CS224n: NLP with Deep Learning **Winter 2019**

Lecture 9 & 10

Course Coordinator: Prof. Chris Manning *Scribes: Akash Gupta*

Dealing with $< UNK >$ -

- Use modest vocabulary size (50K)
- In translation this is a problem since the system is producing rare which are not in the output vocabulary and so it produces $< UNK >$
- Possible approaches - Hierarchical softmax, Noise-contrastive estimation (as seen in word2vec), Train on subsets of vocab at a time and test on an adaptively likely list of words. (Jean et. al ACL 2015).
- Use attention and copying models

Evaluation of MT -

- Adequacy, Fluency, Error categorization, Comparative ranking of translations
- Test in an application with MT as one sub-component(E.g. - QA) and check thru it's performance.
- Automatic metrics - BLEU, TER, METEOR,....
- BLEU is like a weighted geometric average on n-gram precisions.

Doing a research project -

- Step 1 - Define Task. E.g. Summarization.
- Step 1 - Define Dataset (Published or create your own)
- Step 3 - Dataset hygiene (At the beginning, separate off dev and test sets)
- Step 4 - Define your metric(s)
- Step 5 - Establish a baseline (Implement a simple model maybe bag-of-words or logistic regression)
- Step 6 - Implement existing NN model.
- Step 7 - Always be close to your dataset (Visualization, Summary,...)
- Step 8 - Try out different models and model variants.

Question Answering -

- Two parts - Finding documents that (might) contain an answer (by IR/web search), Finding an answer in a paragraph or a document (Reading Comprehension).
- MCTest Reading Comprehension -

$$Passage(P) + Question(Q) -> Answer(A)$$

- Factoid QA - In NLP, the answer for these questions is a named entity having semantic meaning.

Stanford Question Answering Dataset (SQuAD) -

- Evaluation - Exact match (System span is one of n gold answers), F1 (more reliable)
- SQuADv1.1 - All questions have respective answer. Problem with this was that NNs were doing sort of a ranking task of answers rather than learning what the answer means.
- SQuADv2.0 - 1/3 of training questions didn't have answers, and 1/2 of dev/test set didn't have answers.
- SQuAD limitations -
  - Only span based answers (no yes/no, counting, implicit why)
  - Questions were generated after looking at the passage (Problem - Not real world scenario where humans think up questions.)

Stanford Attentive Reader - Simplest neural QA system. Can form baseline for the projects. Training objective is how accurately the model is predicting start and end positions.

Stanford Attentive Reader++ - Uses attention with a randomly initialized vector and calculates weighted sum of hidden states of Bi-LSTM for generating concise vector rep. for the question whereas SAR only took concatenation of forward and backward hidden states.

BiDAF - Bi-directional Attention flow for Machine Comprehension (ICLR 2017). Adds character level processing as well as bi-directional flow of attention i.e. question to passage and vice-versa.

$->$ BERT- top performing network in recent challenges for QA task.