WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval

Daniel Cohen Liu Yang W. Bruce Croft
Center for Intelligent Information Retrieval, University of Massachusetts Amherst, Amherst, MA, USA
{dcohen,lyang,croft}@cs.umass.edu

ABSTRACT

With the rise in mobile and voice search, answer passage retrieval acts as a critical component of an effective information retrieval system for open domain question answering. Currently, there are no comparable collections that address non-factoid question answering within larger documents while simultaneously providing enough examples sufficient to train a deep neural network. In this paper, we introduce a new Wikipedia based collection specific for non-factoid answer passage retrieval containing thousands of questions with annotated answers and show benchmark results on a variety of state of the art neural architectures and retrieval models. The experimental results demonstrate the unique challenges presented by answer passage retrieval within topically relevant documents for future research.

KEYWORDS

Non-factoid Answer Passage Retrieval; Question Answering; Benchmark Dataset

1 INTRODUCTION

Recent advances in deep learning have allowed recent work in numerous fields to achieve state of the art performance on key tasks, with larger networks often outperforming smaller networks after accounting for overfitting. However, these deep neural networks contain millions of parameters even with only a small number of layers that necessitates a large amount of training data compared to more conventional models. As such, high quality openly available benchmark data sets are critical for research progress. Examples include ImageNet [2] for computer vision and SQuAD [11] for machine comprehension. Large, high quality datasets allow the community to not only rapidly develop new models for a task, but also to iteratively learn how a model architecture learn better representations for a specific task.

With the rising popularity of mobile and voice assisted search, where the size of screen and the output length is limited, there is a growing need to develop models for retrieving answer passages. Here, the information need of a query lies between that of a short fact or single sentence, and a document, and cannot be sufficiently answered with either. In terms of question answering, there are existing datasets such as TREC QA [17], WikiQA [21] and InsuranceQA [3] that provide sufficient collections of queries to train a neural network [14, 16, 19, 22]. However, these datasets do not address the answer passage retrieval task since their focus is on retrieving factoids, short snippets, or isolated sentences. In the collection introduced by this paper, the task is not only to retrieve a passage that answers the question, but also to identify where the answer portion of a document begins and ends within a larger topically relevant document.

Currently, there is only one collection specifically created for retrieving answer passages in documents, WebAP [7], where contiguous sentences of a document are labeled as relevant to a query. While addressing the answer passage retrieval task, the WebAP collection suffers from a small number of queries, resulting in poor performance of neural models.

In this paper we present a new collection, WikiPassageQA, containing 4,187 queries created from Amazon mechanical turk[1] over the top 863 Wikipedia documents from the Open Wikipedia Ranking[2]. Each Wikipedia page has multiple queries accompanied with locations of varying length answer passages within the document. As this facilitates numerous representations for evaluating passage retrieval methods, we choose a sliding window method to demonstrate that this collection is sufficient to train deep neural models that outperform standard baselines.

The contributions of this work are as follows: (1) We introduce a new benchmark collection for the research on non-factoid answer passage retrieval[3]. (2) We perform extensive experiments with WikiPassageQA to show benchmark results of various methods including traditional and neural IR models that demonstrate the unique challenges that differentiate answer passage retrieval from past QA tasks.

2 EXISTING RELATED DATASETS

We perform a survey of related question answering and reading comprehension data sets to highlight the differences between them and WikiPassageQA.

Factoid Question Answering: There are several benchmark data sets for the evaluation of factoid question answering, which aim to identity short answer facts such as named entities, numbers and noun phrases. Wang et al. [17] developed a benchmark collection using the Text REtrieval Conference (TREC) 8-13 QA data. They used the questions in TREC 8-12 for training and set aside TREC 13 questions for development (84 questions) and testing (100 questions). This TREC QA data set has become one of the most widely used benchmarks for answer sentence selection [14, 16, 19, 22]. Recently, Yang et al. [21] created the WikiQA dataset using Bing query logs and Wikipedia passages as the source of answers. WikiQA

SIGIR '18, Ann Arbor Michigan, U.S.A.

