Sequence to Sequence Learning with Neural Networks

March 22nd, 2024

Yoojeong Lee

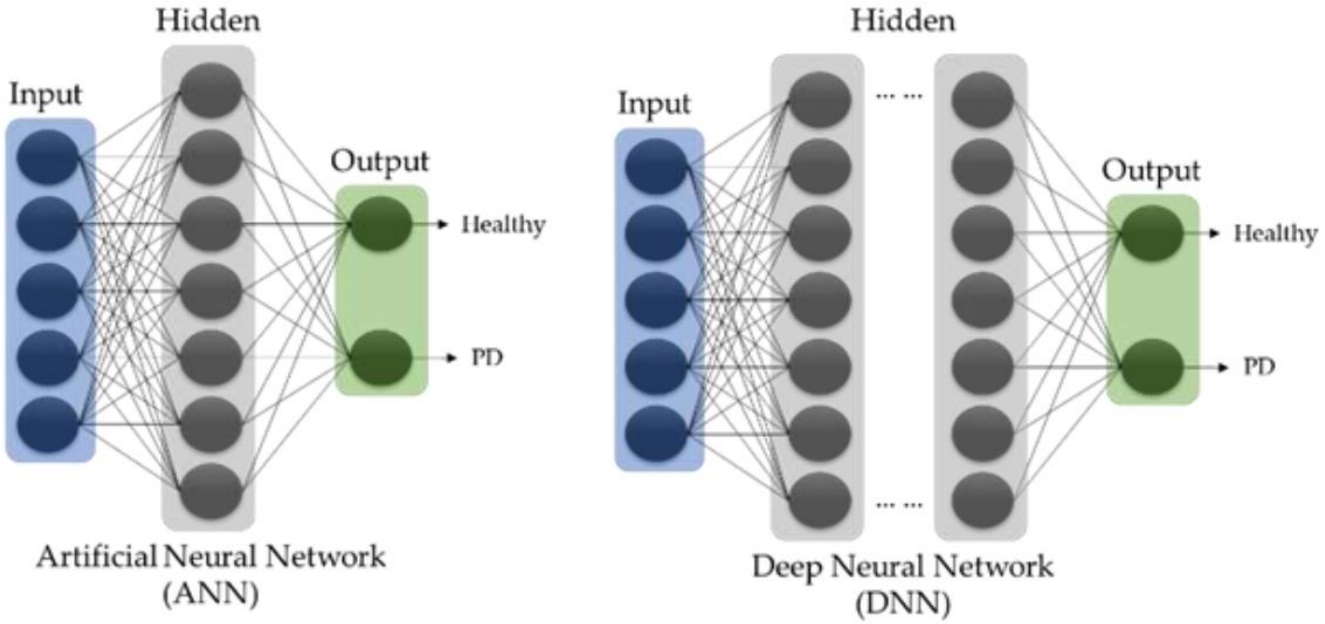
Sookmyung Woman's University

IT Engineering

Table of Contents

- Introduction
- The model
- Experiments

- What's the DNN, RNN, RSTM
  - DNN(Deep Neural Network)
  - Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences.

