MLDS HW2-2

TA
b04901070@ntu.du.tw

Outline

- Timeline
- Task Descriptions
- Requirements
- Q&A

Timeline

Timeline

- **3/26 or 3/30 :**
  - Release HW2-1
- **4/2 or 4/6 : Break**
- **4/9 or 4/13 :**
  - **Present HW2-1**
  - Release HW2-2
- **4/16 or 4/20 :** Break
- **4/23 or 4/29 :** Midterm Break
- **4/30 or 5/4 :**
  - **Present HW2-2**

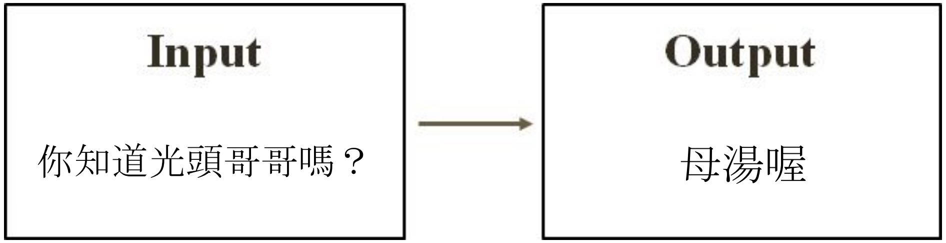
Timeline

- **3/26 or 3/30 :**
  - Release HW2-1
- **4/2 or 4/6 :** Break
- **4/9 or 4/13 :**
  - **Present HW2-1**
  - Release HW2-2
- **4/16 or 4/20 :** Break
- **4/23 or 4/29 :** Midterm Break
- **4/30 or 5/4 :**
  - **Present HW2-2**

Task Descriptions

HW2-2 Introduction

- Chinese Chatbot
  - Input : A sentence
  - Output : The corresponding reply



Input

你知道光頭哥哥嗎？

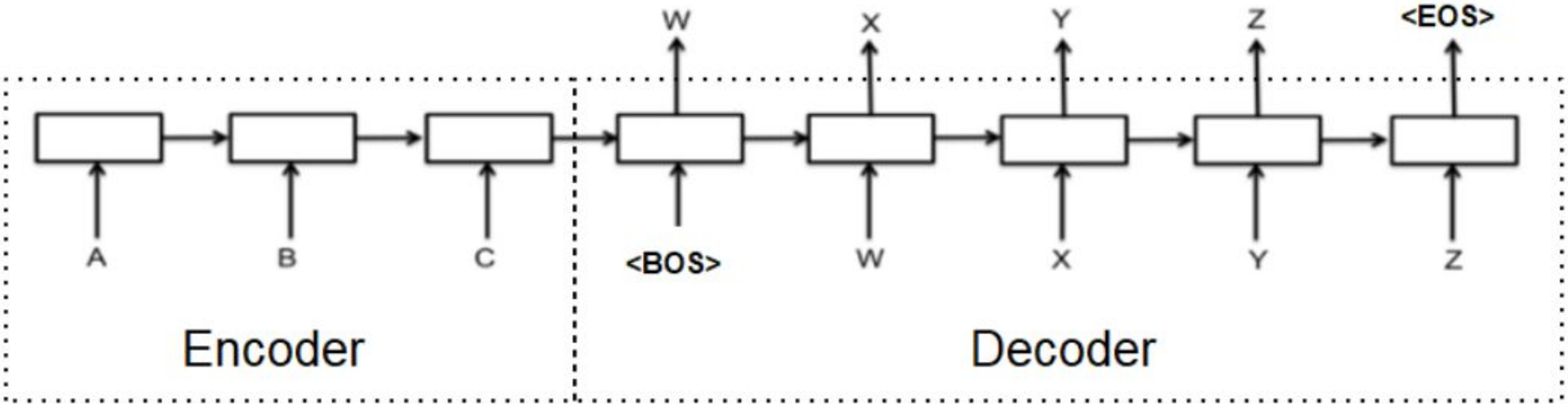
Output

母湯喔

- There are several difficulties including
  - Variable length of I/O
  - Out of Vocabulary

