Evaluating and Comparing Textual Summaries Using Question Answering Models and Reading Comprehension Datasets

Mantas Gavenavicius
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands
m.gavenavicius@student.utwente.nl

ABSTRACT

The currently dominant approaches to automatic evaluation of summaries rely on measuring similarity between a candidate and a reference summary solely through lexical overlap. These methods might be limited in their ability to assess summary factuality, which we address in this work by evaluating summaries by their usefulness for question answering on reading comprehension tasks. We develop a framework for performing these evaluations without reliance on Question Generation models by repurposing existing human crafted datasets. Our experiments show that the scores produced by our method correlate highly with ROUGE when evaluated on the RACE dataset, and have low to medium correlation when evaluated on SQuAD 2, implying that well performing summarization systems (as evaluated by ROUGE) also do well on factual retention, although this is highly varied depending on the particular dataset. Our experiments also indicate that the gap between current state-of-the-art summarization models and simple baselines is still narrow when given out-of-domain text. We further test our method's sensitivity to word order, showing that further adjustments are needed to evaluate the fluency of the summaries.

Keywords

Question Answering, Summarization, Evaluation, Reading Comprehension

1. INTRODUCTION

Summarization is not just an abstract research topic in natural language processing, but a very real task whose product is encountered every day – whether in headlines of news articles, research paper abstracts or book covers.

Since human evaluation of summaries is infeasible at the scale required for deep learning, both model evaluation and training have to rely on using automatic metrics to compare the produced summaries (the output) against the given 'gold references' (the target). However, unlike classification or regression tasks where checking an output is mathematically trivial, summarization has to compare pieces of text that are often hundreds of words in length. Since human language is itself highly complex, capturing an objective level of similarity between texts is a hard task even for humans, and even more so for machines.

33rd Twente Student Conference on IT, July. 3rd, 2020, Enschede, The Netherlands.

Inspired by work in the machine translation field [10] which faces similar challenges, several automatic evaluation metrics have been developed (the most dominant of these, called ROUGE [7], is discussed in depth in section 2.1). While they have achieved great success in research adoption, these metrics tend to rely solely on comparing surface level lexical features of the texts, raising questions whether they can really capture the deeper aspects of language and objectively evaluate the summaries.

To address the limitations inherent in lexical comparisons, there has been work in extrinsic evaluation of summaries, which involves measuring how well machine models perform on some downstream task. Question Answering (**QA**) tasks are particularly interesting in regards to factuality and content selection. The main idea is that asking and answering questions about a text can help evaluate summaries for their ability to select key content without relying on explicit overlap with reference summaries, since the QA models might be able to gather answers even from more abstract summaries. Furthermore, using QA enables evaluation of summary factuality through assessment of answers that are generated using the summary.

In this work, we apply different summarization methods and models to passages from reading comprehension datasets and measure the resulting performance of a question answering model rather than lexical overlaps with reference passages. We formulate three main research questions:

1. How do state-of-the-art summarization methods perform in question answering evaluation settings?
2. Does this method of evaluation correlate with ROUGE scores?
3. Is the method sensitive to sentence and word order in summaries?

To address the possible shortcomings of prior work using question generation we focus specifically on datasets with human-crafted questions. This allows for greater trust that the questions are well written and the answers factually correct. Furthermore, prior work in this area has focused almost exclusively on news articles, which we address by selecting datasets containing passages from different domains.

We discuss related work, including ROUGE and other works using QA as an evaluation mechanism in section 2, our high-level approach in section 3 with a detailed in experimental setup in section 4. We lay out and discuss our results in section 5 and consider the wider implications in section 6. Examples of the datasets and considered outputs can be found in the Appendix.

2. RELATED WORK

We begin our discussion of related work with an in-depth description of the current de-facto standard called ROUGE and its limitations. We then discuss related work using QA as an

evaluation method for summaries and the gaps in the current body of research. We conclude by detailing how our approach deviates in order to address these gaps.

2.1 Current dominant approach

In a landmark paper [7], Chin-Yew Lin introduced the "ROUGE" package, which provided several ways to automatically evaluate textual summaries. It is based on the idea that a given machine-made summary is good if it has a high similarity to a reference summary provided by skilled human summarizers.

