EMBRACING DATA ABUNDANCE

Ondrej Bajgar*, **Rudolf Kadlec*** **& Jan Kleindienst**
IBM Watson, V Parku 4, 140 00 Prague, Czech Republic
{obajgar,rudolf_kadlec,jankle}@cz.ibm.com

ABSTRACT

There is a practically unlimited amount of natural language data available. Still, recent work in text comprehension has focused on datasets which are small relative to current computing possibilities. This article is making a case for the community to move to larger data and is offering the BookTest dataset as a step in that direction.

1 INTRODUCTION

Much work in Text Comprehension (and some other areas of Machine Learning research) focuses on improving models on a few standard datasets, with the CNN/Daily Mail dataset (Hermann et al., 2015) and the Children's Book Test (CBT) (Hill et al., 2015) being among the most popular in the past year. These two datasets managed to avoid the problem of expensive human annotation with their way of automatically generating cloze-style questions (Taylor, 1953) from a suitable text corpus.

Many teams have put effort into improving the accuracy on the above datasets by improving their models' architecture. However there is a more straightforward, well established way of improving model performance: more training data. We do think it is easily possible start using more data in Text Comprehension research: an almost unlimited amount of cloze-style questions can be generated from a suitable corpus and since it takes only about two hours to train some models on the CBT (or two days on the Daily Mail dataset), the computing potential is also not being fully exploited. Since such easy performance gains are possible, we think that our community should study the performance of various architectures also in this realistic scenario of data abundance instead of generally imposing upon ourselves the constraint of small-size training data.

In this brief paper we will use the new BookTest dataset to illustrate this point by showing that the gains from increasing the data size can be surprisingly large compared to the gains from many teams' efforts to improve the models' architecture.

2 BOOKTEST DATASET

The BookTest[1] is a cloze-style question-answering dataset derived from 10,507 copyright-free books available through project Gutenberg (compare to 108 books used for CBT). Except for its size, the dataset is very similar to CBT, and so is the generation procedure: We detect whether each sentence in the corpus contains either a named entity or a common noun that already appeared in one of the preceding twenty sentences. This word is then replaced by a gap tag (XXXXX) in this 21^{st} sentence, which is hence turned into a cloze-style question. The preceding 20 sentences are used as the context document.

We then also add the examples from the CBT CN and NE training datasets. The resulting BookTest dataset hence contains $14,140,825$ training examples[2], as well as a validation set, consisting of

*These authors contributed equally to this work.

[1]BookTest can be downloaded from https://ibm.biz/booktest-v1.

[2]This makes BookTest the largest text comprehension dataset currently available. All other recently introduced datasets like SQuAD (Rajpurkar et al., 2016), Who-did-What (Onishi et al., 2016), NewsQA (Trischler et al., 2016a), Story Cloze Test (Mostafazadeh et al., 2016) and finally MS MARCO (Nguyen et al.) provide less training data. However, these datasets have other qualities that make them still valuable. The only similarly sized dataset is WikiReading (Hewlett et al., 2016), which however provides a much smaller variety of questions than the BookTest, with 20 questions (in their case Wikidata keys) covering 75% of the dataset.

Table 1: Statistics on 4 standard text-comprehension datasets and the BookTest. CBT CN stands for CBT Common Nouns and CBT NE stands for CBT Named Entities.

	CNN	**Daily Mail**	**CBT CN**	**CBT NE**	**BookTest**
# queries	380,298	879,450	120,769	108,719	14,140,825
Avg # tokens	762	813	470	433	522
Vocab. size	118,497	208,045	53,185	53,063	1,860,394

Table 2: Results of various models on CBT validation and test data. For NSE we give results of its variant with the best validation accuracy on each dataset.