HW2-2 Sequence-to-sequence 1/3

- **Two recurrent neural networks (RNNs)**
  Encoder : processes the input
  Decoder : generates the output



HW2-2 Sequence-to-sequence 2/3

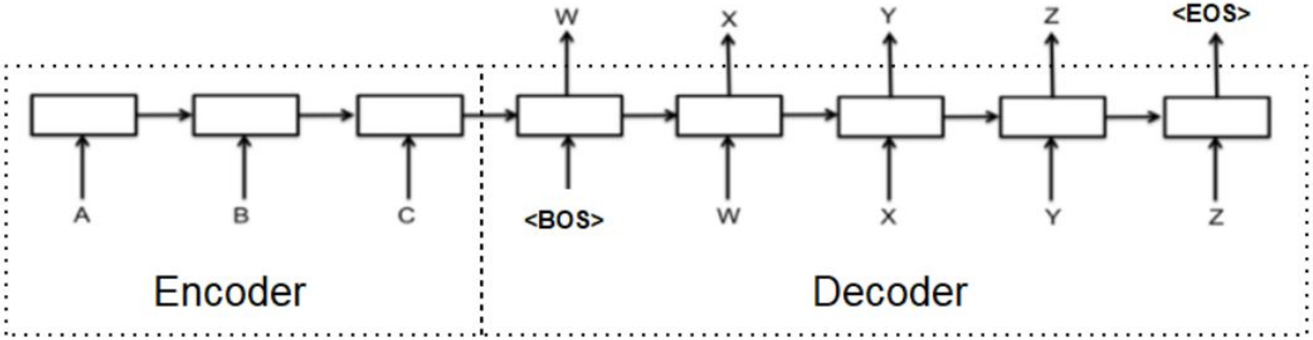
- **Data preprocess:**
  - Dictionary - most frequently word or min count
  - other tokens: <PAD>, <BOS>, <EOS>, <UNK>

    - <PAD> : Pad the sentence to the same length

    - <BOS> : Begin of sentence, a sign to generate the output sentence.

    - <EOS> : End of sentence, a sign of the end of the output sentence.