Inspired by BLEU [10], Lin introduced several ways to compute the similarity between two pieces of text. The most commonly used methods rely on 3 simple syntactic concepts. The first is the recall of *n-grams*, which are simply sequences of words appearing next to each other in a provided order (e.g. 'lazy fox' and 'brown dog sleeps', a 2-gram, and a 3-gram respectively). The second metric is the *longest common subsequence*, with the idea that longer subsequences mean more similar texts overall. The last major metric is the recall of *skip-grams*, which is much like regular n-grams but allows other words to appear in-between (e.g. "lazy brown fox" is a 2-skip-gram of "lazy fox", even though brown is not part of the reference text).

Because of its ease of use and high correlation with human judgements, ROUGE quickly became the de-facto standard when evaluating the performance of summarization systems and remains so to this day [1]. Nonetheless, the method has received some criticism. For one, it relies entirely on syntactic overlap of the two pieces of text. This means that using synonyms would result in a lower score, thus unfairly disadvantaging abstractive summaries, which contain significant amounts of paraphrasing. Secondly, while Lin initially suggested using several reference summaries, modern datasets tend to only have a single summary to compare against. This poses a problem since it inherently assumes that the provided reference (or a summary similar to it) is the only good way to summarize a text. Finally, ROUGE fails to take into account the factual content of the two pieces of text. For example, all given metrics would assign a positive score when comparing 'the fox was brown' and the 'the fox was not brown', although their meanings are exact opposites.

Some previous work [9] has attempted to address the problem of synonyms by extending ROUGE with word embeddings, which "compute the semantic similarity of the words used in summaries". While this circumvents the constraints of having exact word matches, it still relies on matching to the provided 'gold reference' summary and is therefore limited by the quality and availability of these reference summaries. Furthermore, it still fails to capture whether the resulting summary is coherent and factual.

2.2 Question Answering for Summary Evaluation

One promising approach that aims to address the shortcomings of simpler statistical approaches is to evaluate summaries using the Question Answering (**QA**) task [2,3,15], particularly the 'Reading comprehension' subset. In this task, the model is asked to answer questions arising from a specific text, rather than general world knowledge. Therefore, using QA models lets researchers focus on measuring how much of the key content has been retained and whether the produced summary is factually correct, while allowing for summaries that do not necessarily overlap with some given reference

Researchers exploring this approach generally propose a framework in which questions for a passage are generated by a Question Generation (**QG**) model, and the summaries evaluated by a separate QA model. Earliest works [2] generate questions uniformly for the whole passage and evaluate how much has been retained, while latter works use either reference summaries [3] or generate question from the source passage in an unsupervised manner [13]. Other works [15] focus not on content selection, but on content factuality.

To assign scores to output summaries, most proposed metrics compare a QA model's accuracy on the same questions given either the full source text or a 'gold reference' summary, and its accuracy when only the machine-made summary is given.

2.3 Deviation from Prior Work

While using QA tasks as an evaluation metric has shown great promise and high correlations with human judgements, we believe there are some important gaps in the research.

First of all, the existing research performs the summary evaluation using questions that are generated by machines. This has the obvious appeal of allowing the evaluation of any given source text and its summary, provided the question generation model is good enough. On the other hand, the QG model adds additional uncertainty to the evaluation process, because it can be unclear which faults arise from which part of the system.

Secondly, since several of the biggest summarization datasets are collected from daily news websites [4,14], a majority of work has been limited to the evaluation of such summaries. Nonetheless, it must be acknowledged that this is just a tiny facet of what summarization may be used for. Since the News domain has its idiosyncrasies, is important to investigate whether the QA-evaluation approach transfers well to different domains.

To address both of these limitations, this paper details a method based on existing reading comprehension datasets with human crafted questions. We focus on datasets containing general knowledge articles and short stories instead of news articles. This should a) allow researchers to have higher confidence that the questions are of a high quality and that drops in performance can be reduced to one of the other two parts, and b) expand the focus of summarization research.

3. APPROACH

We visualize our high-level approach in Figure 1.

