IBM Watson

Machine Learning Projects at IBM Watson Prague

Rudolf Kadlec & Ondrej Bajgar

12/10/2017

IBM

Our goal

- Use Machine Learning to improve
  - Question answering
    - QA from text documents
    - Structured knowledge bases
  - Human-machine interaction
    - Dialog systems

Text comprehension: Attention Sum Reader (AS Reader)

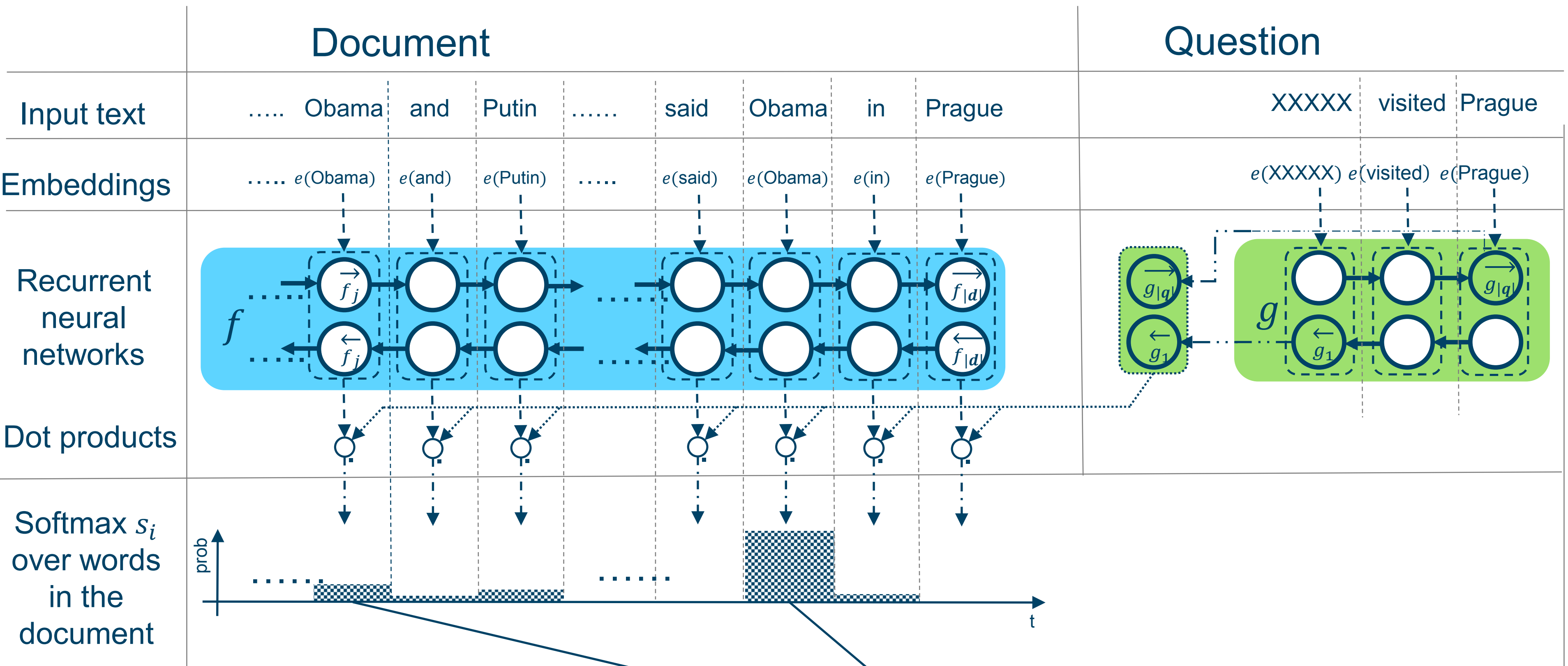
Kadlec, R., Schmid, M., Bajgar, O., & Kleindienst, J. (2016). Neural Text Understanding with Attention Sum Reader. *Proceedings of ACL*. https://arxiv.org/abs/1603.01547

Opensourced: https://github.com/rkadlec/asreader

CNN and Daily Mail (DeepMind)

	Original Version	Anonymised Version
Context		
	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broad-caster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " …
Query		
	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	Producer **X** will not press charges against *ent212*, his lawyer says.
Answer		
	Oisin Tymon	*ent193*

ASReader



Probability of the answer

$$P(\mathrm{Obama}|q, d) = \sum_{i \in I(Obama, d)} s_i = s_j + s_{j+5}$$

This is the key difference compared to DeepMind's models. They use attention to compute weighted sum of word vectors from the document.

CNN and Daily Mail dataset

	CNN		Daily Mail	
	valid	test	valid	test
Deep LSTM Reader †	55.0	57.0	63.3	62.2
Attentive Reader †	61.6	63.0	70.5	69.0
Impatient Reader †	61.8	63.8	69.0	68.0
MemNNs (single model) ‡	63.4	66.8	NA	NA
MemNNs (ensemble) ‡	66.2	69.4	NA	NA
Att-Sum Reader (single model)	68.6	69.5	74.9	73.7
Att-Sum Reader (avg for top 20%)	68.4	69.9	74.5	73.5
Att-Sum Reader (avg ensemble)	73.9	**75.4**	78.0	77.1
Att-Sum Reader (greedy ensemble)	74.5	74.8	78.5	**77.4**

Children's Book Test

	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
Memory Networks (Hill et al., 2015)	70.4	66.6	64.2	63.0
AS Reader (single model)	73.8	68.6	68.8	63.4
AS Reader (avg ensemble)	74.5	70.6	71.1	**68.9**
AS Reader (greedy ensemble)	76.2	**71.0**	72.4	67.5
GA Reader (ensemble) (Dhingra et al., 2016)	75.5	71.9	72.1	69.4
EpiReader (ensemble) (Trischler et al., 2016b)	76.6	71.8	73.6	70.6
IA Reader (ensemble) (Sordoni et al., 2016)	76.9	**72.0**	74.1	**71.0**
AoA Reader (single model) (Cui et al., 2016a)	77.8	**72.0**	72.2	69.4

Models based on IBM's ASReader

Summary

- Easy to implement
- Trains faster than attention blending NNs (e.g., Stanford’s system)

Finding a Jack-of-All-Trades:

An Examination of Transfer Learning in Text Comprehension

Kadlec, R., Bajgar, O., Hrinčár, P., Kleindienst, J.
IBM Watson, Prague lab

Generalization is the key

Cloze style questions

Children's Book Test (Hill et al 2015)

