Machine Learning And Having it Deep and Structured

Homework 3 - Language Models

劉元銘 r02942070@ntu.edu.tw
呂相弘 r03942039@ntu.edu.tw

Outline

- Sentence Completion Challenge
- Language Model
- Other Related Methods
- Data Set
- Homework Requirements
- Additional Rules (Weekly Bonus)
- Grading
- Recommendations

Sentence Completion Challenge

Sentence Completion Challenge

- The task is to complete the sentence with multiple choices given the contextual information.
- Each sentence contains a underline indicating the missing word in the real-world literature.
  - Source: five Sherlock Holmes novels by Sir Arthur Conan Doyle
- Accuracy as evaluation metrics.
- Kaggle: https://inclass.kaggle.com/c/mlsd-hw3

Sentence Completion Challenge

My morning's work has not been ______ , since it has proved that he has the very strongest motives for standing in the way of anything of the sort.

a) invisible

b) neglected

c) overlooked

d) wasted

e) deliberate

Sentence Completion Challenge

My morning's work has not been ______ , since it has proved that he has the very strongest motives for standing in the way of anything of the sort.

a) invisible

b) neglected

c) overlooked

d) wasted

e) deliberate

Language Model

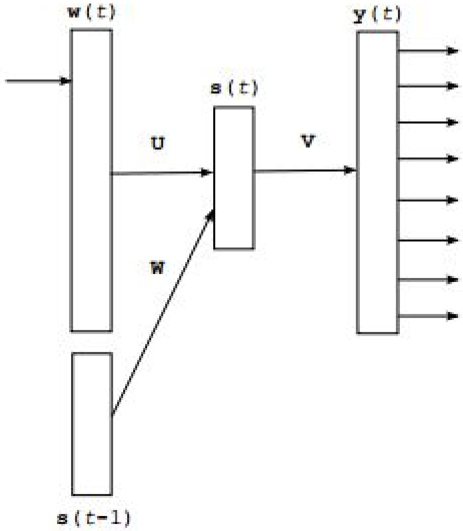
Language Models

- RNNLM (recommended)
- LSTM
- ...

RNNLM

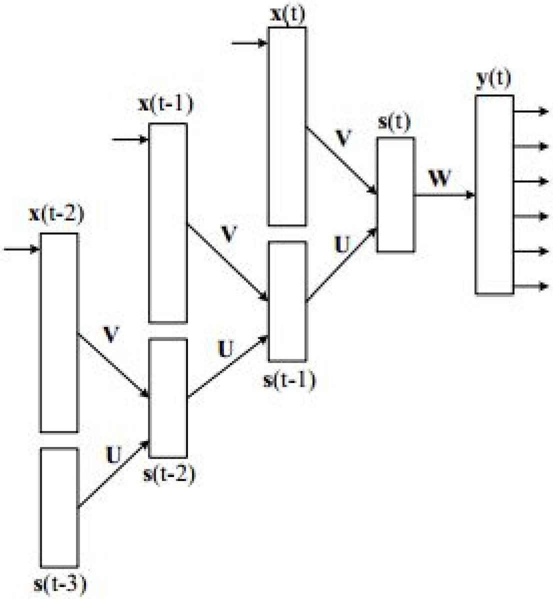
- Recurrent Neural Network Language Model
- Recurrent part
- Store last frame hidden output
- Predict current output

  based on memory
- 133 pages Reference



Training

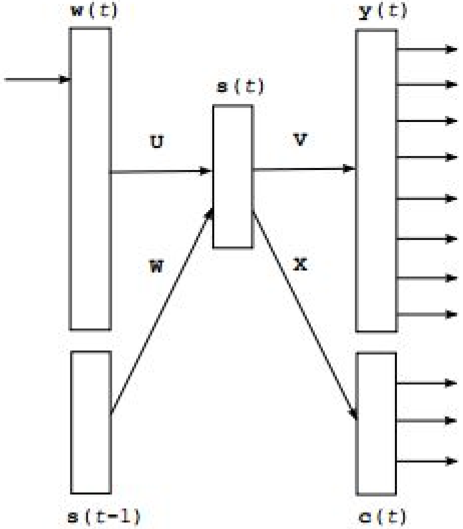
- BackPropagation Through Time(BPTT)
  - Basic training method
  - Buffer history neuron activations
  - Training RNN by **unfolding**
- Noise Contrastive Estimation
  - Advanced training method
  - Acceleration
  - Reference



