Dynamically Generated Question Answering Evidence using Efficient Context-preserving Subdivision

Avi Bleiweiss
BShalem Research, Sunnyvale, U.S.A.

Keywords: Question Answering, Evidence Natural Split, Transformers, Language Model, Deep Learning.

Abstract: Recently published datasets for open-domain question answering follow question elicitation from a fairly small snippet of Wikipedia content. Often centered around an article section, the evidence is further subdivided into context-unaware passages of uniform token-lengths to found the basic retrieval units. In this study we hypothesized that splitting a section perceived as an opaque text fragment may hinder quality of answer span predictions. We propose to dynamically draw content corresponding to an article-section url from the most updated online Wikipedia rather than from an archived snapshot. Hence approaching space complexity of $O(1)$, downward from $O(n)$ for a dataset that is fully populated with static context. We then parse the url bound content and feed our neural retriever with a list of paragraph-like html elements that preserve context boundaries naturally. Using knowledge distillation from a sustainable language model pretrained on the large `SQuAD 2.0` dataset to the state-of-the-art `QuAC` domain, shows that during inference our natural context split recovered answer span predictions by 7.5 F1 and 4.1 EM points over a synthetic distribution of fixed-length passages.

1 INTRODUCTION

Open-domain question answering (QA) is the task of answering assertion queries by searching a large knowledge base, typically unstructured text corpora such as Wikipedia, and finding the answer text span in an article (Chen et al., 2017). Together with recent advances in neural architectures (Vaswani et al., 2017) that let the network learn from massive amounts of unstructured text, the use of self-supervision in language models have proven pivotal to a major qualitative shift for various natural language processing (NLP) tasks. These achievements have pushed state-of-the-art (SotA) open-domain QA systems, replacing traditional information retrieval (IR) methods with dense representations and devise end-to-end training of the context retriever and the machine reader components.

Modern approaches to various NLP tasks are based on pretrained language models with deep contextualized word representation that do not require labeled data. Fine-tuning these pretrained language models was shown to be an effective strategy to reinstate the SotA for literally all open-domain question answering tasks (Li and Choi, 2020). Based on the transformer architecture (Vaswani et al., 2017), the pivotal `BERT` (Devlin et al., 2019) is pretrained on a massive corpus for two unsupervised tasks: a masked language model (MLM) that reconstructs a partially masked segment of input text based on immediate visible context, and next sentence prediction (NSP) in a text sequence. Improving significantly on the performance of `BERT` by applying longer training to the same MLM objective, `RoBERTa` (Liu et al., 2019) utilizes an auxiliary binary classifier to predict whether a question is answerable or not. `AlBERT` (Lan et al., 2019), also a variant of `BERT`, introduces factorized embedding parameterization and cross-layer parameter sharing to achieve a marked network parameter efficiency. `AlBERT` also replaces NSP with sentence order prediction (SOP) and centers on modeling coherence of inter-sentence relation.

Our work followed prior research of dense passage retrieval, however, we challenged the retriever splitting of Wikipedia web pages into text passages of unified lengths. We conjectured that straddling answer spans across non or overlapping passages are harder to predict. In our evaluation, we used the more sustainable `AlBERT` model, [1] pretrained on the `SQuAD 2.0` (Rajpurkar et al., 2018) dataset that was first to

[1]We obtained our `AlBERT` models from the repositories available on https://github.com/huggingface/transformers.

Bleiweiss, A.
Dynamically Generated Question Answering Evidence using Efficient Context-preserving Subdivision.
DOI: 10.5220/0010815700003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 276-284
ISBN: 978-989-758-547-0; ISSN: 2184-433X

Table 1: Our dynamic QA data example: highlighted in gray are system inputs, as the article and section titles are combined to retrieve naturally split context. Paragraph count and text are then fed to the reader that predicts an answer span. Shown are three out of five context preserving paragraphs, along with the answer span that extends the entire passage (check marked).

