

Data Science


Natural Language Processing

NLP Papers Summary

# Day 156: NLP Papers Summary – Asking And Answering Questions To Evaluate The Factual Consistency Of Summaries

By Ryan 4th June 2020 No Comments

## Objective and Contribution

Proposed Question Answering and Generation for Summarisation (QAGS), an automatic evaluation method that is designed to identify factual inconsistencies in a generated summary. QAGS uses question answering on both the source and summary and a factual consistent summary should be able to produce similar answers as the source. We compared QAGS with human judgements and found that there are high correlations. Lastly, QAGS  [ides](#)

interpretability as to which part of the summary are factual inconsistent using the answer and questions generated.

## QAGS Framework

QAGS framework consists of 3 steps:

1. Question generation model to generate questions based on generated summary
2. Question answering models to answer the questions using both the source document and generated summary
3. Answer similarity is computed based on how similar the answers are

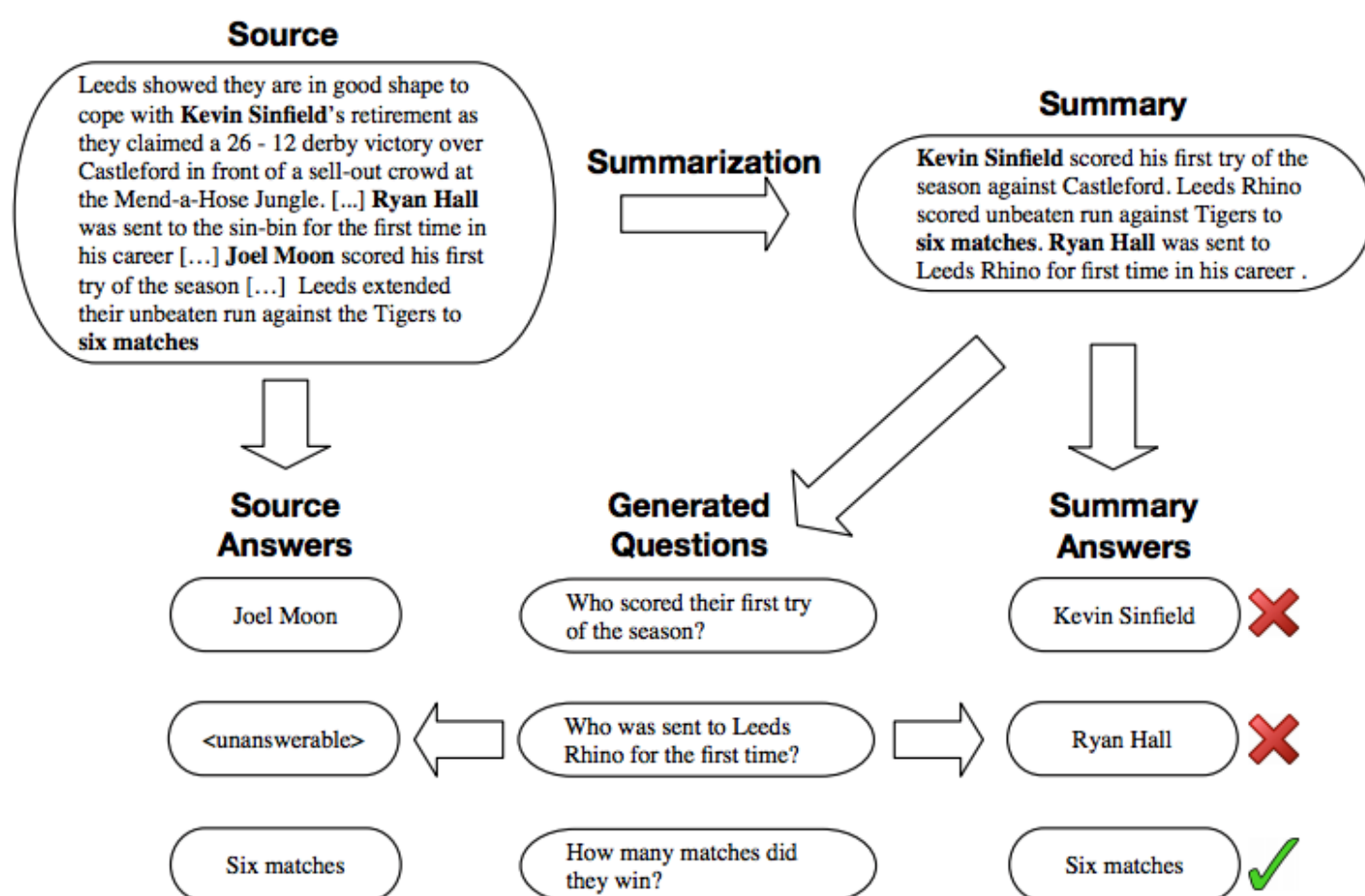


Figure 1: Overview of QAGS. A set of questions is generated based on the summary. The questions are then answered using both the source article and the summary. Corresponding answers are compared using a similarity function and averaged across questions to produce the final QAGS score.

## QUESTION GENERATION

We train a seq2seq model to generate questions based on both the answer and source article. We over-sample questions and use different methods to filter out low-quality questions such as

removing duplicates and questions with three tokens or less. We also feed the question  $\hat{q}$  to the QA model and remove questions that are predicted with no answer. And so, in this step, we generated K questions based on the summary.

## QUESTION ANSWERING

Here, we have an extractive QA models to extract the answers as text spans from the source document and summary. Future work could experiment with abstractive QA models. In this step, we answer those generated questions using both the source and summary to obtain two sets of answers.