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

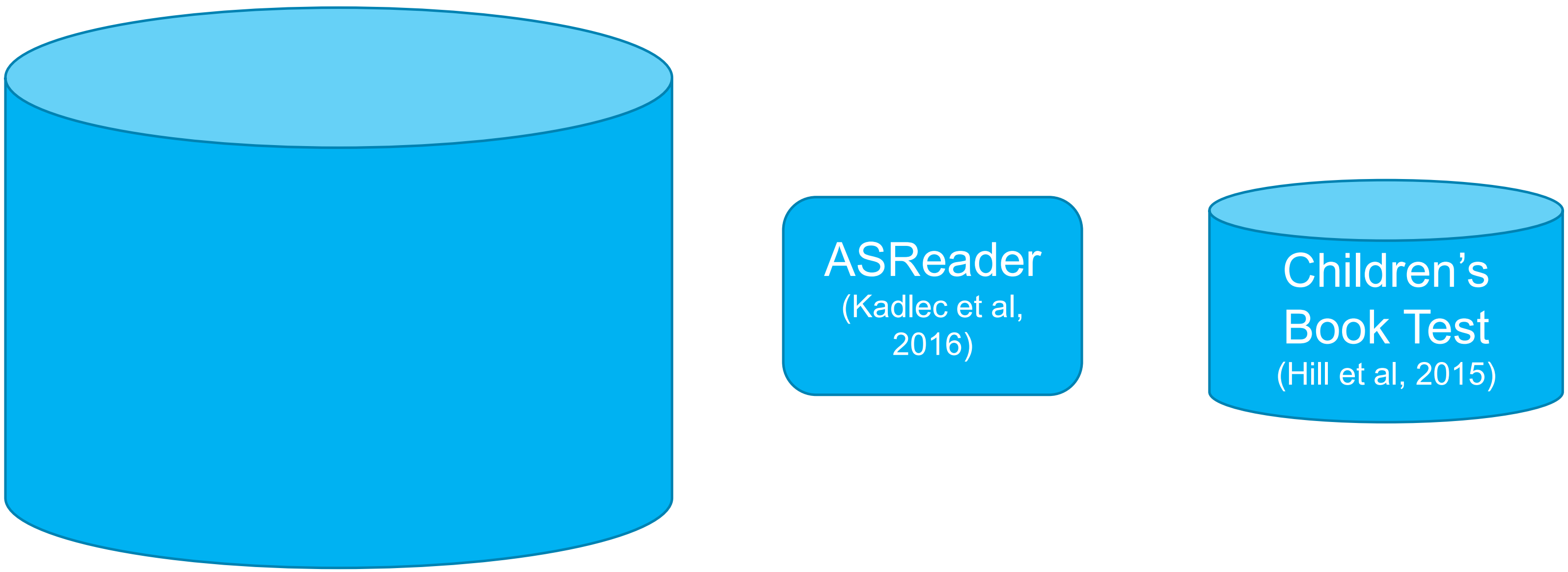
a: Baxter

~ 200k examples (CN+NE)

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations

Starting point

Train | ML Model | Test



BookTest (Bajgar et al, 2016)
14M examples

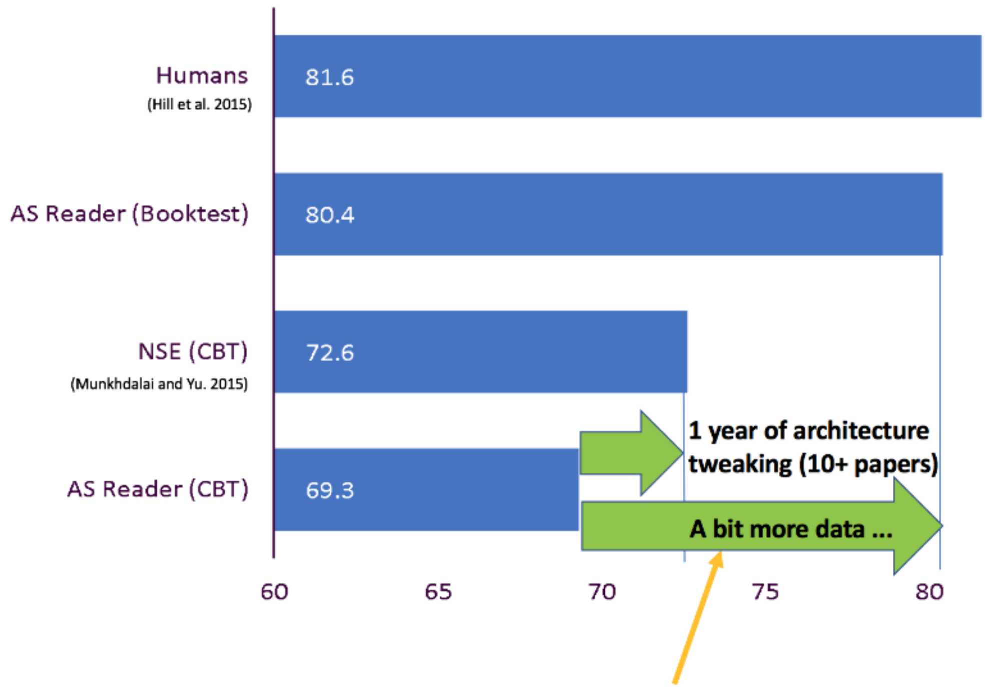
CBT dev/test
2k examples

Bajgar, O., Kadlec, R., & Kleindienst, J. (2016). Embracing data abundance: BookTest Dataset for Reading Comprehension. http://arxiv.org/abs/1610.00956

BookTest

	Named entity		Common noun	
	valid	test	valid	test
Humans (context+query) (Hill et al., 2015)	NA	**81.6**	NA	**81.6**
AS Reader (ensemble) (Kadlec et al., 2016)	76.2	**71.0**	72.4	67.5
GA Reader (ensemble) (Dhingra et al., 2016)	75.5	71.9	72.1	69.4
EpiReader (ensemble) (Trischler et al., 2016b)	76.6	71.8	73.6	70.6
IA Reader (ensemble) (Sordoni et al., 2016)	76.9	**72.0**	74.1	**71.0**
AoA Reader (single model) (Cui et al., 2016a)	77.8	**72.0**	72.2	69.4

Embracing data abundance



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

BookTest

Is there potential for further growth?