Article Title	Ian_McKellen
Section Title	Charity_work
Question	What charity work did he do?
Paragraphs	5
Natural Context	× In April 2010, along with actors Brian Cox and Eleanor Bron, McKellen appeared in a series of TV advertisements to support Age UK, the charity recently formed from the merger of Age Concern and Help the Aged. All three actors gave their time free of charge.
	✓ A cricket fan since childhood, McKellen umpired in March 2011 for a charity cricket match in New Zealand to support earthquake victims of the February 2011 Christchurch earthquake.
	⋮
	× Together with a number of his Lord of the Rings co-stars (plus writer Philippa Boyens and director Peter Jackson), on 1 June 2020 McKellen joined Josh Gad's YouTube series Reunited Apart which reunites the cast of popular movies through video-conferencing, and promotes donations to non-profit charities.

introduce unanswerable questions, and ran inference on the QuAC validation set (Choi et al., 2018). Our contribution is threefold: (1) we propose variable-length context preserving splits of input section sequences rather than fixed segments and show through quantitative analysis plausible quality gains of answer prediction, (2) we motivate a trade-off between a light context-less QA dataset and compelling inline retrieval that sustains only a slight runtime overhead, compared to a costly fully-loaded dataset, and (3) we provide a qualitative empirical review of paragraph-level answerability in natural evidence split.

2 RELATED WORK

Lee et al. (2019) presented end-to-end jointly learned retriever and reader from question-answer string pairs. Using a retriever pretrained to predict the evidence context given a sentence, their method was shown to outperform a BM25 (Robertson and Zaragoza, 2009) baseline by 19 exact match (EM) points, when considering questions of unknown answers. Similarly, rather than capture world knowledge concealed in parameters of ever-expanding networks, Guu et al. (2020) introduced pretrained representations that augment a language model with a learned textual knowledge-retriever to perform reasoning over a corpus of broad-based evidence during inference. By using Maximum Inner Product Search (MIPS) to select top-matching evidence passages, the authors addressed the major computational challenge for incorporating a large-scale neural retrieval module. Leveraging a standard pretrained transformer model along with a question-passage encoder architecture, Karpukhin et al. (2020) devised a dense embedding representation of passages that uses a low-dimensional index for efficient retrieval. While Seo et al. (2019) proposed a hybrid approach using both dense and sparse indexable representations for runtime optimization of training and inference.

Recognizing the need to address questions users want to know the answer to but do not know the answer yet, the research community recently responded with high-quality datasets of information-seeking question-answer pairs, including the Stanford Question Answering Dataset (SQuAD 2.0; Rajpurkar et al., 2018), Question Answering in Context (QuAC; Choi et al., 2018), Conversational Question Answering (CoQA; Reddy et al., 2019), and Natural Questions (NQ; Kwiatkowski et al., 2019). The work on Open-Retrieval Question Answering in Context (OR-QuAC; Qu et al., 2020) extends QuAC by drawing a large passage collection from the entire Wikipedia. All aforementioned datasets present a model that pairs a question along with static content obtained from a Wikipedia article. Given a list of article passages, the QA model predicts an index of the passage that answers the question, as well as a start and end token indices of the minimal span that completely answers the question.

The impact of article splitting into text segments

on the quality of QA systems has been studied to a fairly limited extent. The following review applies mainly to the retriever, as the language model architecture presents a not-to-exceed sequence size constraint the reader module must comply with. Wang et al. (2019) experimented with fixed-size non-overlapping passages and concluded that a basic retrieval unit of 100 token long performed the best. However, they note that answer spans which border passage boundary may lose vital context. Using instead overlapping passages with a sliding window of 100 tokens and a 50-token stride— half-window size— mitigated the shortfall of a non-overlapping passage split and resulted in a performance gain of 4.1 F1 points. On the other hand, Karpukhin et al. (2020) observed inconsequential performance gains when using overlapping passages and resorted to uniform-length distinct passages of 100 tokens each. In a similar vein, Choi et al. (2018) validated the QuAC dataset by dividing the evidence text of a section into twelve chunks of equal size, however, this partition mainly aided multi-turn conversations, as answers to a chain of free-form questions progress through neighboring chunks.

