Investigating the Effect of Background Knowledge on Natural Questions

Vidhisha Balachandran[1*] **Bhuwan Dhingra**[2] **Haitian Sun**[1*]
Michael Collins[2] **William W. Cohen**[2]
[1] Language Technologies Institute, Carnegie Mellon University
[2] Google Research
{vbalacha, haitians}@cs.cmu.edu
{bdhingra, mjcollins, wcohen}@google.com

Abstract

Existing work shows the benefits of integrating KBs with textual evidence for QA only on questions that are answerable by KBs alone (Sun et al., 2019). In contrast, real world QA systems often have to deal with questions that might not be directly answerable by KBs. Here, we investigate the effect of integrating background knowledge from KBs for the Natural Questions (NQ) task. We create a subset of the NQ data, Factual Questions (FQ), where the questions have evidence in the KB in the form of paths that link question entities to answer entities but still must be answered using text, to facilitate further research into KB integration methods. We propose and analyze a simple, model-agnostic approach for incorporating KB paths into text-based QA systems and establish a strong upper bound on FQ for our method using an oracle retriever. We show that several variants of Personalized PageRank based fact retrievers lead to a low recall of answer entities and consequently fail to improve QA performance. Our results suggest that fact retrieval is a bottleneck for integrating KBs into real world QA datasets[1].

1 Introduction

Prior work has shown the benefit of retrieving paths of related entities (Sun et al., 2018; Wang and Jiang, 2019; Sun et al., 2019) and learning relevant knowledge graph embeddings (Sawant et al., 2018; Bordes et al., 2014; Luo et al., 2018) for answering questions on KBQA datasets such as WebQuestions (Berant et al., 2013) and MetaQA (Zhang et al., 2018). But such datasets are often curated to questions with KB paths that contain the right path to the answer and hence are directly answerable via KB. An open question remains whether such approaches are useful for questions not specifically designed to be answerable by KBs. In this paper, we aim to evaluate KB integration for real-world QA settings in the context of the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) which consists of questions naturally posed by users of a search engine. NQ is one of the common benchmarks that is used to test the real-world QA applicability of models, hence motivating our choice.

To study the effect of augmenting KB knowledge, we construct a subset of NQ - *Factual Questions* (FQ). In FQ, answer entities are connected to question entities via short paths (up to 3 steps) in the Wikidata KB (Vrandečić and Krötzsch, 2014). Using FQ, we analyze a simple but effective approach to incorporating KB knowledge into a textual QA system. We convert KB paths to text (using surface forms of entities and relation) and append it to the textual passage as additional context for a BERT-based (Devlin et al., 2019) QA system.

We first establish an upper bound oracle setting by building a retriever that provides the shortest path to an answer. We show that, in the presence of such knowledge, our approach leads to significant gains (up to 6 F1 for short-answers, 8-9 F1 for multi-hop questions). We experiment with several variants of KB path-retrieval methods and show that retrieving good paths is difficult: previously-used Personalized PageRank (Haveliwala, 2003) (PPR)-based methods find answer entities less than 30% of the time, and even our weakly-supervised improvements recall answer entities no more than 40% of the time. As a consequence injecting retrieved KB paths in a realistic QA setting like NQ yields only small, inconsistent improvements.

To summarize our contributions, we (1) identify a new experimental subset of NQ that supports (2) the study of effectiveness of KB path-retrieval approaches. We also (3) describe a simple, model-agnostic method to using oracle KB paths that can significantly improve QA performance and evaluate PPR based path-retrieval methods. To our

[1] Data and Code available at: https://github.com/vidhishanair/fact_augmented_text
[*] Work done at Google Research

Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 25–30
Online, June 10, 2021. ©2021 Association for Computational Linguistics

knowledge this is the first study of such approaches on a QA dataset not curated for KBQA.