    - <UNK> : Use this token when the word is not in the dictionary



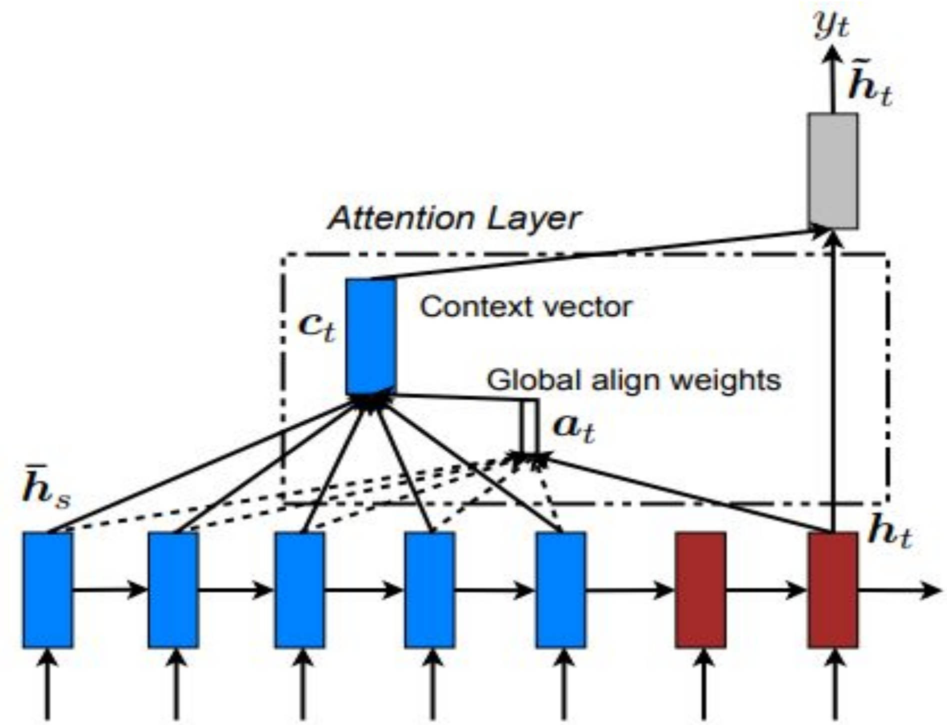
HW2-2 Sequence-to-sequence 3/3

- **Text Input :** Reference Tutorial
  - One-hot Vector Encoding
    - 1-to-N coding, N is the size of the vocabulary in dictionary
    - Usually passing a linear embedding layer
  - e.g.
    - <BOS> = [0, 1, 0, ..., 0, 0, 0, ..., 0, 0, 0]
    - <EOS> = [0, 0, 1, ..., 0, 0, 0, ..., 0, 0, 0]
  - Word to Vector
    - Gensim (Chinese word2vec package)
    - Facebook pre-trained word vectors
- **Text Output :**
  - One-hot Vector Encoding

Training Tips - Attention 1/3

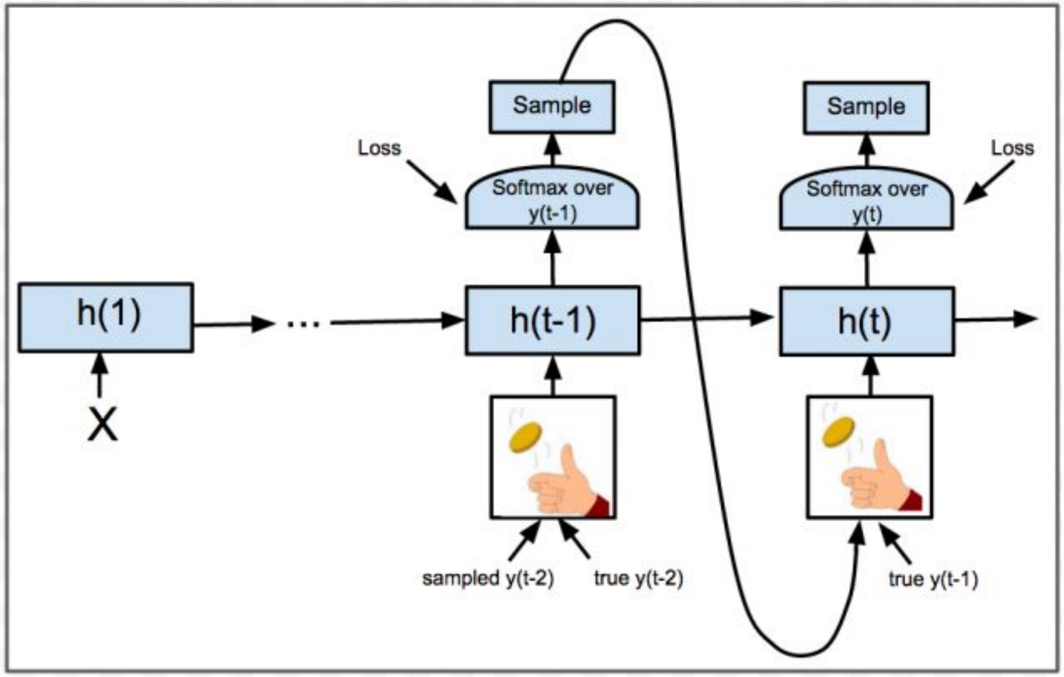
Reference

- Attention on encoder hidden states :
  - Allow model to peek at different sections of inputs at each decoding time step



