NLQuAD: A Non-Factoid Long Question Answering Data Set

Soleimani, A.; Monz, C.; Worring, M.

DOI
10.18653/v1/2021.eacl-main.106

Publication date
2021

Document Version
Final published version

Published in
The 16th Conference of the European Chapter of the Association for Computational Linguistics

NLQuAD: A Non-Factoid Long Question Answering Data Set

Amir Soleimani
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
a.soleimani@uva.nl

Christof Monz
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
c.monz@uva.nl

Marcel Worring
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

Abstract

We introduce NLQuAD, the first data set with baseline methods for non-factoid long question answering, a task requiring document-level language understanding. In contrast to existing span detection question answering data sets, NLQuAD has non-factoid questions that are not answerable by a short span of text and demanding multiple-sentence descriptive answers and opinions. We show the limitation of the F1 score for evaluation of long answers and introduce Intersection over Union (IoU), which measures position-sensitive overlap between the predicted and the target answer spans. To establish baseline performances, we compare BERT, RoBERTa, and Longformer models. Experimental results and human evaluations show that Longformer outperforms the other architectures, but results are still far behind a human upper bound, leaving substantial room for improvements. NLQuAD's samples exceed the input limitation of most pre-trained Transformer-based models, encouraging future research on long sequence language models.[1]

1 Introduction

Over the last few years, there have been remarkable improvements in the area of Machine Reading Comprehension (MRC) and open-domain Question Answering (QA) due to the availability of large scale data sets such as SQuAD (Rajpurkar et al., 2016) and pre-trained language models such as BERT (Devlin et al., 2018). Although non-factoid questions represent a large number of real-life questions, current QA data sets barely cover this area. The reason is that context passages in existing QA data sets are mostly very short and questions mostly factoid, i.e., can be answered by simple facts or entities such as a person name and location (Jurafsky and Martin, 2019). Little attention has been paid to non-factoid and open-ended questions that require complex answers such as descriptions or opinions (Hashemi et al., 2020). Answers to non-factoid questions extend to multiple sentences or paragraphs having few words overlapping with the question (Cohen and Croft, 2016). Non-factoid QA facilitates document assistance systems, where for example, journalists can seek assistance to highlight relevant opinions and interpretations. It can further motivate more research on long sequence

Question: How are people coping in the lockdown?

Headline: China coronavirus: Death toll rises as more cities restrict travel

Document: China has widened its travel restrictions in Hubei province - the centre of the coronavirus outbreak - as the death toll climbed to 26. The restrictions will affect at least 20 million people across 10 cities, including the capital, Wuhan, where the virus emerged. On Thursday, a coronavirus patient died in northern Hebei province - making it the first death outside Hubei. [...] We now know this is not a virus that will burn out on its own and disappear. [...] And we still don't know when people are contagious. Is it before symptoms appear, or only after severe symptoms emerge? One is significantly harder to stop spreading than the other. [...] **One doctor, who requested anonymity, describes the conditions at a hospital in Wuhan. [...] "I was planning to stay in my apartment because I'm scared to go to the gym, and I'm scared to go to out in public, and not many people are willing to go out." (141 words).** Vietnam and Singapore were on Thursday added to the nations recording confirmed cases, joining Thailand, the US, Taiwan and South Korea. [...] Taiwan has banned people arriving from Wuhan and the US state department warned American travellers to exercise increased caution in China. (document length: 921 words)

Figure 1: A question-answer pair in NLQuAD. QA models must predict the answer span within the context document. The correct answer span is bolded. We extract questions and answers, respectively, from the sub-headings and the sub-section bodies from real-word English news articles. Two other questions based on the same article: Can the Coronavirus be stopped? What's the global situation?

[1] Dataset and Models: github.com/asoleimanib/NLQuAD

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1245–1255
April 19 - 23, 2021.

language models. Therefore, a high-quality data set in this area is clearly desired.

To support research towards non-factoid and long QA tasks and to address the existing shortcomings as identified above, we have built NLQuAD, a non-factoid long question answering data set. NLQuAD contains 31k non-factoid questions and long answers collected from 13k BBC news articles. We extract questions and answers from the articles' sub-headings and the following body paragraphs of the sub-headings (see Figure 1).

Questions in NLQuAD are not answerable by a short span of text within the documents. This is in contrast to existing long-context but factoid QA data sets such as NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), NarrativeQA (Kočiský et al., 2018), DuoRC (Saha et al., 2018), HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019). Although these data sets contain long documents, questions are answerable by short entities or a span of entities.

