Question Answering with Long Multiple-Span Answers

Ming Zhu[1], **Aman Ahuja**[1], **Da-Cheng Juan**[2], **Wei Wei**[3], **Chandan K. Reddy**[1]
[1] Department of Computer Science, Virginia Tech, Arlington, VA
[2] Department of Computer Science, National Tsing Hua University, Taiwan
[3] Cloud AI Research, Google Inc.
{mingzhu, aahuja}@vt.edu, x@dacheng.info, wewei@google.com, reddy@cs.vt.edu

Abstract

Answering questions in many real-world applications often requires complex and precise information excerpted from texts spanned across a long document. However, currently no such annotated dataset is publicly available, which hinders the development of neural question-answering (QA) systems. To this end, we present *MASH-QA*[1], a **M**ultiple **A**nswer **S**pans **H**ealthcare **Q**uestion **A**nswering dataset from the consumer health domain, where answers may need to be excerpted from multiple, non-consecutive parts of text spanned across a long document. We also propose *MultiCo*, a neural architecture that is able to capture the relevance among multiple answer spans, by using a query-based contextualized sentence selection approach, for forming the answer to the given question. We also demonstrate that conventional QA models are not suitable for this type of task and perform poorly in this setting. Extensive experiments are conducted, and the experimental results confirm the proposed model significantly outperforms the state-of-the-art QA models in this multi-span QA setting.

What are tips for managing my bipolar disorder?

Along with seeing your doctor and therapist and taking your medicines, simple daily habits can make a difference. Start with these strategies. *(22 words truncated)* Pay attention to your sleep. This is especially important for people with bipolar disorder... *(178 words truncated)* Eat well. There's no specific diet... *(29 words truncated)* Focus on the basics: Favor fruits, vegetables, lean protein, and whole grains. And cut down on fat, salt, and sugar. Tame stress. *(81 words truncated)* You can also listen to music or spend time with positive people who are good company. *(73 words truncated)* Limit caffeine. It can keep you up at night and possibly affect your mood. *(47 words truncated)* Avoid alcohol and drugs. They can affect how your medications work. *(118 words truncated)*

Figure 1: An example of a question and its corresponding answer (highlighted) from MASH-QA. The answer consists of multiple sentences from the context. All the highlighted sentences will form the comprehensive answer. The context here is 632 words long, so we truncate a few portions of it.

1 Introduction

Developing neural networks for question answering (QA) has become an important and fast-growing area of research in the NLP community. Interest in this area is largely driven by the importance and effectiveness of such systems in virtual assistants and search engines. Driven by the development of large-scale datasets such as SQuAD (Rajpurkar et al., 2016, 2018), most of the work in this domain focuses on the task of machine reading comprehension, where the objective is to find a single short answer span—typically ranging from a few words to one sentence in length—given a question and a paragraph context (Xiong et al., 2017; Seo et al., 2017). Natural Questions (Kwiatkowski et al., 2019) makes machine reading comprehension more challenging by providing questions with long contexts. This makes it more suitable for training a typical QA system, which extracts answers from long documents returned by a search engine.

Existing QA datasets mainly consist of questions with short answers—typically ranging from a few words to a sentence—from the context document. Even though Natural Questions dataset

[1]Code: https://github.com/mingzhu0527/MASHQA

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3840–3849

(Kwiatkowski et al., 2019) provided paragraph-length answers for certain questions, these long answers are generally the paragraphs that contain the short answers, making most of the information supplemental (not critical) in nature. Moreover, because of the open-ended nature of many questions, the final comprehensive, succinct and correct answers may need to be extracted from multiple spans or sentences from the document. This problem is exacerbated when several spans that contain the answer are not in the vicinity of each other. Especially, this is often the case in domains such as healthcare, where people seek information regarding their specific health conditions, and the precise answer for their queries usually come from multiple sections or spans of a document.

