Q. Can Knowledge Graphs be used to Answer Boolean Questions? A. It's complicated!

Daria Dzendzik*
ADAPT Centre
Dublin City University
daria.dzendzik
@adaptcentre.ie

Carl Vogel
School of Computer
Science and Statistics
Trinity College Dublin
The University of Dublin
vogel@tcd.ie

Jennifer Foster
School of Computing
Dublin City University
jennifer.foster
@dcu.ie

Abstract

In this paper we explore the problem of machine reading comprehension, focusing on the BoolQ dataset of Yes/No questions. We carry out an error analysis of a BERT-based machine reading comprehension model on this dataset, revealing issues such as unstable model behaviour and some noise within the dataset itself. We then experiment with two approaches for integrating information from knowledge graphs: (i) concatenating knowledge graph triples to text passages and (ii) encoding knowledge with a Graph Neural Network. Neither of these approaches show a clear improvement and we hypothesize that this may be due to a combination of inaccuracies in the knowledge graph, imprecision in entity linking, and the models' inability to capture additional information from knowledge graphs.

1 Introduction

Clark et al. (2019) explore the difficulty of Yes/No questions and introduce the BoolQ dataset which contains 16k questions based on real Google user queries, paired by crowdworkers with passages from Wikipedia. They establish a strong baseline using $BERT_{large}$ (Devlin et al., 2019) and transfer learning from the Multi-Genre Natural Language Inference (MNLI) task (Williams et al., 2018).

In this work, we carry out an error analysis of 200 samples from the $BERT_{large} + MNLI$ baseline model and find out that 77% constitute genuine model errors, almost 6% of samples contain an incorrect answer tag, and 8% do not contain enough evidence to answer the question. The remaining 9% we classified as difficult questions as they involve deep understanding, reasoning, specific knowledge, and sometimes depend on opinion. Due to the unstable behaviour of the model, error samples vary from run to run, where a run refers to the pipeline of MNLI pre-training, BoolQ fine-tuning, and evaluation of the model. We introduce a *stable accuracy* metric to evaluate a system across multiple runs with the same hyperparameters. Stable accuracy over n runs refers to the proportion of questions that are always correctly answered. We observed a 3.3% and an 11% drop of stable accuracy over 2 and 10 runs respectively.

Next we turn our attention to improving machine reading comprehension (MRC) system performance. We hypothesize the system might benefit from additional information about entities and/or relations between the entities, in the question and passage. Consider, for example, (1) where *pei* is an abbreviation of *Prince Edward Island*.

(1) **Question:** *is anne with an e filmed on pei*
Passage: *The series is filmed partially in Prince Edward Island as well as ...*
Gold Answer: Yes **Predicted Answer:** No

A number of works including Mihaylov and Frank (2018); Bauer et al. (2018); Lin et al. (2019); Qiu et al. (2019); Thayaparan et al. (2019); Talmor et al. (2019); Zhao et al. (2020) show successful usage of knowledge graphs (KGs) in several MRC settings.

We propose and evaluate two approaches for augmenting questions and answers with KG information: (1) concatenating the model input with sentences constructed from ConceptNet triples[1] (Speer et al., 2017); and (2) encoding KG entities and relations with the Graph Neural Network (GNN) proposed by Shaw et al. (2019), a model suited to graph-based input. Neither approach shows a significant improvement over the baseline.

*A significant part of this work was done during an internship at Google Research Switzerland in September-December 2019 in collaboration with Massimo Nicosia.

[1] https://conceptnet.io/ – last verified (l.v.) 07/2020

Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 6–14

ISBN 978-1-952148-66-8

Category	#	%	Category	#	%
Factual Reasoning	12	6.0	Paraphrasing	97	48.5
Missing Mention	28	14.0	By Example	7	3.5
Other Inference	16	8.0	Implicit	39	19.5

Table 1: BoolQ errors anlysis by reasoning type.

2 A Closer Look at the BoolQ Baseline

2.1 Error Analysis

We manually analyse 200 errors made by one run of the baseline system (33% of one-run errors) and discover that 6% of them involve an incorrect answer tag and another 8% involve confusing passages which do not give enough support for the answer (see Appendix B for examples).

