Ryan Napolitano

rn2473

Jason Herrera

jjh2210

COMS 4731

HW 4 – Sequence to Sequence Modeling

Written Portion

\*\*\*\* Using 2 Late Days from Ryan Napolitano, rn2473 \*\*\*\*

2 Written Portion

2.1 Group work description

Briefly list the contributions of each group member for this assignment.

As the solutions were released for the majority of the assignment, our group worked collaboratively on the available coding portion of the assignment, nucleus sampling. In addition, we also worked collaboratively on adjusting the solutions to provide material in a manner that facilitated answering the written portions of the assignment.

Our workstyle consisted of working through Zoom with one student listening and providing feedback while the other student wrote code for the assignment. The project was then uploaded using Github, and on the next day the other student would take over the coding duties.

2.2 Error analysis

Describe three kinds of errors the system makes, with examples (input and output) from

the development dataset. You may use greedy decoding for this question. Describe each

error in at least 2 sentences.

1. Greedy Decoding will produce sentences that repeat words and can produce nonsensical statements. This is most likely due to the system choosing the most likely words without any benefit of options with higher probability past an immediate choice.

['name[Wildwood]', 'eatType[coffee shop]', 'food[Chinese]', 'priceRange[£20-25]', 'customer rating[high]', 'near[Ranch]']

GR: 0.5248 <sos> Wildwood is a coffee shop that serves Chinese food near Ranch. food is is <eos>

1. Hallucination occurred in our Beam Search results as some details manifested in our outputs that were not a part of the input. Our model most likely produced these outputs because it was trained on loosely corresponding data.

['name[Wildwood]', 'eatType[coffee shop]', 'food[Chinese]',

'priceRange[£20-25]', 'customer rating[high]', 'near[Ranch]']

BM: 0.3468 <sos> Near the Ranch is a coffee shop that serves

Chinese food at a high price range and a customer rating of 3 out of 5. It is called Wildwood. <eos>

1. Our model also seems to have issues interpreting the subjectivity of user reviews as a 3 of 5 star

rating is often attributed as high. We believe that this has a similar origin to the

hallucinated outputs in that the data is loosely corresponded.

['name[Cocum]', 'eatType[coffee shop]', 'food[English]',

'priceRange[£20-25]', 'customer rating[high]',

'familyFriendly[yes']

BM: 0.4171 <sos> Cocum is a kids friendly coffee shop that offers English food at a high price range and a customer rating of 3 out of 5. <eos>

2.3 Beam search analysis

(a) Show your BLEU scores on the development set for greedy decoding and beam search with different beam sizes (5, 10, 15, 20). For beam search, use the top scoring sequence (i.e. the one with the highest probability) to calculate the BLEU score for that datapoint. Discuss the differences in BLEU score for greedy and beam search decoding.

BLEU for greedy decoding: 0.4696

BLEU for beam size 5 search: 0.5193

BLEU for beam size 10 search: 0.5025

BLEU for beam size 15 search: 0.5069

BLEU for beam size 20 search: 0.5112

Beam Search outperforms Greedy Decoding for all beam sizes. This is expected because Beam Search and Greedy Decoding are equivalent at beam size 1, but once the beam size increases past 1, Beam

Search is able to consider additional sequences that could lead to a higher probability.

There does not appear to be a benefit to increasing the beam size above 5 because, although increasing the beam size results in a higher probability sequence, a higher probability sequence is not necessarily a better match to the human references that the BLEU score is comparing it to.

(b) Provide and discuss three examples of beam search. That is, for three input sequences,

show the input sequence and multiple outputs from the beam search (use beam size=5).