In this work, we introduce *MASH-QA*, a large-scale dataset for question-answering, with many answers coming from multiple spans within a long document. MASH-QA is based on questions and knowledge articles from the consumer health domain, where the questions are generally non-factoid in nature and cannot be answered using just a few words. Fig. 1 shows an example question, and its corresponding context and answer from our dataset, which poses several unique challenges. First, the contexts are comprehensive healthcare articles, which can typically contain tens of paragraphs and hundreds of lines. Context of such length is challenging for existing neural QA models. Second, the answers are typically several sentences long, while current span extraction models usually predict very short spans. Another challenge in this setting raises from the fact that answers can consist of multiple sentences from nonconsecutive parts of a document, which can often be many sentences or even paragraphs apart. This results in sparsely-scattered patterns of semantic relevance in the context with respect to the query. This means that even if the answer comes from different parts of the document, which might be surrounded by the text that have limited relevance to the question, different answer snippets have some form of semantic relevance with each-other, and are centered around the same topic as the question. Although our dataset is from the healthcare domain, we believe that this problem setting can be generalized to other domains, where the questions typically require long and detailed answers.

Considering all these challenges, we formulate our question-answering task as a sentence selection task, which should also model the semantic relevance existing between different answer sentences, even when they are not adjacent to each-other in the context. Hence, we also propose *MultiCo*, a novel neural architecture that can address the challenges discussed above. Our model utilizes XLNet (Yang et al., 2019), which incorporates Transformer-XL units (Dai et al., 2019) to give semantic representations that capture the long-range dependencies existing in the long document context. We also use a sparsified attention mechanism, to ensure that the representations of sparsely scattered answer units are compactly aligned with each-other. The main contributions of this paper can be summarized as follows:

- We present a practical and challenging QA task, where the answers can consist of sentences from multiple spans of the long context. We introduce a new dataset called *MASH-QA* from the consumer health domain, that encompasses the challenges encountered in this task.
- We propose *MultiCo*, a novel neural model that deals with the long context problem, and is able to identify the sentences spanned across the document for forming the answer. MultiCo adapts a query-based contextualized sentence selection approach, combined with a sparse self-attention mechanism.
- Extensive experiments are conducted to evaluate the proposed model on multiple datasets. Our experimental results confirm that our approach outperforms state-of-the-art machine reading comprehension and semantic matching models.

To the best of our knowledge, this is the first work that introduces the QA setting with multiple discontinuous answer spans from a long document.

2 Related Work

Datasets The WikiQA dataset (Yang et al., 2015) contains query-sentence pairs, and their relevance labels, based on articles from Wikipedia. The SQuAD datasets (Rajpurkar et al., 2016, 2018) consist of question-answer pairs based on Wikipedia articles. The questions, however, are generally factoid, the answers are short, and the context is a small paragraph. The Natural Questions dataset (Kwiatkowski et al., 2019) provides a more realistic setting, where the context is a full Wikipedia

page, and the answer is a short snippet from the article. Some of the questions also include a long answer. MS-MARCO (Bajaj et al., 2016), SearchQA (Dunn et al., 2017), and TriviaQA (Joshi et al., 2017) contain questions and a short answer, and the questions are supported by more than one context document, some of which might be irrelevant to the question. CoQA (Reddy et al., 2019) and NarrativeQA (Kočiskỳ et al., 2018) are free-form QA datasets, where the answer is a short, free-form text, not necessarily matching a snippet from the context. ELI5 (Fan et al., 2019) is a long, free-form QA dataset, based on questions and answers from Reddit forums. However, since the evidence documents are collected using web-search, only 65% of supporting documents contain the answer.

Recently, many QA datasets from the medical domain have also been proposed. MedQUAD (Abacha and Demner-Fushman, 2019) and HealthQA (Zhu et al., 2019) are consumer health QA datasets, that contain query-answer tuples, and their relevance labels. emrQA (Pampari et al., 2018) contains rule-based questions constructed from medical records, while questions in CLiCR (Suster and Daelemans, 2018) are based on clinical report summaries.

Techniques Earlier works in QA used similarity based models for classifying answers based on their semantic similarity with the document (Yu et al., 2014; Miao et al., 2016). The public release of SQuAD dataset motivated the development of attention-based neural models (Xiong et al., 2017; Seo et al., 2017; Chen et al., 2017). With the advancements in language modeling (LM) techniques such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), LM-based techniques have gained more popularity in recent times.