2 Dataset

The Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) is a large scale QA dataset containing 307,373 training, 7,830 dev, and 7,842 test examples. Each example is a user query paired with Wikipedia documents annotated with a passage (long answer) answering the question and one or more short spans (short answer) containing the answer. The questions in NQ are not artificially constructed, making the NQ task more difficult (Lee et al., 2019). We use Sling (Ringgaard et al., 2017) (which uses an NP chunker and phrase table for linking entities to Wikidata) to entity link the questions and documents.

To focus on knowledge-driven factoid question answering, we create a subset of NQ having relevant knowledge in the KB. Shortest paths between entities in KB is very often used as a proxy for gold knowledge linking questions to answer (Sun et al., 2019) and we use the same proxy in our setting. Specifically, we select questions whose short answers are entities in the KB and have a short path (up to 3 steps) from a question entity to an answer entity. These paths contain knowledge relevant to the question but are not necessarily the right path to answer the question. We call this subset *Factual Questions (FQ)* containing 6977 training, 775 dev and 264 (83 1-hop, 97 2-hop and 84 3-hop) test samples. FQ being an entity centric subset of NQ, provides a setting to investigate augmenting KB paths for real-world factoid question for which relevant knowledge exists in the KB. Examples of the dataset are provided in Table 4.

3 Model

Given a question Q, our knowledge retriever extracts top facts from a KB. We represent them in natural language form and augment it to a standard BERT model for reading comprehension as additional context along with the passage P.

3.1 Knowledge Retriever

The Knowledge Retriever (KR) uses the input question Q to retrieve relevant facts for augmentation. We use the entities in the question as the set of seed entities denoted as E and use the Personalized PageRank (PPR) algorithm to perform a random walk over the KB to assign relevance scores to other entities around the seed entities.

The *Traditional PPR* algorithm takes the seed entities and iteratively jumps from and expands the seed entities until convergence. At each iteration, a transition with probability γ is made to a new entity in the KB (with all outgoing edges having equal weight) and a transition with probability $1 - \gamma$ is made to the start seed entities. The stationary distribution of this walk gives the relevance scores (PPR weights) of entities (nodes) w.r.t seed entities. Sun et al. (2018) present an improved PPR version, *Question Informed (QI) PPR*, to weigh relations which are semantically closer to the question higher. Specifically, they average the GLOVE (Pennington et al., 2014) embeddings to compute a relation vector $v(R)$ from the relation surface form, and a question vector $v(Q)$ from the question text, and use cosine similarity between them as edge-weights for PPR. For every node, the γ probability is multiplied by the edge-score to weigh entities along relevant paths higher.

We improve on this setting by introducing *Weakly Supervised (WS) PPR*, which uses weak supervision from the QA pairs to train a classifier to discriminate relevant relations from irrelevant ones. We create a classification dataset of questions aligned with relations along the shortest KB path connecting question entities and answer entities as positive relevant examples. Other random relations connected to the question entities form negative examples. We train a simple BERT based classifier to classify relations as relevant or irrelevant conditioned on the question. The trained classifier is used to score relations for every question and used as edge-weights for PPR similar to QI PPR. Examples of the facts retrieved from WS PPR are provided in Table 4.

After running PPR we retain the top-K entities $e_1, \ldots, e_K$ by PPR score, along with any edges between them. To further rank the facts, we compute entity scores as the sum of the PPR score and frequency of the entity in the text and aggregate the subject and object entity scores by taking the maximum score between them.

Oracle Setting: In this upper bound setting for the Knowledge Retriever, the answer entities are known. The facts along the shortest path connecting the question entities and the answer entities are considered as gold or relevant facts to the question and are shuffled and augmented to the input of the

	Factual NQ		Hop 1		Hop 2		Hop 3	
	Short F1	Long F1	Short F1	Long F1	Short F1	Long F1	Short F1	Long F1
Text Only	68.2	77.3	77.3	82.2	60.0	74.3	60.2	73.4
Text + PPR(Q) facts	68.1	77.8	78.3	83.7	57.9	72.8	61.9	75.7
Text + QI PPR(Q) facts	68.2	77.5	79.2	83.9	55.2	72.0	58.9	74.4
Text + WS PPR(Q) facts	67.8	76.3	76.9	81.7	58.1	72.5	60.2	72.1
Text + Clean Oracle	74.9	80.8	79.5	83.0	69.1	80.2	72.4	77.2
Text + Noisy Oracle	75.3	81.3	80.7	84.4	69.7	80.2	71.9	77.2

Table 1: Results on FQ data compared to Alberti et al. (2019). Both Clean and Noisy Oracle setting improve over only text baseline setting. Variants of PPR do not improve over the text only baseline.