Table 1 shows a categorization of the errors according to the reasoning types provided by Clark et al. (2019). The majority of errors belongs to the *Paraphrasing* type (48.5%). In these cases, the answer is in the passage and only a minimum amount of extra knowledge and reasoning is required to answer the question. The *Implicit* and *Missing Mention* types account for 19.5% and 14% of errors respectively. Only about 3.5% of incorrectly answered questions require an understanding of examples given in the passage, 6% requrie factual reasoning, and 8% require other inference.

2.2 Stable Accuracy

We reproduce the results of the baseline $BERT_{large} + MNLI$ model released by Clark et al. (2019).[2] Its accuracy is between 80% and 82% (Fig. 1 (a) ●) with an average 81.41% accuracy over 10 runs (vs. 82.2% reported in Clark et al. (2019)). Our error analysis shows that a significant portion of the correctly answered questions varies from run to run together with around 40% of errors.

We define the ratio of the number of correctly answered questions across n runs to the total number of questions as *stable accuracy*. Formally, if Q is the set of all questions and $Q^i_{correct}$ is the set of correctly answered questions at the i^{th} run, the *stable accuracy* after n runs is defined as (2):

$$StableAccuracy_n = \frac{|\cap_{i=0}^{n} Q^i_{correct}|}{|Q|} \quad (2)$$

The stable accuracy over 10 runs drops to 71% (see Fig 1 (a) ★). Ensembling with a majority voting for

[2] https://github.com/google-research/language/tree/master/language/boolq – l. v. 05/2020

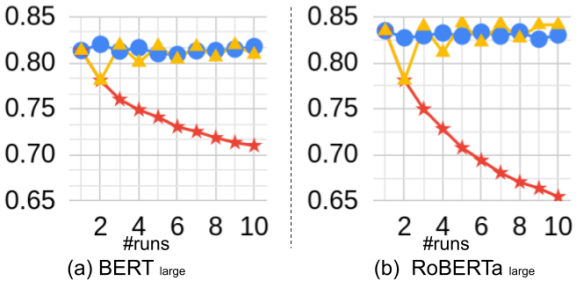

Figure 1: Accuracy (●), stable accuracy (★), and majority voting accuracy (▲) over up to 10 runs of (a) BERT and (b) RoBERTa baselines.

up to 10 runs (Fig. 1 (a), ▲) does not outperform the baseline: the values are within the range of 78.09% and 81.77%.[3]

We repeat the experiment using the robustly optimized $RoBERTa_{large}$ model (Liu et al., 2019) implemented by Wolf et al. (2019) and fine tuned on the MNLI task. This model has a better average accuracy (83.7)% but it is also more unstable: the stable accuracy drops to 64.0% (see Fig. 1 (b)). As with the $BERT$ model, ensembling over 10 runs does not give a performance boost.

This observed behavior means that the system performs well on each run but every time it performs well on a different set of questions. This might be related to the notion of "forgettable" examples described by Toneva et al. (2019). The difference is that they discovered the ability of models to forget the learned examples during the training phase, while we examine stable and unstable examples when the training is finished.

3 Modeling Knowledge Graph Data

Our manual inspection of the results of one baseline system run reveals that approximately 20% of erroneous cases are questions involving some property of an entity or concept, or some hierarchical relationship between entities. An example of the former is (3) and the latter is (4).

(3) *is i 80 in indiana a toll road*

(4) *is college of william and mary an ivy league school?*

We hypothesize that adding knowledge graph data could help in answering such questions, as well

[3] Note that the ensemble performs slightly better with an odd numbers of runs as only the samples with strictly more votes for the correct answer are considered to be answered correctly. This is a very strict evaluation. Alternatively, in the case of a tie, the majority answer (*Yes*) can be selected, but we aim to provide the evaluation with the maximum certainty.

as examples such as (1) and (5) below where the entity in the question is referred to using a different name in the passage.

(5) **Question:** *does smeagol die in lord of the rings* **Passage:** *... Gollum finally ... but he fell into the fires of the volcano, where both he and the Ring were destroyed.* **Answer:** Yes

We use the CloudAPI[4] to annotate text with tokens, part of speech tags, named entities with Freebase[5] KG identifiers (MIDs), numbers, dates and VerbNet[6] roles which can be used for establishing relations between entities.

3.1 Extending Passages with ConceptNet

ConceptNet (Liu and Singh, 2004; Speer et al., 2017) is an open semantic network based on DBPedia, Wiktionary, WordNet, and other resources. It captures common-sense knowledge and was created for computers to understand words and concepts in the same way people do. It was particularly designed to be used by NLP applications and widely used in MRC (Weissenborn et al., 2017; Bauer et al., 2018; Mihaylov and Frank, 2018; Lin et al., 2019; Qiu et al., 2019). Partly inspired by Weissenborn et al. (2017), we convert ConceptNet relations into sentences but instead of embedding them independently, we concatenate them to the baseline model input.