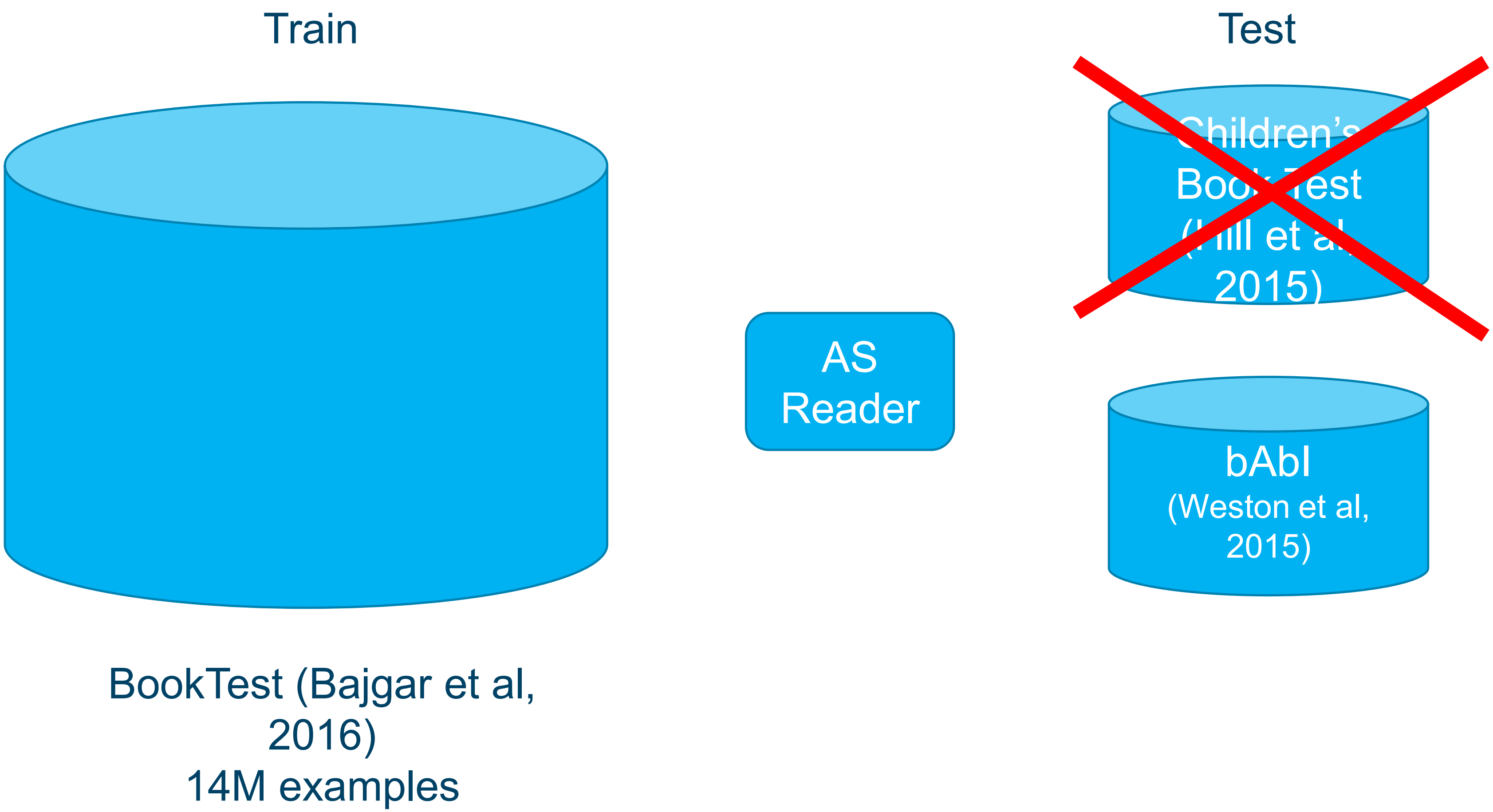
- Human study
  - Performed on the ~20% of examples where AS Reader failed

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

There‘s still plenty of space for improvement!

→ opportunity for other teams to improve on BookTest

Transfer learning?



Simple testing tasks: bAbI tasks

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Simple testing tasks: bAbI tasks

Task 11: Basic Coreference

Daniel was in the kitchen.
Then he went to the studio.
Sandra was in the office.
Where is Daniel? A:studio

Task 12: Conjunction

Mary and Jeff went to the kitchen.
Then Jeff went to the park.
Where is Mary? A: kitchen
Where is Jeff? A: park

Task 13: Compound Coreference

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning

In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Task 15: Basic Deduction

Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of? A:wolves

Task 16: Basic Induction

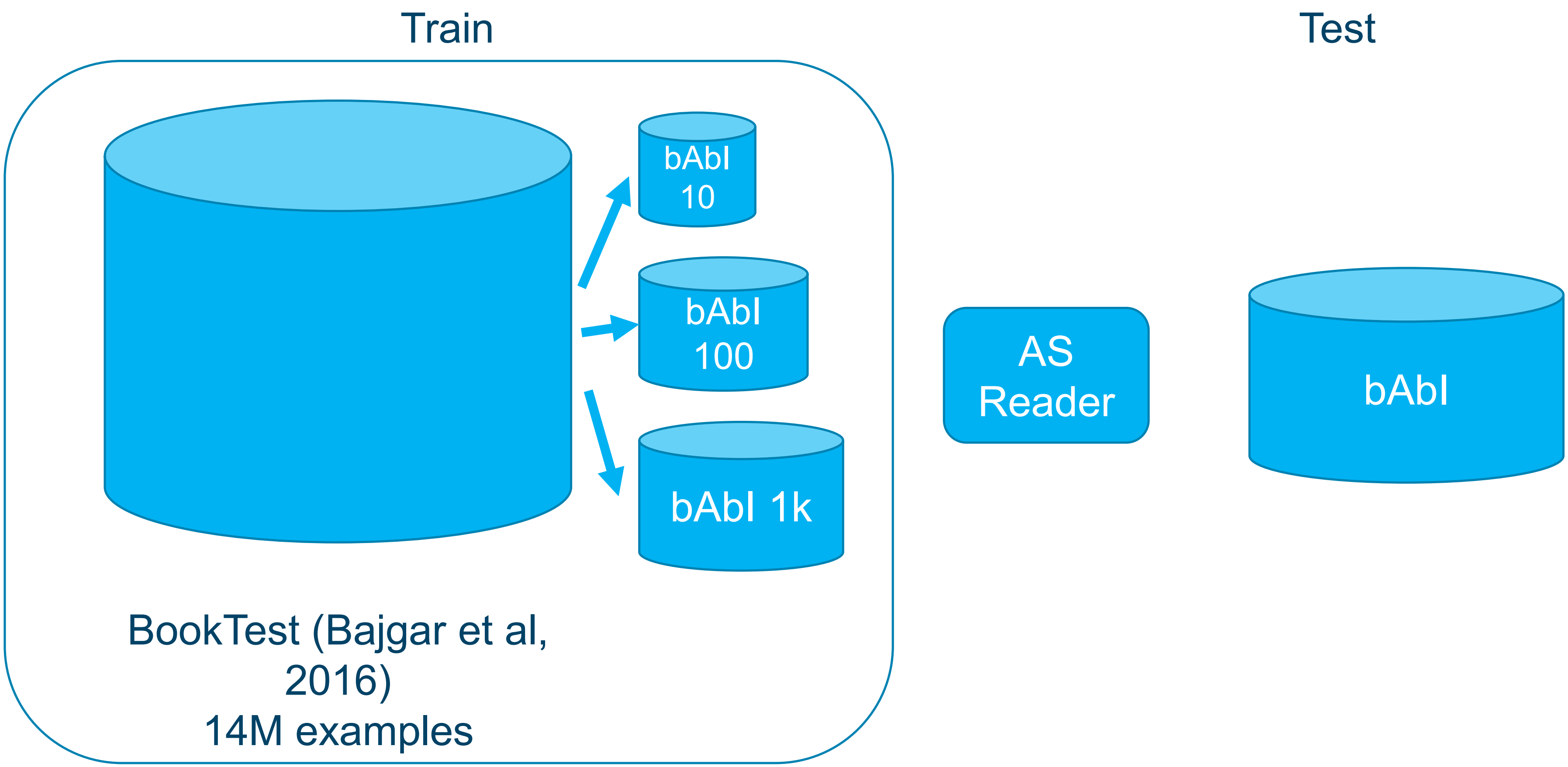
Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg? A:white