## ANSWER SIMILARITY

Here, we have a simple token-level F1 score to compare the answers and measure answer similarity. In this final step, we compare the answers using the similarity metric and averaging the answer similarity score over all questions.

## Experiments

We have two evaluation datasets: CNN/DM and XSUM. We measured the correlations between QAGS and human judgements of factual consistency. For each summary, we collected 3 annotations and obtain a single correctness score per summary by taking the majority vote for each sentence and averaging the binary scores across summary sentences. We compared our QAGS metric with other common summarisation metrics such as ROUGE, METEOR, BLEU, and BERTScore.

## Results

The table below showcase the correlation between different evaluation metrics and human judgements of factual consistency. We show that QAGS achieved the highest correlation by a substantial margin. QAGS performed 2x better than the next best performing metric. QAGS scored significantly lower in XSUM but still outperformed other metrics by a wide margin. This showcase that XSUM dataset is more abstractive.



Metric	CNN/DM	XSUM
ROUGE-1	28.74	13.22
ROUGE-2	17.72	8.95
ROUGE-L	24.09	8.86
METEOR	26.65	10.03
BLEU-1	29.68	11.76
BLEU-2	25.65	11.68
BLEU-3	23.96	8.41
BLEU-4	21.45	5.64
BERTScore	27.63	2.51
QAGS	<b>54.53</b>	<b>17.49</b>

Table 1: Summary-level Pearson correlation coefficients between various automatic metrics and human judgments of correctness for summarization datasets. QAGS obtains substantially higher correlations than all other automatic metrics.

## Ablations

### MODEL QUALITY

We use different models for different steps in our framework. We explore how the quality of these models would affect our evaluation capabilities. For our QA model, we train and fine-tune different version of BERT on SQUAD. The results are showcase below. We show that the best QA model with the highest F1 score does not mean higher correlation with human judgements. In both CNN/DM and XSUM, bert-base QA model achieved the highest correlations with human judgements despite scoring the lowest F1 score.

For our QG model, we use models with increasing perplexity on the NewsQA dataset. The results are showcase below. We show that QAGS is robust to the quality of QG model as we

see no clear trend of higher quality QG model leads to higher correlation with human judgements.