Output Factorization (OF)

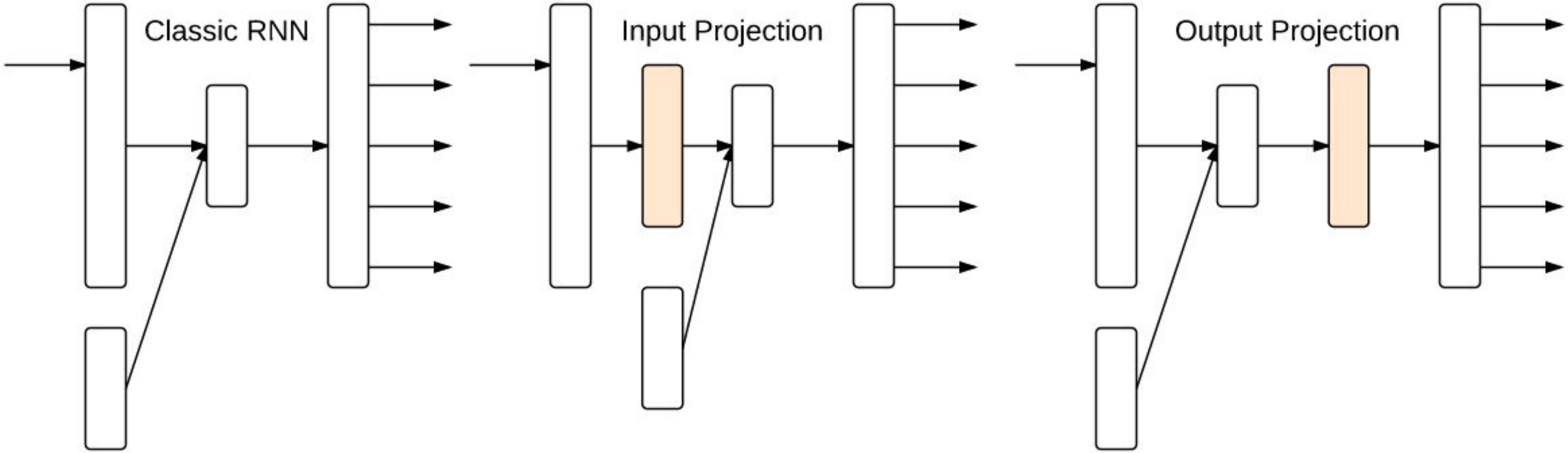
- Predict class first

$$P(w_{t+1}|\mathrm{s}(t)) = P(c_i|\mathrm{s}(t))P(w_i|c_i, \mathrm{s}(t)),$$

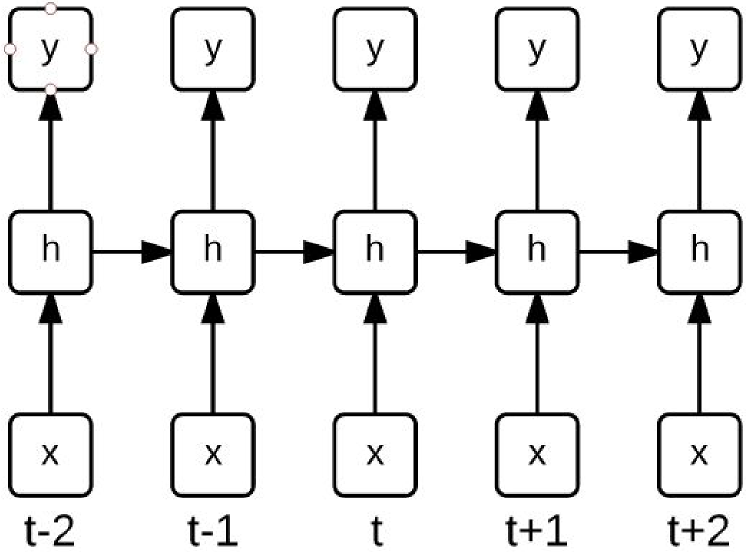


Projection

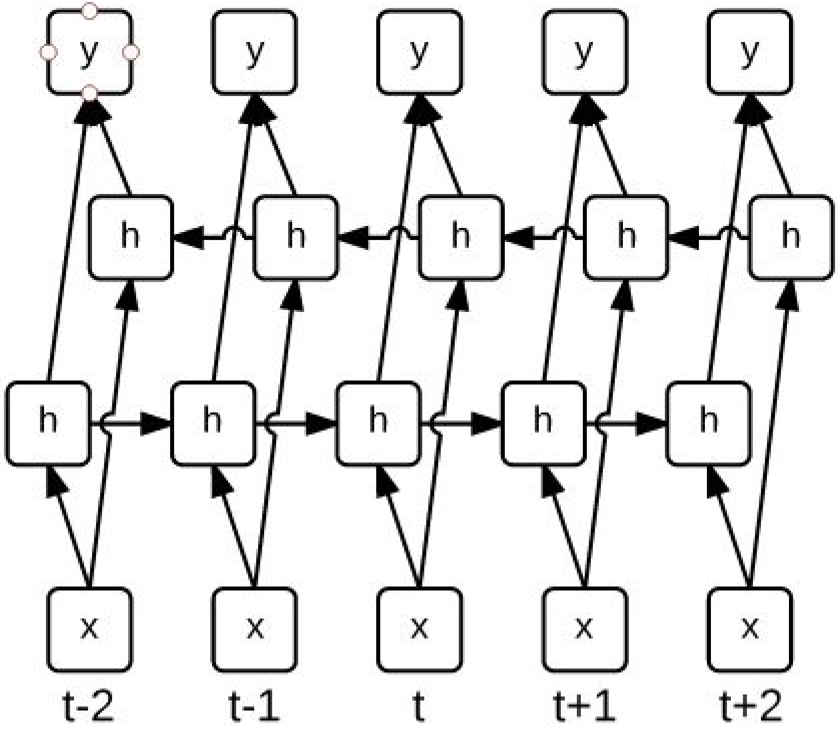
- Output Projection(OP)/Input Projection(IP)