Training Tips - Schedule Sampling 2/3

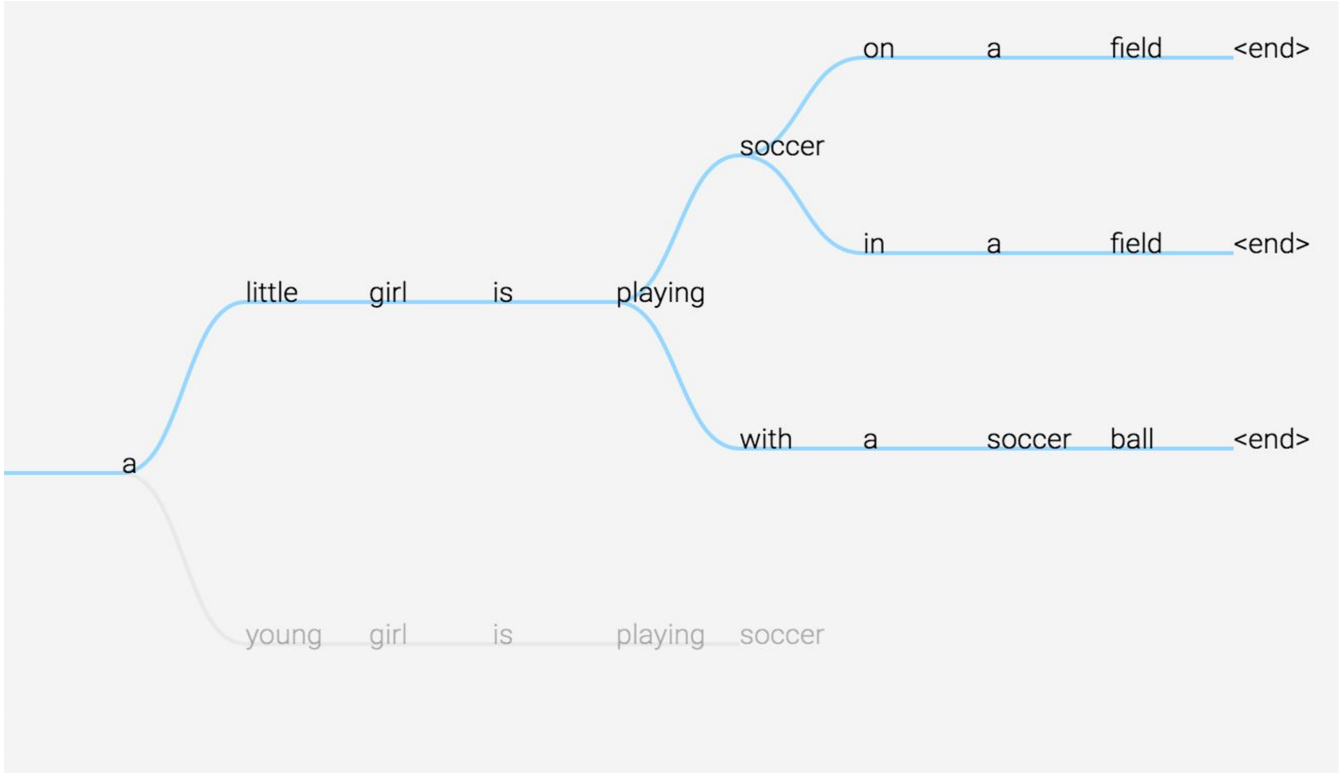
- Schedule Sampling :
  - To solve “exposure bias” problem,
    When training, we feed (groundtruth) or (last time step’s output) as input at odds



https://arxiv.org/abs/1506.03099

Training Tips - Beam search 3/3

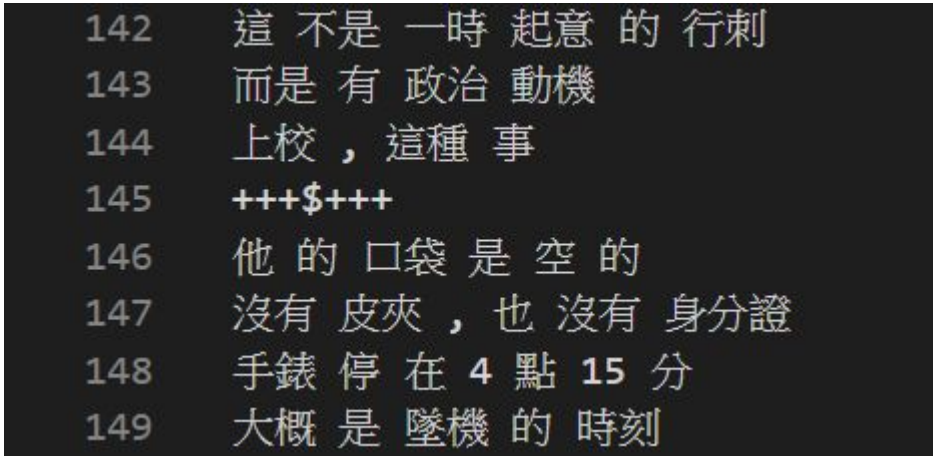
- Beam search :
  - keep a fixed number of paths