In their work, Yu et al. (2019) propose sentence extraction from evidence for the task of multiple-choice machine reading comprehension (MRC). They applied distant supervision to generate imperfect sentence labels that are further used to train the sentence extractor. Labels are denoised by a deep probabilistic framework (Wang and Poon, 2018) that incorporates both sentence-level and cross-sentence linguistic indicators. They used GPT, a pretrained transformer baseline (Radford et al., 2018) shown to improve F1 scores by 2.2 percentage points over the previous highest score. The approach Yu et al. (2019) present fits well multiple-choice MRC tasks that are sentence bound in context. However, span based question answers require multi sentence reasoning and are unlikely to necessarily align with any of a sentence head or tail.

Our approach to apportion input text greatly differs and avoids any context-unaware chunking. We rather abide by the Wikipedia source split of a given article section into html paragraph elements, which leads to a highly efficient retriever that loads at most one section at a time.

3 TASK FORMULATION

Owing to its simple and consistent style, English Wikipedia, an ever growing source of knowledge base, satisfies the task of large-scale open-domain QA. Formally, given a set of n Wikipedia articles $D = \{d_1, \cdots, d_n\}$, each typically divided into m sections $S = \{s_1, \cdots, s_m\}$, and each section over a certain length is split up into l paragraphs, or passages $P = \{p_1, \cdots, p_l\}$. Our language model is presented with a question q along with the content of an article section comprising a list of paragraphs $p_k^{(i,j)}$, where $1 \leq i \leq n$, $1 \leq j \leq m$, and $1 \leq k \leq l$. Our retriever parses a series of variable-length paragraphs p_k that represents a single article section $s_j^{(i)}$, and in this formulation a reference answer span is bound to never straddle across paragraphs. The task of our QA reader is to iterate over the list p_k and predict for each passage a start and end token indices $\in \{1, \cdots, |p_k|\}$ of a minimal span that answers the question fully, or return Null if it is not possible to generate an answer.

We refrained from commonly using over two-year-old static html backup of Wikipedia wikis (Lee et al., 2019; Seo et al., 2019; Karpukhin et al., 2020; Guu et al., 2020), and chose instead to expose our retriever to dynamically load the current most updated article version online, thus obtaining a highly sustainable QA dataset.

4 DATA

The new QuAC dataset (Choi et al., 2018) facilitates learning from information-seeking dialogs. [2] Intended primarily for multi-turn interactions, QuAC is a large-scale dataset comprising 14,000 dialogs and approaching a total of 100,000 questions. QuAC is sought to bridge the gap between dialog and QA, thus answering a series of questions in a conversational manner. To the extent of our knowledge, QuAC is currently the only dataset that provides Wikipedia article and section titles in a QA object, both essential to our retriever for acquiring html paragraphs online. Evidently section content often tend to cover a relatively narrow selection of topics. In his critical review, Yatskar (2019) provides qualitative comparison of SQuAD 2.0, QuAC, and CoQA, and shows pretrained models can transfer between domains while yielding moderate performance gains. Motivated by this analysis, we hypothesized that distilling knowledge from a rich QA domain is more favorable to demonstrate the quality advantage of natural evidence text over a synthetic split. Our language model is thus pretrained on the SQuAD 2.0 dataset (Rajpurkar et al., 2018) and we ran inference on the QuAC development set.

As a preprocessing step, we repurposed QuAC conversational examples of the validation set to fit our retriever that obtains content by inline parsing html

[2] http://quac.ai/

in sections of Wikipedia articles. Our QA data objects are thus considerably more light weight as we dropped the QuAC static context of evidence entirely and only use a handful of meta-data fields. The article and section titles are concatenated to make up a hash tag url, [3] alternatively we search a section ID attribute of an html heading tag. Our reader is provided with the first question of a dialog chain, and the original answer span is used as a gold reference. In Table 1, we show our QA example along with the extracted natural section paragraphs.

Table 2: Statistics summarizing natural passage retrieval.