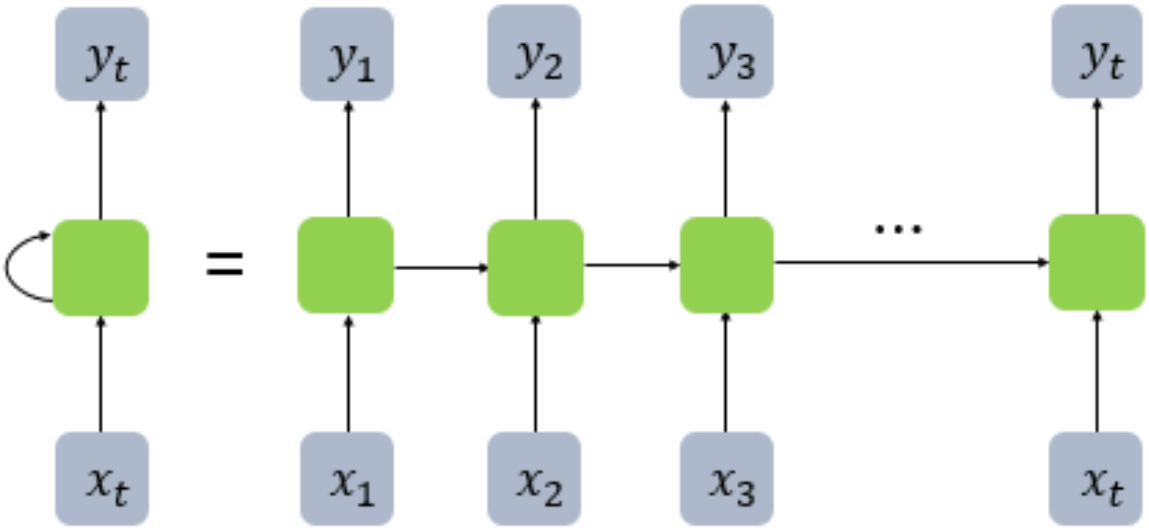


< Compare of ANN & DNN>

- What's the DNN, RNN, RSTM

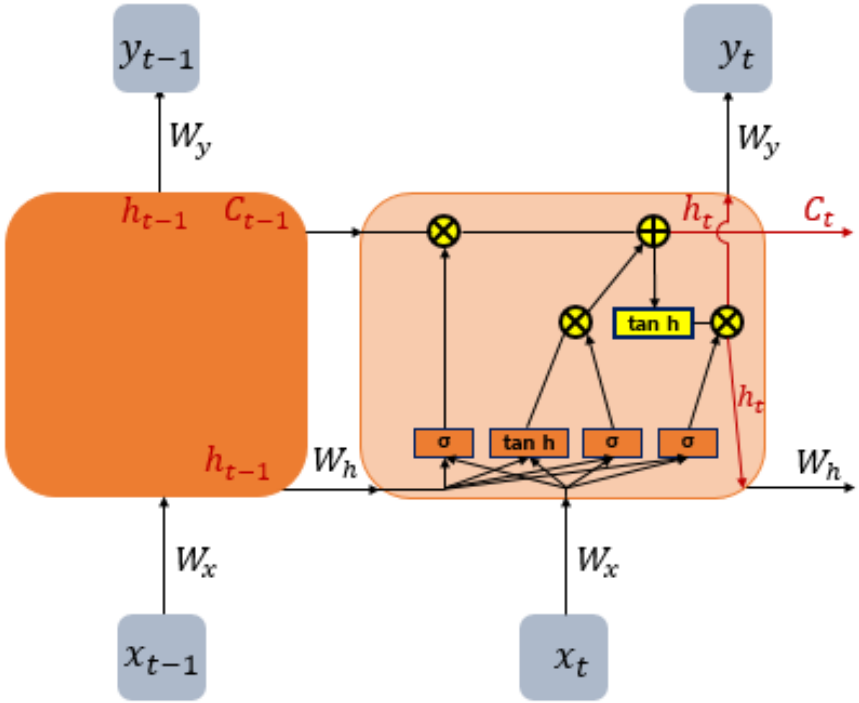
$$h_t = \text{sigm}\left(W^{\text{hx}}x_t + W^{\text{hh}}h_{t-1}\right)$$
$$y_t = W^{\text{yh}}h_t$$

  - RNN(Recurrent Neural Network)
  - Enable to model time-dependent and sequential data problems
  - Suffer from the matter of vanishing gradients.
  - When input and output sequences have different lengths, don't follow a simple pattern.