HW2-2 Data & Format

- **Dataset :**
    - 語音實驗室的電影字幕 - 280 萬句對話
- **Format :**
    - 一行一句話
    - 對話間用+++$+++分隔
    - Download
- **Extra Data** (未整理, 不符合上述格式):
    - 電影data(556M)

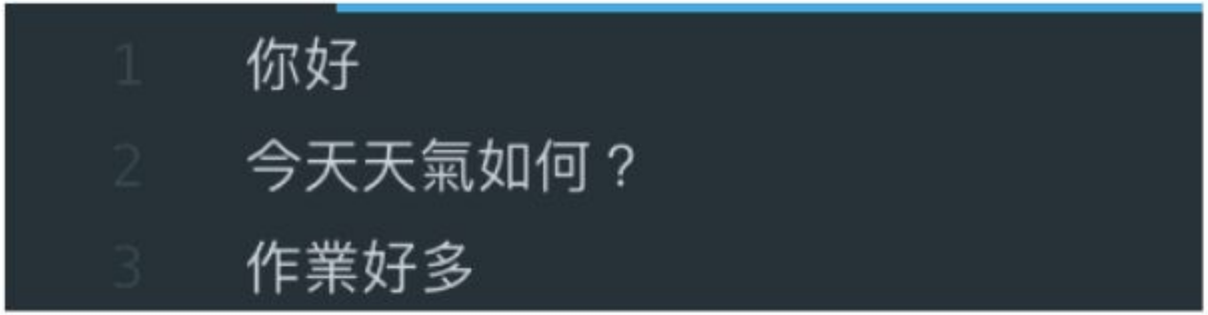
clr_conversation.txt



```
這 不是 一時 起意 的 行刺
而是 有 政治 動機
上校 ， 這種 事
+++$+++
他 的 口袋 是 空 的
沒有 皮夾 ， 也 沒有 身分證
手錶 停 在 4 點 15 分
大概 是 墜機 的 時刻
```