Can it generalize what it learned?
Not really ...

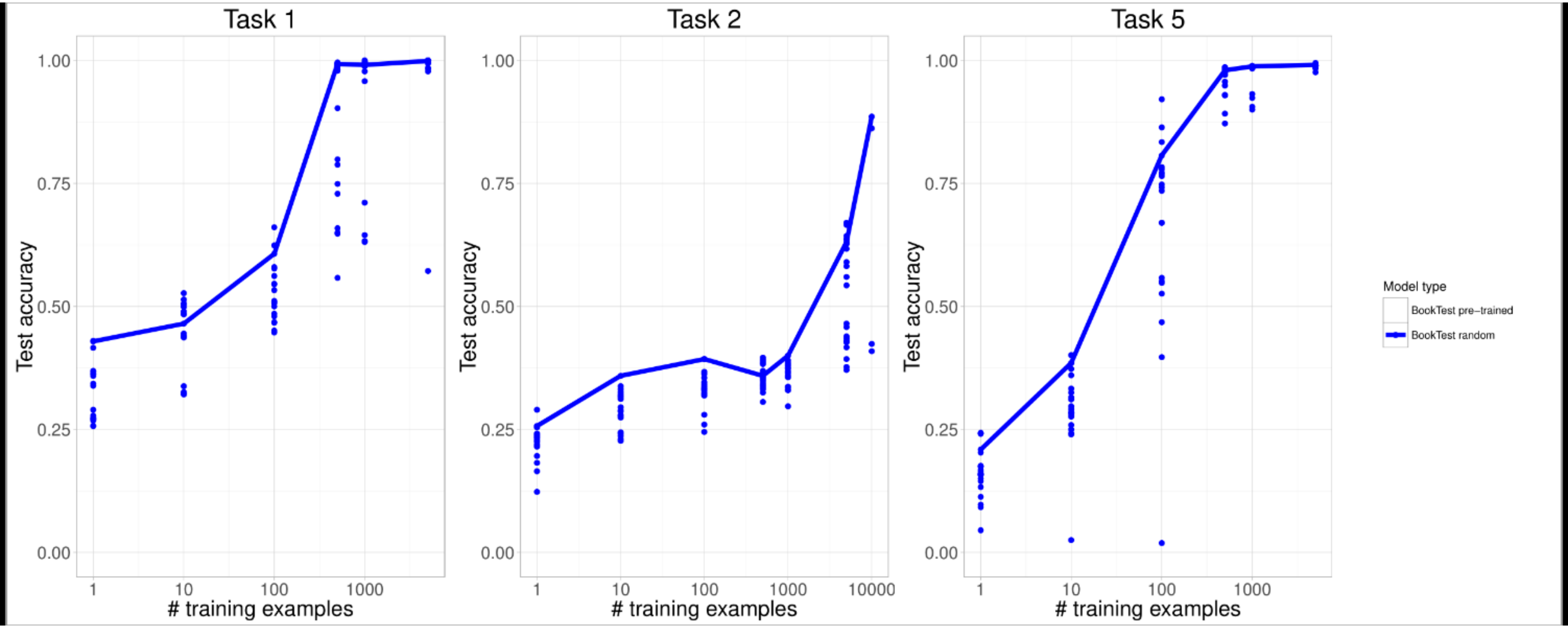
	Model:	Random	Rnd cand.	MemN2N (single) (PE LS RN)	MemN2N (single) (PE LS LW RN)	DMN+ (single)	ASReader	
	Train dataset / Test dataset	not trained	bAbI 10k	bAbI 1k	bAbI 10k	bAbI 10k	bAbI 10k	BookTest 14M
1	Single supporting fact	7.80	31.20	100.00	100.00	100.00	100.00	37.30
2	Two supporting facts	4.40	26.96	91.70	99.70	99.70	91.90	25.80
3	Three supporting facts	3.40	19.14	59.70	97.90	98.90	86.00	22.20
4	Two-argument relations	10.50	33.58	97.20	100.00	100.00	100.00	50.30
5	Three-argument relations	4.40	21.42	86.90	99.20	99.50	99.80	67.60
11	Basic coreference	6.20	30.42	99.10	99.90	100.00	100.00	33.00
12	Conjunction	6.70	27.25	99.80	100.00	100.00	100.00	30.40
13	Compound coreference	5.60	27.73	99.60	100.00	100.00	100.00	33.80
14	Time reasoning	5.00	27.82	98.30	99.90	99.80	95.00	27.60
15	Basic deduction	5.20	37.20	100.00	100.00	100.00	96.70	39.90
16	Basic induction	7.50	45.65	98.70	48.20	54.70	50.30	15.10

BAD!