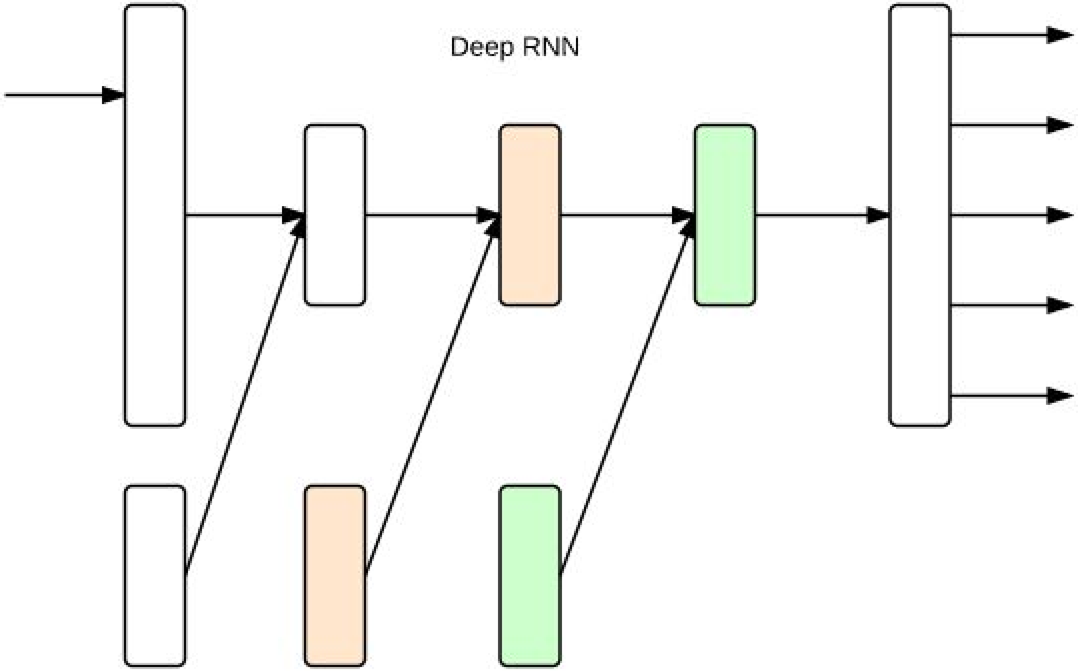
Unidirectional RNN



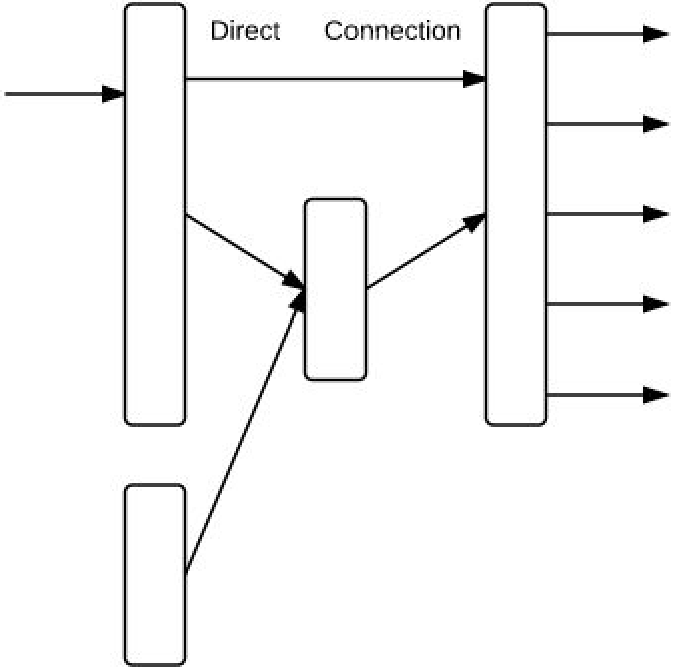
Bidirectional RNN



Deep RNN

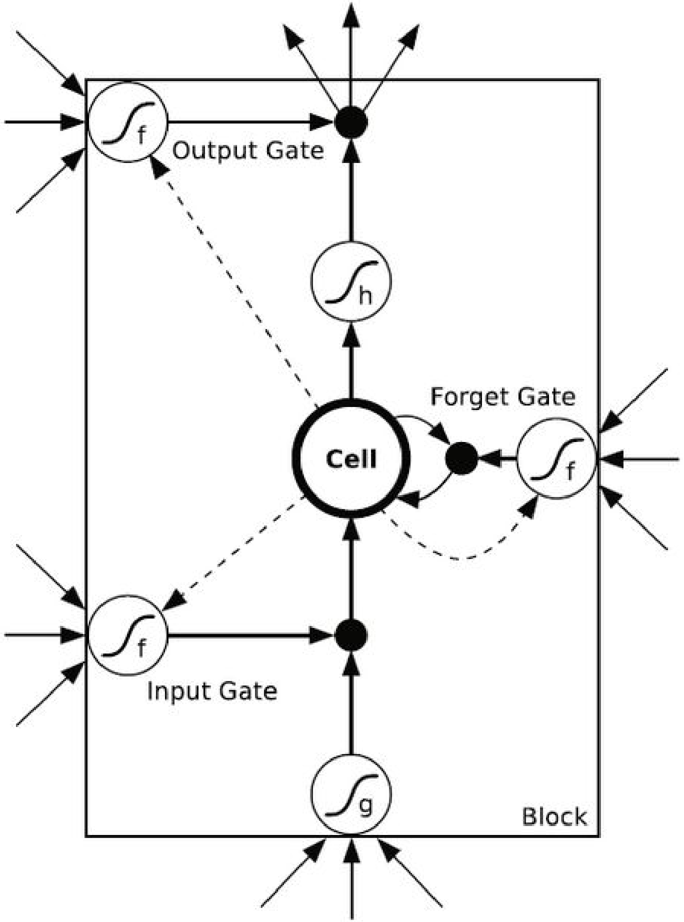


Direct Connection



LSTM

- Long short-term memory
- Replace each neuron in RNN with expanded memory cell
- Very complex
- Similar idea



Other Related Methods

Other Related methods

- You can use toolkit, but only for model combination with language model above
- N-gram language modeling
  - toolkit allowed for this part: SRILM
- Topic models
  - Latent Semantic Analysis(LSA)
  - Non-negative Matrix Factorization(NMF)
- Skip-gram / COBW

N-gram

- For example, n = 3 (trigram)

$$P(W=w_1w_2...w_n)=P(w_1)\,P(w_2|w_1)\,P(w_3|w_1,w_2)\,P(w_4|w_2,w_3)\,P(w_5|w_3,w_4)$$

- Select the sentence of 5 options with the highest probability.