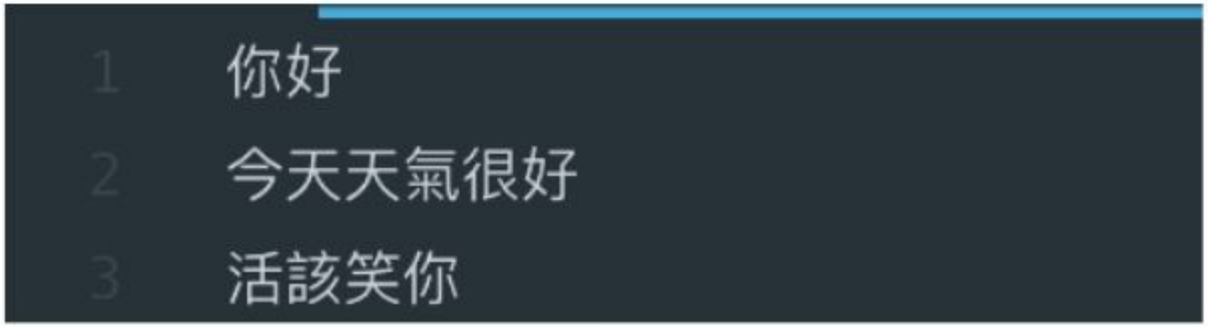
HW2-2 I/O Format

- 一行一句話
  - Input.txt



```
你好
今天天氣如何？
作業好多
```

  - Output.txt



```
你好
今天天氣很好
活該笑你
```

Requirement

HW2-2 Baseline 1/3

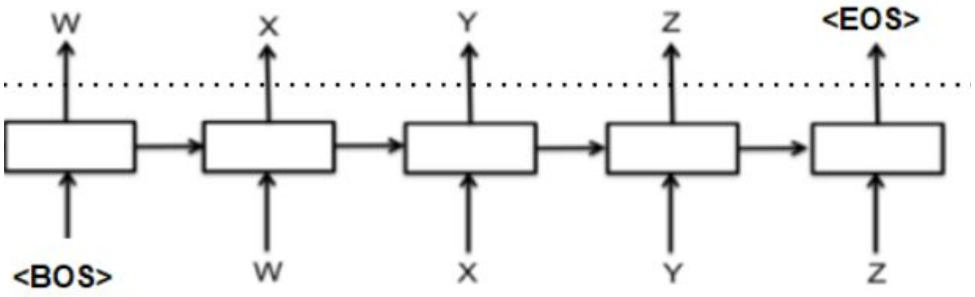
- Evaluation : Perplexity

I love NLP.

$$\prod_i p(w_i) = p(NLP|I\ love) * p(love|I) * p(I)$$

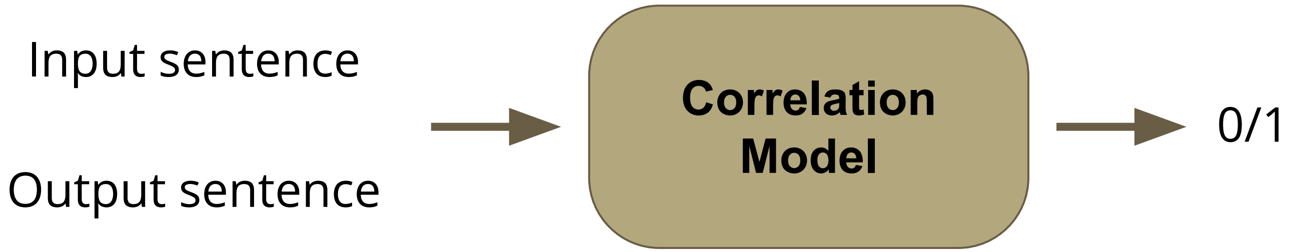
$$\log_2 \prod_i p(w_i) = \sum_i \log_2 p(w_i)$$

$$PP = 2^{-\frac{1}{N}\sum_i \log_2 p(w_i)}$$



HW2-2 Baseline 2/3

- Evaluation : Correlation Score
  - Decided by model.
  - The model is training by given dataset.
  - A kind of discriminator.
- Model detail :
  - Correct I/O pairs scored 1, incorrect pairs scored 0.
  - Activation function sigmoid.



HW2-2 Baseline 3/3

- **Perplexity < 100**
- **Correlation Score > 0.45**
- **Evaluation code**
    - Dependencies:
        - tensorflow 1.6
        - pytorch 0.3.1

Submission & Rules

- Please implement one sequence-to-sequence model (or it’s variant).
- Extra dataset is encouraged to use.
- Recommend toolkit package :
  - Python 3.6
  - **TensorFlow**
  - **Pytorch**
  - Keras
  - matplotlib
  - Gensim

Presentation Requirements

- Model description
  - Describe the seq2seq model of your choice

- How to improve your performance
  (e.g. Attention, Schedule Sampling, Beamsearch...)
  - Describe the method that makes you outstanding
  - Why do you use it?
  - Analysis and compare your model without the method

- Settings, Results, and other Experiments
  - **Demo a few (input, output) pairs from the testing set to show your models performace.**
  - parameter tuning, other improve methods... etc

Advanced Tasks

- Model architecture experiments
  - Compare with different RNN models
  - eg. LSTM, GRU, bidirectonal
  - eg. different ways to calculate scores (attention mechanism)