3.1.1 Sentence Extraction and Filtering

ConceptNet has 34 relation types.[7] Each relation has start and end entities and a strength of relation (relevance weight). We look up every annotated entity from questions and passages in ConceptNet. We extract the top 100 relations according to the relevance weight, and select those where both the start and end entities are in English. We remove relations that are not useful, such as *"External URLs"*, or too broad such as *"FormOf"*. Then we transform ConceptNet relations into simple sentences based on the relation description or, if there is no description, we create a string: `[entity1] [relation] [entity2]`, e.g. the *"panda is near a bamboo forest"* string is created from entites: *"panda"*, *"bamboo forest"* and the relation *"LocatedNear"*. Fig. 2 shows a ConceptNet entity from example (1). The verbalized triples such as *"pei is a synonym of Prince Edward Island"* are prepended to the text passage.

Figure 2: An example of usage ConceptNet entities for answering a Boolean question.

Since such new sentences can add noise (see *polyetherimide* examples in Fig. 2) and a long input might confuse the model (Thayaparan et al., 2019), we aim to add extra sentences to the passages only if it is relevant and can better "explain" the nature of entities. To select those, we rank all extracted sentences S according to the sum of their similarities with the question q and passage p as shown in (6):

$$\forall s \in S : score(s) = g(k(s), k(q)) + g(k(s), k(p)) \quad (6)$$

where $g \in \{correlation, cosine\}$ are similarity measures, k is a semantic embedding function. We use the semantic textual similarity model[8] proposed by Yang et al. (2018). To filter more examples, we add an empirically tuned threshold for similarities[9] and select only those sentences which were ranked as the most similar to the question and passage by both correlation (inner product) and cosine similarity, and each score is higher than the established thresholds. Another method of selecting relevant sentences is to consider only the relations which connect an entity in the question to an entity in the passage. We then combine these two strategies: we add sentences only to the examples which meet both criteria (Intersection) or all that meet at least one of the criteria (Union).

3.1.2 Results

Table 2 shows the results averaged over 5 runs. With threshold filtering we add sentences to 21.84%

[4] https://cloud.google.com/apis/docs/overview – l.v. 07/2020

[5] https://en.wikipedia.org/wiki/Freebase_(database) – l.v. 07/2020

[6] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html – l.v. 07/2020

[7] Based on https://github.com/commonsense/conceptnet5/wiki/Relations – l.v. 07/2020. We found a few more like *"language"* or *"occupations"*.

[8] Available via TensorFlowHub (Cer et al., 2018): https://www.tensorflow.org/hub/ – l.v. 07/2020

[9] We used: correlation > 220; cosine similarity > 1.38.

of passages, obtaining an average accuracy of 81.23% (see Table 2: SentEmb). Using entity relations from questions and answers, 22.58% of QA pairs are affected but the performance is slightly worse (see Table 2: Q&P Match).

The intersection gives the best performance. By affecting only 1.23% of the data, we obtain 81.46% average accuracy and 82.05% accuracy for the ensemble majority voting scenario. The Union criterion does not show any improvement on accuracy. The Intersection improvement, as well as the disimprovement of SentEmb, Q&PMatch, and Union, are not statistically significant with respect to the baseline.[10]

	Base line	Sent Emb	Q&P Match	Intersection		Union
Data Coverage (%)	-	21.84	22.58	1.23		38.57
AVG	**81.26**	81.23	80.86	81.23	**81.46**	80.72
Stable	73.84	73.19	72.61	73.25	73.74	72.40
Ensemble	81.62	81.89	81.37	**81.92**	**82.05**	81.10

Table 2: Percentage of data changed and accuracy over 5 runs: average (**AVG**), **Stable**, and **Ensemble**.