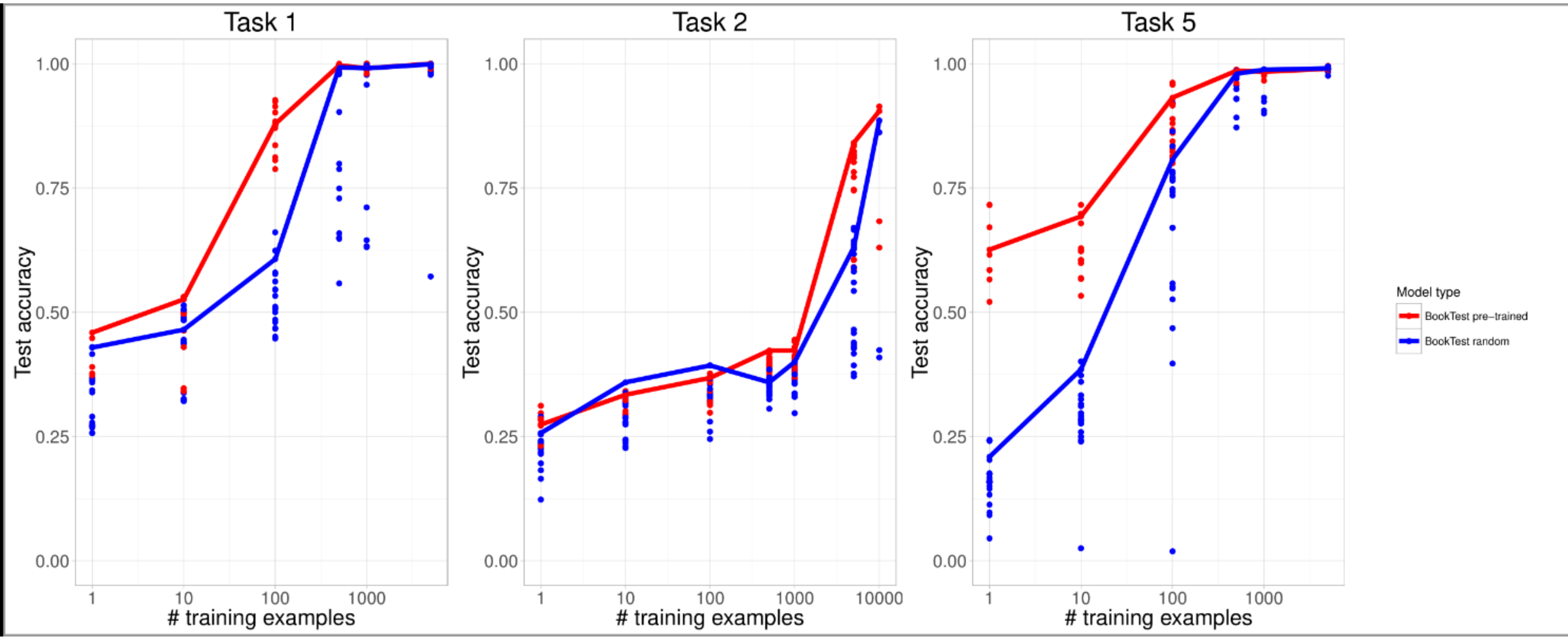
Finetuning - bAbI

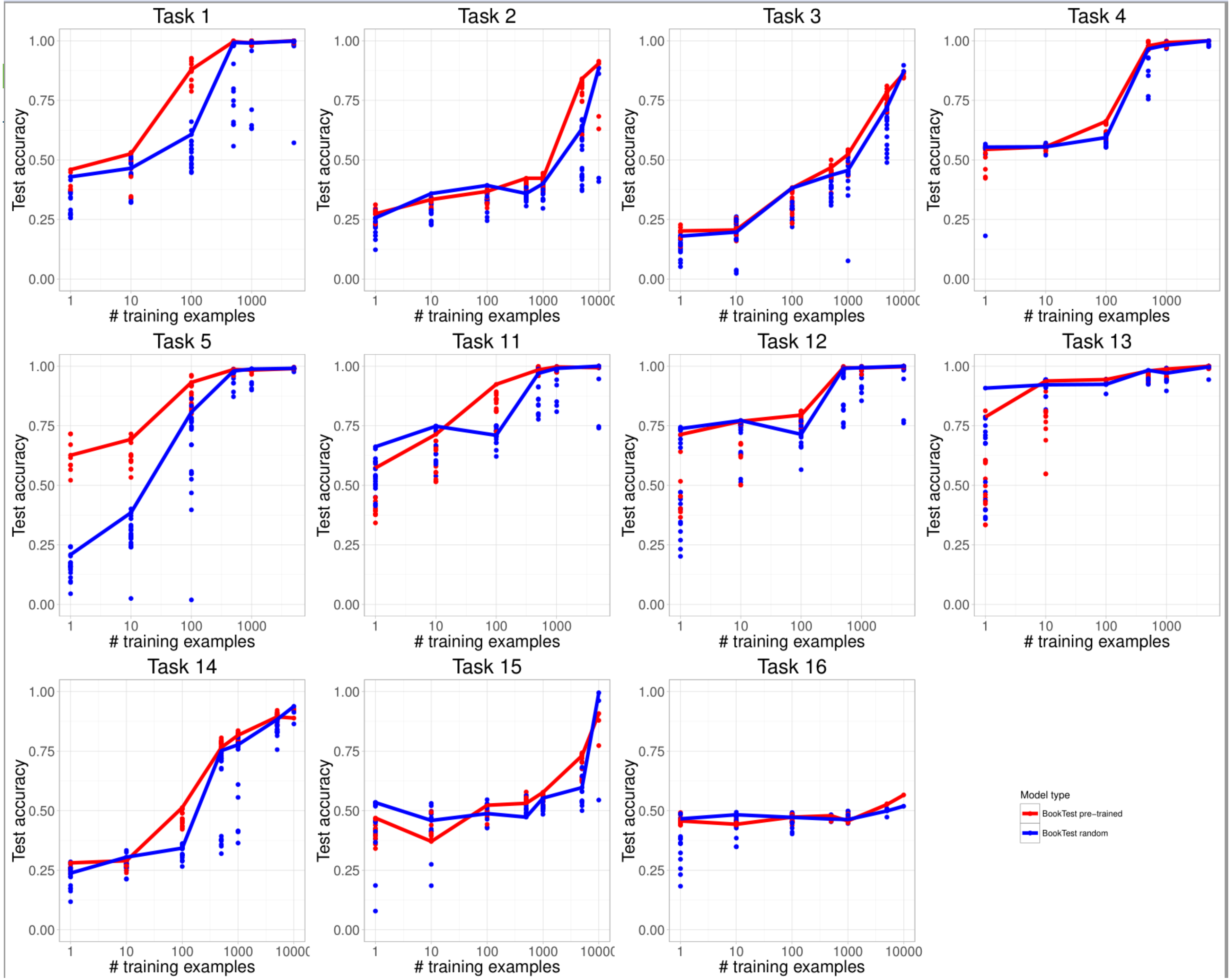


2nd Experiment:
It does better with target-adjustment!