	Min	Max	Mean	SD
paragraphs / section	2	10	4.5	1.3
tokens / section	136	1495	377.3	140.4
tokens / paragraph	2	449	84.5	55.8

QuAC validation set (Choi et al., 2018) contains one thousand questions obtained from 1,000 unique Wikipedia sections, of which 67 questions, a share of 6.7 percentage points, are unanswerable, and the remaining 933 questions are answerable. In addition, the QuAC set has 68 affirmation questions, or 6.8%, with 63 and 5 yes and no answers, respectively. [4] Our model expects to return a Null answer for affirmation questions. Table 2 reviews summary statistics for our retrieval of natural paragraphs from one thousand distinct sections, suggesting a maximum of 449 tokens per paragraph that is well under the reader maximal position-embeddings constraint of 512 elements. In Figure 1, we show the distribution of natural context with about eighty percent apportioned to sections comprising three to five paragraphs, [5] in a range of two to ten paragraphs per section.

5 SECTION SAMPLING

In the most prevalent practice, the size of a section base-unit split is an ad hoc choice based entirely on heuristic shortcuts, and thus the reader is often prone to predict suboptimal answer spans. To make the evaluation reproducible across QA systems and impartial to a fixed or natural subdivision choice, we reformulated the selection of a split base-unit as a section-paragraph sampling problem. In this framework, the size of a split base-unit is declared private and defined distinctly for each section, rather than specifying a global split parameter which affects the entire Wikipedia domain. In addition, the QA dataset provides in each example a section sampling-window s_w property with a size that amounts to the token-length of the shortest section paragraph p_k

$$s_w = \underset{1 \le k \le l}{\operatorname{argmin}} f(|p_k|),$$

where f is an optional user function that defaults to a division by two to sustain a minimal sampling rate. Figure 2 presents visualization of a uniformly sampled section showing the progression of a non-overlapping sliding window across the section paragraphs and their corresponding reference answer-spans.

Given a fixed length split and a variable number of non-overlapping section splits n, the time complexity for searching a unified split section to derive the answer span is linear with the number of non-overlapping splits $O(n)$. Although the computational path is identical, the alternative method of a sliding window with a fifty percent overlap incurs nonetheless a runtime cost that doubles compared to the non-overlapping approach. Space complexity is $O(1)$ for a split unit that fully contains a reference answer span, however an answer span might straddle two or more splits that the retriever has to concatenate and avoid partial answer spans, thus leading to a worst case space complexity of $O(n)$.

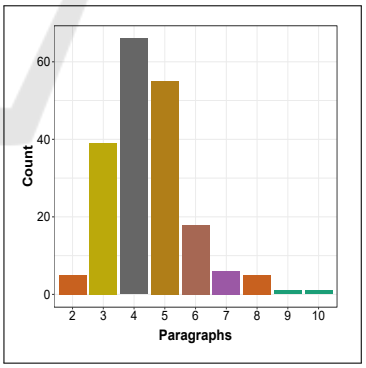


Figure 1: Distribution of natural paragraphs in a section.

Conversely, natural split of a section is more efficient, as searching an html element by ID can safely assume $O(1)$ on a modern browser with a hash table— a perfect data structure for element mapping. The time complexity of simple queries for looking up a class or a tag name is not worse than $O(n)$, and requesting an html element by a tag name is in many cases $O(1)$. Space complexity is consistently $O(1)$ as

[3] A section hash-tag url: https://en.wikipedia.org/wiki/Ian_McKellen#Charity_work.

[4] We note that the QuAC paper (Choi et al., 2018) cites larger portions of both unanswerable and affirmation questions— 20.2 and 22.1 percent, respectively.

[5] QuAC crowdworkers often use a subset of a section content, however we consistently acquire the entire section.