	Shortest Path Fact R	Ans R
BM25	19.1	29.8
PPR(Q)	33.0	28.8
QI PPR(Q)	31.2	25.2
WS PPR(Q)	**51.0**	**40.0**

Table 2: Answer Recall and Shortest Path Fact Recall metrics for the different Retrieval Methods. Traditional and QI PPR methods have very low recall and WS PPR method improves the recall significantly.

QA model in place of the KB retrieved facts. As the oracle setting uses gold KB links, this setting is tested on the FQ subset where such links exist and is called the Clean Oracle. To establish a harder upper bound setting, random facts about the question are added in addition to the oracle shortest path facts using PPR, forming a Noisy Oracle setting.

3.2 Knowledge Augmented Text for QA

Given a ranked set of triples from the retriever, a natural language statement is constructed from each fact using the surface form of the entities e_s and e_o and the natural language description of R (e.g. “Washington D.C capital of United States”) similar to Lauscher et al. (2020). These form the background knowledge to be injected F. We then tokenize them using the standard BERT tokenizers and augment them to the input of QA model as X = “ [CLS] Question tokens [SEP] Passage tokens [SEP] Fact tokens”.

Following Alberti et al. (2019), we use a simple BERT architecture by training two linear classifiers independently on top of the output representations of X for predicting the answer span boundary (start and end). We assume that the answer, if present, is contained only in the given passage, P, and do not consider potential mentions of the answer in the background F. For instances which do not contain the answer, we simply set the answer span to be the special token [CLS]. We use a fixed Transformer input window size of 512, and use a sliding window with a stride of 128 tokens to handle longer documents. We use 256 tokens each for document passage input and KB facts.

	Long F1	Short F1
DecAtt + DocReader	54.8	31.4
BERT$_{\text{joint}}$	64.7	52.7
BERT$_{\text{joint}}$ *	**68.1**	54.0
Traditional PPR	66.7	54.3
QI PPR	65.8	53.6
WS PPR	67.5	**54.4**

Table 3: Results on Full NQ. Baselines: DecAtt (Parikh et al., 2016), DocReader (Chen et al., 2017), and Bert$_{\text{Joint}}$ (Alberti et al., 2019). *- our reimplementation. WS PPR improves over previous baseline on Short F1 and has comparable performance to Bert$_{\text{Joint}}$ on Long F1.

4 Experiments and Results

Setup: As every passage that doesn’t contain the answer is a potential negative, we sample a subset of negatives to balance the dataset. For the Factual NQ subset, we sample 2% of the negatives as in Alberti et al. (2019) to enable faster training. We find that increasing the negatives to 10% improves results by $\sim$2 points and hence for a fair comparison, we sample 10% of the negatives for our models and the reimplemented baseline on the Full NQ dataset. We use the same preprocessing steps and all other hyperparameter settings as in Alberti et al. (2019).

4.1 Retriever Results

While the KB retriever’s effect can be measured in the downstream QA model, it is beneficial to