3.2 Modeling Knowledge Graphs with GraphNNs

Facing instability of the BERT-based baseline and low coverage of ConceptNet (see Section 4) we experiment with a new architecture and knowledge graph. To better model graph-based input, such as entities and their relations, we tried a transformer-based seq2seq GNN (Shaw et al., 2019). Entities, relations and input tokens are embedded and fed to a GNN sub-layer that incorporates edge representations extending the self-attention mechanism. The encoder-decoder attention layer considers both encoder output token and entity representations, jointly normalizing attention weights over tokens and entities. In our case, the GNN decoder simply outputs our expected answers: "Yes" or "No" (see Fig. 3). In this case, we initialize the GNN with a pre-trained $BERT_{large}$ model and only fine tune on BoolQ.

As an alternative to ConceptNet we also tried the Google Knowledge Graph. It has more than 500 billion facts about 5 billion entities.[11] The entities describe real-world objects and concepts like people, places, events, and things. Entities are represented as nodes and connected by relations. The latter can simply indicate that a relation is present, or they may encode the type of relation. We try the first three of the following possible experiments:

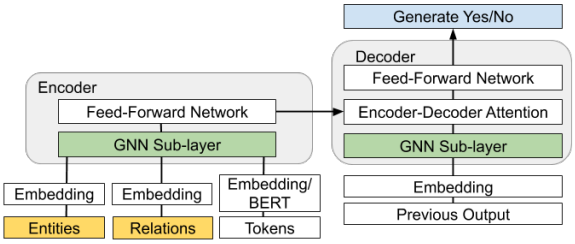

Figure 3: The GNN architecture based on Shaw et al. (2019) without action selection and copy mechanism.

1. adding a relation between different entities which have the same MID;
2. only adding connections between entities across the QA pair, as in the ConceptNet Q&P Match experiment;
3. distinguishing different types of relations;
4. adding a relation between different mentions of the same entity;
5. adding entities not mentioned in the text but linked to the mentioned entities.

3.2.1 Results

The results are presented in Table 3. The first row shows the baseline $BERT$ model with no KG data and the remaining rows show the $BERT + GNN$ system with no KG data, with ConceptNet or with the Google Knowledge Graph. Adding KG information does not outperform the baseline result. None of the differences between the baseline are statistically significant.

	No KG	+ConceptNet	+GKG
BERT_{large}	**78.09**	-	-
GNN + BERT	77.37	77.4	76.80
+ Same MID	-	-	77.60
+ Relation Type	-	-	77.75
+ Q&AMatch	-	-	76.95

Table 3: **GNN** accuracy results on a development set using ConceptNet or Google KG (**GKG**).

4 Analysis

ConceptNet Even after the filtering described in Section 3.1.1, we observe that often the relations from ConceptNet are too general and do not add new information, e.g. *"cookie jar is a type of jar"*.

[10] According to the two sample proportion Z-Test the maximum difference: $z = -1.3674$, $p = 0.17068$

[11] https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/ – l.v. 07/2020

Such relations are already part of the language model. Petroni et al. (2019) show that $BERT$ contains relational knowledge and has a strong ability to recall factual knowledge without fine-tuning.

Furthermore, some entities are missing, e.g. there is a *"Tom Hanks"* entity but no *"Meg Ryan"* entity, or the entity *"dragon ball"* contains only non-English connections, confirming the general coverage issue of KGs. [12]

Sensitivity We observe that the GNN is sensitive to the learning rate and hyper-parameters. Better tuning may compensate for the difference in performance wrt to the $BERT$ baseline.

Entity recognition and linker We found issues with the entity linker. Named entities are often not covered or the MID is missing. In some cases, the entity has a wrong MID, e.g. in (7) the entity "*northern ireland*" is not recognised but the entity "*ireland*" (Republic of Ireland) is mentioned instead, while the entity "*great britain*" is recognised with the MID of "*United Kingdom*".

(7) **Question:** *is northern ireland part of the great britain* **Passage:** *... Great Britain is part of the United Kingdom of Great Britain and Northern Ireland ...* **Answer:** No

The questions in the BoolQ dataset are lowercased, and this may have affected the entity recognition.

Do KGs affect stable accuracy? We observe a positive tendency towards stable correct answers in the ConceptNet experiments (Table 4). The number of new stable correct answers is higher than the number of new stable errors for all settings except Q&AMatch. Also, for all scenarios except Intersection, the number of questions where the predicted answer fluctuates from incorrect to correct is higher than the number of questions where the predicted answer fluctuates from correct.

Is a KG necessary? The BoolQ dataset was not originally created to be used with a KG, and the passages were selected such that they contain the information required to answer a question. For some questions, such as (1) the additional information provided by a KG is helpful, and for questions like (7), even though the passage has all the required information, a KG could highlight the relation between entities and help answer the question. However, there are also cases where a KG is not needed or cannot be applied, e.g. (8) and (9).

