

Submission Due: 11:59pm on March 15 (Mon)

Submission: Send iPython notebook (or Python codes) and a report in PDF to ai6127.assignments@gmail.com with "AS2-[YourName]"

2. Question Two [50 marks]

- a. Rerun the implementations of machine translation (MT) of Tutorial 6 with the multiple parallel datasets at <http://www.statmt.org/wmt14/training-parallel-nc-v9.tgz> as follows: **(20 marks)**
 - i. Refer to <http://www.statmt.org/wmt14/translation-task.html#download> (News Commentary) for details
 - ii. Randomly split a dataset into 5 subsets (S_1, \dots, S_5). Run training/testing for 5 times as follows: **(5 marks)**
 1. Select S_i ($i=1, \dots, 5$) as test dataset and the other 4 subsets as training dataset. Train a model with the training dataset and evaluate the model against the test dataset in terms of BLEU (BLEU-1, BLEU-2, BLEU-3).
 2. Report the average of the 5 BLEU scores for each dataset.
 - iii. Rerun the implementations for Question 3 **(5 marks)** and Question 6 **(10 marks)** of Tutorial 6.
 1. Do not use the filters of Question 1.b of Tutorial 6 (i.e. MAX_LENGTH, prefixes)
 2. For Question 6, if you do not find parameters that outperform the MT implementation of Question 3 for a dataset, specify which parameters you have tested and discuss why they were not effective for the dataset.
 3. Note that the file has 4 parallel datasets (CS-EN, DE-EN, FR-EN, RU-EN). Rerun the implementations for all the 4 datasets, English as target language.
- b. The word embeddings of Tutorial 6 are randomly initialized. Initialize them with the word embeddings that are pre-trained by word2vec or Glove using the monolingual training data at <http://www.statmt.org/wmt14/translation-task.html#download>. Rerun the implementations for Question 2.a.iii of this assignment (for Question 3 of Tutorial 6) with the pre-trained word embeddings. **(10 marks)**
- c. Attention variants **(20 marks)**
 - i. The Attention Decoder of Tutorial 6 is different from the attention decoder of the lecture (Sequence-to-sequence with attention). Explain the difference **(5 marks)**
 - ii. Replace the dot-product attention of Question 2.c.ii of this assignment to the following attention variants: **(10 marks)**
 1. Multiplicative attention: $e_i = s^T W h_i \in \mathbb{R}$
 - a. Where $W \in \mathbb{R}^{d_2 \times d_1}$ is a weight matrix
 2. Additive attention: $e_i = v^T \tanh(W_1 h_i + W_2 s) \in \mathbb{R}$
 - a. Where $W_1 \in \mathbb{R}^{d_3 \times d_1}, W_2 \in \mathbb{R}^{d_3 \times d_2}$ are weight matrices and $v \in \mathbb{R}^{d_3}$ is a weight vector. d_3 (the attention dimensionality) is a hyper-parameter.

3. Refer to <https://ruder.io/deep-learning-nlp-best-practices/index.html#attention> for more details of the variants
4. Compare evaluation results with the results from Questions 2.b and 2.c.ii of this assignment, and discuss why. **(5 marks)**