	Named entity		Common noun		
	valid	test	valid	test	
Humans (context+query) (Hill et al., 2015)	NA	***81.6***	NA	***81.6***	
AS Reader (avg ensemble) (Kadlec et al., 2016)	74.5	70.6	71.1	68.9	CBT training
AS Reader (greedy ensemble) (Kadlec et al., 2016)	76.2	71.0	72.4	67.5	CBT training
NSE (Munkhdalai & Yu)	78.2	73.2	74.3	71.9	CBT training
AS Reader (single model)	80.5	76.2	83.2	80.8	BookTest training data
AS Reader (ensemble)	81.9	77.5	85.5	83.3	BookTest training data

$10,000$ named entity (NE) and $10,000$ common noun (CN) questions; one test set for NEs and one for CNs, each containing $10,000$ examples. The training, validation and test sets were generated from non-overlapping sets of books. Statistics of the BookTest are summarized in Table 1.

When generating the dataset we removed all editions of books used to create CBT validation and test sets from our training dataset. Therefore the models trained on the BookTest can be evaluated on the original CBT data and they can be compared with recent text-comprehension models utilizing this dataset (Hill et al., 2015; Kadlec et al., 2016; Sordoni et al., 2016; Dhingra et al., 2016; Trischler et al., 2016b; Weissenborn, 2016; Cui et al., 2016; Munkhdalai & Yu).

3 EXPERIMENTS

We tested the performance gains from using more data using the Attention Sum Reader (AS Reader), a simple text-comprehension model whose central idea – using a sum of attention given to each unique word in the text to select an answer to the query – has inspired many other recent models. We hence consider this model a good representative of the many recent text-comprehension architectures listed earlier. A detailed description of the architecture can be found in (Kadlec et al., 2016).

We simply trained the model on the BookTest data and compared the results to models trained on the CBT.

Results. Table 2 summarizes the improvement in accuracy thanks to using more data. It shows the human baseline provided in (Hill et al., 2015), the AS Reader and Neural Semantic Encoder (NSE) (Munkhdalai & Yu) (current state-of-the-art) trained on the original CBT dataset and then the result for an AS Reader ensemble trained on the BookTest but evaluated on CBT.

While improving the model architecture as in (Sordoni et al., 2016; Dhingra et al., 2016; Trischler et al., 2016b; Weissenborn, 2016; Cui et al., 2016; Munkhdalai & Yu) while still using the original CBT training data lead to improvements of around $1 - 4\%$ absolute compared to the AS Reader's performance, inflating the training dataset provided a boost of $6.5 - 17.4\%$ while using the same model. Our ensemble even exceeded the human baseline provided by Hill et al. (2015) on the CN dataset.

The model takes approximately two weeks to converge on a single Nvidia Tesla K40 GPU.

4 DISCUSSION

The gain from increasing the dataset size is hence considerable (as was previously pointed out also by Banko & Brill (2001); Halevy et al. (2009)). At least some real-world systems may want to benefit from this so we believe current research should start focusing model design more in this direction, since this may mean focusing on other model aspects than with smaller data. Here are some of the new challenges that we need to face.

Firstly, since the amount of data is practically unlimited – we could even generate them on the fly resulting in continuous learning similar to the Never-Ending Language Learning (Mitchell et al., 2015) – it is now the speed of training that determines how much data the model is able to see. Since more training data significantly help the model performance, focusing on speeding up the algorithm may be more important than ever before. For instance Chen et al. (2016) achieves better single-model performance on CBT compared to the AS Reader, probably partly thanks to regularization, however the training is about 7 times slower.

Another challenge is to generalize the gains from large data to a specific target domain. While there are huge amounts of natural language data in general, it may not be the case in the domain where we may want to ultimately apply our model. Hence we may be facing not a scenario of simply using a larger amount of the same training data, but rather extending training to a related data domain, hoping that some of what the model learns on the added data will still help it on the original task.

This is highlighted by our observations from applying a model trained on the BookTest to CBT test data. If we move model training from joint CBT NE+CN training data[3] to a subset of the BookTest of the same size (230k examples), we see a drop in accuracy of around 10% on the CBT test datasets. Hence even though the CBT and BookTest datasets are as close as two disjoint datasets can get, the transfer is still very imperfect. Rightly choosing data to augment the in-domain training data is certainly a problem worth exploring in future work.