DOI: 10.11.23/-----

[1] https://www.mturk.com/mturk/welcome
[2] http://law.di.unimi.it/
[3] The data set can be downloaded from T.B.A.

data is more than an order of magnitude larger than the previous TREC QA data. Feng et al. [3] created InsuranceQA, which is a data set in the insurance domain. It consists of questions from real world users and answers composed by professionals with deep domain knowledge about insurance. These data sets either only include very short answers and answer sentences for factoid questions, or only for a closed domain like insurance. However, the WikiPassageQA data proposed in this paper includes many long passages for non-factoid questions and there are no restricted domains for these questions and answers.

Non-Factoid Question Answering: There have been previous efforts on developing benchmark data sets for non-factoid question answering or answer passage retrieval [4, 7, 20]. Perhaps the closest prior research to our work is the WebAP data set created by Keikha et al. [7, 20]. Compared to WebAP, WikiPassageQA has a two significant differences: (1) the number of questions in WikiPassageQA is significantly larger than that of WebAP (4187 v.s. 82). (2) WikiPassageQA has different properties on the specificity of queries. WebAP used previous TREC tropical queries whereas WikiPassageQA has questions with more focused information needs.

There are also non-factoid QA data built from community question answering (CQA) data. The most commonly known of these are the Yahoo L4 "manner" questions and a filtered non-factoid collection from the entire Yahoo L6 Webscope collection (nfL6)[1]. While both CQA collections and WikiPassageQA target non-factoid questions, there are two significant differences between them.

(1) The candidate answers from the CQA collections either come from other questions, which may not have any semantic relationship to the target query, or come from "non-best" answers submitted in response to the query. These candidate answers have unreliable and generally missing labels. This noise in the relevance judgements leads to unreliable training and testing.

(2) As opposed to WikiPassageQA, these CQA collections consist of answer passages without surrounding text. This results in a much easier task due to the greater difference between candidate CQA answers than neighboring passages within a Wikipedia document.

Reading Comprehension: The other related data sets are reading comprehension data sets including MCTest [12], CNN /Daily News [5], Children's Book Test [6], SQuAD [11], MS MARCO [9], BAbI [18], etc. Unlike answer sentences or passages in the question answering datasets, these reading comprehension data sets mostly involve selecting a specific short span within a sentence, selecting an answer from predefined choices, or predicting a blanked-out word of a sentence given previous context sentences. WikiPassageQA stands apart by using only user annotated answer passages rather than synthetic data, and most accurately reflects the task of finding raw answer passages within a larger document.

In summary, WikiPassageQA is the only large data set with long passages as answers for thousands of non-factoid questions in the open domain.

3 THE WIKIPASSAGEQA DATASET

3.1 Query And Answer Passage Synthesis

The dataset was created using Amazon's mechanical turk platform, where we sourced high quality crowd workers to create questions based on a Wikipedia document. We restricted workers to have over 1000 assignments completed as well as having over a 98% approval rating to ensure quality submissions. While workers were able to work on multiple human intelligence tasks (HITs), no worker was able to submit twice on the same Wikipedia page. In a similar manner to the creation of the SQuAD collection [11], each worker was asked to create five non-factoid questions and indicate location of their respective answer passages within the document. "Who", "Where", and "When" questions were explicitly prohibited to prevent factoid answers. A relevant passage was deemed to be more than one contiguous sentence, with no additional information that doesn't address the query. In order to prevent low quality submissions, workers were able to submit less than five queries if the document was not suitable for the task. Workers were paid $0.65 per HIT and the total cost of data annotation was $638.

Table 1: WikiPassageQA collection statistics. "P" in the first column denotes "Passages".

Data	Train	Dev	Test	Total
Questions	3349	419	419	4187
CandidateP	194231	25807	23981	244019
PosCandidateP	5556	705	700	6961
NegCandidateP	170505	24746	23043	218294
% of PositiveP	0.029	0.027	0.029	0.029
CandidateP/Query	58.318	62.036	57.647	58.616
PosCandidateP/Query	1.659	1.683	1.671	1.663
AvgLenOfQuestion	10.752	10.852	10.420	10.729
AvgLenOfAnswerP	141.793	147.885	144.732	142.698