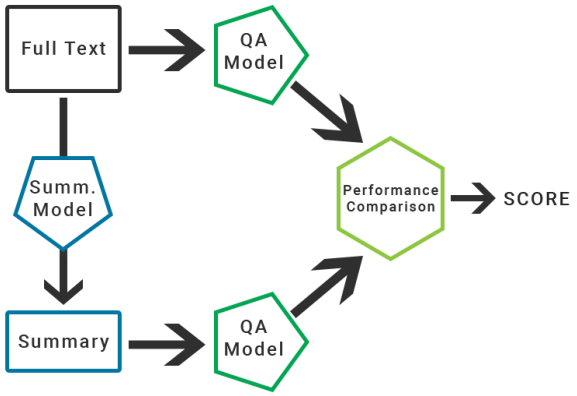

Figure 1 - Overview of the approach

The core idea is to repurpose an existing reading comprehension dataset, and evaluate it in the following two ways:

1) On the original articles, wherein we give the full source text to the QA model to establish the plausible performance ceiling.
2) On the summaries of the articles, wherein we first feed the source texts into a summarization engine, and

then substitute the produced summaries in place of the original articles when evaluating on a QA model.

The two scores are then compared, with the idea that a higher performance drop correlates to higher content losses.

We achieve component independence by exchanging the machine-made summaries in place of source texts using the same formatting as the original dataset. This allows the QA models to use the same interface when evaluating both full texts and summaries, and the process is done in sequence - first producing the entire set of summaries, and then proceeding to evaluation. The score for a summarization engine is then taken as the performance of the QA model over the entire test set.

4. EXPERIMENTAL SETUP

Datasets. We evaluate our summarization models on two different datasets – RACE [5] and SQuAD 2 [12]. The former is constructed from reading comprehension tasks as used in Chinese high school examinations, where each passage has four candidate answers with a single correct choice. The latter is constructed from Wikipedia articles which have human crafted questions that are either answerable by an exact span from the text, or unanswerable from the given passage (both types are 50%). Both of the datasets are interesting since they are outside of the news domain, involve reading comprehension, and have different question types. RACE has more varied types of questions – involving inference, attitude detection, and summarization/paraphrasing. On the other hand, SQuAD 2 has the interesting property of containing a large number of unanswerable questions, and is also based on span selection.

We add examples of questions, original passages, baselines and summaries from RACE and SQuAD 2 in appendix A and appendix B respectively.

Summarization models. For summarization comparison we select 2 different transformer architectures (BART large [6] and T5 [11] in various sizes, all pretrained on the CNN/DM [14] dataset), 2 statistical methods (TextRank [8], a graph based sentence ranking model, and SMMRY [17] – a paid API that utilizes tf-idf to rank sentences). An important point of note is that we do not finetune the transformer models because of computational constraints and thus they are producing summaries on out-of-domain articles. We instead rely on pretrained models distributed by the HuggingFace [16] open model repository, since they provide models that are already finetuned on a different summarization dataset (CNN/DM [14]), however this limits our selection to the aforementioned two architectures.

Baselines. We also construct 6 different baselines for the RACE dataset, and 5 for SQUAD 2.0. The high number of baselines is used to establish not only common comparison points (Lead 3 sentences), but also to investigate whether our evaluation method is sensitive to underlying properties of the summaries (sentence and word level ordering, whether the texts themselves exhibit uneven content distribution etc.). The baseline types differ between the two datasets to better reflect their idiosyncrasies. For one, we swap out Lead 3 sentences for "first 20% of the words" since the Squad dataset is split into passages and are on average much shorter in length (RACE averages 350 tokens per article, whereas SQUAD 2 is less than half that at 144 tokens). We also pass in an unrelated nonsensical passage as a substitute for the "no text" RACE baseline, so as to signal the UnifiedQA model that this is a Squad type question (i.e. that "no answer" can be the desired output).

Evaluation. We evaluate our summaries using a pretrained UnifiedQA question answering model that is based on the more general T5 model. UnifiedQA is trained on a large variety of different question answering datasets, which allows it to transfer knowledge gained from one dataset to another, and is able to answer question from different datasets without any new task specifications. UnifiedQA achieves state-of-the-art performance on multiple datasets, comes in multiple different model sizes, and paired with its generality makes for a strong choice for our evaluation. It is also notable that UnifiedQA has been exposed to context passages of varying lengths, which is important in our particular case. Because of computational constraints, we do no further finetuning on this model, and simply choose to use the

Table 1 - Results for summaries on RACE dataset using the 3B version of UnifiedQA as the Answering model. We report the accuracy over all questions (each question has the same weight). We also report normalized scores, wherein the full text performance is held to be the ceiling (100%) and all others are percentages in respect to it. Best results in each column (aside from the full text) and the name of the overall best performing model are marked in bold.