Given enough data the AS Reader was able to exceed human performance on CBT CN reported by Hill et al. (2015) which raises the question whether there is still room for improvement.

5 HUMAN STUDY

After reaching the mentioned human baseline, we have decided to examine whether there is space for further improvement on the CBT by testing humans on a random subset of 50 named entity and 50 common noun validation questions that the AS Reader ensemble could not answer correctly. The results are summarized in Table 4. They show that a majority of questions that our system could not answer so far are in fact answerable. Hence the CBT may still be used for tracking the improvements of machine learning models (possibly in contrast to CNN/DM where Chen et al. (2016) pointed out that most questions left unanswered by models may in fact be unanswerable even by humans).

Table 3: Accuracy of humans on validation examples answered incorrectly by AS Reader trained on BookTest.

Dataset	% correct answers
Named Entities	66%
Common Nouns	82%

6 CONCLUSION

Few ways of improving model performance are as solidly established as using more training data. Yet we believe this principle has been somewhat neglected by recent research in text comprehension. As a gentle reminder to the community we have shown how large the performance gains from using more data can be. Yes, experiments on small datasets certainly can bring useful insights. However we believe that the community should also embrace the real-world scenario of data abundance. The BookTest dataset we are proposing gives the reading-comprehension community an opportunity to make a step in that direction.

[3]Note that while here we are using joint CBT NE+CN data to create an equivalent of a 230k subset of our BookTest, for most other experiments on the CBT teams used NE and CN as two separate training datasests.

REFERENCES

Michele Banko and Eric Brill. Scaling to Very Very Large Corpora for Natural Language Disambiguation. *Proceedings of ACL*, 2001.

Danqi Chen, Jason Bolton, and Christopher D. Manning. A Thorough Examination of the CNN / Daily Mail Reading Comprehension Task. In *Proceedings of ACL*, 2016.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-Attention Neural Networks for Reading Comprehension. 2016. URL http://arxiv.org/abs/1607.04423.

Bhuwan Dhingra, Hanxiao Liu, William W. Cohen, and Ruslan Salakhutdinov. Gated-Attention Readers for Text Comprehension. 2016. URL http://arxiv.org/abs/1606.01549.

Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. ISSN 1541-1672. doi: http://doi.ieeecomputersociety.org/10.1109/MIS.2009.36.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pp. 1684–1692, 2015.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. WIKI READING : A Novel Large-scale Language Understanding Task over Wikipedia. *Acl 2016*, pp. 1535–1545, 2016.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. Neural Text Understanding with Attention Sum Reader. *Proceedings of ACL*, 2016.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories. *Proceedings of NAACL*, 2016. URL http://arxiv.org/abs/1604.01696.

Tsendsuren Munkhdalai and Hong Yu. Reasoning with Memory Augmented Neural Networks for Language Comprehension. In *Proceedings of ICLR 2017*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. URL https://arxiv.org/abs/1611.09268.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. Who did What: A Large-Scale Person-Centered Cloze Dataset. (3), 2016. URL http://arxiv.org/abs/1608.05457.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *EMNLP*, 2016. URL http://arxiv.org/abs/1606.05250.

Alessandro Sordoni, Phillip Bachman, and Yoshua Bengio. Iterative Alternating Neural Attention for Machine Reading. 2016. URL https://arxiv.org/abs/1606.02245.

Wilson L Taylor. Cloze procedure: a new tool for measuring readability. *Journalism and Mass Communication Quarterly*, 30(4):415, 1953.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA Dataset. 2016a. URL https://arxiv.org/abs/1611.09830.

Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. Natural Language Comprehension with the EpiReader. 2016b. URL http://arxiv.org/abs/1606.02270.

Dirk Weissenborn. Separating Answers from Queries for Neural Reading Comprehension. 2016. URL http://arxiv.org/abs/1607.03316.