Question	Hops	Clean Oracle Facts	WS PPR Facts
Who is the existing prime minister of pakistan ?	1	Prime Minister of Pakistan *officeholder* Imran Khan .	Pakistan *office held by head of government* Prime Minister of Pakistan . Imran Khan *position held* Prime Minister of Pakistan . Pakistan *head of government* Imran Khan . Prime Minister of Pakistan *officeholder* Imran Khan . Imran Khan *instance of* human . Pakistan *instance of* country .
What emperor took over france after the reign of terror	3	Reign of Terror *part of* French Revolution . French Revolution *significant event* 18 Brumaire . 18 Brumaire *participant* Napoleon .	Napoleon *participant of* French Revolution . Absolute Monarchy *subclass of* Monarchy . First French Empire *head of state* Napoleon . Seven years ' war *instance of* war .
Who plays the bad guy in looney tunes back in action ?	1	Looney Tunes: Back in Action *cast member* Steve Martin .	Heather Locklear *instance of* human . Heather Locklear *occupation* actor . Looney Tunes: Back in Action *cast member* Heather Locklear . Stan Freberg *occupation* actor . Looney Tunes: Back in Action *cast member* Stan Freberg . Looney Tunes: Back in Action *cast member* Steve Martin . Steve Martin *instance of* human . Steve Martin *sex or gender* male .
Where does the book of daniel take place	2	The Burning Fiery Furnace *narrative location* Babylon. Book of Daniel *derivative work* The Burning Fiery Furnace .	The Burning Fiery Furnace *based on* Book of Daniel . Book of Daniel *derivative work* The Burning Fiery Furnace . Belshazzar's Feast *based on* Book of Daniel . Book of Daniel *derivative work* Belshazzar's Feast . Suzanne bathing *based on* Book of Daniel . Belshazzar's Feast *based on* Book of Daniel . Book of Daniel *derivative work* Suzanne bathing .

Table 4: Examples of Clean Oracle facts and WS PPR retrieved facts. Relations are highlighted in *Italics*.

directly measure the quality of the top retrieved facts. As we consider the shortest path between the question and answer entities as gold facts, we evaluate our retriever using recall of answer entities and shortest path facts in a set of 200 questions from FQ. We compare our retriever with BM25 (Robertson and Zaragoza, 2009), traditional PPR and QI PPR (Sun et al., 2018) as baselines. Table 2 shows the retriever recall results. BM25, traditional PPR and the QI PPR have very poor recall of answers and facts. The low recall of QI PPR shows that questions in NQ do not have similar predicates to relations in the KB, and hence do not benefit from pretrained word vectors. In WS PPR answer entity recall improves by 15 points and Shortest Path fact recall improves by 20 points showing significant improvement. This shows that retrieval methods need question supervision to work in real-world settings and that heuristic methods do not adapt well to it. We show qualitative examples of oracle and retrieved facts in the Appendix.

Additionally, Table 5 (Top) shows that the question independent knowledge (passage entities as seeds PPR(P)) version is slightly worse than question dependent knowledge (question entities as seeds PPR(Q)), showing the benefit of a question dependent factual knowledge retriever.

	Text Only	PPR(Q)	PPR(P)
Short F1	68.2	68.1	67.6
Long F1	77.3	77.8	76.8

	Text Only	Aug Facts	Sep Facts
Short F1	52.7	54.4	52.3
Long F1	64.7	67.5	62.5

Table 5: Top: Comparing different seeds for PPR on FQ. Using question entities as starting seeds is better than passage specific entities. Bottom: Comparing Facts as Augmented Input (Aug Facts) v/s as Separate Input (Sep Facts) on NQ. Augmenting Facts as additional context is significantly better than embedding them via an independent module.

4.2 QA Performance

Factual Questions: Table 1 shows the results of our Knowledge Augmented QA system on the FQ subset[2]. The clean oracle setting improves over the text only baseline and when segregated along the number of hops in the gold shortest path, it has significantly large gains for 2 and 3 hop questions. These questions are generally more complex involving multiple steps of reasoning and augmenting gold facts linking the question to the answer entities significantly helps in the model's performance.

[2] As NQ and FQ rely on span based evaluation, we do not consider KB only baselines for fair comparison.