$$P(w^i)=\frac{N(w^i)}{\sum_{j=1}^{V} N(w^j)} \qquad P(w^j|w^k)=\frac{N(<w^k,w^j>)}{N(w^k)} \qquad P(w^j|w^k,w^m)=\frac{N(<w^k,w^m,w^j>)}{N(<w^k,w^m>)}$$

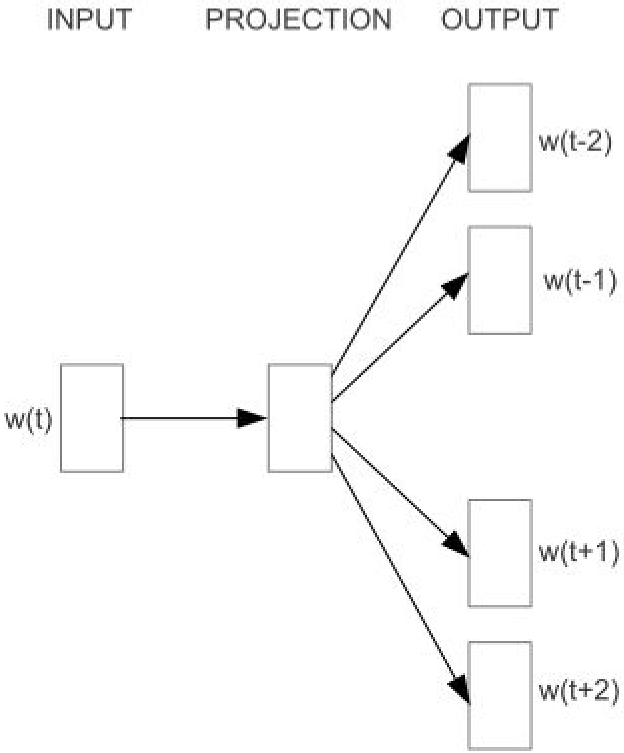
w^i : a word in the vocabulary

V : total number of different words in the vocabulary

$N(\cdot)$ number of counts in the training text database

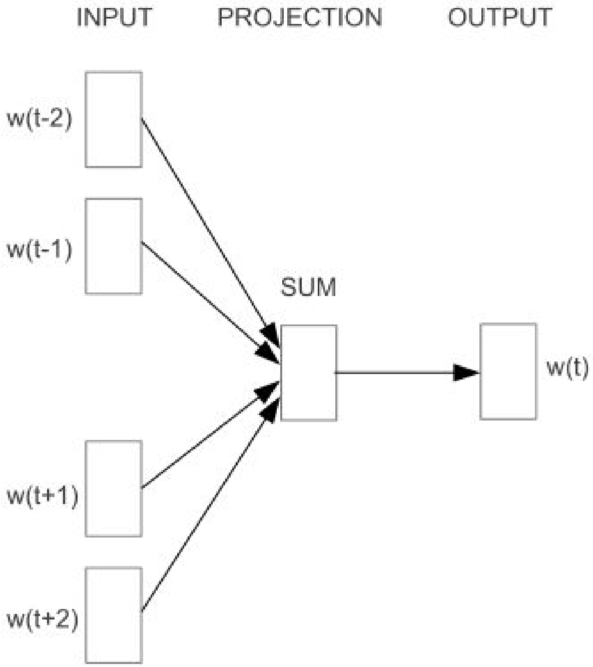
Skip-Gram

- No activation function
- Taught in class
- Reference



CBOW

- No activation function
- Taught in class
- Reference



CBOW

Data Set

Data Set

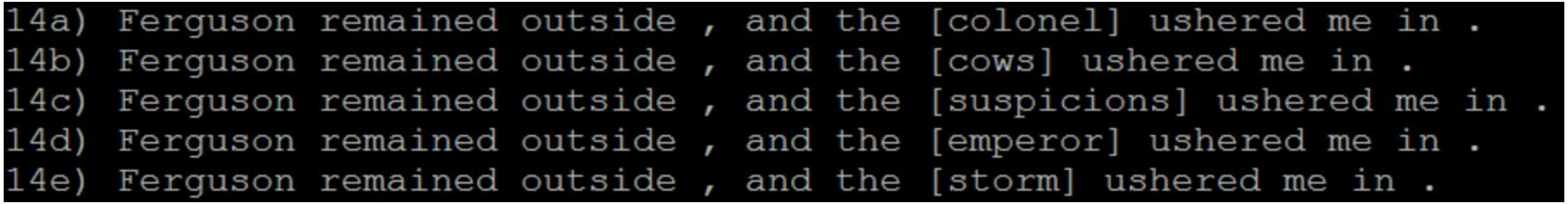
- Training data
- Testing data
- Please download from Kaggle

Training Data

- 19th century novels
- Extremely noisy with punctuations, headers and other annotated notes that may or may not convey language information
- **Data preprocessing is crucial in this task**
- What should be removed and what shouldn’t?

