IBM Watson

Ondřej Bajgar*, Rudolf Kadlec*, Jan Kleindienst
{obajgar, rudolf_kadlec, jankle}@cz.ibm.com
* equal contribution

Embracing Data Abundance

There is a practically unlimited amount of natural language data available. Still, most text-comprehension research is done on relatively small datasets. Isn't it time to also embrace the real-world scenario of data abundance?

Task: Reading Comprehension

Context document (a story of 20 sentences):
Alice and Bob were playing on the lawn while their dog Rex was running around.
[...]
Bob threw a ball into his friend's face.

Question (cloze-style in our case)
„You're mean!", **XXXXX** said to Bob.

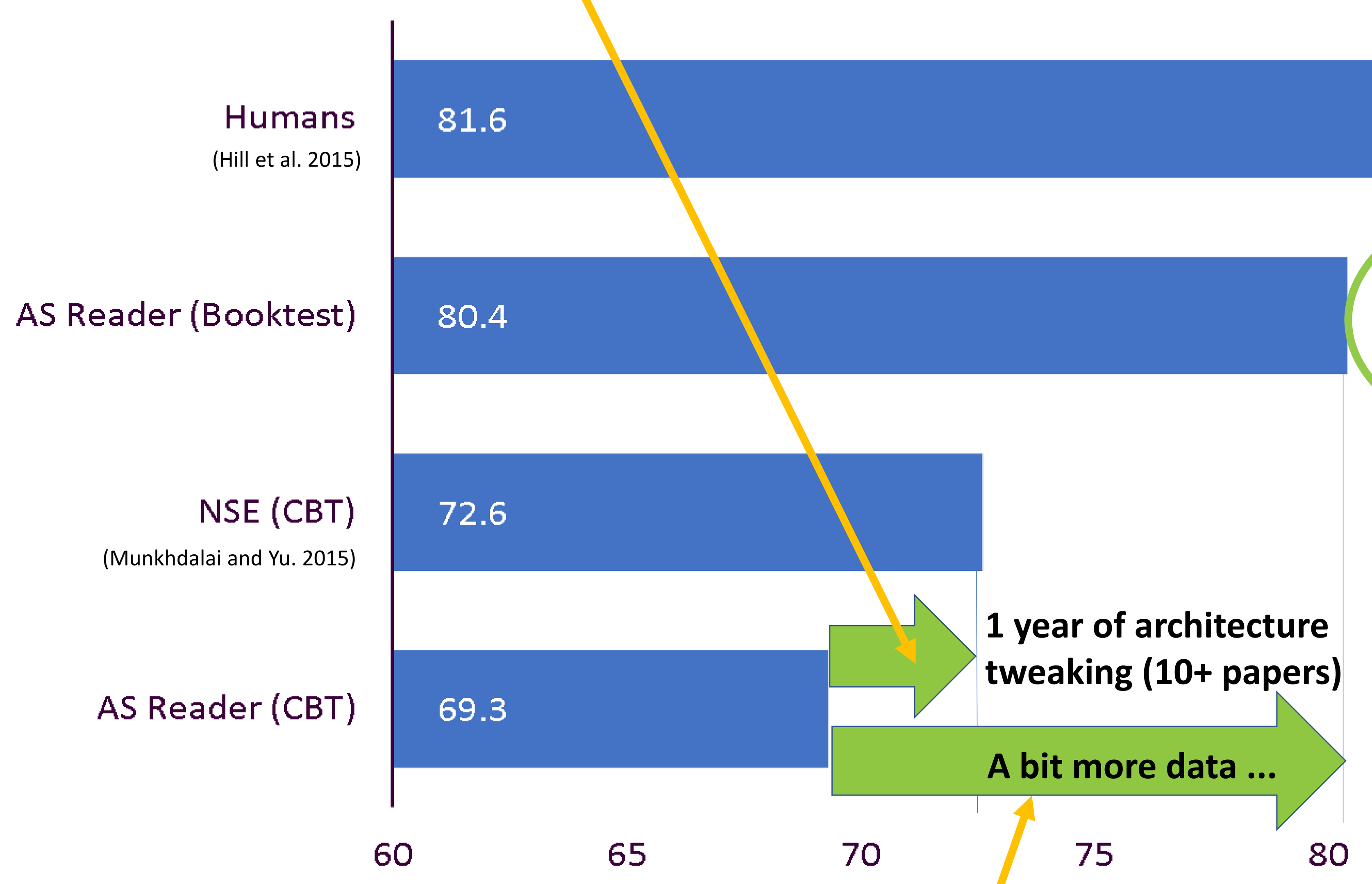
Answer candidates (10)
Bob **Alice** Rex friend dog stone lawn eye threw the

- Examples can be generated automatically from a suitable corpus
- Hence we can get an **unlimed amount of training data**
- Still **most research is done on small datasets**
  - (e.g. Children's BookTest (CBT) (Hill et al. 2015) where training the AS Reader takes only ~2 hours)
- Aren't teams focusing on an **unrealistic scenario**?