In particular, Natural Questions covers two types of short and long answers. However, due to its factoid questions, most long answers are still sections containing exactly the short answers and so are trivial (e.g., "Where is the world's largest ice sheet...?", Short: "Antarctica"; Long: "The Antarctic ice sheet is the largest single mass of ice on Earth..."). Furthermore, although a small portion (13%) of Natural Questions samples have only long answers, they are still spans of simple facts. For example, "Who is the author of the book Arabian Nights?" has no short answer simply because there are multiple authors: "The work was collected over many centuries by various authors, translators...". In contrast, we address non-factoid questions requiring complex answers like opinions and explanations. NLQuAD's answers are open and not predefined. Figure 3 and Table 3 present our question types. NLQuAD's questions are also not self-contained. For example, "How are people coping in the lockdown?" or "What's the global situation?" cannot be answered without the context from the document (see Figure 1). Section 3.2 discusses our question types in detail.

In most existing QA data sets such as SQuAD, crowd-workers generate questions based on provided short passages and extract answers from the passages (Rajpurkar et al., 2016). This method of question generation can make QA samples trivial because models can simply detect the most related span to the question by guessing based on shallow pattern matching (Kočiský et al., 2018). In contrast, all annotations in NLQuAD are done automatically and directly based on the news articles themselves. NLQuAD, unlike MS MARCO (Bajaj et al., 2016) and ELI5 (Fan et al., 2019), does not use information retrieval (IR) methods to collect supporting documents. Retrieved documents in these data sets are not guaranteed to contain all facts required to answer the question or they occasionally just contain information related to the question but no answers.

NLQuAD requires document-level language understanding. With an average document length and answer length of 877 and 175 words, respectively, it exceeds the maximum input length of the state of the art QA models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) due to their memory and computational requirements. Thus, training and evaluating the (document, question, answer) tuples is impossible using such models in an end-to-end manner. It is worth noting that it is also harder to perform pre-selection methods before the final span detection because our answers are long. Meanwhile, most of our questions are not self-contained. For example, to answer the question "How are people coping in the lockdown?" (Figure 1), the system needs to read the document to interpret the concept of "lockdown" and then locate the information regarding the people's behaviour.

We also show the shortcomings of the F1 score and ROUGE-N scores in evaluating long sequences. There is a higher chance of overlap between the word N-grams in two long sequences causing F1 and ROUGE-N to over-estimate the performance. Therefore, we propose to use Intersection over Union (IoU) measuring position-sensitive overlap between two spans.

In summary, our contributions are as follows: (1) We introduce a new data set for non-factoid long QA that to the best of our knowledge is the first data set requiring long answer span detection given non-self-contained and non-factoid questions; (2) We show the limitations of the F1 score in evaluating long answers and propose a new evaluation metric; (3) To establish baseline results, we experiment with three state-of-the-art models: BERT, RoBERTa, and Longformer, and compare them with human performance. To handle the input length limitations of BERT and RoBERTa, we pro-

data sets	Avg # Words Que.	Doc.	Ans.	QA Type	Samples
SQuAD	10	117	3	Factoid Span Detection	150k
NewsQA	8	616	4	Factoid Span Detection	100k
TriviaQA	14	2895	2	Factoid Span Detection	95k
NarrativeQA	10	656	5	Factoid Span Detection	47k
DouRC-Self	9	591	3	Factoid Span Detection	186k
DouRC-Pharaphrase	9	1240	3	Factoid Span Detection	186k
HotpotQA	18	917	2	Factoid Span Detection	113k
Natural Questions	9	7360	192	Factoid Span Detection	307k
DuReader	5	396	67	Factoid & Non-Factoid Span Detection	200k
DQA	N/A	N/A	54	Factoid & Non-Factoid Span Detection	17k
MS MARCO	6	56	14	Answer Generation	183k
ELI5	42	858	131	Answer Generation	272k
NLQuAD	7	877	175	Non-Factoid Span Detection	31k

Table 1: Comparison of NLQuAD with SQuAD, MS MARCO, and long-context QA data sets.

pose to train these models in a sliding-window approach; (4) We finally show that the state-of-the-art models have limited performance in the non-factoid long QA task.

2 Existing data sets

Existing large-scale QA data sets can be categorized based on their context passage length in two groups: short-context QA, i.e., data sets with paragraph-level context, and long-context QA, i.e., data sets with multiple-paragraph or document-level context. Long-context QA can potentially include questions demanding long answers. In this section, we only review QA datasets. However, it is worth noting that very recently, (Tay et al., 2020a) introduced a unified benchmark using different tasks for evaluating model quality under long-context scenarios.