Testing Data

- Five of Conan Doyle’s Sherlock Holmes novels
  - *The Sign of the Four (1890)*, *The Hound of the Baskervilles (1892)*, *The Adventures of Sherlock Holmes (1892)*, *The Memoirs of Sherlock Holmes (1894)*, and *The Valley of Fear (1915)*
- 1040 sentences, Each with five options(a)(b)(c)(d)(e)



```
14a) Ferguson remained outside , and the [colonel] ushered me in .
14b) Ferguson remained outside , and the [cows] ushered me in .
14c) Ferguson remained outside , and the [suspicions] ushered me in .
14d) Ferguson remained outside , and the [emperor] ushered me in .
14e) Ferguson remained outside , and the [storm] ushered me in .
```

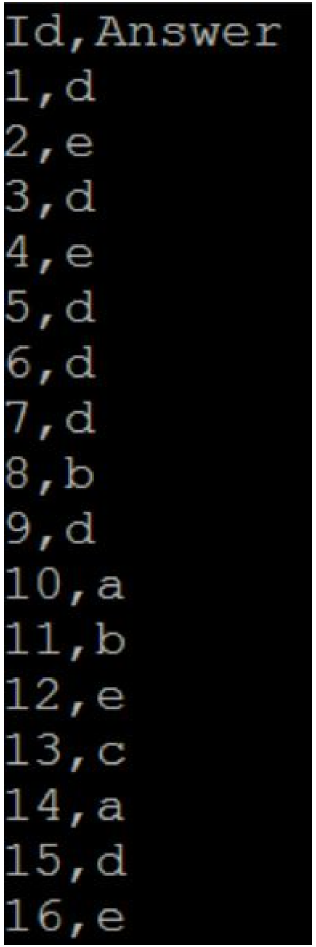
Homework Requirement

Homework Requirements

- You have to at least implement one of:
  - RNNLM/LSTM
  - Basically no toolkit allowed, but you can make a request if some toolkit (library) can be used
- Can use toolkits of other methods for model combination
- Language: C++/C, Python, Matlab...
- Kaggle Submission & Ceiba Submission

Kaggle Submission

- .csv file
- question id
- comma
- choice
- Kaggle



```
Id,Answer
1,d
2,e
3,d
4,e
5,d
6,d
7,d
8,b
9,d
10,a
11,b
12,e
13,c
14,a
15,d
16,e
```

Ceiba Submission

- Code
  - Detail environment setting
  - Code documentation
- Report
  - Basic information
  - Data structures and algorithms
  - Experiments settings and results
  - Division of teamwork

Additional Rules (Weekly Bonus)

Additional Rules (Weekly Bonus)

- Every Friday before 23:59, every group can explain the detail of your method and setting for your best uploaded score
  - Which model?What features & training parameters?
  - Bonus 2 points would be granted if properly described (Random guess gain 1 point @first week, but gain 0 point @second week)
  - First week: https://goo.gl/IMsdzP
  - Second week: https://goo.gl/1G73A5

Grading

Grading

- Kaggle Accuracy 60%
- Report 40%
- Implementation 20%
- Bonus
  - First Place 15%
  - First Runner-up 10%
  - Second Runner-up 5%

Grading - Kaggle Accuracy

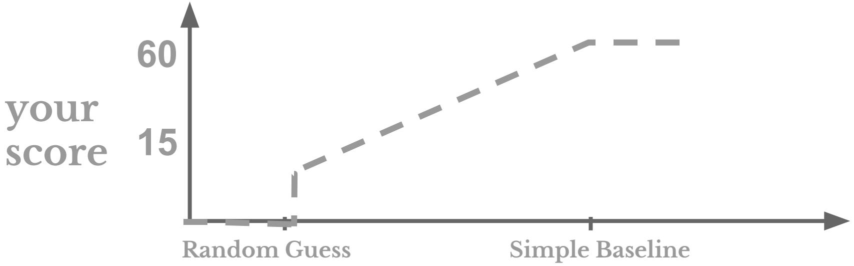
You have to use RNN/LSTM alone to achieve below performance:

Baseline 1 - Random Guess (15%) :

- Better than random guess: accuracy 20%.
- You must achieve the baseline 1 or you will receive 0.

