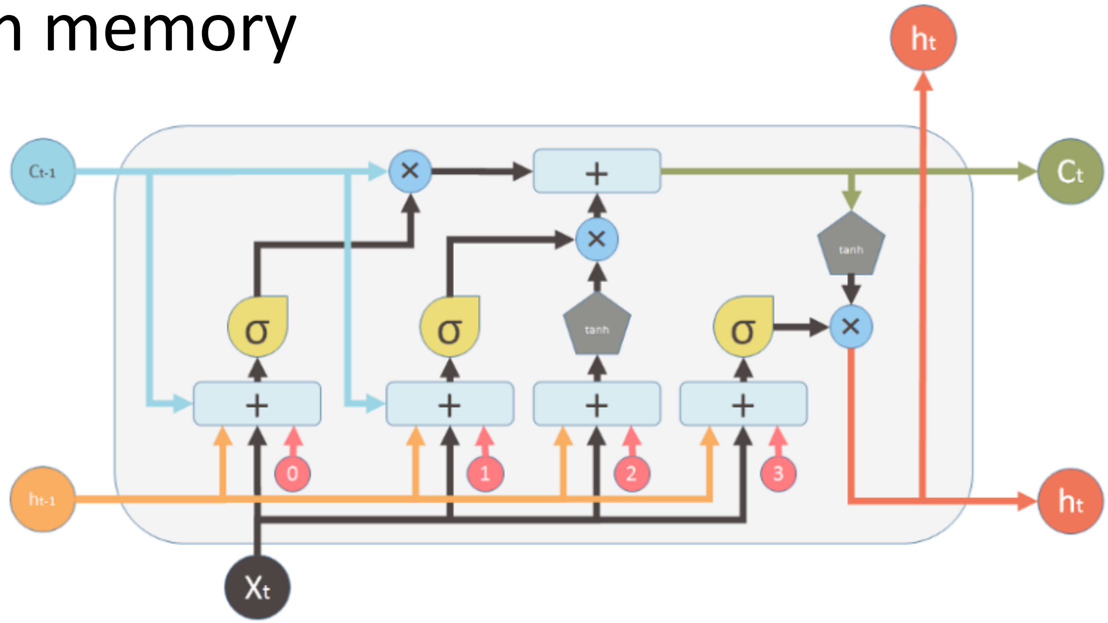
=> can map sequence to sequence

=> but, long-range dependency!!

$$
\frac{\partial h_T}{\partial h_t} = \left(W^{hh} \right)^{T-t} * \prod_{i=t}^{T-1} sigm'\left(W^{hh} h_i + W^{hx} x_{i+1} \right)
$$

2.2 LSTM

- Long-Short term memory



Add cell state that can memorize (I,F,O,G gate)

=> partial solution of 'long-range dependency'

$$\frac{\partial C_T}{\partial C_t} = \prod_{i=t+1}^{T} f_i$$

$$\begin{aligned}
f_t &= \sigma(W_{xh_f} x_t + W_{hh_f} h_{t-1} + b_{h_f}) \\
i_t &= \sigma(W_{xh_i} x_t + W_{hh_i} h_{t-1} + b_{h_i}) \\
o_t &= \sigma(W_{xh_o} x_t + W_{hh_o} h_{t-1} + b_{h_o}) \\
g_t &= \tanh(W_{xh_g} x_t + W_{hh_g} h_{t-1} + b_{h_g}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}$$

2.2 LSTM

- Conditional Probability Distribution

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

- Pros and Cons
  1. Back Propagation
  2. Memory circuit + Neural Network
  3. Long-Term dependency
  4. Exploding Gradient

