Lecture 15: Natural Language Generation

Shubham Gupta

March 29, 2020

1 Introduction

- Explore neural approaches for NLG
- NLG is subcomponent of:
  - MT
  - Abstractive Summarization
  - Dialogue
  - Createive writing
  - Freeform Q and A
  - Image captioning
- Language Modelling: Predicting next word given words so far. $P(y_t|y_1, y_2, ..)$
- Conditional Language Modelling: Predicting next word given words so far and another input x. $P(y_t|y_1, y_2, .., x)$
  - MT
  - Summarization
  - Dialogue
- During training of RNN-LM, we feed gold(true) target sentence to decoder, regardless of it's output. This is called **Teacher Enforcing**.

2 Decoding algorithms

- Greedy decoding: Pick most probable word at each step. Use that as next work. Output can be poor
- Bean search: Find high probability sequence. Keep track of k most probable sequences. After stopping, choose sequence with highest probability.

2.1 Sampling based decoding

- Pure sampling: Randomly sample from P_t to obtain next word.
- Top-n sampling: Random sampleafrom P_t, restricted to top-n most probable words. More efficient than beam search. Increase n to get more diverse/risky output. Decrease to get generic/safe output
- Softmax temperature
  - At t, compute scores by using softmax function
  - Apply temperature hyperparameter τ to the softmax

$$P_t(w) = \frac{exp(s_w/\tau)}{\sum_{\hat{w}\epsilon V} exp(\frac{\hat{s_w}}{\tau})} \tag{1}$$

  - Increasing τ =¿ Dist becomes uniform. More diverse output
  - Decreasing τ =¿ Dist becomes spikes. Less diverse output
  - This is **not a decoding algorithm**. This is a technique that can be applied during testing time along with regular decoding algorithms

3 NLG tasks and neural approaches to them

3.1 Summarization

- Given input x, summary y which is shorter and contains main information of x
- *Sentence Simplicifation*: Rewrite source text in simplier way

3.2 Stratergies

- Extractive: Extract original text
- Abstractive: generate new text with NLG

3.3 Metrics

- ROUGE: Recall Oriented Understudy for Gisting Evaluation
- Like BLEU. Based on n-gram overlap. Differences are:
  - No brevity penalty
  - ROUGE based on recall, BLEU based on precision
- Commonly reported ROUGE scores are:
  - ROUGE-1: unigram overlap
  - ROUGE-2: bigram overlap
  - ROUGE-L: longest common subsequence overlap

3.4 Neural Summarization: copy mechanisms

- seq2seq + attention systems good at writing output BUT bad at copying details
- Solution: Use attentiion to enabe seq2seq system to easily copy words and phrases from input to the output.
  - More useful in summarization
  - Allows hybrid extractive/abstractive approach
  - Copy too much. Mostly copy whole sentences
  - Collapses into mostly extractive system
  - Bad at content selection. Fails for long documents
  - Solution: use bottom up summarization. Use word masking, which tells the model that certain words cannot be included in the summary. Better and less copying

3.5 RL for summarization

- Optimize ROUGE-L
- Improves ROUGE scores but makes readability much worse
- RL + ML model produces higher ROUGE scores and higher readability score

4 Dialogue

- Task-oriented dialogue
  - Assistive: Customer support, giving recos, QA
  - Co-operative: Solve task together
  - Advesarial: Compete
- Social Dialogue
  - Chit-chat
  - Therapy/mental wellbeing
- seq2seq + attention has problems
  - Generic
  - Irrelevant responses
  - Repetition
  - Lack of context
  - Lack of consistent persona

4.1 Irrelevant response

- Unrelated response.
- Solution: Maximum Mutual Info between input S and response T.

$$T = argmaxlogp(T|S) - logp(T) \tag{2}$$

4.2 Generic/Boring response problem

- Easy fixes: Directly upweight rare words during beam search
- Use a sampling decoding algorithm rather than beam search
- **Conditioning fixes**: Condition decoder on additional context
- Train retrieve-and-refine model instead of generate-from-scratch model

4.3 Repetition problem

- Simple: Directly block repeating n-grams during beam search. Very effective
- Complex solutions
  - Train a coverage mechanism. Prevent attention from attending to same words multiple times.
  - Training objective to discourage repetition. Non-differentiable function. Will have to be solved with RL.

5 Storytelling

- generate paragraph given input
- Shared embedding space.
- Use skip-thought vectors to create sentence embedding
- Learn mapping from images to skip-thought encoding of their captions
- Train RNN-LM to decode skip-thought vector to the original text
- Put the two together

6 NLG evaluation

- Word overlap metrics not ideal for MT. Worse for summarization.
- Find focused automatic metrics
  - FLuency(compute prob wrt LM)
  - Style
  - Diversity

– Relevance
– Length and repetition
– Task specific metrics