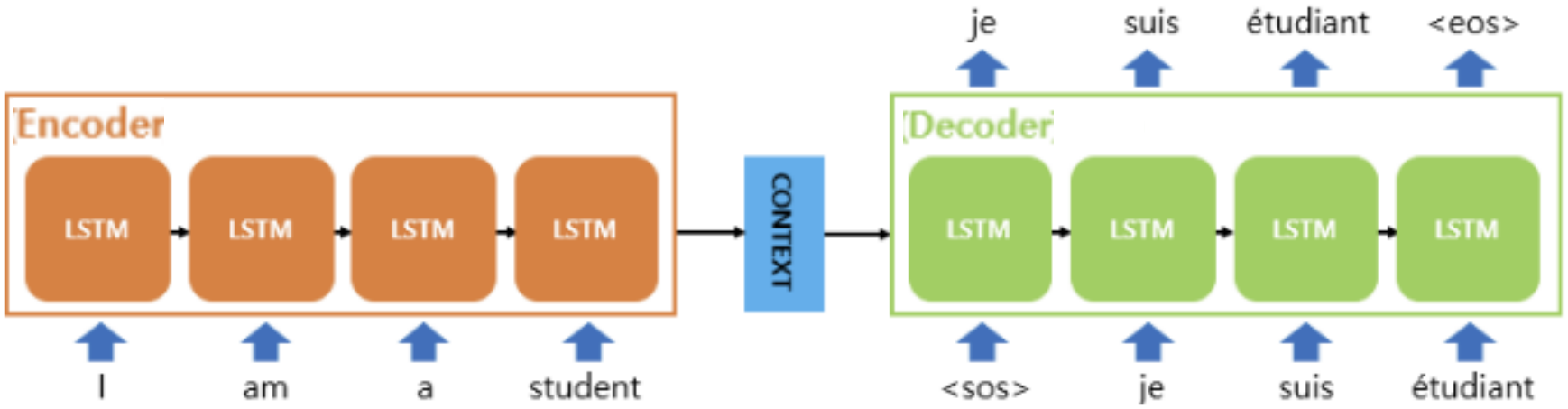
< RNN hidden layer >

- What's the DNN, RNN, LSTM
  - LSTM(Long Short-Term Memory)
  - Can capture long-term dependencies and handle sequential data well.
  - Computationally expensive and require a large amount of training data.



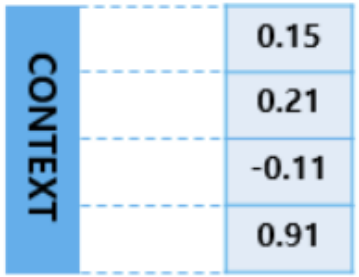
< RSTM structure>

Sequence to Sequence learning



- Context Vector

The encoder sequentially processes all words in the input sentence, compressing all this information into a single vector.



- Dataset

  - Source Language : English (Use the most frequent words 80000)
  - Target Language : French (Use the most frequent words 160000)

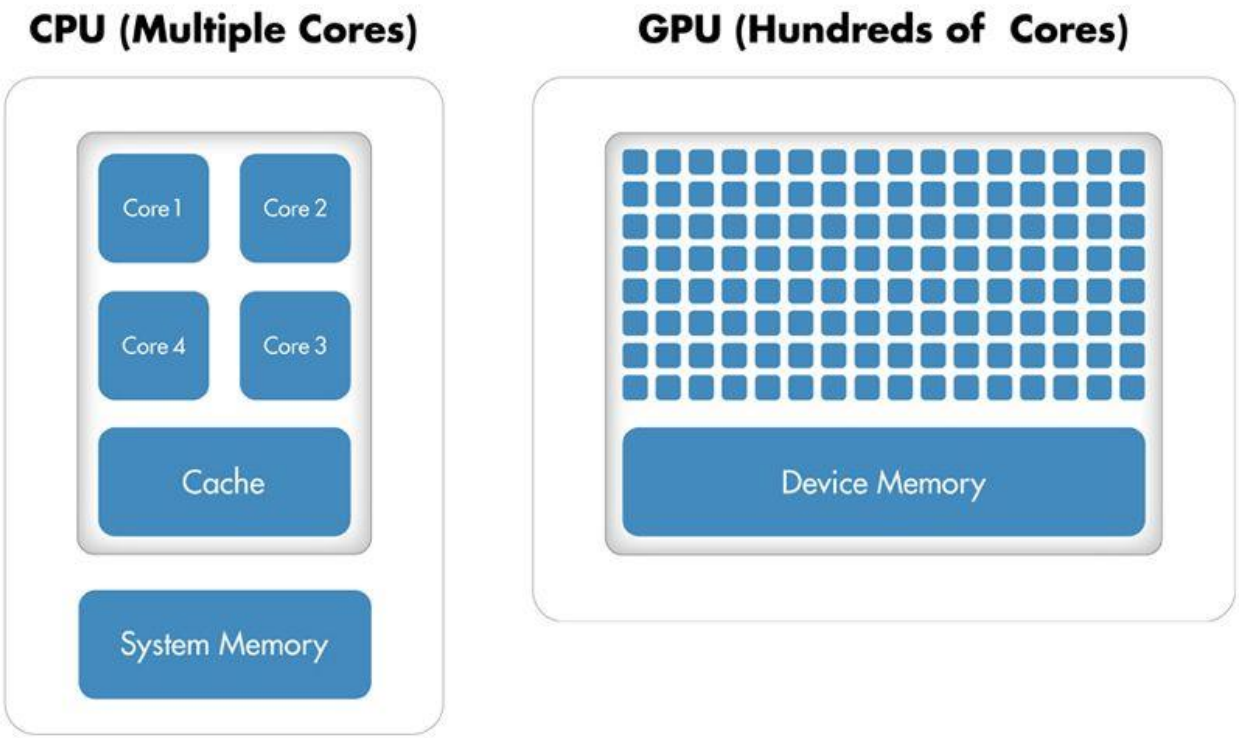
  - out-of-vocabulary word was replaced with a special “UNK” token.