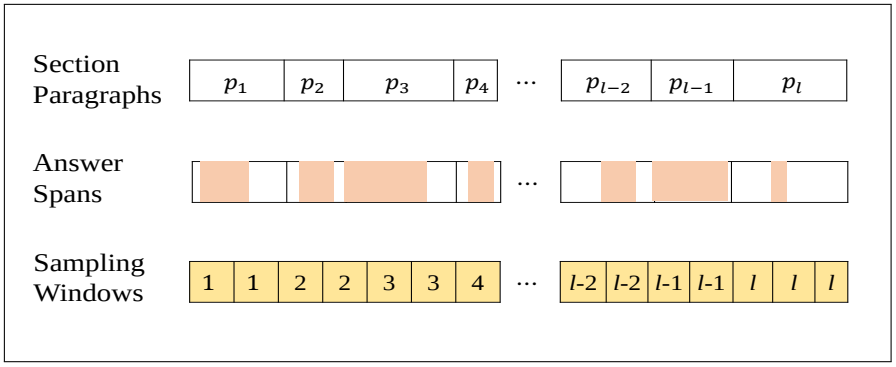


Figure 2: A uniformly sampled section by the retriever: shown are variable-size paragraph distribution (top), reference answer spans each confined to a distinct paragraph and may overlap either partially or entirely with the paragraph extent (middle), and a non-overlapping sampling window s_w sliding across paragraph spans identified by their indices (bottom).

the retriever retains a paragraph-worth text sequence of variable length.

6 MODEL

Our QA neural architecture uses an `AlBERT` (Lan et al., 2019) language model pretrained on a large Wikipedia knowledge base that is shared between an effective inline retriever and a reader. We chose `AlBERT` for its inter-sentence modeling that benefits a downstream QA task with multi-sentence inputs, and moreover, `AlBERT` addresses `BERT` scalability shortfall by lowering memory cost and improving runtime efficiency via parameter-reduction techniques. Following standard `BERT`-style transformers (Devlin et al., 2019), we concatenated spans of text by applying tokenization, delimiting sequences with `[SEP]` tokens, prefixing a `[CLS]` token, and appending a `[SEP]` token. Formally, the BERT function takes one or two string arguments x_i

$$\text{BERT}(x_1) = [\texttt{CLS}]\,x_1\,[\texttt{SEP}]$$
$$\text{BERT}(x_1, x_2) = [\texttt{CLS}]\,x_1\,[\texttt{SEP}]\,x_2\,[\texttt{SEP}]$$

and returns vector embeddings corresponding to representations of the `[CLS]` pooling token and each of the input words.

In a span-based QA system, the model is fed with a question q and a paragraph p_k retrieved from a Wikipedia article section $s_j^{(i)}$. Presuming p_k contains the answer, our task is to predict the answer as a contiguous sequence of tokens confined to the paragraph p_k. To `AlBERT` we represent the input question and the paragraph as a single packed sequence preceded by the `[CLS]` pooled token, and the underlying transformer outputs two hidden vectors $\in \mathbb{R}^H$

$$h_{start} = \text{BERT}(q, p_k)\,[START]$$
$$h_{end} = \text{BERT}(q, p_k)\,[END],$$

where $START$ and END are token indices of the span endpoints, and $START \leq END$. Applied to each section paragraph individually, candidate spans are scored with a multi-layer perceptron (MLP)— the simplest form of a feed-forward neural network— configured with four layers and 64 hidden units over the concatenation $[h_{start}; h_{end}]$ (Lee et al., 2019; Guu et al., 2020), and the highest scoring span is used as the answer prediction. During inference, the language model outputs the answer string a_s of the highest scoring derivation $f(\theta, q)$

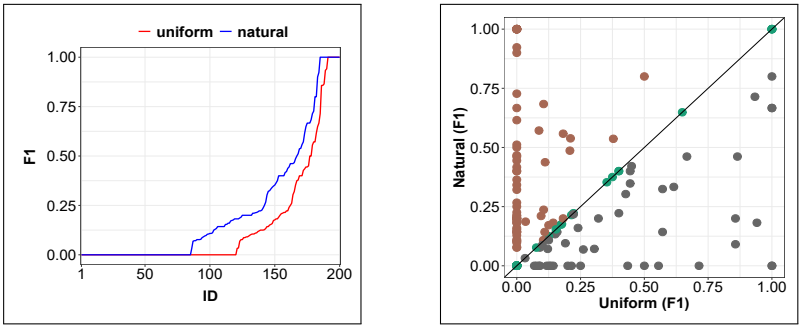
$$a_s = \text{text}(\underset{\theta}{\text{argmax}}\, f(\theta, q)),$$

where θ is a three-tuple construct of the indices $\{k, START, END\}$ — identifying a section paragraph and span endpoints — that we iterate over the prediction scoring function.