2.1 Short-Context Question Answering

SQuAD (Rajpurkar et al., 2016) is a factoid span detection data set with short answers. Crowd-workers generated the questions given a set of articles. DROP (Dua et al., 2019) makes the problem more challenging by adversarially-created questions requiring discrete reasoning over the text. SQuAD and DROP use Wikipedia pages as context passages whereas SearchQA (Dunn et al., 2017) uses IR approaches to collect context passages.

Answer generation based on a set of passages is another approach to address this task. MS MARCO (Bajaj et al., 2016) consists of real-world search queries and retrieved documents corresponding to the queries.

There are also different types of QA data sets such as Antique (Hashemi et al., 2020), which is a data set for answer retrieval for non-factoid questions. There is also a range of multiple-choice QA tasks such as RACE (Lai et al., 2017), ARC (Clark et al., 2018), SWAQ (Zellers et al., 2018), and COSMOS QA (Huang et al., 2019) that are clustered together with the short-context QA data sets.

2.2 Long-Context Question Answering

Factoid QA has been applied to longer documents, however, the nature of factoid questions limits answers to short texts. NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), NarrativeQA (Kočiský et al., 2018), and DuoRC (Saha et al., 2018) fall into this category and their documents are extracted from news articles, stories, and movie plots, respectively. On the other hand, DQA (ter Hoeve et al., 2020) is a document-centred QA data set aimed at document assistance systems. Along with Yes/No questions, it also includes non-factoid questions with relatively long answers. However, the questions are generated by crowd-workers based on a small set of documents.

DuReader (He et al., 2018) consists of real-word Chinese queries and corresponding retrieved documents. It contains both factoid and non-factoid (40%) questions and consequently has longer average answer length than pure factoid datasets.

The multi-hop QA task, requiring multi-hop reasoning over multiple paragraphs, can also be considered as long-context QA if models process paragraphs together. HotpotQA (Yang et al., 2018) is a multi-hop data set, but the answer length of its factoid questions is as limited as that of short-context QA data sets.

Natural Questions (Kwiatkowski et al., 2019) is a factoid QA task with much longer documents and two types of answer lengths. It consists of factoid questions, retrieved Wikipedia pages, and short

Number of QA pairs	31k
Number of Documents	13k
Number of Unique Questions	24k
Avg. Document Length (Word)	876.8
Avg. Answer Length (Word)	174.6
Avg. Question Length (Word)	7.0
Avg. Document Length (Sentence)	38.7
Avg. Answer Length (Sentence)	7.5
Avg. Question Length (Sentence)	1.0
Avg. Question per Document	2.4

Table 2: NLQuAD: data set statistics.

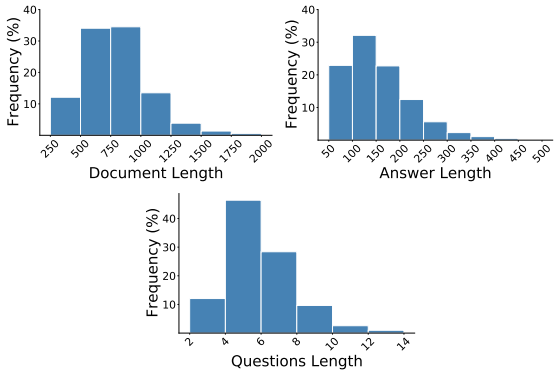

Figure 2: Distribution of the number of words in document, question and answer.

answers (yes/no, entities) as well as long answers (bounding boxes with the information to infer the answer). However, due to the nature of factoid questions, the majority of long answers are sections containing exactly the short answer or simple facts.

ELI5 (Fan et al., 2019) consists of real-world questions with answers provided by the Reddit community. The task is to generate answers given a set of documents retrieved from the Web. However, the documents are not guaranteed to completely address the questions. Furthermore, evaluation metrics for sequence generation tasks such as the ROUGE score (Lin and Och, 2004) are far from perfect to assess the quality of generated answers.

Table 1 compares existing long-context question answering data sets along with SQuAD and MS MARCO. We report the average length for data sets with different types of answers.

3 Data Set Design

NLQuAD consists of news articles as context documents, interrogative sub-headings in the articles as questions, and body paragraphs corresponding to the sub-headings as contiguous answers to the questions. We automatically extract target answers because annotating for non-factoid long QA is rather challenging and costly. To ensure the quality of answers in addition to the initial investigations, we perform human evaluations (Section 5.3). We choose the BBC news website as the resource of our documents and the question-answer pairs, mainly because its articles contain a considerable amount of high-quality question-like sub-headings which are suitable for the QA task.