- How they Used LSTM in Seq2Seq learning?
  - Used two different LSTMs.
  - Increase model parameter number
  - Train naturally multiple language pairs

  - LSTM with four layers
  - Deeper LSTM significantly outperformed

  - Reverse order of the words
  - C , B, A -> α, β, γ
  - Helps SGD better understand and grasp the relationship between the input and output of the model.

- Parallelization
  - CPUs typically handle tasks sequentially
  - GPUs have a large number of cores that allow them to process multiple tasks simultaneously



- Experimental Results - BLEU score : Table 1 & 2
  - BLEU (BiLingual Evaluation Understudy) is a metric for automatically evaluating machine-translated text.

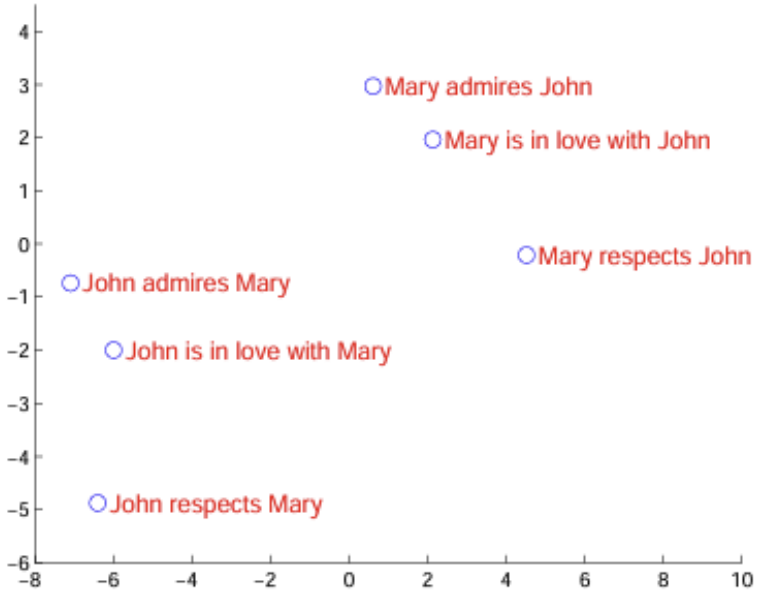
Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size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

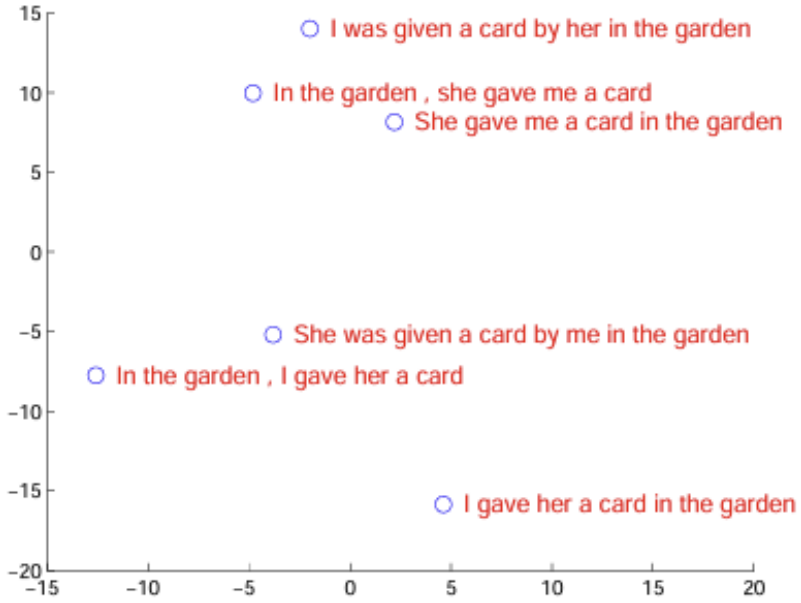
Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