7 EXPERIMENTS

We analyzed the quality impact of natural context split on span-based QA systems, and compared our performance with the more established fixed-length and non-overlapping passage approach. Our results found no appreciable performance advantage to overlapped context splits, and concur with a similar observation made by Karpukhin et al. (2020). We note that in unified split mode the sampling window w_s is set to half the number of tokens provided in the context field of a `QuAC` example, and the base unit size resolves to $\min\{100, s_w\}$.

In our evaluation, we combined `AlBERT` language models pretrained on the large-scale `SQuAD 2.0` (Rajpurkar et al., 2018) corpus with knowledge distillation to a downstream inference task on our adaptation of the lower-resourced `QuAC` development set. An approach that was recently upheld by Yatskar (2019) ex-



(a) . (b) .

Figure 3: Inference F1 scores for (a) tracking and (b) correlating uniform and natural-grain context retrieval.

periments of knowledge transfer from a model pretrained on SQuAD 2.0 to the QuAC validation set, showing scores improving by about 2 F1 points compared to a BiDAF (Seo et al., 2017) baseline model augmented with self-attention (Clark and Gardner, 2018). Posed on a set of Wikipedia articles, the latest version of the SQuAD dataset has 150,000 questions of which 50,000 are open-ended and unanswerable. Presenting a much more realistic task, QA systems cross-evaluated on SQuAD 2.0 must not only answer questions when possible, but also indicate that no answer is possible by a single section paragraph and refrain from answering the question altogether.

We used token-based F1 score as our primary evaluation metric, where precision and recall are computed by considering the intersection of answer tokens in the prediction and reference text sequences, after removing stop-words. Our system assigns no answer questions an F1 of one if it correctly predicts no answer, and zero otherwise. We compute the maximum F1 among all paragraphs of a Wikipedia article section, and return an average section F1 over the entire dataset examples.

In our experiments, we report QA performance results using AlBERT-base-v2 and AlBERT-xlarge-v2 models. The network configurations, short of added customization, point to a defined contrast between 12 layers with 768 hidden units and 24 layers with 2,048 hidden units for the base and xlarge network architectures, respectively. The AlBERT models we used constrain the length of input text sequences to 512 tokens. However, both a retrieval of a fixed split into 100 token-long passages and natural paragraphs that top 449 words (Table 2) are clearly immune to this limitation. To improve visualization clarity of our results, we resorted to rendering a subset of randomly selected 200 samples from our flattened QuAC development set that comprises a total of 1,000 examples.

Table 3 summarizes our experimental results in percentage points providing F1, exact match (EM), and unanswerability rates. The pretrained xlarge model configuration consistently improves scores by about 3.5 F1 points compared to the base model. Moreover, a natural context split gains on average a marked 7.5 F1 points over a fixed passage retrieval and across both AlBERT architectures. Similarly, EM scores improve by up to 4.1 points using a natural context split, with the observation of a fine drawn performance edge to favor the xlarge model. Natural paragraph subdivision on AlBERT-xlarge achieved our best scores of 26.5 F1 and 8.5 EM.

In contrast to a generic AlBERT model pretrained on SQuAD 2.0 with cross domain transfer to a low-resourced QuAC validation set of 1,000 question-answer pairs, when trained on its native training set of 100,000 pairs using a highly customized BiDAF model, QuAC achieved a reasonable modest F1 score of 51.8 for no dialog context (Choi et al., 2018). Whereas Yatskar (2019) attained a cross transfer F1 score of 34.3 without fine-tuning on the QuAC train set, more in line with both our setup and results.