NLQuAD's characteristics make it an appealing and challenging data set for the non-factoid long QA task: Its context documents are long, and its questions are non-factoid in a way that cannot be answered by single or multiple entities. The questions are addressed by more than seven sentences on average. Meanwhile, it covers a wide range of topics, making it an open-domain QA data set.

The BBC news articles typically follow a specific template. They begin with an introductory section consisting of news summaries (Narayan et al., 2018) and one or more sections accompanied by sub-headings. Each section contains multiple short to medium-length paragraphs. We remove the template and section break-lines to prevent revealing possible answer boundaries.

3.1 Data Curation

We exploit Wayback Machine,[2] a digital archive of the Web, and Wayback Machine Scraper[3] to scrape the article archives. Links in the scraped pages are used to collect additional pages from the original website. We scraped the English BBC news website from 2016 to 2020 as a limited number of questions can be found in articles before 2016. Only textual information is kept and we strip away multimedia objects and hyperlinks outside of the body of the articles. Duplicate documents are removed and questions with bullet list answer types are discarded. We detect interrogative sub-headings by checking if they end with a question mark.

3.2 Data Set Statistics

NLQuAD contains 31k non-factoid questions based on 13k supporting documents from news articles. Table 2 shows the data set statistics. We randomly partition the data set into training (80%), development (10%), and evaluation (10%) sets.

While NLQuAD has long documents and long-answer QA pairs, the histograms in Figure 2 indicate the wide range of samples. Figure 3 presents a visualisation of the distribution of question types

[2] archive.org/web

[3] github.com/sangaline/wayback-machine-scraper

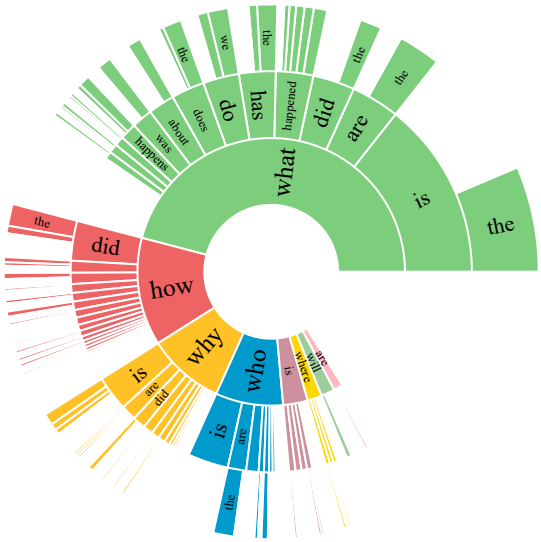

Figure 3: Distribution of trigram prefixes of questions in NLQuAD. Empty portions indicate suffixes with small percentages. NLQuAD covers a wide range of non-factoid question types.

What	How	Why
is the background...	did the attack...	is the US...
is the latest...	did we get...	is this happening...
is the reaction...	did it come...	is there a...
is happening in...	does the US/UK...	are there protests...
is in the...	does it work...	are the fires...
are the allegations...	has the government...	did the US...
did the court...	have the authorities...	was the vote...
happened in the...	do you know...	does this matter...
has the reaction...	is the shutdown...	has the US...
do we know...	many people are...	do not we...

Table 3: Top 4-grams prefixes of questions in NLQuAD. Even 'What' questions are non-factoid and need longer answers (descriptions or opinions)

in terms of their first three tokens. Table 3 also lists high frequency examples of "what", "how" and "why" questions. NLQuAD has a large percentage of "how" and "why" question types where also the "what" examples are non-factoid and consequently require longer explanations as answers.

We manually investigated 100 randomly sampled question-answer pairs from the NLQuAD training set and find that 87% of the questions are not self-contained and require additional contextual information to be understood or disambiguated. Most of the answers consist of explanations, descriptions, or opinions, and only 2% of the questions can be answered by a short span of text.

4 Baseline Models

To investigate the difficulty level of NLQuAD for state-of-the-art QA systems and to establish baseline results, we evaluate the performance of BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and Longformer (Beltagy et al., 2020). Longformer is a scalable model for processing long documents and has been used for long sequences such as document classification (Beltagy et al., 2020) and document re-ranking (Sekulić et al., 2020). We refer readers to Tay et al. (2020b) for a detailed survey on efficient transformers. We train these Transformer-based (Vaswani et al., 2017) models to predict the span of the answer in a context document given a question and document.