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	~45

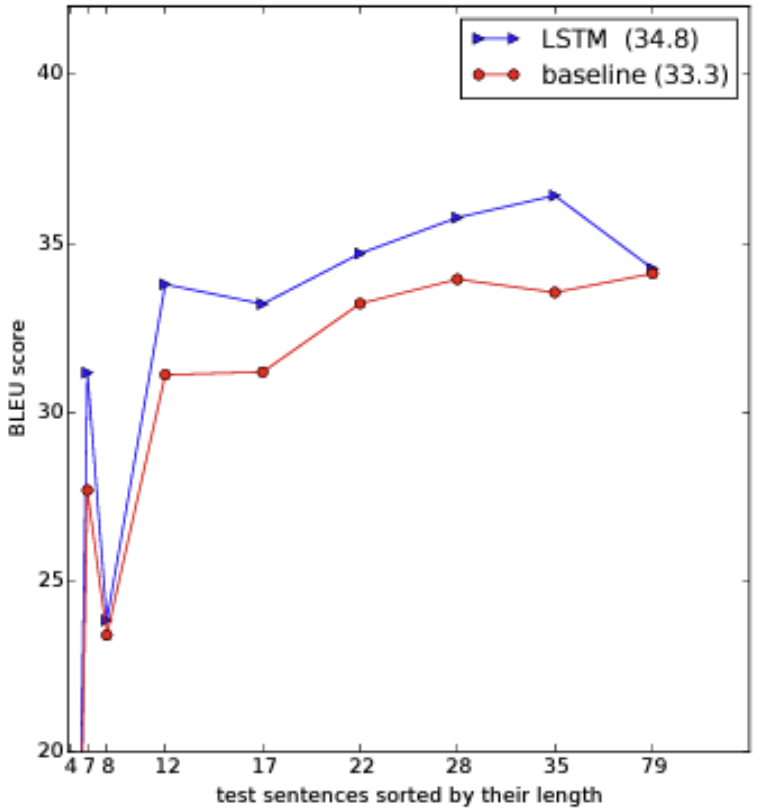
Table 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

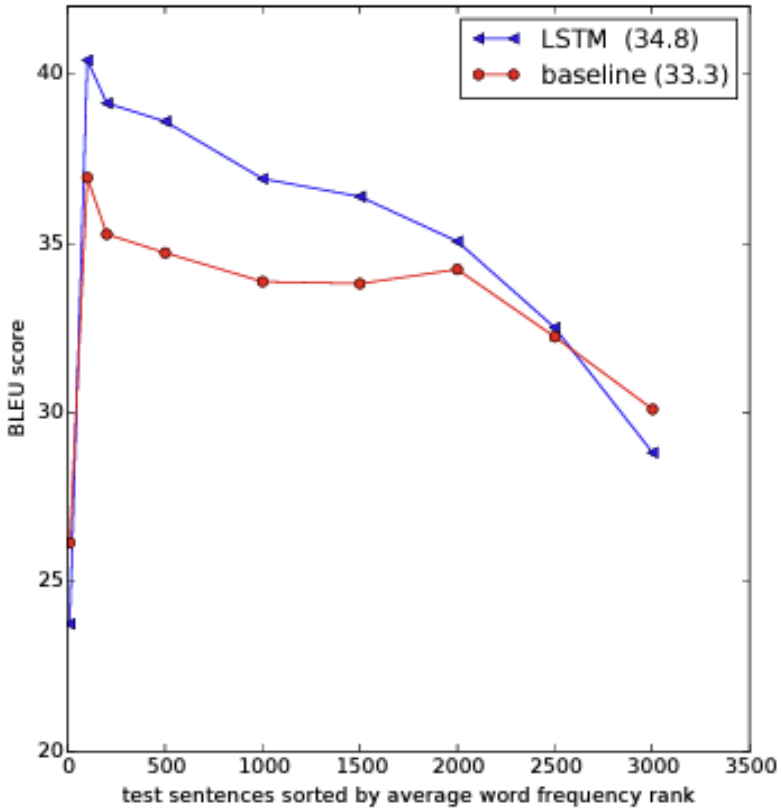
- Experimental Results - Model Analysis : Figure 2
  - Influenced by word order but are not significantly affected by changing from active to passive voice.
  - Turn a sequence of words into a vector of fixed dimensionality.