3.2 Evaluating Answer Passage Quality

Once a batch of question and answer passages was completed, they were resubmitted to the Amazon mechanical turk platform in a verification poll. For each question and assignment passage from an assignment, five workers were asked to provide two ratings: (1) rate the question as *factoid*: 0, *non-factoid*: 1 and (2) the answer passage as *Excellent, Great, Fair, Poor* with point values 3, 2, 1, 0 respectively. The Kappa coefficient of question type was 0.930 and 0.659 for factoid/non-factoid and answer passage quality during this evaluation process, which indicates good agreement score among different annotators. Question-answer passage pairs were removed if mean scores for these two ratings were, respectively, less than 0.66 and 2 to ensure quality. This filtering process reduced the original collection of question-answer passage pairs from 4908 to 4187 pairs.

3.3 Collection Characteristics

As seen in Table 1, the filtered collection possesses annotated answer passages significantly longer than previous QA datasets. Breaking down the queries by the first word of the question, "what", "how" and "why" make up 43.8%, 36.6%, and 14.0% of the collection. The next most common start word is "in" at 1.2%, acting as a prepositional phrase for the question. Across all question words, Figure 1 shows that the answer passages have a similar length distribution with 99.9% of all passages having less than 400 words. As there is only one relevant passage for each question, there is a risk of

false negative passages. However, due to the specific prompt of requiring the information need of the query spread over multiple sentences, the relevant passages are highly likely to be unique to the Wikipedia page. A comparison of sample question and annotated answers is provided in Table 2 between WikiPassageQA and TREC QA data.

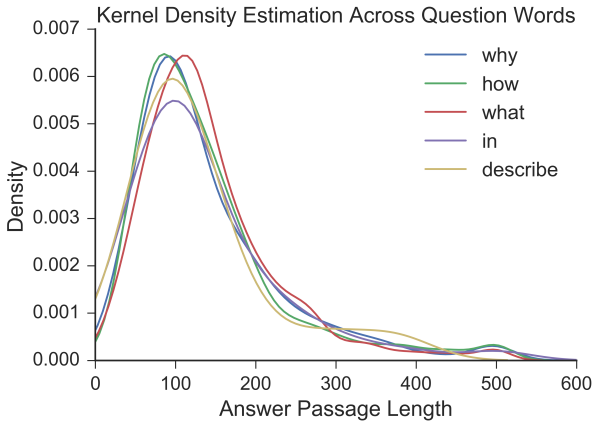
Table 2: Comparison of example questions and answers in TREC QA Track data and WikiPassageQA data.

Sample Questions and Answers in TREC QA Track Data
Query 201: Question: What was the name of the first Russian astronaut to do a spacewalk? Answer: Aleksei A. Leonov Answer Document ID: LA072490-0034
Query 202: Question: Where is Belize located? Answer: Central America Answer Document ID: FT934-14974
Query 203: Question: How much folic acid should an expectant mother get daily? Answer: 400 micrograms Answer Document ID: LA061490-0026
Sample Questions and Answer Passages in WikiPassageQA Data
Query 4114: Question: Why is Japan so densely populated? Document ID: 496 Document Name: Japan.html Answer Passages: The main islands, from north to south, are Hokkaido, Honshu, Shikoku and Kyushu. The Ryukyu Islands, which include Okinawa, are a chain to the south of Kyushu. Together they are often known as the Japanese archipelago. About 73% of Japan is forested, mountainous, and unsuitable for agricultural, industrial, or residential use. As a result, the habitable zones, mainly located in coastal areas, have extremely high population densities. Japan is one of the most densely populated countries in the world.
Query 2402: Question: What is the structure of Australia's members of parliament? Document ID: 400 Document Name: Member_of_parliament.html Answer Passages: **Passage 1** A Member of Parliament is the representative of the voters to a parliament. In many countries with bicameral parliaments, this category includes specifically members of the lower house, as upper houses often have a different title. Members of parliament tend to form parliamentary groups with members of the same political party. The Westminster system is a democratic parliamentary system of government modelled after the politics of the United Kingdom. This term comes from the Palace of Westminster, the seat of the Parliament of the United Kingdom. A member of parliament is a member of the House of Representatives, the lower house of the Commonwealth parliament. Members may use "MP" after their names; "MHR" is not used, although it was used as a post-nominal in the past. **Passage 2** A member of the upper house of the Commonwealth parliament, the Senate, is known as a "Senator". In the Australian states of New South Wales, Victoria and South Australia, a Member of the Legislative Assembly or "lower house," may also use the post-nominal "MP." Members of the Legislative Council use the post-nominal "MLC." Members of the Jatiyo Sangshad, or National Assembly, are elected every five years and are referred to in English as members of Parliament. The assembly has directly elected 300 seats, and further 50 reserved selected seats for women. The Parliament of Canada consists of the monarch, the Senate , and the House of Commons.