3 MASH-QA Dataset

3.1 Dataset Description

Since we focus on the task of multi-span question-answering from long documents, our dataset consists of *(question, context, [answer sentences])* tuples. Each tuple consists of a natural language question, which can be answered using one or more sentences from the context. Context here is a long document, a typical web article with multiple paragraphs. Each answer consists of several sentences, which can either belong to one single span, or multiple spans from the context document. Since questions in our dataset can have multiple sentences that form the answer, we provide the index of all correct answer sentences with each tuple. We refer to the single-span answer subset of our dataset as MASH-QA-S, and the multi-span answer subset as MASH-QA-M. Some of the basic statistics of our dataset are shown in Table 1.

	MASH-QA-S	MASH-QA-M	MASH-QA
# Contexts	5,210	3,999	5,574
# QA pairs	25,289	9,519	34,808
# Train QA	19,989	7,739	27,728
# Dev QA	2,614	879	3,493
# Test QA	2,686	901	3,587

Table 1: Basic statistics of MASH-QA dataset.

3.2 Data Collection and Processing

Our dataset consists of consumer healthcare queries sourced from the popular health website WebMD[2]. The website contains articles from a diverse set of domains related to consumer healthcare. Each healthcare section on the website also consists of questions related to common healthcare problems faced by people. The answers to these queries consist of sentences or paragraphs from the article associated with the relevant healthcare condition. These answers have been curated by healthcare experts, and can accurately answer the corresponding query. Because of the nature of the domain, correctness of the answer is especially important, as in domains such as healthcare, an incorrect answer to a consumer can have dire consequences.

For each question, we first split the answer into sentences. We also split each of the context documents into the constituent sentences. Next, for every answer, we map each of its sentences to the corresponding sentence from the context. We notice that some of the answer sentences have been manually edited by the healthcare experts who answered the question. In such cases, we select a set of candidate sentences from the context that are similar to the answer sentence using tf-idf match, and then manually select the sentence that corresponds to the answer.

3.3 Dataset Characteristics

A comparison of our dataset with other QA datasets from general and healthcare domains is shown in Table 2. Table 3 shows some of the common question types from our dataset. We discuss some of

[2]https://www.webmd.com/

	Dataset	#QA	Context Source	QA Type	Answer Span	Context Length	Answer Length
Generic	WikiQA	3K	Wikipedia	Extractive	Single	238.4	11
	SQuAD-1.1	108K	Wikipedia	Extractive	Single	117.2	3.1
	Natural Questions	307K	Wikipedia	Extractive	Single	7320.3	85.2 (long)
	ELI5	270K	Web Search	Abstractive	Multiple	857.6	130.6
Healthcare	CLiCR	105K	Clinical Reports	Abstractive	Single	1385.4	2.7
	emrQA	400K	Medical Records	Extractive	Single	955.4	10.2
	MedQUAD	47K	Health articles	Ranking	Single	N/A	123.9
	HealthQA	8K	Health articles	Ranking	Single	N/A	233.4
	MASH-QA	35K	Health articles	Extractive	Multiple	696.2	67.2

Table 2: Comparison of MASH-QA dataset with other Question Answering datasets.

Starts With	%age	Example
What	46.09	What are the symptoms of gastritis? What are tips for treating acne?
How	31.03	How can I prevent blisters? How does exercise help stress?
Can, Is, Are Do, Does	11.01	Can I prevent sinusitis? Is scalp psoriasis common?
When	3.65	When do I need eye protection? When is flu season in the U.S.?
Why	2.05	Why do we have tears? Why do I need dental exams?

Table 3: Common question types and their examples from the MASH-QA dataset.

the key observations below:

Answers with Multiple Spans A key characteristic of our dataset is that, for many questions, the answers are obtained using information from multiple, discontinuous spans from the document, making the task more challenging in nature. The existing multi-document or multi-span QA datasets are abstractive in nature, and the support documents were curated using automatic techniques, such as web search. Because of this nature, the answer is not guaranteed to be found in the context, and the documents are often noisy, with limited relevance to the question. In contrast, our dataset contains multi-span answers that are curated by experts, which ensures that the different answer spans have information that is required to answer the question. Moreover, for a domain such as healthcare, we believe the extractive setting is ideal, since abstractive answers can introduce unpredicted errors resulting from answer generation.