INPUT SEQUENCE A:

['name[Wildwood]', 'eatType[coffee shop]', 'food[Chinese]', 'priceRange[£20-25]', 'customer rating[high]', 'near[Ranch]']

1. GR: 0.5248 <sos> Wildwood is a coffee shop that serves Chinese food near Ranch. food is is <eos>
2. BM: 0.8235 <sos> Wildwood is a coffee shop providing Chinese food at a high price range. It is near Ranch. Its customer rating is high. <eos>
3. BM: 0.8042 <sos> Wildwood is a coffee shop providing Chinese food at a high price range. It is located near Ranch. Its customer rating is high. <eos>
4. BM: 0.3468 <sos> Near the Ranch is a coffee shop that serves Chinese food at a high price range and a customer rating of 3 out of 5. It is called Wildwood. <eos>
5. BM: 0.3608 <sos> Near the Ranch is a coffee shop that serves Chinese food at a high price range and a customer rating of 3 out of 5 is Wildwood. <eos>
6. BM: 0.3849 <sos> Near the Ranch is a coffee shop that serves Chinese food at a high price range and a customer rating of 3 out of 5. <eos>

INPUT SEQUENCE B:

['name[Cocum]', 'eatType[coffee shop]', 'food[English]', 'priceRange[£20-25]', 'customer rating[high]', 'familyFriendly[yes]']

1. GR: 0.3026 <sos> Cocum is a coffee shop that shop coffee shop food is and is range of is <eos>
2. BM: 0.4171 <sos> Cocum is a kids friendly coffee shop that offers English food at a high price range and a customer rating of 3 out of 5. <eos>
3. BM: 0.4791 <sos> Cocum is a kids friendly coffee shop that offers English food at a high price range and is highly rated by customers. <eos>
4. BM: 0.3652 <sos> Cocum is a kids friendly coffee shop with a price range of £20-25 and a customer rating of 3 out of 5. <eos>
5. BM: 0.3221 <sos> Cocum is a kids friendly coffee shop with a price range of £20-25 and a customer rating of 3 out of 5 and serves English food. <eos>
6. BM: 0.5058 <sos> Cocum is a kids friendly coffee shop that offers English food at a high price range and a customer rating of customer rating. <eos>

INPUT SEQUENCE C:

['name[The Punter]', 'eatType[coffee shop]', 'food[Chinese]', 'priceRange[£20-25]', 'customer rating[high]', 'familyFriendly[no]', 'near[Café Sicilia]']

1. GR: 0.3684 <sos> The Punter is a coffee shop that is Chinese food near Café Sicilia. It is not is not a customer and <eos>
2. BM: 0.2251 <sos> Near Café Sicilia, there is a coffee shop that offers Chinese food at a high price range called The Punter. It is not kid friendly and has a high customer rating. <eos>
3. BM: 0.2574 <sos> Near Café Sicilia, there is a coffee shop that offers Chinese food at a high price range and a customer rating of 3 out of 5 named The Punter that is not kid friendly. <eos>
4. BM: 0.2502 <sos> Near Café Sicilia, there is a coffee shop that serves Chinese food at a high price range and a customer rating of 3 out of 5 named The Punter that is not kid friendly. <eos>
5. BM: 0.1635 <sos> Near Café Sicilia, there is a coffee shop that offers Chinese food at a high price range and a customer rating of 3 out of 5 named The Punter. <eos>
6. BM: 0.1570 <sos> Near Café Sicilia, there is a coffee shop that serves Chinese food at a high price range and a customer rating of 3 out of 5 named The Punter. <eos>
7. How do the multiple outputs differ?

The outputs differ at times omitting different pieces of information given in the inputs, like the case in Input A, output 6 where the restaurant name is not given at all. The outputs also differ in their placing of the information as sometimes the name of the restaurant is given in the beginning of the sentence, and at other times at the end, as seen in Input A, outputs 4 and 2.

1. Will beam search always include the greedily decoded output? (Why or why not?)

In our implementation of Beam Search, the greedily decoded output will not always be included. Since we are using beams of size 5, 10, 15, and 20 the model will consider many alternatives rather than the immediate highest probability option. If we used a beam size of 1, we would see the greedily decoded output *always* included.

1. What is one advantage and one disadvantage of beam search, compared to greedy decoding?

One advantage of Beam Search is that it allows the model to look at other more highly probable sequences that can result in a more human-like sequence more often than a Greedy Decoder. A disadvantage of Beam Search is that it “tends to degrade with large beam widths” (Cohen). With a larger beam width the model tends to select lower probability selections first, then higher probability selections later, which leads to higher probabilities overall. This, however, generally lead to lower evaluation scores for the sequences.