The noisy oracle setting which has additional facts with oracle facts maintains the QA performance showing that random facts with oracle are still useful to the QA model. This shows that the presence of relevant knowledge from the KB helps QA performance and establishes a strong upper bound for our KB integration. The performance drops when the QA model is given only the PPR facts, without the oracle facts. Both Short and Long answer F1 are similar to the text only setting showing that the retrieved facts are not providing any relevant knowledge to the QA model. Though the weakly supervised setting improves recall of answer entities and shortest path facts, it doesn't improve on the downstream QA task showing that this improved recall is still insufficient for the model to leverage. Comparing the oracle and no-oracle settings, we believe that better KB retrieval methods that have bery high recall of answer entities and relevant facts could lead to improved QA performance, even in real-world complex questions.

We also validate that our performance gains in oracle settings were not due to trivial entity overlap between the text and retrieved facts. We measure the entity overlap in the entire dev set and found that on average, correct predictions had 3.67 entities in common while incorrect predictions had 3.28, and the overall dev set had about 3.54. The small difference in overlap indicates that the oracle setting doesn't leverage any hidden bias.

Natural Questions: Table 3 show the performance of incorporating KB facts in the Full NQ task. Though we see improvements to previously published results, careful ablations reveal that the baseline achieves similar results with more (10%) negative examples. This confirms that even in the full dataset PPR methods fail to retrieve relevant knowledge for the model to leverage for QA.

Facts as Augmented Input: To understand the benefit of augmenting facts as input, we compare against a baseline where the retrieved facts are separately represented by a Transformer. We use a stacked Transformer with the same architecture as BERT as a fact encoder. We feed top retrieved facts in natural language form to it, use a multi-head attention layer between the text only BERT representation and the fact representation and use the new fact-attended text representation for prediction similar to Section 3.2. Results on NQ in Table 5 shows that the separate fact representation has lower performance than our approach showing the benefit of our augmented input approach.

4.3 Qualitative Examples:

Table 4 shows examples of facts from clean oracle and retrieved facts from WS PPR for questions of varying difficulty. The first two examples shows a question where the oracle KB path (shortest path connecting question entities to answer entities) is the correct reasoning path for answering the question. The third and fourth examples shows a case where the oracle KB path contains relevant knowledge for the question but is not the right path for answering the question. WS PPR in all cases retrieves relevant facts about the question entity, and some oracle KB facts. For the first and the third examples, WS PPR retrieves the entire KB path. In the second and last example, WS PPR retrieves part of the oracle KB path but not the entire path.

5 Conclusion

We investigate incorporating KB facts into a real-world QA - Natural Questions. We create a subset of NQ, Factual Questions, to facilitate evaluation of KB integration. We present an oracle setting, where the gold KB path is provided and establish a strong upper-bound. We experimentally show that PPR based retrievers have low recall of answer entities and do not improve downstream QA showing that path-retrieval is a bottleneck for KB integration.

Acknowledgements

We thank the anonymous reviewers for their feedback and suggestions. We are grateful to Mandar Joshi for detailed regular discussions and feedback during the course of this work. We thank Chris Alberti, Kenton Lee and Matthew Siegler for feedback and help with implementation.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

A. Bordes, S. Chopra, and J. Weston. 2014. Question answering with subgraph embeddings. In *EMNLP*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *ACL*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.

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.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

Kangqi Luo, F. Lin, Xusheng Luo, and K. Q. Zhu. 2018. Knowledge base question answering via encoding of complex query graphs. In *EMNLP*.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Michael Ringgaard, Rahul Gupta, and Fernando CN Pereira. 2017. Sling: A framework for frame semantic parsing. *arXiv preprint arXiv:1710.07032*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

U. Sawant, S. Garg, S. Chakrabarti, and G. Ramakrishnan. 2018. Neural architecture for question answering using a knowledge graph and web corpus. *Information Retrieval Journal*, 22:324–349.

Haitian Sun, Tania Bedrax-Weiss, and William W. Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. In *EMNLP/IJCNLP*.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *EMNLP*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Chao Wang and Hui Jiang. 2019. Explicit utilization of general knowledge in machine reading comprehension. In *ACL*.

Y. Zhang, Hanjun Dai, Zornitsa Kozareva, A. Smola, and L. Song. 2018. Variational reasoning for question answering with knowledge graph. In *AAAI*.