[12] https://conceptnet.io/c/en/jar, https://conceptnet.io/c/en/tom_hanks – An English term in ConceptNet 5.8, https://conceptnet.io/c/en/meg_ryan – 'meg ryan' is not a node in ConceptNet, https://conceptnet.io/c/en/dragon_ball,– l.v. 07/2020

New		Sent Emb	Q&P Match	Inter section	Union
Stable	**Correct**	27	27	4	54
	Error	18	28	0	32
Fluct.	**Err→Corr**	34	44	1	65
	Corr→Err	19	23	5	42

Table 4: **New Correct (Error)** corresponds to the number of new stable (wrt to baseline) correct (incorrect) predictions, **New Fluct.** is the number of new questions where answer fluctuates: **Err→Corr** (**Corr→Err**) is the number of questions where answer was a stable error (correct), becoming correct (error) sometimes.

(8) **Question:** *do all ni numbers have a letter at the end* **Passage:** *The format of the number is two prefix letters, six digits, and one suffix letter. The example used is typically QQ123456C. ...* **Answer:** Yes

(9) **Question:** *was the movie insomnia based on a book* **Passage:** *Robert Westbrook adapted the screenplay to novel form, which was published by Alex in May 2002.* **Answer:** No

In (8) a question is asked about a number format and the information about the specific last symbol is unlikely to be a part of a KG. (9) contains a very short passage explicitly saying there is a book but it was adapted from the screenplay. In this case, a KG could provide potentially confusing information simply stating that there is a book.

5 Conclusion

In this work, we take a closer look at a $BERT$ baseline system on the BoolQ dataset, which reveals some inconsistencies in the data and some instability in the model. We try two approaches to integrating knowledge graph information, one based on augmenting the passage text and another using a Graph Neural Network. Neither are successful. One culprit is the lack of coverage of ConceptNet and another is related to accuracy of the entity recognition. We also suggest that the number of questions where suitable KG data is needed and could be found might just not be enough for the models to learn from.

Acknowledgments

We are extremely gratefully to Massimo Nicosia from Google Research Switzerland without whom this work would not be possible. We thank the anonymous reviewers for their constructive and helpful feedback. Finally, a big thank you to Andrew Dunne, Lauren Cassidy, and Meghan Dowling.

This research is partly supported by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and the European Regional Development Fund.

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7(0):452–466.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

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

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Delai Qiu, Yuanzhe Zhang, Xinwei Feng, Xiangwen Liao, Wenbin Jiang, Yajuan Lyu, Kang Liu, and Jun Zhao. 2019. Machine reading comprehension using structural knowledge graph-aware network. 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 5896–5901, Hong Kong, China. Association for Computational Linguistics.

Peter Shaw, Philip Massey, Angelica Chen, Francesco Piccinno, and Yasemin Altun. 2019. Generating logical forms from graph representations of text and

entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 95–106, Florence, Italy. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, Viktor Schlegel, and André Freitas. 2019. Identifying supporting facts for multi-hop question answering with document graph networks. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 42–51, Hong Kong. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Dirk Weissenborn, Tomáš Kočiský, and Chris Dyer. 2017. Dynamic integration of background knowledge in neural NLU systems. *arXiv:1706.02596*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

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

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

Chen Zhao, Chenyan Xiong, Xin Qian, and Jordan Boyd-Graber. 2020. Complex factoid question answering with a free-text knowledge graph. In *The Web Conference 2020 (formerly WWW conference)*.

A BoolQ Dataset Details

The BoolQ dataset (Clark et al., 2019) is a part of the SuperGLUE benchmark[13] (Wang et al., 2019). About 3000 question and passages come from NaturalQuestion (Kwiatkowski et al., 2019). The main statistics about the dataset is collected in Table 5.

Size	Length in Tokens					
	Question			Passage		
	Min	Max	Avg	Min	Max	Avg
15942	3	21	8.9	6	813	108

Table 5: The basic statistics for the BoolQ dataset.

Clark et al. (2019) showed the $BERT_{large}$ model (Devlin et al., 2019) outperforming recurrent models with attention (Wang et al., 2018), both in their vanilla version and in combination with deep contextualized word representation (Peters et al., 2018).