Category	Model	Word count Average (± std. deviation)	Accuracies				
			Full Dataset	Full Dataset Norm.	Summary Questions	Inference Questions	Attitude Questions
-	Original full text	349.16 (±94.86)	81.62	100.00	83.28	71.53	91.84
Transformers	T5 Small	70.76 (±15.83)	64.32	78.80	71.92	63.19	**85.71**
	T5 base	70.41 (±14.29)	64.89	79.50	**78.86**	57.64	73.47
	T5 large	68.73 (±14.58)	65.49	80.24	77.92	61.81	81.63
	BART large	68.49 (±19.99)	**65.72**	**80.52**	78.23	**66.67**	79.59
Statistical	SMMRY	69.49 (±31.90)	64.27	78.74	72.87	60.42	75.51
	Textrank	69.32 (±30.67)	63.95	78.35	75.39	63.89	**85.71**
Baselines	Lead 3 sentences	71.47 (±44.89)	63.69	78.03	73.50	63.19	77.55
	Random 3 sentences (ordered)	69.21 (±44.09)	63.32	77.58	71.50	65.74	80.27
	Random 3 sentences (shuffled)	69.33 (±42.95)	63.90	78.29	71.30	61.11	83.67
	Random 20% of words (Ordered)	70.26 (±19.07)	59.46	72.85	73.50	60.42	71.43
	Random 20% of words (shuffled)	70.29 (±19.09)	57.40	70.33	68.45	61.81	67.35
	No text	0	52.86	64.76	59.62	54.17	69.39
Question Counts			3498		317	144	49

Table 2 - Results for summaries on SQUAD 2.0 dataset using the 3B version of UnifiedQA as the Answering model. The exact columns are a percentage of question answers that were answered with the exact string, the f1 scores are calculated on a per word basis and then averaged over the entire dataset. We also report normalized scores, wherein the full text performance is held to be the ceiling (100%) and all others are percentages in respect to it. Best results in each column (aside from the full text) and the name of the overall best performing model are marked in bold.

Category	Model	Word count Average (± std. deviation)	Accuracies						
			Total exact	Total f1	Total f1 Norm.	HasAns exact	HasAns f1	NoAns exact	NoAns f1
-	Original full text	144.66 (±65.76)	84.43	87.65	100	85.7	92.14	83.16	83.16
Transformers	T5 small	29.45 (±15.41)	53.49	56.17	64.08	24.76	30.12	82.14	82.14
	T5 base	29.12 (±13.06)	54.24	56.91	64.93	**26.43**	**31.79**	81.97	81.97
	T5 large	28.69 (±12.86)	53.2	55.79	63.65	24.75	29.93	81.58	81.58
	BART large	28.44 (±13.74)	**56.45**	**58.47**	**66.71**	25.59	29.64	**87.22**	**87.22**
Statistical	Textrank	32.13 (±16.26)	55.58	57.46	65.56	24.71	28.47	86.36	86.36
Baselines	Lead 20% of words	28.54 (±13.16)	55.66	57.66	65.78	24.58	28.57	86.66	86.66
	Lead 20% of words (shuffled text, ordered)	28.53 (±13.16)	52.14	54.26	61.91	19.32	23.55	84.88	84.88
	Lead 20% of words (shuffled text, unordered)	28.53 (±13.16)	52.39	54.57	62.26	20.02	24.4	84.66	84.66
	Random 20% of words	28.54 (±13.17)	35.75	39.49	45.05	6.21	13.69	65.21	65.21
	Unrelated article (no text)	30.00 (±0.00)	49.31	49.46	56.43	0.51	0.82	97.96	97.96
Question counts			8862			4532		4332	

general 3B variant. When producing answers, we set the model temperature parameter to 0 (taking the most likely outputs, rather than sampling according to a predicted distribution), which is extremely important for attaining stable outputs and scores.

5. RESULTS

RACE. Our experimental results on RACE are summarized in Table 1. Transformer based summarization models consistently outperform all baselines and statistical methods, albeit by a relatively low margin. The performance of T5 models grows with model size as expected, and the score difference between small and large is comparable to their ROUGE differences on CNN/DM dataset (1.17 and 1.38 percentage point differences on this work and CNN/DM respectively), with the caveat that these might not be linearly comparable metrics. BART and T5-large score similarly, mirroring their close performance on ROUGE evaluations. Some surprising results arise from our baselines, which are discussed more in depth in the next section.