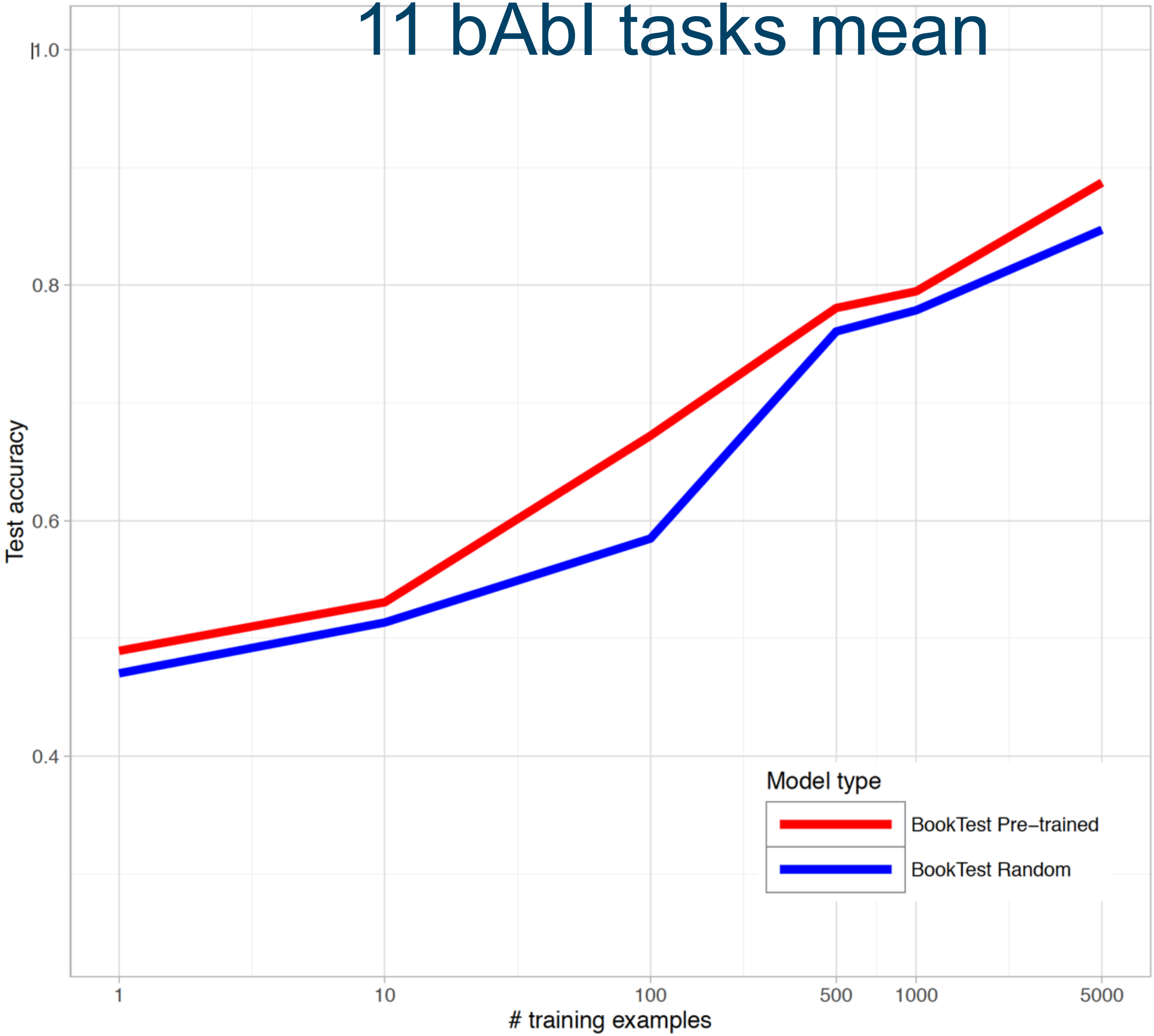
2nd Experiment:
It does better with target-adjustment!




Task 1
Task 2
Task 3
Task 4
Task 5
Task 11
Task 12
Task 13
Task 14
Task 15
Task 16
Test accuracy
training examples
0.00
0.25
0.50
0.75
1.00
1
10
100
1000
10000
Model type
BookTest pre-trained
BookTest random

11 bAbI tasks mean

b



Knowledge Base Completion

Kadlec, R., Bajgar, O., & Kleindienst, J. (2017). Knowledge Base Completion: Baselines Strike Back. *Repl4NLP Workshop at ACL 2017*.

Knowledge base completion

- Goal
  - Understand structured data
  - Given KG train NN model that can predict missing information
  - Entity prediction:
    - given query (subject, predicate, ?)
    - predict the correct object

KBC: Our work

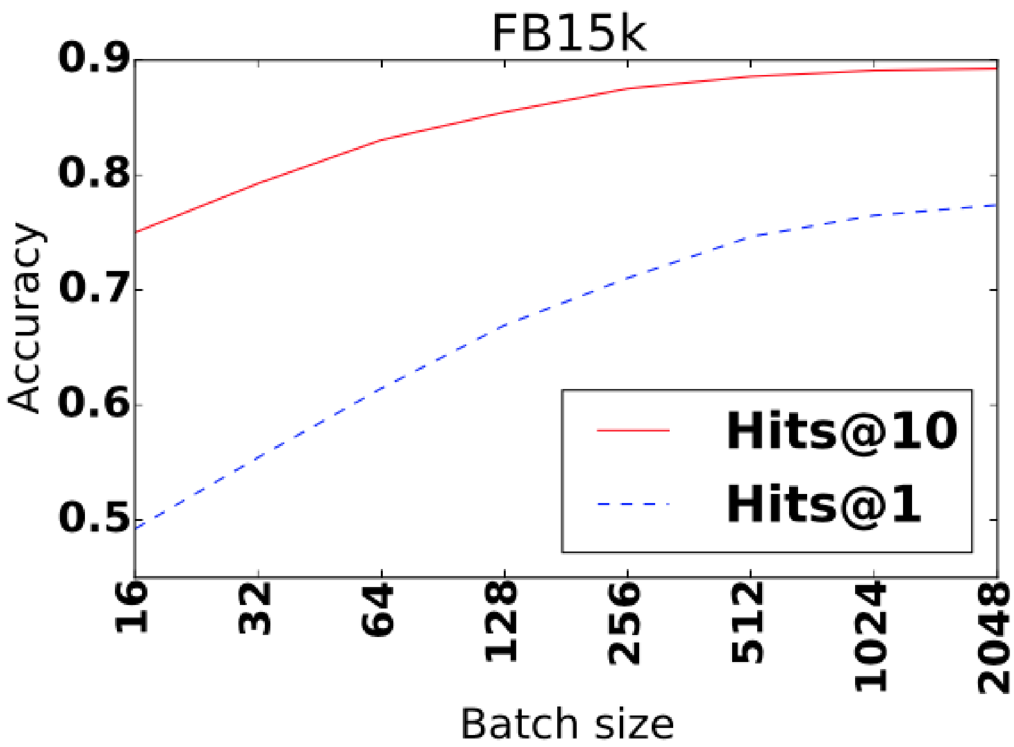
- We evaluated performance of baseline models on standard datasets
  - FB15k (derived from Freebase)
  - WN18 (derived from WordNet)

- To our surprise a simple baseline --- DistMult model (Yang et al. 2015) with proper training objective scored competitively

Our work: Results

- DistMult is in top 3 results for 4 out of 6 commonly reported metrics!