Table 3: Scores comparing fixed-length with natural context splits. The higher the score the better for F1 and EM metrics and conversely for unanswerability.

	Fixed Split		Natural Split	
AlBERT Model	base	xlarge	base	xlarge
F1	15.5	19.2	23.1	**26.5**
EM	4.2	4.5	8.3	**8.5**
Unanswered	26.3	24.1	8.5	**7.2**

We compared the distribution of F1 scores between uniform and naturally-split context retrieval. The plots in Figure 3a visually identify natural split to outperform uniform subdivision with an area-under-curve of 0.67 vs 0.55, thus improving scores on average by about 22 percentage points. Additionally, we conducted qualitative error analysis and in Figure 3b, we review F1 score correlation between fixed and

natural section splits on the AlBERT baseline model. Notably there are 52 correct answerable questions in a natural split predicted unanswerable with fixed-size text sequences. On the other hand, there are only 17 answered questions in a uniform split deemed unanswerable in a context preserving split.

8 DISCUSSION

We compared our system performance with external baselines at two levels. First, we ran inference using a large BERT model that has 334M network parameters and was pretrained on SQuAD 2.0. This proved to fairly match our performance on an xlarge AlBERT model with only 60M parameters. Second, we sought to contrast our performance against existed open QA systems. However, ORQA (Lee et al., 2019), Multi-Passage BERT (Wang et al., 2019), DPR (Karpukhin et al., 2020), and REALM (Guu et al., 2020), all use SQuAD 1.1 that has no generalization of unanswerable questions. Unlike SQuAD 2.0 that incurs a much larger gap between humans and machine comprehension (Rajpurkar et al., 2018), thus making it challenging to compare performance evenhandedly. For example, combining BERT with QANet (Yu et al., 2018) in (Wang et al., 2019) achieved a 27.8 F1 score compared to 26.5, our best.

Table 4: Statistics summarizing paragraph answerability in a section context.

Paragraph	Min	Max	Mean	SD
answerable	0	10	1.6	1.5
unanswerable	1	12	5.6	2.2

In this section, we follow with a brief review of paragraph answerability. Our proposed natural split approach suggests at least theoretically a more simpler task for distinctly selecting the section paragraph that completely answers the question. Ideally, our system would only point to a single paragraph that contains the answer span, along with the rest of the section paragraphs predicted as no answer and thus returning [CLS]. Table 4 shows statistics of paragraph answerability to support our conjecture with 1 to 2 and 5 to 6 answerable and unanswerable paragraphs per section, respectively. Figure 4 expands visually on the distribution of paragraph answerability.

Our work motivates inline retrieval of article sections from Wikipedia rather than scraping an outdated dump. To parse a section html we only make use of the article and section title fields defined in a QuAC QA object, but except the dialog context text altogether. [6] This on its own reduced the memory footprint of our QA dataset by about 89% to found effective communication between the retriever and the reader. Runtime wise, we observed on average 1.59 and 1.72 iterations per second for reading context from local files and issuing online Wikipedia requests, respectively. Although the section paragraphs we retrieved never exceeded the neural reader input constraint of 512 token-length, our system provides paragraph simplification in the event it does, to avoid further chunking.

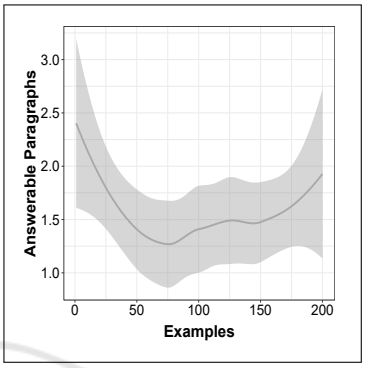


Figure 4: Distribution of answerable paragraphs.

Using natural section paragraphs is markedly more intuitive in the challenging task of multi-turn question answering. In this scenario, each answer to a conversational question expects span-based reasoning in a context-preserving passage, and not only that, retrieved section paragraphs present a dependency conditioned on previously obtained paragraphs. Hence models must render a well-defined pattern to access the entire dialog context (Choi et al., 2018).