2.3 Sequence to Sequence Model

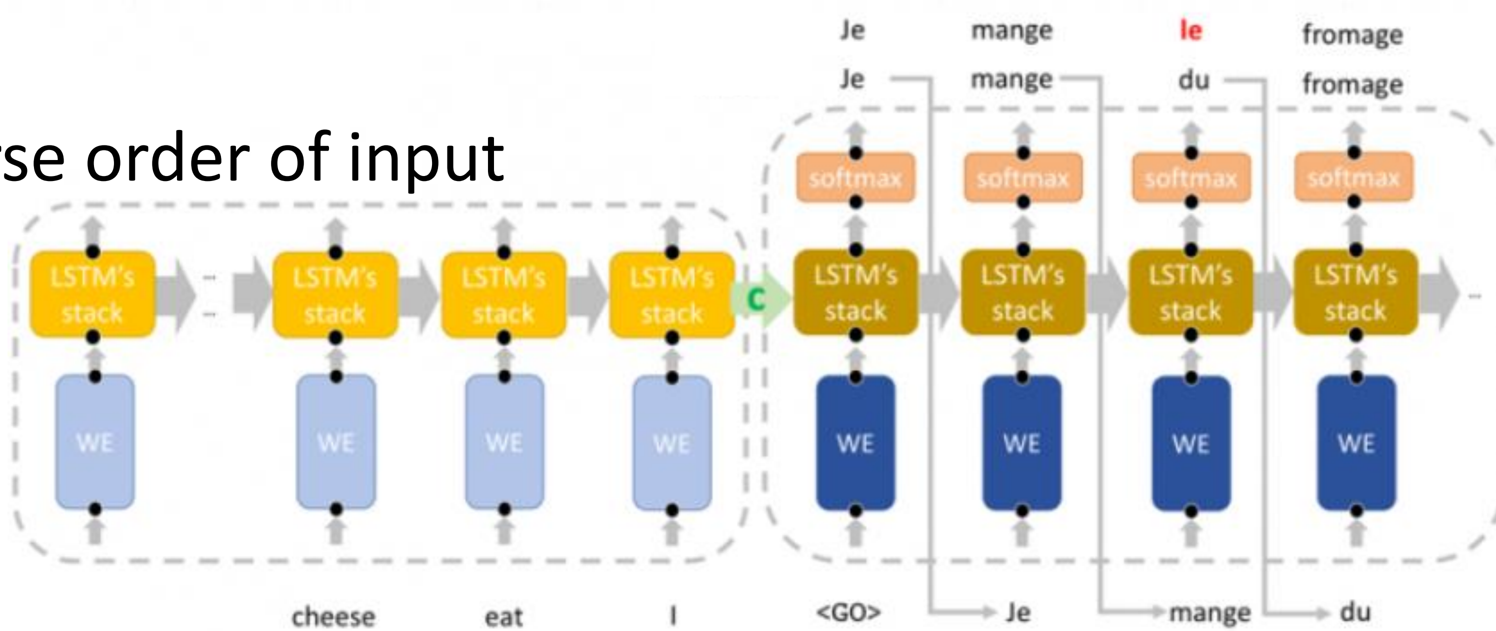
- Encoder – Decoder

  1) negligible computational cost

  2) train multiple language pairs simultaneously

- LSTM with4 layers

- Reverse order of input



3. Experiment

- 3.1 Training Details
- 3.2 Metric
- 3.3 Results

3.1 Training Details

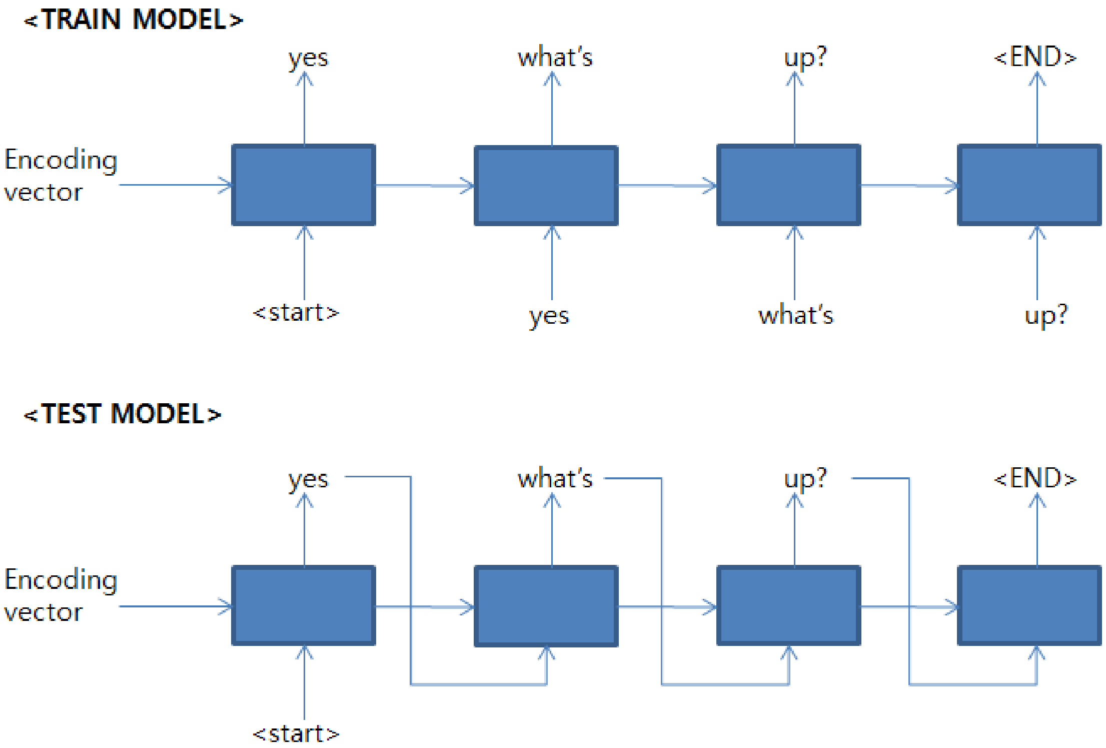
- Decoding and Rescoring

1. Train :

$$1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

2. Test :

$$\hat{T} = \arg\max_{T} p(T|S)$$