4.1 BERT and RoBERTa

The BERT QA model concatenates question and document pairs into a single sequence and predicts the answer span by a dot product between the final hidden vectors, a start vector and an end vector (Devlin et al., 2018). Due to the memory and computational requirements, BERT can encode sequences with a maximum length of 512 tokens that is less than the average sample length in NLQuAD. Therefore, we adopt a sliding window approach. We split the samples into segments using a sliding window of 512 tokens and a stride of 128 tokens. Each segment is augmented with its corresponding question. The segments can include no answer, a portion of the answer, or the entire answer. We train BERT on the segments independently. Finally, the predicted spans corresponding to a single sample are aggregated to predict the final span that is the span between the earliest start position and the latest end position. The output is considered empty when all segments have empty spans.

RoBERTa has the same model architecture and input length limitation as BERT but with a robustly optimized pre-training scheme allowing it to generalize better to downstream tasks such as QA (Liu et al., 2019). We apply the same sliding window approach for RoBERTa.

4.2 Longformer

In order to process the question and entire documents at the same time, we use the Longformer model. It employs an attention mechanism scaling linearly with the sequence length which enables Longformer to process up to 4,096 tokens. It uses multiple attention heads with different dilation configurations to attend to the entire sequence and includes global attention to question tokens in the sequence. Question and document pairs are packed together into a single sequence without having to

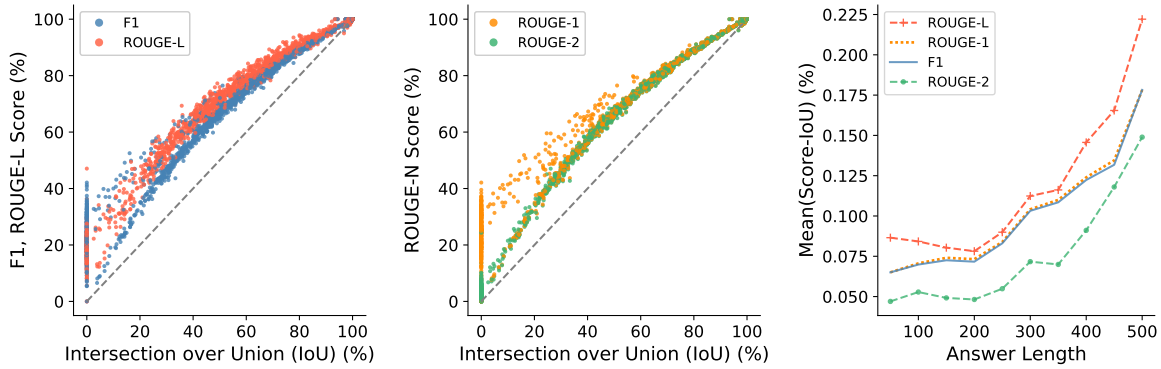

Figure 4: Comparing F1, ROUGE-N and IoU. Left/Middle: All scores behave similarly in the higher values, but F1 and ROUGE-N over-estimate the performance in the lower IoU values due to a higher chance of overlap between the bag of words, n-grams, or longer LCSs in the prediction and target spans. The dashed line shows $y = x$. Right: F1 and ROUGE-N over-estimate more in samples with longer answers. Results are plotted for the Longformer on the development set.

use sliding windows and the answer span is calculated by a dot product (Beltagy et al., 2020).

5 Experiments

5.1 Evaluation Metrics

Exact Match (EM) and the macro-averaged F1 score are the two main evaluation metrics in the span detection QA task (Rajpurkar et al., 2016). Exact Match determines if the prediction exactly matches the target which can be a too strict criterion for long answers. The F1 score measures the overlap between the words in the prediction and the target. It treats sequences as a bag of words. Unfortunately, in long answers, it is highly likely that a random, long span shares a considerable number of tokens with the target span.

The ROUGE-N scores (Lin and Och, 2004), which are primarily used for sequence generation evaluation, have the same drawback in long sequences. ROUGE-N measures the N-gram overlap between the prediction and target. High chances of overlap of unigrams and bigrams in long sequences cause ROUGE-1 and ROUGE-2 to over-estimate performance. The same holds for ROUGE-L with the Longest Common Sub-sequence (LCS) because of a high chance of longer LCSs between two long sequences.

To better take sequence similarities into account, we propose to evaluate models with the Intersection over Union (IoU) score, also known as Jaccard Index. IoU is defined as follows:

$$IoU = \frac{|p \cap t|}{|p \cup t|}$$

Question: How did we get here?