We further seek to perform a finer-grained analysis on RACE by subdividing questions into categories. Since the original dataset does not provide explicit question type labels, we rely on simple heuristics. For example, the summarization type questions are tagged by matching phrases "best title", "main idea", "mainly about". We find that the Transformer models perform exceedingly well on these types of question, with the best model falling fewer than 5 percentage points below the performance on full passages. We showcase two other question subtypes, however the results have a lot of variance because of the low question counts and are thus deemed unreliable.

Squad 2.0 Our results on the Squad 2.0 dataset are summarized in table 2. Transformer models once again outperform all other chosen methods by about 1-2 percentage points. It seems apparent that the formulation of this dataset makes it much more resilient to "nonsensical" baselines, such as those sampling random words, or providing unrelated text. This is likely due to the fact that answer must be text spans coming from the article, rather than explicit options that can be guessed. This increases the gap not only between the nonsensical baselines and guided summarization models, but also between the original performance and all other measurements. We find that BART has the overall best performance, with the T5 outperforming it in questions that have answers but dropping its performance during unanswerable questions. In fact, both T5 (all sizes) and the random words baselines perform worse on unanswerable questions than full passages, thereby seemingly confusing the QA model into believing some of these questions are in fact answerable. This is similar to the "Random 20% of words" baseline, wherein the QA model guessed that there was no answer less frequently than if it had the full passage.

Correlation with ROUGE. We investigate the hypothesis that model scores as evaluated by our method are correlated with their ROUGE scores (albeit on a different dataset). To do this, we independently calculate ROUGE scores for a subset of methods on the CNN/DM dataset, the results of which are shown together with our scores on RACE and SQuAD 2 in Table 3.

Table 3 - ROUGE scores on CNN/DM compared to our scores on RACE and SQuAD2 for a subset of our evaluation models. In particular, we omit all of the random based baselines, the no text/random article baselines, and SMMRY.

Model	ROUGE-1	ROUGE-2	RACE	SQuAD2
T5 Small	38.96	17.23	64.32	56.17
T5 Base	40.1	17.78	64.89	56.91
T5 Large	41.59	18.85	65.49	55.79
BART large	44.16	21.28	65.72	58.47
TextRank	40.2	17.56	63.95	57.46
Lead 3	40.05	17.48	63.69	-
Lead 20%	36.51	17.21	-	57.66

We showcase Pearson Correlations between the measurements in table 4. From this we can conclude that QA as an evaluation metric is highly correlated with ROUGE measurements when performed on the RACE dataset, and has a low to medium correlation when performed on SQuAD 2. In both cases, ROUGE-2 has a higher correlation than ROUGE-1, indicating that correct word sequences are more important than raw words. As a point of note, these measurements might be thrown off by the counterintuitive behavior of T5 in our SQuAD 2 measurements (namely, that the performance peaks at the "base" size, and the large model has lower overall performance than the

small model). Further investigation with more models is thus required before drawing strong conclusions.

Table 4 - Pearson correlations between ROUGE and our measurements.

	RACE	SQuAD2
ROUGE-1	0.771072	0.213752
ROUGE-2	0.814147	0.459183

6. DISCUSSION

The unreasonable effectiveness of empty passages (in MCQ datasets). As evidenced by our evaluation on RACE, Multiple Choice Questionnaires provide large surface areas for educated guessing – not only do simple baselines perform well (Lead 3, Random 3), but also ones that provide almost no information to humans, such as random words and "No text". This suggest that a) the QA model is sometimes picking up on the overall word usage patterns, rather than meaningful sequences (as suggested by random shuffled words), and b) that it has gotten exceptionally good at guessing what choice is likely even while entirely missing the context of the question. For this reason, open span selection datasets seem to be a better choice provided all other parts stay constant.

"No Answer" options are desirable, as evidenced by our results on SQUAD 2.0. Having the "no answer" option allows us to more clearly see which information actually went missing, whereas the RACE model still takes a short at guessing 1 of the 4 options. Further research could look into augmenting MCQ with "no answer" or "none of the above" options, thereby extending its resilience to nonsensical guessing.

Low sensitivity to sentence order, as evidenced by our "shuffled" sentences baselines. They perform extremely closely to their ordered counterparts, thus indicating that additional metrics are needed to assess the fluency and overall flow of the summaries.