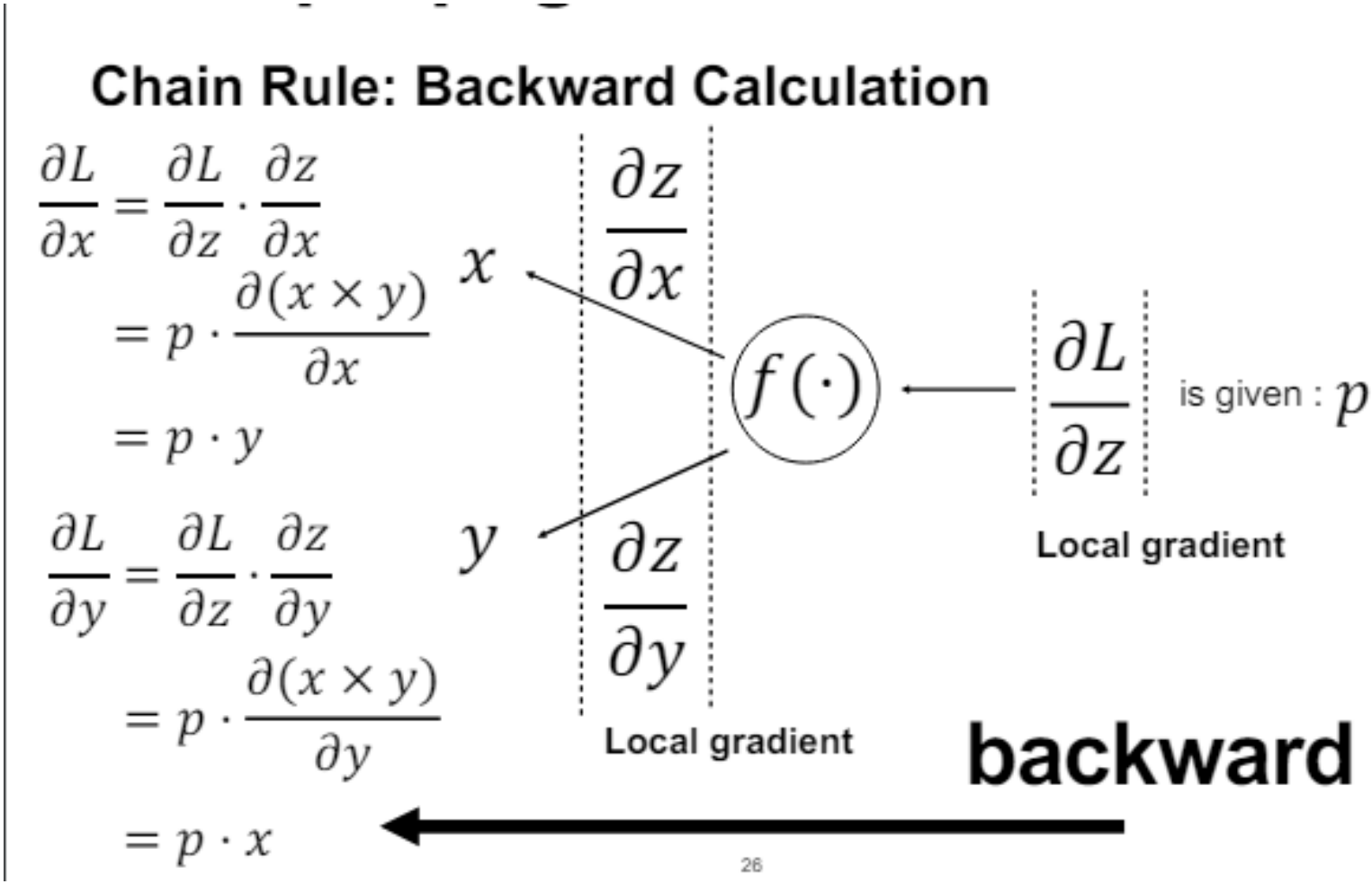


- Experimental Results – BLEU Score : Figure 3
  - Sentence legth
  - Word frequency





- Chain Rule


Chain Rule: Backward Calculation
$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial x}$
$= p \cdot \frac{\partial (x \times y)}{\partial x}$
$= p \cdot y$
$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial y}$
$= p \cdot \frac{\partial (x \times y)}{\partial y}$
$= p \cdot x$
x
y
$\frac{\partial z}{\partial x}$
$\frac{\partial z}{\partial y}$
Local gradient
$f(\cdot)$
$\frac{\partial L}{\partial z}$
is given : p
Local gradient
backward
26