Headline: Eta disarms: French police find 3.5 tonnes of weapons

Target Answer: Slowly, and with many false starts. Eta used parts of south-western France as a base, even though most of its operations were against Spanish targets in Spain. The group has, however, killed some French policemen, but mostly during police raids on members of the group. Etaś first ceasefire was in 1998, but collapsed the following year. A similar declaration in 2006 only lasted a matter of months, ending when Eta bombed an airport car park, killing two people. Four years later, in 2010, Eta announced it would not carry out further attacks and in January 2011, it declared a permanent and "internationally verifiable" ceasefire but refused to disarm. In recent years, police in France and Spain have arrested hundreds of Eta figures and seized many of its weapons. Etaś political wing, Herri Batasuna, was banned by the Spanish government, which argued that the two groups were inextricably linked.

Prediction: The group was set up more than 50 years ago in the era of Spanish dictator General Franco, who repressed the Basques politically and culturally. Eta's goal was to create an independent Basque state out of territory in south-west France and northern Spain. Its first known killing was in 1968, when a secret police chief was shot dead in the Basque city of San Sebastian. France and Spain refuse to negotiate with Eta, which is on the EU blacklist of terrorist organisations.

Figure 5: A prediction span that is semantically different from the target span but has a F1=30% (Prec.=43%, Rec.=23%) and IoU=0. Red shows the overlapping words in the prediction span with the target. Articles (a, an, the) and punctuations are discarded before overlapping calculation. (ROUGE-1=32%, ROUGE-2=4%, ROUGE-L=24%)

where p and t and are the predicted and target contiguous intervals over the context document, containing the positions of the tokens. Intersec-

Method	EM	Prec.	Rec.	F1	IoU
BM25L	0.03	29.66	83.37	41.86	23.28
BM25L-oracle	**12.03**	**50.44**	51.18	**50.30**	**29.16**
Random Span	0.00	28.43	78.40	39.91	20.67
First Span	0.00	25.40	72.30	36.02	15.70
Last Span	0.03	29.38	**83.90**	41.77	23.63

Table 4: Ranking results on the development set. BM25L performs similar to selecting the last 512 tokens in the context document as the answer. BM25L-oracle knows the target answer span size.

Method	EM	Prec.	Rec.	F1	IoU
BERT-base e=2,s=128	23.27	60.28	**84.10**	64.34	54.04
BERT-base e=1,w,s=128	23.33	59.79	81.50	63.12	53.11
BERT-base e=2,w,s=128	**24.53**	**61.78**	83.46	**64.90**	**54.81**
BERT-base e=3,w,s=128	22.77	60.24	83.73	63.89	53.49
BERT-base e=2,w,s=256	24.09	61.64	79.08	63.38	53.41
BERT-base e=2,w,s=512	17.87	58.06	66.35	55.98	46.01
RoBERTa-base e=2,s=128	26.18	62.59	82.87	65.25	55.47
RoBERTa-base e=1,w,s=128	25.32	61.76	84.36	65.22	55.28
RoBERTa-base e=2,w,s=128	27.21	**62.71**	**85.34**	**66.17**	**56.33**
RoBERTa-base e=3,w,s=128	26.65	61.83	84.78	65.55	55.79
RoBERTa-base e=2,w,s=256	**27.33**	62.21	82.33	66.08	56.23
RoBERTa-base e=2,w,s=512	17.17	62.16	64.71	57.11	47.17
BERT-large e=2,w,s=128	28.54	63.83	84.68	66.95	57.24
RoBERTa-large e=2,w,s=128	**30.92**	**66.74**	**87.47**	**69.85**	**60.56**

Table 5: BERT and RoBERTa results on the development set. e=#epoch, w=warm-up over the first 1,000 steps, s=stride.

tion ($p \cap t = \{x | x \in p \text{ and } x \in t\}$) measures the overlapping interval and union ($\cup$) is defined as $p \cup t = \{x | x \in p \text{ or } x \in t\}$.

Figure 4 (left/middle) compares the F1 and ROUGE-N scores and IoU for the Longformer model on the development set. The F1 and ROUGE-N scores are always higher than IoU, but the metrics perform similarly in their higher values. Somewhat surprisingly, the F1 score can be up to 40% while there is no overlap between the two spans and IoU=0. We manually inspected the spans with F1>0 and IoU=0 and saw no significant semantic similarity between the predicted answer span and the target span. The same pattern repeats for the ROUGE-N scores. ROUGE-1 similar to F1 can reach 40% while IoU=0, but ROUGE-2 and ROUGE-L are less prone to such over-estimation due to lower chance of overlap of bigrams than unigrams and shorter LCSs in two random non-overlapping sequences. Figure 4 (right) indicates that the F1 and ROUGE-N scores are higher than IoU for longer answers reiterating the fact that these scores over-estimate more for longer sequences. Figure 5 shows two spans in a document with high F1 and ROUGE-N percentages, but different meanings.