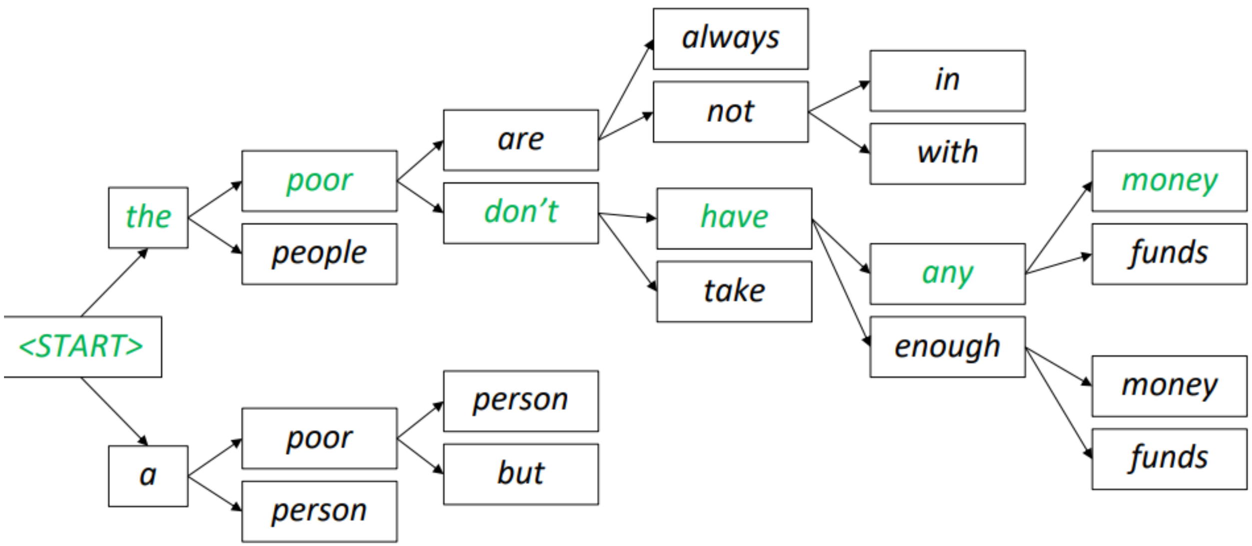
3.1 Training Details

- Beam Search (Test)

  Extension of Greedy search

  : discard all but B most likely hypothesis

Beam size = 2



3.1 Training Details

- Reversing order of words in source sentence

$$c \;\; b \;\; a \longrightarrow \alpha \;\; \beta \;\; \gamma$$

Easy to establish connection between En-Decoder

Why?

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

maybe Beam search + conditional probability?