Comprehensive and Compact Answers The answers in our dataset are generally comprehensive, and all the sentences in an answer contribute information that is important to answer the question. In existing datasets with long answers, majority of the information in the long answer is supplemental in nature. Natural Questions, for example, provides a short answer for the question, and a long answer that was created by selecting the entire paragraph containing the short answer. The answers in our dataset, on the other hand, have multiple sentences, each of which contains a unique piece of information about the subject in the query. We believe that comprehensiveness and compactness of answers are vital in the healthcare domain, since answers with missing information can potentially mislead people, while answers with extra information can be overwhelming.

Question Types A majority of the questions in our dataset are non-factoid and open-ended in nature, and seek for detailed information about the health condition. A significant proportion of the questions are "How" type, and such questions generally tend to be open-ended. Although questions starting with "What" generally ask for specific facts, we find that many of these questions, such as the ones shown in Table 3, are in fact open-ended, and require long answers. Our dataset also contains many "Yes/No" type questions, which often require explanations.

4 The Proposed MultiCo Model

Given a query and a document, the goal of our MultiCo model is to select the sentences that can accurately answer the query. An intuitive way to solve this problem would be to use a text matching model that takes the query and a sentence as the input, and predicts their relevance. However, as shown later, this approach does not capture the overall context of the sentences. Hence, in our problem setting, where multiple sentences from a document can belong to the answer, it gives poor results. Hence, our proposed approach uses the concept of *query-based contextualized sentence selection* from the document.

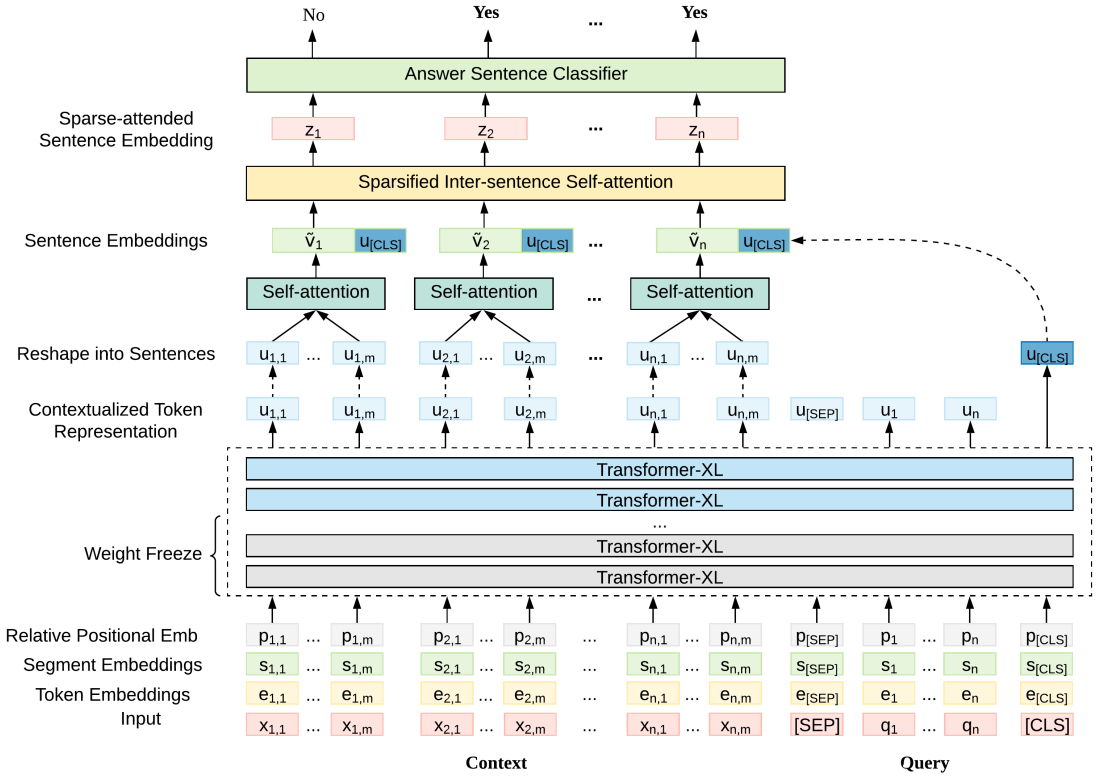

Figure 2: Architecture of the proposed MultiCo model.

