Query Focused Abstractive Summarization Error Analysis

Kailash Karthik S

March 5 2020

1 Introduction

An experiment was conducted to verify the salience of keyphrase-based queries in abstractive summarization using the model proposed by (1). The evaluation scores of the summaries was presented in a prior document and an error analysis is presented here.

For a TL;DR version, please skip to section 5 which contains the conclusion and an illustrative example.

2 Goals

The ROUGE scores shows a significant improvement in the recall scores (for example, the ROUGE-1 recall increases by 50% from 0.35 to 0.53. However, the precision of the summaries decreases, negating the gain in recall and thereby producing a similar F-score. Please refer to Appendix A for the results of the summary evaluation.

Thus the goal of the analysis is to perform a manual evaluation to understand these results to explain them and to use those insights to build a better query focused summarization model.

The initial hypothesis - under the assumption that the query focusing technique works, is that perhaps the model is retrieving longer summaries since there are more sentences in the document that are relevant to the query than mentioned in the highlights. The second goal is to validate this claim.

Finally, a validation of the assumption made in the previous sentence is also required - that the model indeed uses the queries to add focus to the summaries as opposed to ignoring their presence.

3 Experimental Setup

A sample set of 51 documents are chosen at random and the summaries are generated using both the query-focused and query-agnostic models. Two separate queries are used for decoding using the query-focused model to verify that

changing the query indeed changes the summaries generated. These summaries are used for the following analyses. This set of document is called the *Analysis Set* from now on.

4 Observations and Analysis

1. **Summary Length**

   As mentioned above, it is observed that lengths of the summaries produced by the query focused model are higher than the other model. In the analysis set, it is observed that while the average length of the summary generated by our model is 4.82, the summaries generated by the query-agnostic model are much shorter and have an average length of 1.49. This could be a possible explanation for the low precision of our model as it contains more words than the target.

2. **Effect of Queries**

   While not all the summaries generated are strictly query-focused, it is observed that in many cases the model uses the query to generate sentences that are relevant to the query. However, the model tends to generate a large number of sentences related to the query and perhaps loses the idea of sentence salience. A few examples are below:

   (a) **Query** : jones

   Target Summary:
   "amelia rose jones allegedly murdered by her grandfather , mark jones , 45 .", "alleged that fatal blows came from jones while she was in his sole care ."

   Generated Summary:
   "mark jones , 45 , is accused of attacking newborn amelia jones three times .", "infant was taken to hospital in 2012 after grandfather jones rang 999 .", "jones disliked amelia 's father , ian skillern , who had finished relationship with his daughter sarah jones , 25 .", "mark jones is being tried before a jury at newport crown court .", "amelia rose jones was murdered by her grandfather mark jones .", "a grandfather allegedly murdered his granddaughter as he babysat her because he disliked her father ."

   (b) **Query**: terminal pancreatic cancer last year

   Target Summary:
   "barbara carter was diagnosed with terminal pancreatic cancer last year ."

Generated Summary:
"barbara carter , 56 , was diagnosed with terminal pancreatic cancer in november .", "barbara carter was given the wedding of her dreams to partner graham .", "barbara , from bury , mentioned the wedding to her daughter joanne , 31 whilst enduring chemotherapy treatment .", "barbara was diagnosed with pancreatic cancer last year .", "barbara says she has taken her mind off the cancer ."

3. **Uninformative Summaries**

The query focus of the model is not perfect by any means as in cases, the model generates summaries that are neither query focused nor salient. This could be due to the fact that the extractive model does not know how to handle such queries. A possible solution to this problem - which is also our next step - is to train the model using multiple queries for each document so that the model learns to generate different summaries on the same article for different queries. An example of such a summary with uninformative and not salient content is as follows:

(a) **Query**: sylvester stallone

Target Summary:

"it was claimed sylvester stallone had to intervene in the row ."

Generated Summary:

"liverpool cruiserweight climbed out of the ring and started punching his corner man .", "sylvester stallone , who is reprising his role as rocky balboa , was alleged to have broken the fight .", "tony bellew films the spin-off to the rocky franchise .", "bellew denies reports that he was involved in a bust-up ."

4. **Simulation of Conceptual Queries**

The query focusing mechanism seems to summarize using the queries as abstract concepts and looking for content in the source articles that have a conceptual similarity to it. The following is an illustrative example of the same as the model learns the relationship Batter Pack - Tesla - Elon Musk. Though the summary in itself is not very query focused, it is shown as evidence of this concept hopping ability of the model.

(a) **Query**: new home battery pack

Target Summary:

"launch date coincides with the arrival of tesla 's new home battery pack ."