3.1 Training Details

- Exploding Gradient

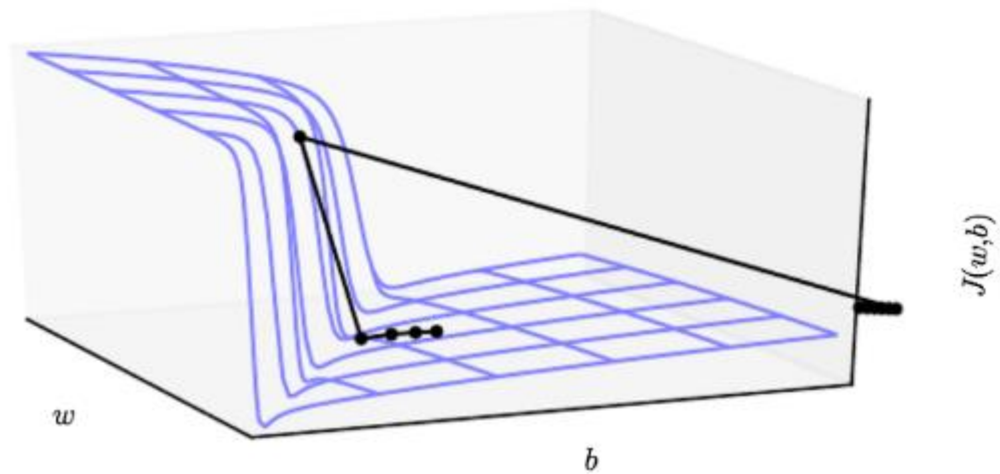
  Solution: Norm Clipping (scaling gradient)

Just scaling :

gradient direction is unchanged

Algorithm 1 Pseudo-code for norm clipping

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|} \hat{\mathbf{g}}$
end if



3.1 Training Details

- Sentences within a minibatch were roughly of the same length :

Why?

with randomly chosen training sentences,
computation in the minibatch is wasted

How to?

extra preprocessing?

3.2 Metric

- BLEU(Bilingual Evaluation Understudy)

$$\text{BLEU}_{wN}\ \text{score} = BP \times \sqrt[N]{\prod_{n=1}^{N} p_n}$$

$$BP = \begin{cases} 1 & \text{if } |\mathcal{C}| > |\mathcal{R}_{\text{closest}}| \\ e^{1-r/c} & \text{if } |\mathcal{C}| \leq |\mathcal{R}_{\text{closest}}| \end{cases}$$

$$p_n = \frac{\sum_{w_1 \ldots w_n \in \mathcal{C}} \min\left(|\mathcal{C}|_{w_1 \ldots w_n}, \max_{\mathcal{R}} \left(|\mathcal{R}|_{w_1 \ldots w_n} \right) \right)}{\sum_{w_1 \ldots w_n \in \mathcal{C}} |\mathcal{C}|_{w_1 \ldots w_n}}$$

C: candidate , R: reference (translation)

C_w : # of co-occurrence w in candidate

3.2 Metric

- BLEU’s limitation

1. BLEU : product of n-grams

   -> likely to have a score of 0 in the sentence unit

   -> can only be evaluated in corpus units.