3.4 Data Overview and Experimental Settings

As the collection consists of queries and relevant contiguous sentences, there are a variety of ways to evaluate models. In order to benchmark the dataset on common IR models for answer passage retrieval, we segment each Wikipedia article into passages of six sentences each, which is the average number of sentences in all annotated answer passages. As relevant passages can be split between windows, we deem a candidate passage as relevant if greater than 15% of the bigrams within the annotated answer passage occurs in a candidate passage. This results in an average of 1.66 relevant passages for each query. As each Wikipedia document is distinct, at retrieval time only candidate passages from the target query's Wikipedia page were used in training and evaluation rather than

Figure 1: Distribution of Answer Passage Lengths.



all passages in the entire collection. Training of the neural models were done with a 0.8/0.1/0.1 split for training, development, and testing sets resulting in 3349, 419, and 419 queries for each set. As this is an IR task, we partition the queries rather than Wikipedia articles common in reading comprehension tasks [11].

3.5 Learning Models and Evaluation Metrics

We benchmark our dataset on two naive baselines, three traditional IR methods, and five deep neural models as shown in Table 3.

Baselines. These two methods (WC, WC.IDF) examine the performance using overlapped word count statistics between the question and the candidate answer passage to provide a reference point for other methods. WC.IDF is the overlapped word count statistics weighted by IDF. It can be viewed as an unnormalized TF-IDF summation.

Traditional IR Models. The traditional IR models include the TF-IDF Vector Space Model (VSM), BM25 [13], and Query Likelihood (QL) [10] with Dirichlet smoothing. These models will show the performances of traditional IR baselines for answer passage retrieval.

Neural IR Models. Five neural models are used to evaluate answer passage retrieval with this collection: **(1)** A standard two layer **LSTM** network [1, 16] is used as a simple model to benchmark a strong non factoid neural model. **(2) CNN+TF** adopts siamese convolutional neural networks to learn representations of questions and candidate answer passages. The QA pairs are concatenated along with *tf* information after the CNN subnetwork, and passed through a feedforward network to produce a scalar relevance score, which is the approach proposed by Severyn and Moschitti [14]. **(3) LSTM-CNN+TF** adds a LSTM layer for the long term dependency modeling prior to a CNN [15]. This approach reflects the impact of explicitly modeling the passage as a temporal structure on ranking. **(4) Char+Word-CNN-LSTM** possesses the same structure as (2), but utilizes character embeddings to deal with out of vocabulary instances. **(5) Memory-CNN-LSTM** model uses a doc2vec [8] representation as its starting memory tensor, and iteratively reads and writes from it at each sentence within a passage. This includes

Table 3: Benchmark results of different methods on WikiPassageQA. Numbers in bold font mean the result is better compared with the best baseline.

Type	Method	MAP	MRR	P@5	P@10	nDCG	Recall@5	Recall@10	Recall@20
Base	WC	0.3456	0.4004	0.1370	0.0923	0.5096	0.4618	0.6079	0.7615
	WC.IDF	0.3417	0.3898	0.1351	0.0928	0.5049	0.4518	0.6129	0.7526
Traditional IR	VSM	0.3970	0.4588	0.1476	0.0921	0.5490	0.4837	0.5979	0.7464
	BM25	0.5373	0.6258	0.1947	0.1151	0.6659	0.6334	0.7311	0.8309
	QL	0.5436	0.6338	0.1947	0.1151	0.6715	0.6353	0.7275	0.8426
Neural IR	LSTM	0.3352	0.3947	0.1197	0.0780	0.4912	0.3915	0.5894	0.7169
	CNN+TF	0.4009	0.4581	0.1572	0.1099	0.5577	0.5212	0.7024	0.8412
	LSTM-CNN+TF	0.3577	0.4156	0.1351	0.0942	0.5196	0.4538	0.6187	0.7608
	Char+WordCNN-LSTM	0.4385	0.5534	0.1728	0.1104	0.5837	0.5709	0.6931	0.8326
	Memory-CNN-LSTM	**0.5608**	**0.6792**	**0.2083**	**0.1228**	**0.6791**	**0.6522**	**0.7329**	**0.8592**

tf information at the sentence level, and takes into account the probability of the sentence generating each term as well.