Cohen, Eldan and Beck, Christopher. (Unconstrained) Beam Search is Sensitive to Large Search Discrepancies. Accessed at: <https://openreview.net/forum?id=BkE8NjCqYm>

2.4 Nucleus sampling analysis

(a) Show your BLEU score on the development set using nucleus sampling for decoding

(use p=0:95) and compare it with greedy decoding and beam search (beam size=5). Discuss the differences.

(b) How does the average sequence probability compare to greedy decoding and beam

search? Give an explanation for differences observed.

(c) Use nucleus sampling with different values of p (0.1, 0.5, 0.95) to generate 5 sequences

per sample. You can randomly sample 5 examples from the development

set for this comparison.

i. How does the value of p affect the quality of the generated outputs?

ii. How does the value of p affect the average number of tokens in the nucleus?

(d) Use nucleus sampling (p=0:95) to generate 5 sequences per sample and compare

the output with beam search (beam size=5). You can randomly sample 5 examples

from the development set for this comparison.

i. How does the quality of the sequences generated using nucleus sampling compare to those generated by beam search?

ii. What is one advantage and one disadvantage of nucleus sampling, compared

to beam search?

iii. When would you prefer nucleus sampling? When would you prefer beam

search?

2.5 Beam search with nucleus sampling

One natural extension would be to combine beam search and nucleus sampling, so that we can get the benefits of both approaches.

(a) Write pseudocode for a modified beam search where rather than considering the top k next words per sequence, you sample k words from the nucleus of the distribution, at each decoding step.

(b) What’s wrong with combining them in this way?

(c) Now write a modified version of the pseudocode you wrote above, that fixes this issue. Feel free to cite existing work that may help you address the problem, but you may not copy the solution. You should be able to explain it in your own words, and write your own psuedocode.

2.6 What’s wrong with BLEU?

Although several new metrics have been proposed, BLEU remains a common evaluation metric used to evaluate models.

(a) What are some problems with BLEU as a metric? List two and explain why these

are a problem.

1. Machine translation systems can over generate “reasonable” words, that create high precisions translations while also being improbable (Papeneni, 2). These translations still lead to high BLEU score metrics. Since some forms pf BLEU do not take into account frequency, this is a disadvantage as high scoring sequences may be un-readable but produce high BLEU scores.
2. Another shortcoming of using BLEU as a metric is that human translations do not perform to near perfect scores using BLEU as a metric. This makes it difficult to use BLEU as a measure when generating text since it does not measure how human-like a text is, but rather compares similar unigrams between the two.

(b) What are the benefits of BLEU as a metric?

1. A major benefit of using BLEU as a metric is that since it compares n-grams between a refence and output, it can be used across multiple languages. This allows us to judge a model across multiple languages using a single metric.
2. BLEU scores have been demonstrated to show high correlation with human judgment for translations into English from many different languages (Papeneni, 8). This allows for insightful measures into how a model performs in translations tasks.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. BLEU: a Method for Automatic Evaluation of Machine Translation. Accessed at <https://www.aclweb.org/anthology/P02-1040.pdf>

2.7 How could we deal with unseen restaurant names?

Try generating a sentence with a restaurant name that the model has never seen before. Show the result. How could you change the model to deal with this? Be specific enough that one could try to implement this. (You will not actually implement this, so it could be quite complicated. As long as it’s within the realm of possibility, dream big.)

2.8 Paper analysis

You read two papers that discussed the problem of hallucination of neural models in abstractive summarization (https://arxiv.org/pdf/1509.00685.pdf) and for language generation

(https://arxiv.org/abs/1911.03373). Compare the methods proposed in each paper to alleviate hallucination by:

1) providing a brief, paragraph-length description of the method,

2) discuss evidence on how well it works as indicated in the paper and

3) providing your opinion on whether it can improve your model’s generations and why.