Similarly, context-aware retrieval has the advantage of seamless integration into structured reasoning using knowledge graphs (KG), where entities are represented as nodes and relations between them as edges. To this extent, our QA data example (Table 1) facilitates KG representation by implicitly retaining the links of a section to both its parent article and down to the paragraph siblings at the leaf nodes. Reasoning over KG has been recently explored by Asai et al. (2020), who uses recursive retrieval of reasoning paths over Wikipedia hyperlinks, and Yasunaga et al. (2021) leverage both MLM and KG to address negation in questions.

[6] The section title in a QuAC QA object often overlooks annotations that we edited manually. For example, replacing a hyphen with an n-dash in a year-range expression.

9 CONCLUSION

In this work, we presented the more theoretically-sound context-aware partition of Wikipedia article sections for span-based question answering. We showed favorable quality gains to context-preserving over incidental context-unaware passages, even after recasting the latter for robust sampling. Our empirical analysis of an information-seeking QA task substantiated our contention that a pivotal answerable paragraph is distinctly identified among the remaining section paragraphs predicted unanswerable. Using a sustainable context-less QA dataset proved to incur an inconsequential runtime cost of inline queries.

We look forward to the research community for a broader acceptance of data split as a core retriever functionality, and follow with the endorsement of natural context retrieval that deems essential for dialog and multi-hop QA. We envision the incorporation of neural text simplification to further improve the efficacy of QA tasks.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

REFERENCES

Asai, A., Hashimoto, K., Hajishirzi, H., Socher, R., and Xiong, C. (2020). Learning to retrieve reasoning paths over wikipedia graph for question answering. In *Learning Representations, (ICLR)*, Addis Ababa, Ethiopia.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1870–1879, Vancouver, Canada.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2174–2184, Brussels, Belgium.

Clark, C. and Gardner, M. (2018). Simple and effective multi-paragraph reading comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 845–855, Melbourne, Australia.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186, Minneapolis, Minnesota.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020). Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning (ICML)*, Online.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942. http://arxiv.org/abs/1909.11942.

Lee, K., Chang, M.-W., and Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6086–6096, Florence, Italy.

Li, C. and Choi, J. D. (2020). Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5709–5714, Online.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692. http://arxiv.org/abs/1907.11692.

Qu, C., Yang, L., Chen, C., Qiu, M., Croft, W. B., and Iyyer, M. (2020). Open-retrieval conversational question answering. In *Conference on Research and Development in Information Retrieval (SIGIR)*, page 539–548, New York, NY. Association for Computing Machinery.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2018). Language models are unsupervised multitask learners. Technical report.

Rajpurkar, P., Jia, R., and Liang, P. (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)*, pages 784–789, Melbourne, Australia.

Reddy, S., Chen, D., and Manning, C. D. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Seo, M., Lee, J., Kwiatkowski, T., Parikh, A., Farhadi, A., and Hajishirzi, H. (2019). Real-time open-domain

question answering with dense-sparse phrase index. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy.

Seo, M. J., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. In *Learning Representations, ICLR*, Toulon, France.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008. Curran Associates, Inc., Red Hook, NY.

Wang, H. and Poon, H. (2018). Deep probabilistic logic: A unifying framework for indirect supervision. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1891–1902, Brussels, Belgium.

Wang, Z., Ng, P., Ma, X., Nallapati, R., and Xiang, B. (2019). Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., and Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 535–546, Online.

Yatskar, M. (2019). A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2318–2323, Minneapolis, Minnesota.

Yu, A. W., Dohan, D., Le, Q., Luong, T., Zhao, R., and Chen, K. (2018). Fast and accurate reading comprehension by combining self-attention and convolution. In *Learning Representations (ICLR)*.

Yu, D., Wang, H., Sun, K., Chen, J., Yu, D., McAllester, D., and Roth, D. (2019). Evidence sentence extraction for machine reading comprehension. In *Computational Natural Language Learning (CoNLL)*, pages 696–707, Hong Kong, China.