2. BLEU : score a single translated reference.

   -> can not deal with substitute expressions

   -> In real life, sentences can be translated in many different ways

3.3 Results

- Performance of LSTM on WMT's 14 Eng to Frch

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

ensemble's power

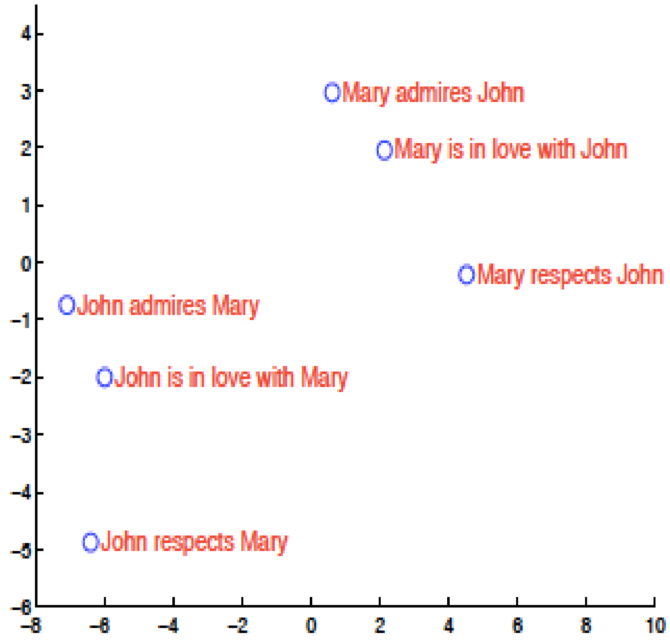
3.3 Results

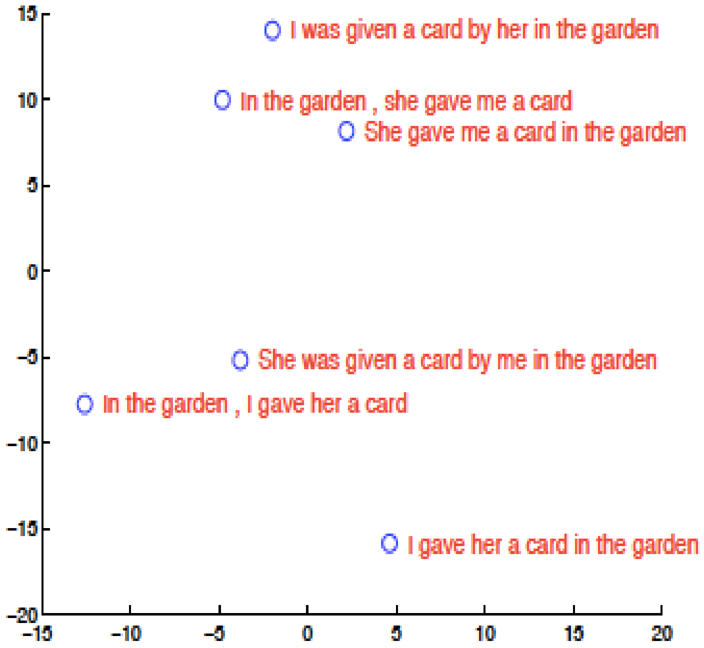
- Method comparison NN with SMT

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	$\sim$45

3.3 Results

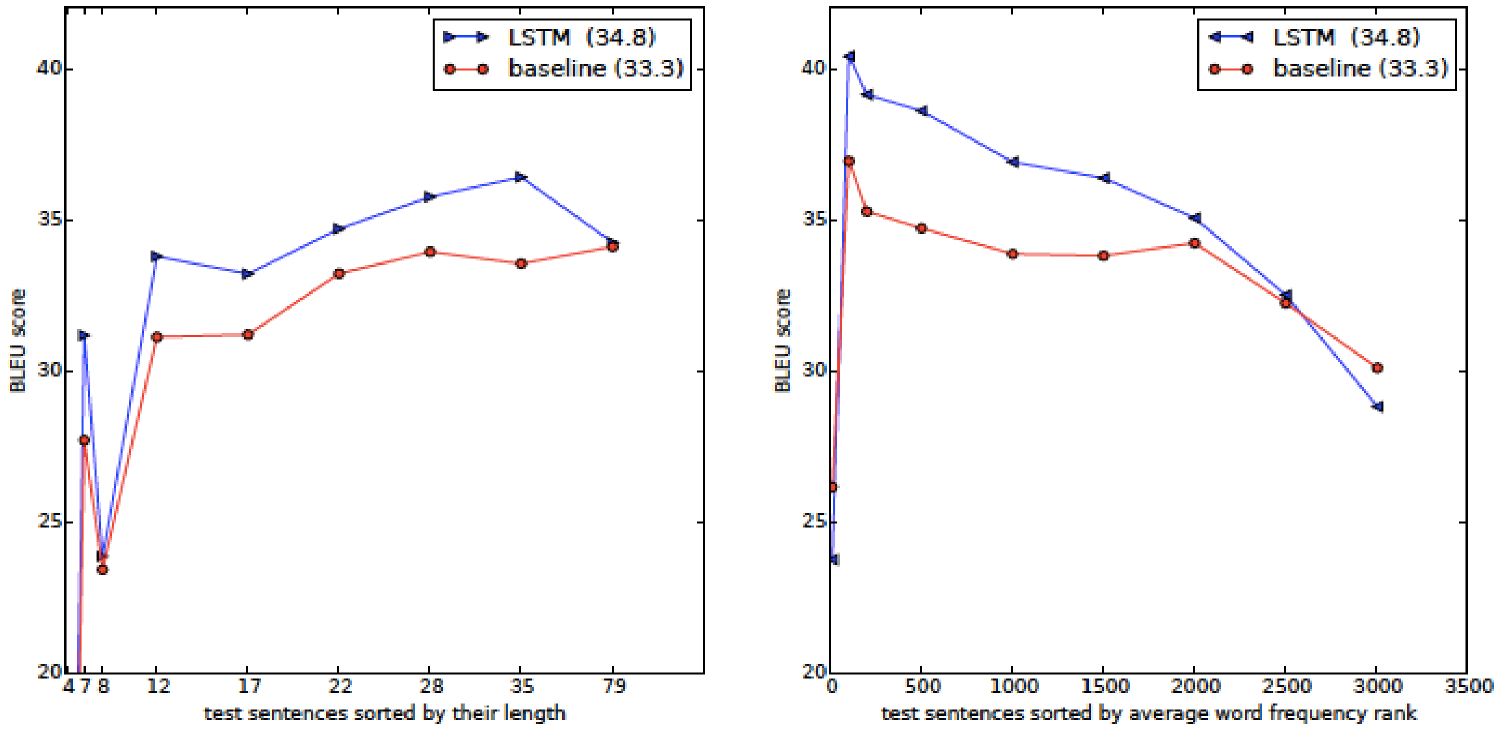
- Sentence representation by 2-dimensional PCA