Generated Summary:

"tesla 's billionaire founder fuelled speculation by saying it ' not a car '", "elon musk unveils new line of products .", "launch date coincides with arrival of stationary battery pack .", "musk fans making lighthearted guesses ranging from iron man suit to an inflatable helicopter .", "elon musk has announced the launch of a ' major line of products '"

5. **Decoding with Alternate Query**

   When the model was used to summarize articles with focus on queries that it was not trained on, it is observed that the model fails to generate a completely different summary. Though in some cases, one sentence in the generated summary is different, it is largely the same. The reason for this behavior can possibly be attributed to the fact that the model was trained on a single query for each document and thus has not learned to distinguish between different queries on the same document. The rationale for training on a single query for each document was to be able to compare it with the query-agnostic model and thus, this inability of the model should not be alarming at the moment. If the model, however, fails to distinguish even after training on multiple query-summary pairs for each article, then we would indeed have to rethink our approach to this problem.

5 Conclusion

From the analysis of the summaries generated by the query-focused model, I am of the opinion that the model does learn to focus on the query in our architecture. The main drawback is that it does not learn to distinguish between queries and the solution proposed is to train the model on multiple query-summary pairs for each article. I conclude from the example outputs analyzed that the direction might be correct. An illustrative example to support my claim is shown below. The model generates additional sentences not present in the target summary as it uses the query focusing technique to understand that those sentences are indeed relevant to the query:

Query : tonga prop taione vea

Target Summary: newcastle have signed tonga prop taione vea from london welsh .

Generated Summary: newcastle have announced the signing of **tonga** international **prop taione** . the 26-year-old has won six caps for **tonga** .

taione vea signed for newcastle falcons on a two-year deal . newcastle director of rugby dean richards said he is ‘ a quality **prop** ’ . **vea** has six caps for **tonga** and previously played for wasps .

References

[1] Yen-Chun Chen and Mohit Bansal. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting, 2018; arXiv:1805.11080.

Summary Evaluation Results

A Rouge Scores

A.1 Generic Summary

```
---------------------------------------------
1 ROUGE-1 Average_R: 0.35180 (95%-conf.int. 0.34780 - 0.35552)
1 ROUGE-1 Average_P: 0.24725 (95%-conf.int. 0.24428 - 0.25013)
1 ROUGE-1 Average_F: 0.26934 (95%-conf.int. 0.26675 - 0.27199)
---------------------------------------------
1 ROUGE-2 Average_R: 0.15029 (95%-conf.int. 0.14717 - 0.15355)
1 ROUGE-2 Average_P: 0.10382 (95%-conf.int. 0.10164 - 0.10602)
1 ROUGE-2 Average_F: 0.11308 (95%-conf.int. 0.11081 - 0.11549)
---------------------------------------------
1 ROUGE-L Average_R: 0.32161 (95%-conf.int. 0.31776 - 0.32524)
1 ROUGE-L Average_P: 0.22637 (95%-conf.int. 0.22350 - 0.22908)
1 ROUGE-L Average_F: 0.24640 (95%-conf.int. 0.24378 - 0.24901)
```

A.2 Query Focused Summary

```
---------------------------------------------
1 ROUGE-1 Average_R: 0.53736 (95%-conf.int. 0.53352 - 0.54095)
1 ROUGE-1 Average_P: 0.19016 (95%-conf.int. 0.18808 - 0.19204)
1 ROUGE-1 Average_F: 0.26501 (95%-conf.int. 0.26263 - 0.26709)
---------------------------------------------
1 ROUGE-2 Average_R: 0.24484 (95%-conf.int. 0.24088 - 0.24854)
1 ROUGE-2 Average_P: 0.08259 (95%-conf.int. 0.08123 - 0.08393)
1 ROUGE-2 Average_F: 0.11601 (95%-conf.int. 0.11427 - 0.11774)
---------------------------------------------
1 ROUGE-L Average_R: 0.50013 (95%-conf.int. 0.49621 - 0.50366)
1 ROUGE-L Average_P: 0.17671 (95%-conf.int. 0.17469 - 0.17851)
1 ROUGE-L Average_F: 0.24634 (95%-conf.int. 0.24415 - 0.24835)
```

B Meteor Scores

B.1 Generic Summary

	Test Matches			Reference Matches		
Stage	Content	Function	Total	Content	Function	Total
1	47662	56714	104376	47662	56714	104376
2	707	22	729	709	20	729
3	2437	701	3138	2522	616	3138
4	4392	5055	9447	4436	4687	9123
Total	55198	62492	117690	55329	62037	117366

```
Test words:              395448
Reference words:         292277
Chunks:                  74590
Precision:               0.25792009148063005
Recall:                  0.34729165042860527
f1:                      0.29600712625446174
fMean:                   0.3301325782108259
Fragmentation penalty:   0.5478467664588851

Final score:             0.14927051273528996
```

B.2 Query Focused Summary

	Test Matches			Reference Matches		
Stage	Content	Function	Total	Content	Function	Total
1	74074	80225	154299	74074	80225	154299
2	1430	58	1488	1435	53	1488
3	3400	1036	4436	3504	932	4436
4	6311	7642	13953	6543	6210	12753
Total	85215	88961	174176	85556	87420	172976

```
Test words:              799038
Reference words:         292277
Chunks:                  104734
Precision:               0.19391897927234603
Recall:                  0.5244754377118569
f1:                      0.2831473606419651
fMean:                   0.41767851762424385
Fragmentation penalty:   0.5423390215732409

Final score:             0.1911551590437498
```