4.1 Problem Formulation

Given a query Q and a context document $D = \{s_1, .., s_n\}$, where s_i refers to the i^{th} sentence in the document, the objective of our model is to classify each sentence s_i as relevant or not for the given query, conditioned on other sentences present in the document. Let $c_i \in \{0, 1\}$ be the relevance label that depicts whether sentence s_i belongs to the answer or not. Mathematically, we want to model the probability $P(s_i = c_i|Q, D)$ for $i \in \{1, .., n\}$.

4.2 Model Architecture

Figure 2 shows the architecture of our proposed model. The main components of our model are described in detail below:

Query and Context Encoder To encode the query and the long document context, we use XLNet (Yang et al., 2019) as the encoder. One of the main advantages of XLNet is that it is based on the Transformer-XL framework (Dai et al., 2019), which is specifically designed to deal with long documents. This makes it an ideal choice in our setting, as it can effectively encode the long context. Moreover, using a large pre-trained language model also allows us to obtain high quality token representations.

In our model, we first tokenize the query and each context sentence, and then pad each sentence upto a pre-defined maximum sentence length m. Let $\{X_1, .., X_n\}$ represent the sentences, where $X_i = \{x_{ij}\}_{j=1}^{m}$ be the tokens in sentence s_i, and let $Q = \{q_j\}_{j=1}^{n}$ represent the query. Following (Yang et al., 2019), we concatenate a [CLS] token to the query, and a [SEP] token at the end of the last sentence. The encoded representations can be obtained by the equation below:

$$\begin{aligned} \boldsymbol{U}_1; ..; \boldsymbol{U}_n, \boldsymbol{u}_{[SEP]}, \boldsymbol{U}_q, \boldsymbol{u}_{[CLS]} = \\ \text{XLNet}(X_1; ..; X_n, [SEP], Q, [CLS]) \end{aligned} \tag{1}$$

Sentence Embeddings To obtain a fixed dimensional vector for each sentence s_i, we use self-attention (Lin et al., 2017) over the encoded representations $\boldsymbol{U}_i$ obtained in the previous step, to get the intermediate sentence embedding $\tilde{\boldsymbol{v}}_i$.

$$\begin{aligned} \boldsymbol{h}_{ij} &= \mathbf{w}_a \tanh(\mathbf{W}_a \boldsymbol{u}_{ij}) \\ \alpha_{ij} = \underset{j}{\text{softmax}}(\boldsymbol{h}_{ij}) \quad \tilde{\boldsymbol{v}}_i &= \sum_{j=1}^{m} \alpha_{ij} \boldsymbol{u}_{ij} \end{aligned} \tag{2}$$

Here, α represents attention weights. Next, to add the overall context and query representations to

the sentence representation, we concatenate the embedding of [CLS] token returned by XLNet, to get the final sentence vector $\boldsymbol{v}_i = [\tilde{\boldsymbol{v}}_i; \boldsymbol{u}_{[CLS]}]$.

Sparsified Inter-Sentence Attention The multi-span nature of answers in our dataset requires us to have a mechanism to link the different answer sentences with each-other. Moreover, the number of relevant sentences in the context is much less than the total number of sentences in the context. Hence, we use a sparsified inter-sentence attention layer based on α-entmax (α = 1.5) (Peters et al., 2019; Correia et al., 2019) to introduce sparsity.

$$\begin{aligned} \boldsymbol{g}_{ij} &= \mathbf{w}_b \tanh(\mathbf{W}_b[\boldsymbol{v}_i; \boldsymbol{v}_j]), \\ \beta_{ij} = \underset{j}{\alpha\text{-entmax}}(\boldsymbol{g}_{ij}) \quad \boldsymbol{z}_i &= \sum_{j=1}^{n} \beta_{i,j} \boldsymbol{v}_j \end{aligned} \tag{3}$$