3.3 Results

- Performance on long sentence



4. Conclustion

- 4.1 Conclusion
- 4.2 Related Work

4.1 Conclusion

- LSTM with a limited vocabulary can outperform a standard SMT-based system whose vocabulary is unlimited

- Improvement by reversing the words

- Good at long sentences

4.2 Attention Mechanism

- LSTM's long-term dependency

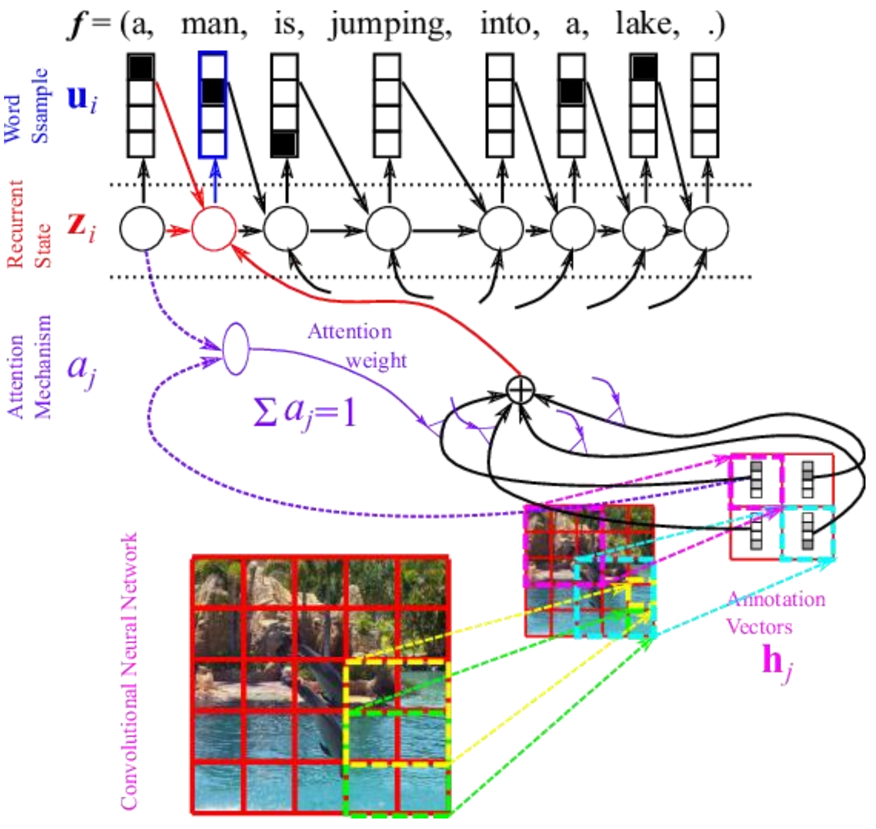
“BottleNeck” problem
:Need to know all the information in the input sentence to single vector (context vector)

But the information needed for each word will be different.