## DOMAIN EFFECTS

The QAGS framework requires labelled data to train both QG and QA models. This might be effective in a data rich domain but in niche domains, we might not have access to labelled data. In those situations, we are forced to use out-of-domain data to train our models which may negatively impact our QAGS quality due to domain shift. We assess the impact of this domain shift by training our QG model using SQUAD which it's a collection of wikipedia articles rather than CNN articles. The new correlations score with SQUAD-QG model is 51.53 and 15.28 on CNN/DM and XSUM dataset respectively. This is lower than the correlation scores when using NewsQA-QG model but it still significantly outperformed other evaluation metrics.

## NUMBER OF QUESTIONS

Lastly, we explore how the number of questions would affect the correlation with human judgements. The results are showcase below and it shows that as the number of questions increase, we see a consistent increase in correlation scores in both evaluation datasets. We also observed that a) with only 5 questions, we are able to achieve correlations higher than other evaluation metrics and b) there is only a small increase in correlation scores when increasing number of questions from 20 to 50, showcasing decreasing marginal benefit of including more than 50 questions.



## ANSWER SIMILARITY METRICS

There are many methods to measure similarity between two answers. An alternative to token-level F1 score, we can use exact match (EM) which it's more restrictive. With EM, we obtain correlation scores of 45.97 and 18.10 on CNN/DM and XSUM. Different answer similarity metrics are open for exploration.

## Re-ranking with QAGS

We explore the sentence ranking experiment where we have around 400 triplets, consisting of source sentences from CNN/DM and two generated summary sentences, where one is factually consistent and the other is inconsistent. We used QAGS and other baseline comparisons to measure how often it ranks the consistent sentence higher than the inconsistent sentence. The results are displayed below and QAGS outperformed all previous NLI methods.



## Qualitative Analysis

The QAGS framework provides high interpretability as the questions and answers generated allow us to directly highlight errors in summaries. The figure below showcase example of questions and answers generated. As we can see, by using questions and answers, we can detect several factual inconsistencies in our generated summary. For example, the attacker's name is Usman Khan but was changed to Faisal Khan in the summary. Our QG model can generate appropriate questions and our QA model focuses mainly on named entities and noun phrases. In the future, we can expand the answer candidates, allowing us to detect different kinds of errors. The second example showcase the weakness of QAGS where sometimes, different answers are correct but might not have common tokens, resulting in false error.



To determine the quality of our generated questions, article and summary answers, we manually annotated 400 triplets on the XSUM summaries and label them by their quality. We found that 8.75% of generated questions are nonsense and 3% are well-formed but couldn't be answer by the generated summary. This shows that a large proportion of our generated questions are easy to understanding, meaningful, and relatable. 8.25% of questions are well-formed but couldn't be answer by the source document, largely due to non-sensical facts that QG model turns into questions.

We have a large 32.50% of incorrectly answers using the source article, indicating that our QA model is weak. Finally, 8% of questions are answered correctly using both source article and summary but due to little or no overlap in tokens, it was identify as incorrect.

## Conclusion and Future Work

Potential future work could be to improve the question answering models, apply the metrics to different types of data or fields such as translation and image captioning.

Source: <https://arxiv.org/pdf/2004.04228.pdf>

Ryan

Data Scientist

Previous Post

Day 155: NLP  
Papers Summary -  
TRAIN ONCE,

Next Post

Day 157: NLP  
Papers Summary -  
Explaining





TEST ANYWHERE:  
ZERO-SHOT  
LEARNING FOR  
TEXT  
CLASSIFICATION

EXPLAINABLE

Prediction of  
Medical Codes  
from Clinical Text

[Data Science](#) [Natural Language Processing](#) [NLP Papers Summary](#)

## Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan

30th December 2020





[Data Science](#) [Implementation](#) [Natural Language Processing](#)

## Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

**Ryan**

29th December 2020





[Data Science](#) [Ryan's PhD Journey](#)

## Day 363: Ryan's PhD Journey – Literature Review – Knowledge Acquisition – 1st Passes XIV

**Ryan**

28th December 2020