Baseline 2 - Simple Baseline (45%) :

- Simple Baseline in Kaggle (released day 7)
- Once achieve the baseline, you can get the full credit in this part
- If you didn't make it ^_^:



Please add [pure RNN] or [pure LSTM] on the Kaggle submission description and note on the report for this part of grading!

Grading - Report

- Report (40%)
  - Group Information
  - Preprocessing/Data structure/Algorithm
  - Division of teamwork
  - What have you done? (including other methods)
  - Experiments and **Results**
  - No more that 4 A4 pages with font size 12

Grading - Implementation

- Implementation(20%)
  - Upload your code
  - Environment setting
  - Compilation instructions
  - Package dependencies
  - Code documentation on how to reproduce your work and change the parameters.

Recommendation

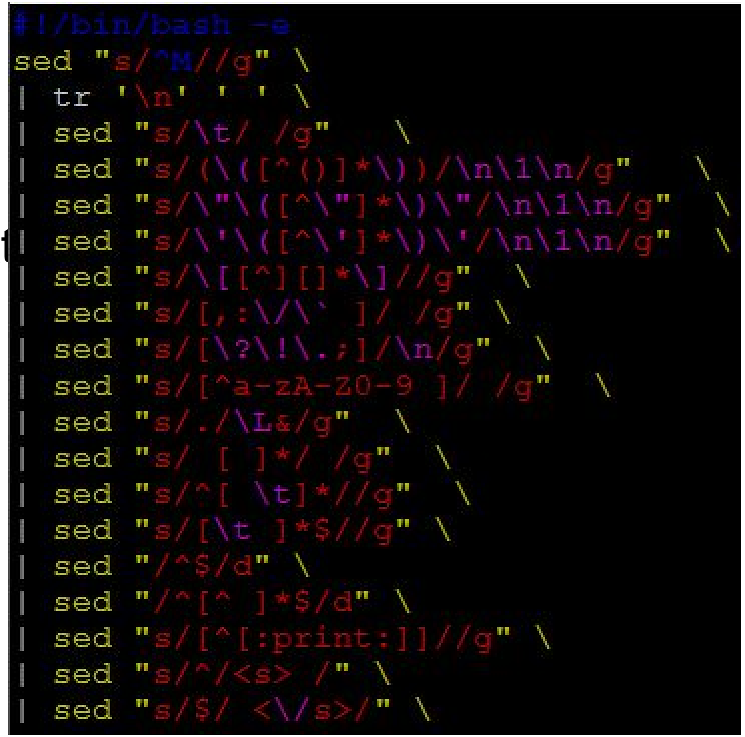
Recommendation

- RNN usually not fits in GPU acceleration
  - However, with tricks it can apply
- Start earlier!
- Model combination
- Prepare detailed manual for your code
  - Environment setting and compilation
  - Application Programming Interface

Preprocessing sample codes

- bash script
- usage

cat training/*.TXT ./preprocessing.sh > training.txt



```
#!/bin/bash -e
sed "s/^M//g" \
| tr '\n' ' ' \
| sed "s/\t/ /g"    \
| sed "s/(\([^()]*\))/\n\1\n/g"    \
| sed "s/\"\([^\"]*\)\"/\n\1\n/g"   \
| sed "s/\'\([^\']*\)\'/\n\1\n/g"   \
| sed "s/\[[^][]*\]//g"   \
| sed "s/[,:\/\` ]/ /g" \
| sed "s/[\?\!\.;]/\n/g"   \
| sed "s/[^a-zA-Z0-9 ]/ /g"   \
| sed "s/./\L&/g"   \
| sed "s/ [ ]*/ /g"   \
| sed "s/^[ \t]*//g"   \
| sed "s/[\t ]*$//g" \
| sed "/^$/d" \
| sed "/^[^ ]*$/d" \
| sed "s/[^[:print:]]//g" \
| sed "s/^/<s> /" \
| sed "s/$/ <\/s>/" \
```

Toolkits can use/cannot use

Can use:

- word2vec, topic model, N-gram, cvxopt

Cannot use:

- blocks, lasagne