β_{ij} here represents attention weights of sentence i with respect to sentence j. For any given sentence, α-entmax above gives sparse attention weights over other sentences in the context. This makes the final representation only conditional on a small number of other sentences with similar semantic nature, and zeroes out the effect of other sentences, unlike the standard softmax. For any given vector $\boldsymbol{g}$, it can be calculated as follows.

$$\alpha\text{-entmax}(\boldsymbol{g}) = \text{ReLU}[(\alpha - 1)\boldsymbol{g} - \tau\mathbf{1}]^{1/\alpha - 1} \tag{4}$$

Here, τ is the threshold, which can be computed as per Peters et al. (2019). As we can see, the function will give a zero probability for all values of $g \leq 1/(\alpha-1)$, hence resulting in a sparse probability distribution.

Answer Classifier After computing the representation of each sentence with respect to the query and the overall context, we pass the sentence vector $\boldsymbol{z_i}$ through a multi-layer dense network, followed by softmax, to get the final answer probability distribution $\hat{\boldsymbol{y}}_i$.

$$\hat{\boldsymbol{y}}_i = \text{softmax}(\mathbf{W}_{out}\boldsymbol{z}_i + \mathbf{b}_{out}) \tag{5}$$

4.3 Optimization

Since we model the question-answering task as a sentence classification task, we use binary cross entropy as the loss function to train our model. Let $\boldsymbol{y}_i$ be the true binary labels for sentence s_i. The loss for each sentence can be computed as follows:

$$\mathcal{L} = -\sum_{j \in \{0,1\}} y_{ij} \log(\hat{y}_{ij}) \tag{6}$$

5 Experiments

5.1 Implementation Details

We implemented our model in TensorFlow (Abadi et al., 2016). The model was trained using Adam optimizer (Kingma and Ba, 2015), with a learning rate of 2×10^{-5}. The maximum length for query and context sentences was set to 64 and 32 tokens respectively, and the maximum number of sentences in one segment was set to 13. For longer contexts, we split them into multiple segments of 13 sentences each, and append query to each segment. The maximum input length, including context, query and other tokens, was set to 512 tokens. We used a pre-trained version of XLNet (24 layers, 340M parameters), and allow only the top 12 layers to be trainable, as previous research (Jawahar et al., 2019) suggests that the semantic features are learned mainly by the top layers. All the experiments were run on servers with single Tesla K80 GPUs.

5.2 Performance against Answer Sentence Classification Based Methods

In our first set of experiments, we would like to observe the performance of our model (which computes the probability of sentence being in the answer conditional on both the query and the full context) compared to pairwise models (which only use the query and the sentence under consideration) that classify the query-sentence as relevant or not using semantic matching. As suggested earlier, this is an intuitive way to solve the sentence classification task. Hence, for this task, we compare the performance of our model against other semantic matching baselines, that predict the relevance label for each sentence individually, given the *(query, sentence)* pair as the input.

Baselines and Evaluation Metrics We compare our model against various semantic matching models for this task. The semantic matching models which are used for our experiments were based on **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), and **XLNet** (Yang et al., 2019). For all these models, we use the standard 24-layer pre-trained versions of their LARGE models, and fine-tune them to do semantic matching on *(query, sentence)* pairs. We also use **TANDA** (Garg et al., 2020), which utilizes a BERT-based architecture to answer questions using pairwise (query, sentence) classification approach, as a baseline model.

We evaluate all the models on two levels: Sentence-level evaluation computes the **P**recision, **R**ecall, and **F1**-score based on the predicted label (relevant or not) of each sentence. These set of metrics will reward a model, even if the answer is partially correct. We also evaluate Answer-level Exact Match (EM), which computes the percentage of answers, whose predicted label matches the true label, for all the sentences in the answer. This will help us evaluate if the model can get the entire answer correct.

	Sentence			Answer
Model name	**P**	**R**	**F1**	**EM**
TANDA	56.48	16.42	25.44	8.95
BERT	56.18	16.25	25.21	8.89
RoBERTa	57.70	19.06	28.65	9.40
XLNet	56.05	19.73	29.19	9.09
MultiCo	**58.16**	**55.90**	**57.00**	**22.05**

Table 4: Comparison of MultiCo with other baseline Classification models on MASH-QA dataset.