- Better dialogue generation
  - Do some clean-up on the corpus
  - Reinforcement Learning

- BLEU score on chatbot model
  - Calculate the BLEU score on this task (nltk package)
  - Analyze about the results

Submission

- Deadline: **2019/04/30 or 2019/05/04 23:59 (GMT+8)**
- Submit your presentation files to: google drive
- Your github must have several files under the directory: **hw2/hw2_2/**
  - Readme.*
  - other implementation code (Files for training is required)
  - trained model (exempt if model is uploaded to another cloud space)
- In your Readme:
  - Specify the toolkits/libraries and their corresponding version you used.
  - Describe how to download the trained model.
  - State clearly how to run your program to generate the results in your report.

Q&A

b04901070@ntu.edu.tw

How to reach the baseline ?

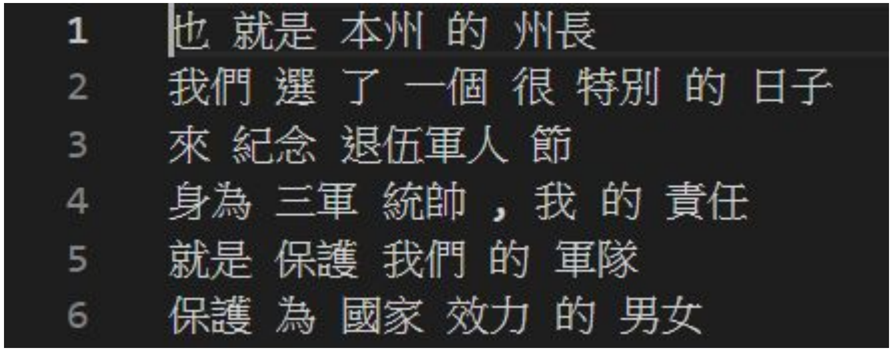
- Gensim word to Vector: Vocabulary size is 50,000, vector dimension is 250
- Select sentences with 2~15 words and at most 2 <UNK> tokens.
- Dataset size is around 800,000.
- Model description
  - Batch size = 50
  - Hidden units = 1024 (two layers LSTM cell)
  - Adam optimizer with initial lr=0.001
  - Gradient clipping=5
  - Schedule Sampling and Attention

Data preprocessing

- Third party library
- Use ASCII to remove English and punctuation mark
- Custom-made rules

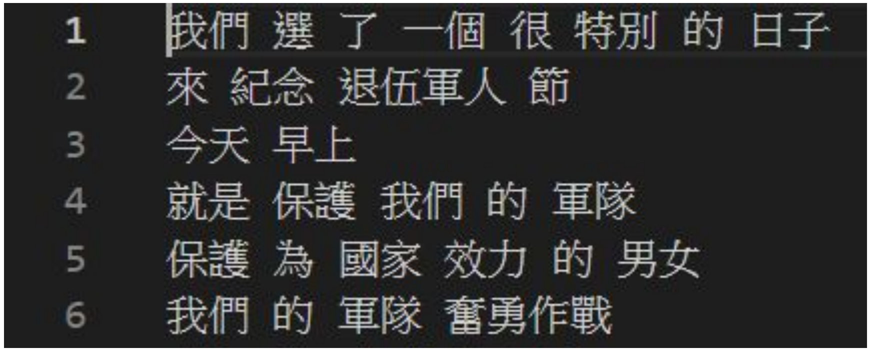
- Dataset(after certain clean-up) is around 900,000 pairs (link)

question.txt



```
也 就是 本州 的 州長
我們 選 了 一個 很 特別 的 日子
來 紀念 退伍軍人 節
身為 三軍 統帥 ， 我 的 責任
就是 保護 我們 的 軍隊
保護 為 國家 效力 的 男女
```

answer.txt



```
我們 選 了 一個 很 特別 的 日子
來 紀念 退伍軍人 節
今天 早上
就是 保護 我們 的 軍隊
保護 為 國家 效力 的 男女
我們 的 軍隊 奮勇作戰
```