3.6 Experimental Results and Analysis

As seen in Table 3, traditional IR models like QL achieve a very competitive baseline, outperforming all but one of the neural models. Memory-CNN-LSTM outperform all other methods including traditional IR models and neural IR models. Only Memory-CNN-LSTM was developed for answer passage retrieval of this length, where it sequentially iterates through each sentence while updating a memory tensor. All other neural models were designed for retrieving either sentences or passages with a mean approximate length of 50 tokens. This contrasts sharply with the characteristics of WikiPassageQA, shown in Table 1, where the mean length of an answer passage is 142.7 tokens. Similar to the results shown in [1], CNN+TF fails to outperform a standard BM25 baseline, indicating the difficulty of neural IR architectures generalizing to new tasks at a different text granularity. The relatively poor performance of these conventional neural networks indicates the unique challenges present in the non-factoid answer passage retrieval task. The WikiPassageQA collection provides an open benchmark data with answer correctness judgments to the research community for non-factoid answer passage retrieval. We will make our dataset freely available to encourage exploration of more expressive models.

4 CONCLUSIONS AND FUTURE WORK

Answer passage retrieval within topically relevant documents shows unique challenges not present in other QA collections. Until this collection, there were no previous answer passage retrieval collections available that were suitable for the exploration of deep learning models. We presented this new collection and benchmarks to provide an openly available resource so that others can extend our research on non-factoid answer passage retrieval. For the future work, we will study more different neural architectures for non-factoid answer passage retrieval. Answer summarization for non-factoid QA is also an interesting direction to explore.

5 ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF #IIS-1160894, and in part by NSF grant #IIS-1419693. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Daniel Cohen and W. Bruce Croft. End to End Long Short Term Memory Networks for Non-Factoid Question Answering. In *ICTIR '16*.
[2] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.
[3] Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying Deep Learning to Answer Selection: A Study and An Open Task. *CoRR* abs/1508.01585 (2015).
[4] Ivan Habernal, Maria Sukhareva, Fiana Raiber, Anna Shtok, Oren Kurland, Hadar Ronen, Judit Bar-Ilan, and Iryna Gurevych. New Collection Announcement: Focused Retrieval Over the Web. In *SIGIR '16*.
[5] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. *CoRR* abs/1506.03340 (2015).
[6] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *CoRR* abs/1511.02301 (2015).
[7] Mostafa Keikha, Jae Hyun Park, and W. Bruce Croft. 2014. Evaluating Answer Passages Using Summarization Measures. In *SIGIR '14*.
[8] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014). http://arxiv.org/abs/1405.4053
[9] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016).
[10] Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR '98*.
[11] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *CoRR* abs/1606.05250 (2016).
[12] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *EMNLP '13*.
[13] Stephen Robertson and Stephen Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR '94*.
[14] Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *ACM SIGIR (SIGIR '15)*. ACM, New York, NY, USA, 373–382. DOI: http://dx.doi.org/10.1145/2766462.2767738
[15] Ming Tan, Bing Xiang, and Bowen Zhou. 2015. LSTM-based Deep Learning Models for non-factoid answer selection. *CoRR* abs/1511.04108 (2015). http://arxiv.org/abs/1511.04108

[16] Di Wang and Eric Nyberg. A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering. In *ACL '15*.
[17] Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. In *EMNLP '07*.
[18] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *CoRR* abs/1502.05698 (2015).
[19] Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Bruce Croft. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In *CIKM '16*.
[20] Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W. Bruce Croft, Jiafeng Guo, and Falk Scholer. Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval. In *ECIR '16*.
[21] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *EMNLP '15*.
[22] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep Learning for Answer Sentence Selection. *CoRR* (2014).