5.2 Results and Discussion

We use the BM25L ranking function (Trotman et al., 2014) to investigate how a basic IR approach can detect answer spans using TF-IDF features. We adopt a sliding window approach with a window size of 512 and a stride of one sentence. We compare BM25L with random window (span) selection and the first and last window selection in the documents. Table 4 presents the results of the ranking functions. In the BM25L-oracle, we set the window size to the target answer span size. BM25L-oracle outperforms the other methods but the results are far from perfect. There is no significant difference between BM25L and other methods. The results restate the fact that there is little word overlap between non-factoid questions and their answers.

We analyze the performance of BERT and RoBERTa with different hyper-parameters on the development set in Table 5. Smaller strides, i.e., higher overlap between the segments, and warm-up contribute to better performances. RoBERTa constantly outperforms BERT, which is to be expected as RoBERTa is optimized robustly during the pre-training. We use the HuggingFace's Transformers (Wolf et al., 2019) code[4] and train the base and large models on 2 and 4 GPUs, respectively. We

[4] github.com/huggingface/transformers

Method	#Param.	EM	Prec.	Rec.	F1	IoU
BERT-base	110M	25.03	60.60	82.48	63.96	53.75
BERT-large	340M	30.29	64.87	84.62	67.91	58.39
RoBERTa-base	125M	29.07	64.02	84.79	67.19	57.65
RoBERTa-large	355M	33.40	67.79	**87.56**	71.10	62.39
Longformer	149M	**50.30**	**83.92**	85.17	**81.38**	**73.57**

Table 6: NLQuAD evaluation set results. Longformer surpasses the other models in all the metrics except recall.

have to use a batch size of 12 and 8, respectively, for the base and large models because of the long input sequence size and memory limitations.

We use the official AllenAI Longformer code[5] to train Longformer on NLQuAD. We use the same batch size of 12 (batch size of 1 and gradient accumulation over 12 batches) and learning rate warm-up for the first 1,000 steps. Due to memory requirements, we limit the experiments to only the Longformer base model (the large model cannot fit on our GPUs even with a batch size of 1). We ran the experiments on 2 NVIDIA P40 (24GB GPU memory) for about one day for 5 epochs. Similarly, we choose the best epoch based on the performance on the development set.

Table 6 summarizes the scores obtained by the baseline systems on the NLQuAD evaluation set. While Longformer significantly outperforms BERT and RoBERTa, its performance, particularly in terms of IoU and EM, is far from perfect. This demonstrates that NLQuAD and non-factoid QA is still an open problem for state-of-the-art models.

5.3 Human Evaluation

To ensure that the samples are of high quality, in addition to the initial investigation and pre-processing steps, we asked four volunteers to investigate 50 random samples from the evaluation set. They rated the goodness of answers on a 3-point scale: (1: Irrelevant answer; 2: Good answer after adding or removing some sentences; 3: Perfect answer). The average score is 2.56 indicating the high quality of NLQuAD's QA samples.

In order to benchmark human performance, we asked the four volunteers to answer 50 questions, a randomly sampled subset of evaluation set. They were given unlimited time to detect the answers, but on average, it took them about 270 seconds to answer a question. Table 7 compares human performance with Longformer and RoBERTa-large on the same subset. Similar to HotpotQA (Yang et al., 2018), we estimate the human upper bound by taking the best human answer in terms of our primary evaluation metric (IoU) for each sample. While NLQuAD is a challenging task both for humans and the state of the art QA models, the human upper bound performance significantly outperforms the models. We suspect that the mediocre average of human performance, considering the high score of the target answers, might be because volunteers are not familiar with the articles' writing style or they might have become exhausted by reading long articles.

Furthermore, we asked another volunteer to compare the target answers with the predicted answers in a pairwise comparison for 100 samples. Figure 6 shows that the target answers are preferred in 37% and 64% of cases over the Longformer and RoBERTa predictions, respectively. The human evaluation is in line with the results shown in Table 6 and Table 7.

[5] github.com/allenai/longformer

Method	EM	Prec.	Rec.	F1	IoU
RoBERTa-large	36.00	61.78	87.00	66.41	57.09
Human-AVG	35.50	86.67	68.66	72.52	62.94
Longformer	56.00	78.55	83.69	78.27	70.64
Human-UB	**74.00**	**97.79**	**94.88**	**95.49**	**92.63**

Table 7: Comparing human performance with Longformer and RoBERTa-large on a subset of evaluation set. UB=upper bound, AVG=average.