Summarization type questions[1] could be an interesting direction of further research, as the gap between Transformer produced summaries and the Original text are a lot lower than the full dataset, while the gap between baselines and large transformers increases. Further work could look into constructing such questions using semi-supervised means.

7. CONCLUSION

Our work shows that the latest Transformer models generally outperform simpler statistical methods and various baselines when using QA as an evaluation mechanism, albeit by a relatively low margin.

Furthermore, the scores produced by our method on the RACE dataset have a high correlation with the models' ROUGE scores on the CNN/DM dataset (this being the dataset that the transformer models were trained on), and a low one when compared to our evaluations on SQuAD 2. This implies that the ROUGE method inherently captures at least some of the factual details that are being asked about in RACE, whereas SQuAD might not necessarily focus on the same parts.

Finally, we show that further additions to the method are needed to properly evaluate sentence and word level ordering, since the gaps between shuffled and ordered texts is narrow to non-existing.

[1] such as those in RACE

8. ACKNOWLEDGEMENTS

I would like to thank Christin Seifert for her invaluable advice and insight as a thesis supervisor that helped shape this research, and for the numerous detailed reviews she gave while I was writing the paper. Additional thanks to Daniel Khashabi, one of the original UnifiedQA authors, for answering my questions on how to use their work.

9. REFERENCES

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. *arXiv:1707.02268 [cs]* 84, (July 2017). Retrieved April 23, 2020 from http://arxiv.org/abs/1707.02268

[2] Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2018. A Semantic QA-Based Approach for Text Summarization Evaluation. *arXiv:1704.06259 [cs]* (April 2018). Retrieved April 27, 2020 from http://arxiv.org/abs/1704.06259

[3] Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question Answering as an Automatic Evaluation Metric for News Article Summarization. *arXiv:1906.00318 [cs]* (June 2019). Retrieved April 26, 2020 from http://arxiv.org/abs/1906.00318

[4] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 708–719. DOI:https://doi.org/10.18653/v1/N18-1065

[5] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, 785–794. DOI:https://doi.org/10.18653/v1/D17-1082

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]* 24, (October 2019). Retrieved April 23, 2020 from http://arxiv.org/abs/1910.13461

[7] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, 74–81. Retrieved April 23, 2020 from https://www.aclweb.org/anthology/W04-1013

[8] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Barcelona, Spain, 404–411. Retrieved April 27, 2020 from https://www.aclweb.org/anthology/W04-3252

[9] Jun-Ping Ng and Viktoria Abrecht. 2015. Better Summarization Evaluation with Word Embeddings for

ROUGE. *arXiv:1508.06034 [cs]* 18, (August 2015). Retrieved April 23, 2020 from http://arxiv.org/abs/1508.06034

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. DOI:https://doi.org/10.3115/1073083.1073135

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]* (October 2019). Retrieved June 20, 2020 from http://arxiv.org/abs/1910.10683

[12] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Melbourne, Australia, 784–789. DOI:https://doi.org/10.18653/v1/P18-2124

[13] Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. *arXiv:1909.01610 [cs]* (September 2019). Retrieved May 3, 2020 from http://arxiv.org/abs/1909.01610

[14] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *arXiv:1704.04368 [cs]* (April 2017). Retrieved April 27, 2020 from http://arxiv.org/abs/1704.04368

[15] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. *arXiv:2004.04228 [cs]* (April 2020). Retrieved April 26, 2020 from http://arxiv.org/abs/2004.04228

[16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]* 71, (February 2020). Retrieved April 19, 2020 from http://arxiv.org/abs/1910.03771

[17] SMMRY - About. Retrieved June 20, 2020 from https://smmry.com/about

APPENDIX.

A. EXAMPLES FROM RACE DATASET

Questions: (correct answers are marked in bold)

Q1: according to this article, why are you more likely to see the northern lights in the winter? (a) they move from rural areas to cities. (b) the weather is more stable at that time. (c) they're a special feature of many festivals. **(d) the sky is dark for longer periods then**.

Q2: according to the article, how do visitors to iceland have fun in the winter? (a) they learn how to ski down the mountains. (b) they photograph famous historic sites. **(c) they explore the countryside in well made vehicles.** (d) they spend a week at one of the seaside resorts.

Q3: what is true about the volcanoes of iceland? **(a) the majority of them are quiet.** (b) their age hasn't been determined. (c) all but one of them are extinct. (d) citizens aren't affected by them.