Results on MASH-QA As we can see from the results in Table 4, MultiCo significantly outperforms the classification baselines on the MASH-QA dataset, on both the sentence-level and answer-level metrics. Since we model the sentence conditional on both the query and other sentences in the context, our model can take into account the semantic dependencies that exist between multiple sentences in a document, and their relationship with the query. Other techniques only use the query and the sentence under consideration, and do not take into account the association between different answer sentences, which leads to lower performance.

Model name	**P**	**R**	**F1**
TANDA	**68.47**	45.00	54.31
BERT	48.10	56.32	51.89
RoBERTa	56.23	53.92	55.05
XLNet	48.54	51.19	49.83
MultiCo	56.79	**56.92**	**56.86**

Table 5: Comparison of MultiCo with other baseline classification models on WikiQA dataset.

Results on other QA datasets We also evaluate the performance of our proposed model on other QA datasets, to observe its generalizability to other settings. Since there are no existing datasets that contain multi-span answers, the only dataset that can resemble our problem setting is WikiQA. Here, we only calculate the sentence-level metrics, as most of the answers in WikiQA contain only one sentence. The results presented in Table 5 show that our model outperforms all other baselines. A paired t-test indicates that our model outperforms RoBERTa with more than 95% confidence level (experimented with 5 different random seeds). The baselines have a better performance on WikiQA as compared to MASH-QA, which can be attributed to two factors: shorter context length, and fewer sentences per answer. Because of this, the techniques used in our model to handle these factors have minimal effect. Nonetheless, our model still outperforms the baselines, which shows that our technique can be generalized to other QA settings as well.

5.3 Performance against Span Extraction Based Methods

In this setup, we show the comparison of our proposed model with other span extraction based methods. This setup allows us to evaluate how the sentence selection/classification approach performs in contrast to approaches that predict the start and end indices of the answer span. Since such methods are designed only to predict a single start and end index, the applicability of such approaches is only limited to cases where the answer can only have one span from the context. Hence, for this setup, we only use the subset MASH-QA-S of our dataset that contains questions with single span answers.

Baselines and Evaluation Metrics We use the following baseline techniques in this experiment task: **DrQA Reader** (Chen et al., 2017) uses an RNN-based architecture, along with context-to-query attention, to compute the answer. **BiDAF** (Seo et al., 2017) uses bidirectional attention (query-to-context and context-to-query) for answer span prediction. We also use the QA versions of **BERT**, **SpanBERT** (Joshi et al., 2020), and **XLNet**, as the baselines. For the former three models, we use the standard pre-trained versions of LARGE models, and fine-tune them on our dataset.

Since our objective here is to predict the answer span for the single answer, we use F1 and Exact Match (EM) as the evaluation metrics. F1 measures the overlap between the predicted and the true answers, and EM measures the percentage of overall predicted answers that exactly match the true answer.