Effort well spent?

This is where most research effort goes.

Shouldn't we put more focus here?



Is there still space for improvement?

What we did: We took the successful **AS Reader** model (Kadlec et al. 2016) and examined how big an improvement more data can bring by training it on BookTest and evaluating it on CBT which allows us to compare it to the many models previously tested on CBT

Human study

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%

Yes, machines can still do much better!

Results

	Named entity		Common noun		
	valid	test	valid	test	
Humans (query) (Hill et al., 2015)	NA	52.0	NA	64.4	
Humans (context+query) (Hill et al., 2015)	NA	**81.6**	NA	**81.6**	
LSTMs (context+query) (Hill et al., 2015)	51.2	41.8	62.6	56.0	CBT training data
Memory Networks (Hill et al., 2015)	70.4	66.6	64.2	63.0	CBT training data
AS Reader (single model)	73.8	68.6	68.8	63.4	CBT training data
AS Reader (avg ensemble)	74.5	70.6	71.1	**68.9**	CBT training data
AS Reader (greedy ensemble)	76.2	**71.0**	72.4	67.5	CBT training data
GA Reader (ensemble) (Dhingra et al., 2016)	75.5	71.9	72.1	69.4	CBT training data
EpiReader (ensemble) (Trischler et al., 2016b)	76.6	71.8	73.6	70.6	CBT training data
IA Reader (ensemble) (Sordoni et al., 2016)	76.9	**72.0**	74.1	**71.0**	CBT training data
AoA Reader (single model) (Cui et al., 2016a)	77.8	**72.0**	72.2	69.4	CBT training data
NSE (Munkhdalai and Yu, 2016)	78.2	73.2	74.3	71.9	
AS Reader (single model)	80.5	76.2	83.2	80.8	BookTest training data
AS Reader (greedy ensemble)	82.3	**78.4**	85.7	**83.7**	BookTest training data

Table 2: Results of various architectures on the CBT test datasets.

New Dataset: The BookTest

- Very similar to the Children's Book Test, but more than 60x larger
- Each example consists of a context document and a cloze-style question the answer to which is one word from the document
- Generated from **10,507 copyright-free** books available thanks to project Gutenberg

	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

Available at **https://ibm.biz/booktest-v1**

Conclusions

- We believe the text comprehension community should put more focus on exploiting the vast amount of data that are available in cloze-style question answering instead of focusing chiefly on small datasets.
- Architecture improvements from small data may not help in the large-data world
- Most research is done in the setting **limited data – unlimited training time**
- We think we should pay more attention to the case **unlimited data – constrained time**
- The **Booktest** dataset is an opportunity to make a small first step in exploring these directions

References

- Long version of this paper:
  - Ondrej Bajgar, Rudolf Kadlec & Jan Kleindienst, Embracing Data Abundance: The BookTest Dataset for Reading Comprehension, arXiv:1610.00956
- Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-Attention Neural Networks for Reading Comprehension. 2016. http://arxiv.org/abs/1607.04423.
- Bhuwan Dhingra, Hanxiao Liu, William W. Cohen, and Ruslan Salakhutdinov. Gated-Attention Readers for Text Comprehension. 2016. http://arxiv.org/abs/1606.01549
- 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, Ondrej Bajgar, and Jan Kleindienst. Neural Text Understanding with Attention Sum Reader. Proceedings of ACL, 2016.
- Tsendsuren Munkhdalai and Hong Yu. Reasoning with Memory Augmented Neural Networks forLanguage Comprehension. In Proceedings of ICLR 2017
- Alessandro Sordoni, Phillip Bachman, and Yoshua Bengio. Iterative Alternating Neural Attention for Machine Reading. 2016. https://arxiv.org/abs/1606.02245.
- Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. Natural Language Comprehension with the EpiReader. 2016b. http://arxiv.org/abs/1606.02270.