Q4: what does this article explain? (a) some of iceland's urban cultural attractions. **(b) a way to reduce the cost of a trip to iceland.** (c) reasons for visiting iceland in june and july. (d) the average price for a short tour of iceland.

Original Article

iceland, an island just south of the arctic circle, has fairly mild winters, thanks to warm ocean currents. time your vacation here during the winter months to take advantage of off-season deals. off-season means good deals on flights, hotels and tours. you may also find that the locals are a bit friendlier and more welcoming when tourists aren't arriving in crowds. in the winter months, there are less than seven hours of daylight; thus, chances are good you'll catch sight of the northern lights. sunsets are also beautiful at this time, making for some great photo opportunities. in iceland, winter is the perfect time to hike glaciers, go ice climbing, explore caves made out of hardened lava and much more. one of the most popular activities is off-roading in a specially-equipped "super jeep". before booking your trip, be sure to check for volcano alerts. there are about 130 volcanoes on or around iceland. thirty-five of them are active. in 2010, a volcano named eyjaallajokull exploded, sending clouds of ash up to four kilometers into the atmosphere. the ash drifted toward the uk and europe. because the ask could damage aircraft engines, airlines operating in the region were forced to cancel flights for six days. as a result, thousands of people were stuck in airports. recently, another volcano named bardarbunga has become active, erupting ash into the air. such events, if large enough, could prevent your trip from going ahead. so check the latest volcano news prior to making your reservations.

BART

iceland has fairly mild winters, thanks to warm ocean currents. time your vacation during the winter months to take advantage of off-season deals. there are about 130 volcanoes on or around iceland, and 35 are active. in 2010, a volcano named eyjaallajokull

T5 large

iceland has mild winters, thanks to warm ocean currents . off-season means good deals on flights, hotels and tours . in the winter months, there are less than seven hours of daylight; see the northern lights . before booking your trip, be sure to check for volcano alerts. there are about 130 volcanoes on or around iceland. thirty-five of them are active .

SMMRY

time your vacation here during the winter months to take advantage of off-season deals . in iceland, winter is the perfect time to hike glaciers, go ice climbing, explore caves made out of hardened lava and much more . recently

TextRank

time your vacation here during the winter months to take advantage of off-season deals. there are about 130 volcanoes on or around iceland. recently, another volcano named bardarbunga has become active, erupting ash into the air.

Lead 3

iceland, an island just south of the arctic circle, has fairly mild winters, thanks to warm ocean currents. time your vacation here during the winter months to take advantage of off-season deals. off-season means good deals on flights, hotels and tours.

Random3 jumbled

before booking your trip, be sure to check for volcano alerts. the ash drifted toward the uk and europe. time your vacation here during the winter months to take advantage of off-season deals.

Random words ordered

island arctic, fairly mild the to . may that more in months are seven hours at time for iceland, ice made of hardened lava much . one in"your sure volcano alerts . thirty-five are exploded of . the toward europe . damage the flights . airports another become, such, could prevent making

B. EXAMPLES FROM SQUAD 2. DATASET

Questions (subset):
Q1: "Which name is also used to describe the Amazon rainforest in English?"
Answers: [“also known in English as Amazonia or the Amazon Jungle”, "Amazonia or the Amazon Jungle", “Amazonia”]

Original Article

"The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America. This basin

encompasses 7,000,000 square kilometres (2,700,000 sq mi), of which 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Venezuela, Ecuador, Bolivia, Guyana, Suriname and French Guiana. States or departments in four nations contain \"Amazonas\" in their names. The Amazon represents over half of the planet's remaining rainforests, and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species."
BART "The Amazon rainforest is a moist broadleaf forest that covers most of the Amazon basin of South America. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the"
T5 base "the Amazon rainforest is a moist broadleaf forest that covers most of the Amazon basin of South America . the majority"
T5 Large "the amazon rainforest is a moist broadleaf forest that covers most of the amazon basin of south america . states or departments in four nations contain \"Amazonas\" in their names . the amazon represents over half of the planet"
TextRank "The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America."
Lead 20% of words "The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the"
Random words "and remaining over minor by Amazônica nine as rainforest, :, the Brazil``encompasses Guyana Amazônia Amazonia Forêt of% is or sq Jungle Guiana This Amazon covers their that rainforest Venezuela: Colombia square majority 2,700,000"