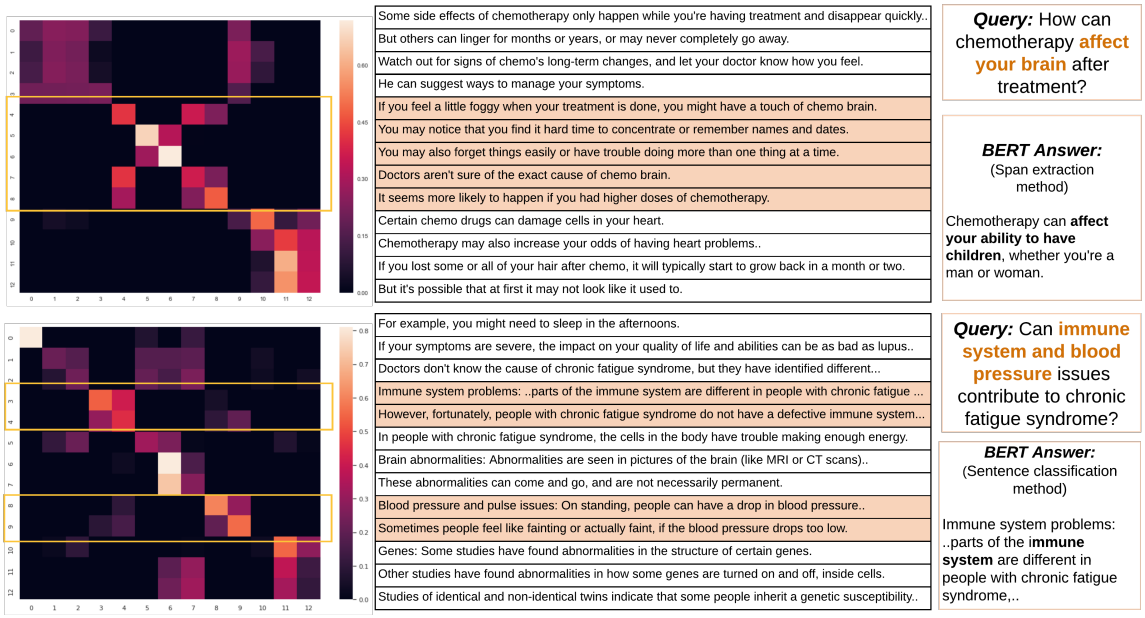

Figure 3: Heatmap of attention weights from the inter-sentence attention layer for two QA pairs. The matrices show the attention weights of each sentence with respect to every other sentence from the context. The high values of diagonal elements represent the weight of a sentence with respect to itself. Answers from BERT are shown on the right.

Model name	F1	EM
DrQA Reader	18.92	1.82
BiDAF	23.19	2.42
BERT	27.93	3.95
SpanBERT	30.61	5.62
XLNet	56.46	22.78
MultiCo	**64.94**	**29.49**

Table 6: Comparison of MultiCo with other baseline Question Answering models on MASH-QA-S dataset.

Results The results for the span prediction task on single-span MASH-QA are shown in Table 6. As we can see, MultiCo outperforms all the other baselines by a wide margin. This can be attributed to the fact that most of the QA models proposed so far in the literature are mainly focused on the extractive QA datasets with short answers, that typically range upto a few words. The answers in MASH-QA on the other hand, are longer, making the task more challenging. For long answers, where the minimum answer unit is a sentence, models trained with sentence-level objective are likely to perform better than those with word-level objectives.

5.4 Qualitative Results

For qualitative analysis, we analyze the effect of using sparse attention on the model performance. In Fig. 3, we plot the heatmap of the attention weights obtained from the sparse attention layer, for two query-context pairs from our dataset. The first example here contains an answer with four consecutive sentences. As we can see, the attention weights for these sentences are high with respect to each other, and zeroed out with respect to non-answer sentences. Similarly, non-answer sentences only attend to other non-answer sentences. A similar trend is observed in the other example, that contains four answer sentences from two non-consecutive spans.

The answers obtained from the baseline BERT model using the two QA approaches are also shown. Using the span extraction approach, BERT gives an incorrect short answer, while with the pairwise query-sentence classification approach, it only predicts one answer sentence correctly. We observe that these answers have been selected based on superficial cues. By linking semantically similar sentences, the sparse attention ultimately helps to link the query with answer sentences that have limited similarity with the query, but are similar to other answer sentences.

6 Conclusion

We proposed a novel form of question-answering, where answers to a question consist of multiple sentence-level spans from a long document. To

support this task, we introduce MASH-QA, a novel and challenging QA dataset from the consumer health domain. MASH-QA consists of questions that can be answered using information from multiple spans from the document. To motivate further research in multi-span QA, we also propose a novel QA architecture called MultiCo, that uses query-based contextualized sentence selection approach for finding multi-span answers from long documents. By using a sentence-selection based objective, our model outperforms the existing state-of-the-art QA models by a wide margin.

Acknowledgements

This work was supported in part by the US National Science Foundation grants IIS-1619028, IIS-1707498 and IIS-1838730.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):511.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. 2019. Adaptively sparse transformers. 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 2174–2184.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

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 preprint 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.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Eunsol Choi, Daniel S 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.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Tomáš Kočiskỳ, 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, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *International Conference on Learning Representations (ICLR)*.

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 preprint arXiv:1907.11692*.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368.

Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations (ICLR)*.

Simon Suster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *International Conference on Learning Representations (ICLR)*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pages 2472–2482.