Method	Filtered WN18 MR	WN18 H10	WN18 MRR	FB15k MR	FB15k H10	FB15k MRR	Extra features
SE (Bordes et al., 2011)	985	80.5	-	162	39.8	-	None
Unstructured (Bordes et al., 2014)	304	38.2	-	979	6.3	-	
TransE (Bordes et al., 2013)	251	89.2	-	125	47.1	-	
TransH (Wang et al., 2014)	303	86.7	-	87	64.4	-	
TransR (Lin et al., 2015b)	225	92.0	-	77	68.7	-	
CTransR (Lin et al., 2015b)	218	92.3	-	75	70.2	-	
KG2E (He et al., 2015)	331	92.8	-	59	74.0	-	
TransD (Ji et al., 2015)	212	92.2	-	91	77.3	-	
lppTransD (Yoon et al., 2016)	270	94.3	-	78	78.7	-	
TranSparse (Ji et al., 2016)	211	93.2	-	82	79.5	-	
TATEC (Garcia-Duran et al., 2016)	-	-	-	58	76.7	-	
NTN (Socher et al., 2013)	-	66.1	0.53	-	41.4	0.25	
HolE (Nickel et al., 2016)	-	94.9	**0.938**	-	73.9	0.524	
STransE (Nguyen et al., 2016)	**206**	93.4	0.657	69	79.7	0.543	
ComplEx (Trouillon et al., 2017)	-	94.7	**0.941**	-	84.0	0.692	
ProjE wlistwise (Shi and Weniger, 2017)	-	-	-	**34**	88.4	-	
IRN (Shen et al., 2016)	249	**95.3**	-	**38**	**92.7**	-	
RTransE (García-Durán et al., 2015)	-	-	-	50	76.2	-	Path
PTransE (Lin et al., 2015a)	-	-	-	58	84.6	-	
GAKE (Feng et al., 2015)	-	-	-	119	64.8	-	
Gaifman (Niepert, 2016)	352	93.9	-	75	84.2	-	
Hiri (Liu et al., 2016)	-	90.8	0.691	-	70.3	0.603	
R-GCN+ (Schlichtkrull et al., 2017)	-	**96.4**	0.819	-	84.2	0.696	
NLFeat (Toutanova and Chen, 2015)	-	94.3	**0.940**	-	87.0	**0.822**	Text
TEKE_H (Wang and Li, 2016)	**114**	92.9	-	108	73.0	-	
SSP (Xiao et al., 2017)	**156**	93.2	-	82	79.0	-	
DistMult (orig) (Yang et al., 2015)	-	94.2	0.83	-	57.7	0.35	None
DistMult (Toutanova and Chen, 2015)	-	-	-	-	79.7	0.555	
DistMult (Trouillon et al., 2017)	-	93.6	0.822	-	82.4	0.654	
Single DistMult (this work)	655	94.6	0.797	42.2	**89.3**	**0.798**	
Ensemble DistMult (this work)	457	**95.0**	0.790	**35.9**	**90.4**	**0.837**	


FB15k
Accuracy
0.9
0.8
0.7
0.6
0.5
16
32
64
128
256
512
1024
2048
Batch size
Hits@10
Hits@1

KBC: Our work - Implications

- DistMult assumes all relations are symmetric!
- =>
- Either
  - The datasets are odd, or
  - Current standard metrics are improper, or
  - Previous models weren't pushed to their limits

$$s(\mathbf{h}, \mathbf{r}, \mathbf{t}) = \mathbf{h}^T \cdot W_{\mathbf{r}} \cdot \mathbf{t} = \sum_{i=1}^{N} h_i r_i t_i$$