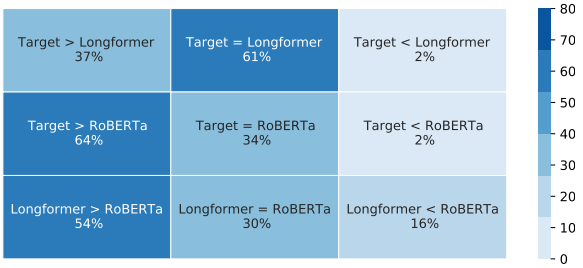

Figure 6: Pairwise comparison between the target spans, Longformer, and RoBERTa's predicted spans. X>Y means X is more preferable.

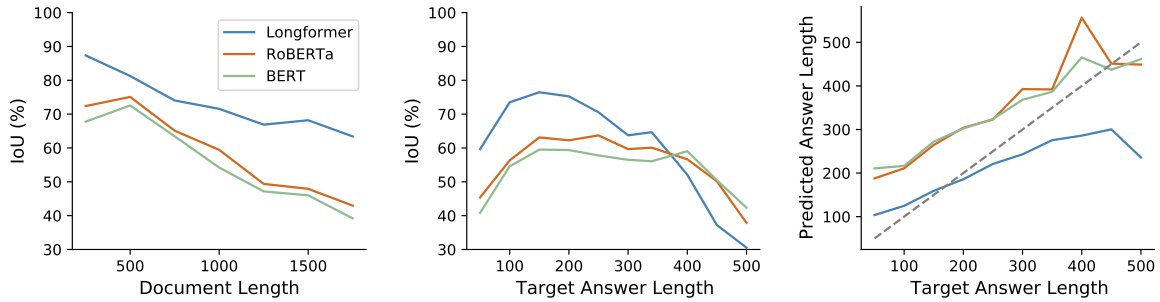

Figure 7: Effect of document and answer length on the performances. Left: IoU drops in all models for longer documents. Middle: RoBERTa and BERT outperform Longformer in longer answers. Right: Longformer has a bias to predict shorter answers while RoBERTa and BERT predict longer answers. The dashed line means $y = x$.

5.4 Error Analysis

Figure 7 compares the performance of BERT, RoBERTa, and Longformer for instances with different document and answer lengths. As expected, both longer documents and longer answers are harder for the models. Surprisingly, BERT and RoBERTa outperform Longformer for longer answers. The same pattern occurs for F1 and EM (not shown in the figure).

Figure 7 (right) shows that RoBERTa and BERT behave completely differently compared to Longformer for longer answer lengths. The former models have a bias to predict longer spans while Longformer under-estimates the length of the answer span. This different behaviour might be due to the sliding window approach and the prediction aggregation in the RoBERTa and BERT models and the attention dilation strategy in Longformer.

6 Conclusion

We introduce NLQuAD, a non-factoid long question answering data set from BBC news articles. NLQuAD's question types and the long lengths of its context documents as well as answers, make it a challenging real-world task. We propose to use Intersection over Union (IoU) as an evaluation metric for long question answering. To establish a baseline performance, we experimented with the BERT, RoBERTa, and Longformer question answering models. Longformer outperforms the other methods with an IoU of 73.57%, but the results show that the performance of state-of-the-art question answering systems is far from perfect. We hope NLQuAD will inspire more research in the area of document-level language understanding and question answering.

Acknowledgments

This research was partly supported by VIVAT. We thank the BBC for giving permission to publish our extracted data for non-commercial, research purposes. We also thank our volunteers for providing human assessments.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv:1611.09268*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv:1803.05457*.

Daniel Cohen and W. Bruce Croft. 2016. End to end long short term memory networks for non-factoid question answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR 16, page 143–146, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv:1704.05179*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. ANTIQUE: A non-factoid question answering benchmark. In *European Conference on Information Retrieval*, pages 166–173. Springer.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White. 2020. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, page 43–52, New York, NY, USA. Association for Computing Machinery.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. COSMOS QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2019. Speech and language processing, 3rd edition.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 605–es, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Ivan Sekulić, Amir Soleimani, Mohammad Aliannejadi, and Fabio Crestani. 2020. Longformer

for MS MARCO document re-ranking task. *arXiv:2009.09392*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020a. Long range arena: A benchmark for efficient transformers. *arXiv:2011.04006*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020b. Efficient transformers: A survey. *arXiv:2009.06732*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HOTPOTQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.