B Erroneous and Confusing Examples

Some questions in BoolQ are formulated in a certain context which might change given time. For example (10) which is asking about a movie released ***this year***. As the dataset was released in 2019 the data could be collected in 2018 so then the answer is *yes* but if this question would be asked in 2015 or today (2020) the answer should be *no*. Another example (11) where a passage provides the information about United States citizens border crossing requirements but the question does not specify what kind of citizenship the person asking the question holds. In contrast with example (12) where the question and passage provide an unconditional outcome as a holder of the Schengen visa (information from question) can enter Montenegro for 30 days (information from the passage). So, in such cases like examples (10) and (11), the passage information is not enough to answer the questions unconditionally.

(10) **Question:** *is there a star wars movie this year*
Passage: *The first film was followed by two successful sequels, The Empire Strikes Back (1980) and Return of the Jedi (1983); ... A prequel trilogy was released between 1999 and 2005, albeit to mixed reactions from critics and fans. A sequel trilogy concluding the main story of the nine-episode saga began in 2015 with The Force Awakens. ... Together with the theatrical spin-off films The Clone Wars (2008), Rogue One (2016) and Solo: A Star Wars Story (2018), Star Wars is the second highest-grossing film series ever.*
Answer: Yes (true)

(11) **Question:** *Can I get into Canada with a military ID?*
Passage: *(Title: American entry into Canada by land) Canadian law requires that all persons entering Canada must carry proof of both citizenship and identity. A valid U.S. passport or passport card is preferred, although a birth certificate, naturalization certificate, citizenship certificate, or another document proving U.S. nationality, together with a government-issued photo ID (such as a driver's license) are acceptable to establish identity and nationality.*
Answer: Yes

(12) **Question:** *Can I go to Montenegro with a Schengen visa?*
Passage: *Nationals of any country may visit Montenegro without a visa for up to 30 days if they hold a passport with visas issued by Ireland, a Schengen Area member state, ...*
Answer: Yes

Some passages looked unrelated or do not contain enough information to obtain the answer, e.g. (13 - 14). The passages are related to the questions but specific information is missing the answer "Yes" cannot be confirmed by the passages. We observe, around 8% of questions we confusing or have certain assumptions.

(13) **Question:** *is daisy the director of shield in the comics*
Passage: *Daisy Johnson, ... The daughter of the supervillain Mister Hyde, she is a secret agent of the intelligence organization S.H.I.E.L.D. with the power to generate earthquakes.*
Answer: Yes

(14) **Question:** *is chicken cordon bleu made with blue cheese*

[13] https://super.gluebenchmark.com/ – l.v. 07/2020

Passage: *A cordon bleu or schnitzel cordon bleu is a dish of meat wrapped around cheese (or with cheese filling), then breaded and pan-fried or deep-fried. Veal or pork cordon bleu is made of veal or pork pounded thin and wrapped around a slice of ham and a slice of cheese, breaded, and then pan fried or baked. For chicken cordon bleu chicken breast is used instead of veal. Ham cordon bleu is ham stuffed with mushrooms and cheese.*
Answer: Yes

There are a few examples of errors (15 - 17) from the dataset. The first error example is asking if shower gel can be used instead of shampoo in a negative form (*"is it bad to ..."*) and the passage says that they are perfectly substitutable so the answer should be *No (it is not bad)*. In the second example (16) the passage explicitly says India does not have a national language so the answer should be *No*. And in the third example (17) there is nothing that should make the reader believe there were any games outside of Russia, so the answer should be *Yes*. According to our analysis 6% of samples have the wrong answer tag.

(15) **Question:** *Is it bad to wash your hair with shower gel?*
Passage: *... This means that* ***shower gels can also double as an effective and perfectly acceptable substitute to shampoo****, even if they are not labelled as a hair and body wash.*
Answer: Yes Should be **No**

(16) **Question:** *Is Hindi is our national language of India?*
Passage: *The Constitution of India designates the official language of the Government of India as Hindi written in the Devanagari script, as well as English.* ***There is no national language as declared by the Constitution of India.*** *Hindi is used for official purposes ...*
Answer: Yes Should be **No**

(17) **Question:** *are all world cup matches played in russia*
Passage: *The 2018 FIFA World Cup was the 21st FIFA World Cup, an international football tournament contested by the men's national teams of the member associations of FIFA once every four years. It* ***took place in Russia*** *from 14 June to 15 July 2018. ...*
Answer: No Should be **Yes**