IBM Watson

Hybrid Dialog State Tracker

M Vodolán, R Kadlec, J Kleindienst

EACL 2017

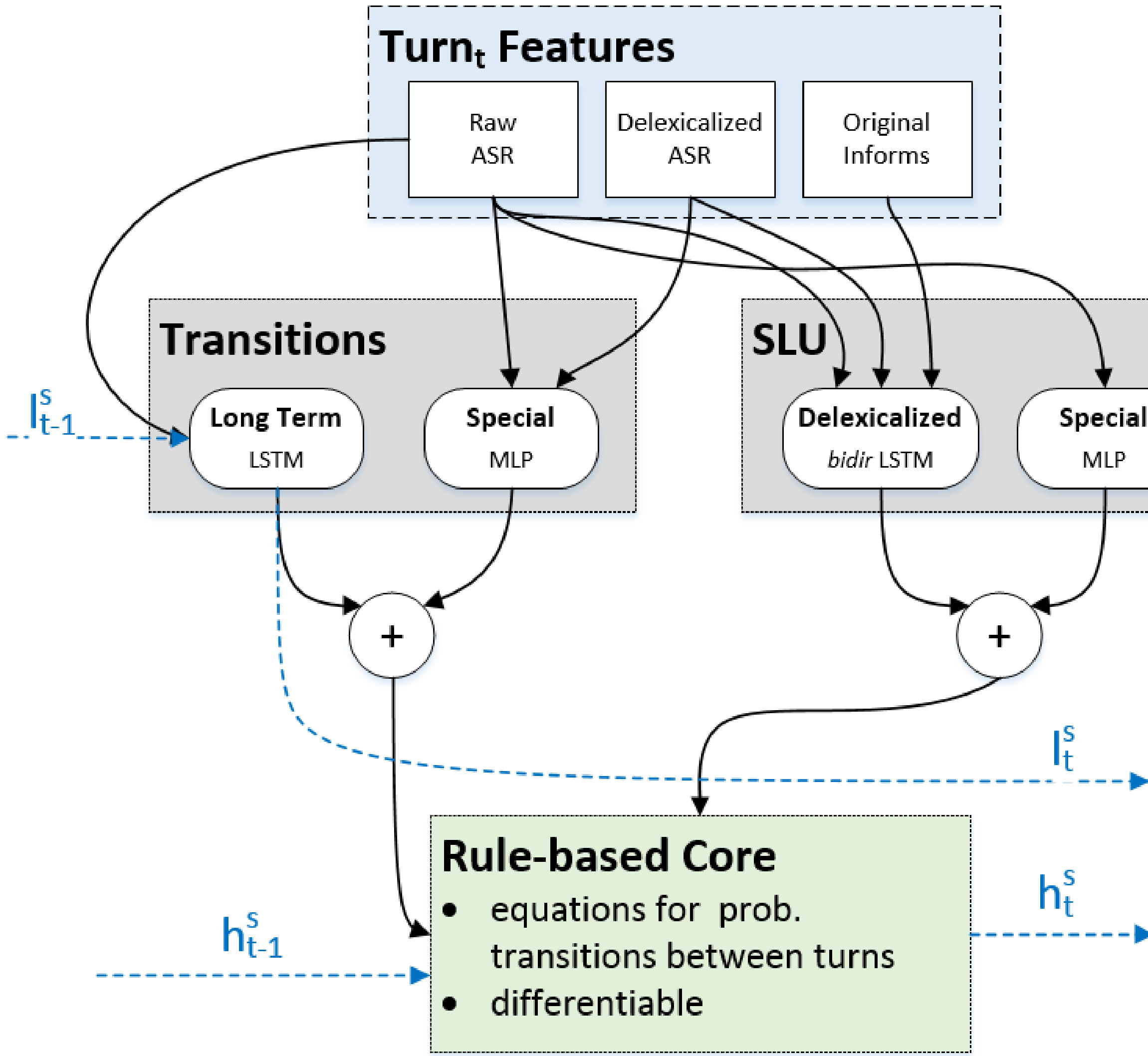
IBM

Belief Tracking

- Accumulation of evidence about user goal
- Helps to improve ASR misunderstandings during dialog

	Belief state
U: I would restaurant with indian food ***SLU:*** *italian ~ 0.6, indian ~ 0.4*	***italian ~ 0.6*** *indian ~ 0.4*
M: What type of food would you like?	
U: Indian ***SLU:*** *indonesian ~ 0.6, indian ~ 0.4*	***indian ~ 0.6*** *italian ~ 0.2* *indonesian ~ 0.2*

HDST with ASR Features – Architecture



HDST with ASR Features – SLU Motivation

- Delexicalized unit

  I don’t want *%value% %slot%*

- Specialized unit

  It all an food please

  ASR error

  Italian

HDST with ASR Features – Results

- **DSTC2** (2014)
  - restaurant search
  - 2 000 training dialogs

	ASR	Batch ASR	Accuracy (dstc2_test)	L2 (dstc2_test)	post DSTC	test validated
Hybrid Tracker – this work	√	√	**.810**	.318	√	√
DST2 stacking ensemble [11]	√	√	.798	**.308**	√	√
Hybrid Tracker – this work	√	√	**.796**	**.338**	√	
Williams [4]	√	√	.784	.735		
Hybrid Tracker – this work	√		**.780**	.356	√	
Williams [4]	√		.775	.758		
Henderson et al. [5]	√		.768	**.346**		
Yu et al. [12]	√		.762	.436	√	

Table 1: Joint slot tracking results for various systems reported in the literature. The trackers that used ASR/Batch ASR have √ in the corresponding column. The results of systems that did not participate in DSTC2 are marked by √ in the "post DSTC" column. The first group shows results of trackers that used dstc test data for validation. The second group lists individual trackers that use ASR and Batch ASR features. The third group lists systems that use only the ASR features.

Quantitave evaluation of Deep Learning models

Ongoing work

How do we tell which architecture / algorithm is better?

Quantitative evaluation

- Need to choose:
  - **Metric**
    - E.g. Accuracy, BLEU, cross-entropy, Hits@10
    - Each covers a different aspect of performance
  - **Dataset**
    - ImageNet, SQuAD, Penn Treebank
    - Again measures only some subskills
  - **Comparison methodology**
    - Comparison criterion
    - Statistical technique

Ideally an architecture should be evaluated across multiple datasets/metrics

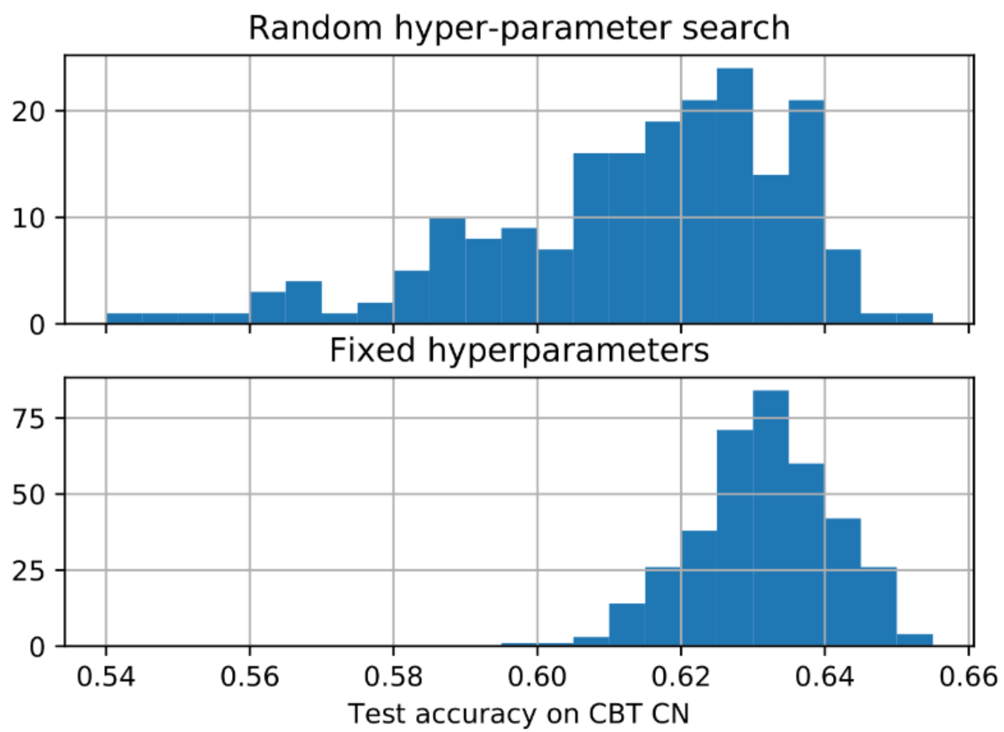
Current standard in Deep Learning

- 1 metric
- 1 dataset
- sometimes probably cherry-picked from among several

Problems

- Usually the result of the best single model is reported

- Does not account for random variation in metric scores



Thank you!
Any questions?