4.2 Attention Mechanism

"Neural Machine Translation by Jointly Learning to Align and Translate" present attention mechanism

[-Bahdanau](), [Cho]() (2015, ICLR)



- Details : To be continued…..

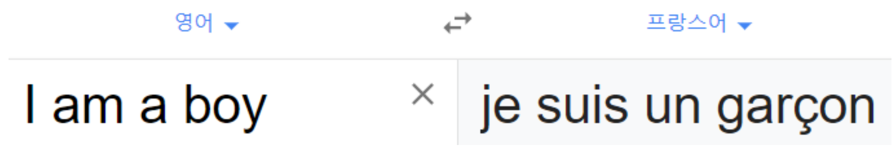
(Week5 Paper LOL)

5. Question

Question

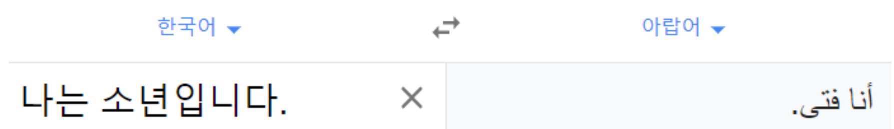
- Reversing Words in source sentence

- “French;English” :similar structure



- How about language pairs which grammatical structures are quite different.

(ie. Korean vs Arabic)



First Arabic Words : 소년

Second : 나는

Question

- Beam Search

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

- As the beam size increases, the change of BLEU score is not significant.

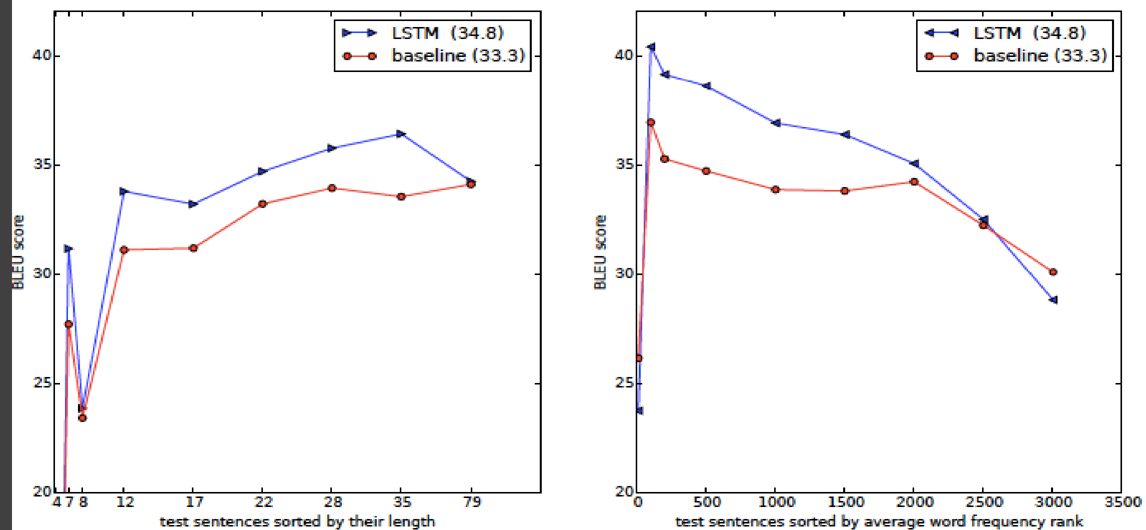
  Is there an economic benefit to maintaining the larger B while paying a larger cost?

- From another perspective,

  Can Beam Search evaluate appropriate alternate translations? (finding the right pair of candidates)

Question

- Performance Analysis



- Why Figure 3 left, the accuracy is reduced sharply around length 8 - 12
- Why the BLEU score increases to a certain level as sentence length increases.
- Reversal occurs between baseline model and about 2500 words in Figure 3 right