Supplementary for "Learning Answer-Entailing Structures for Machine Comprehension"

Mrinmaya Sachan[1] **Avinava Dubey**[1] **Matthew Richardson**[2] **Eric P. Xing**[1]
[1]Carnegie Mellon University [2]Microsoft Research
[1]{mrinmays, akdubey, epxing}@cs.cmu.edu [2]mattri@microsoft.com

1 Results on MCTest dataset

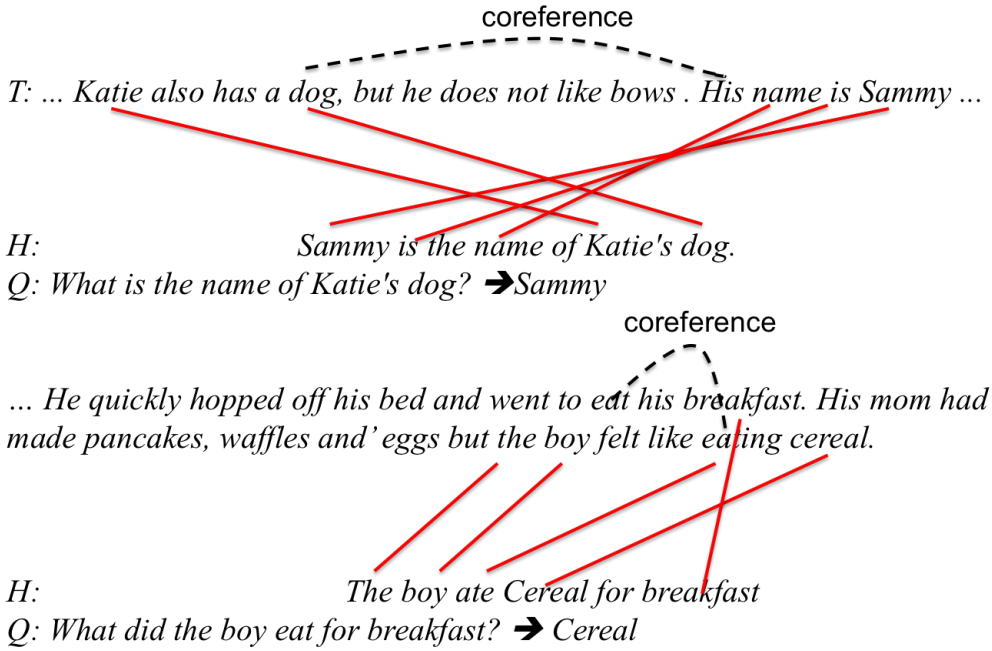
Table 1 shows the detailed numbers in the Figure 2 of the main paper.

2 Results on bAbl dataset

Table 2 shows the complete results of various LSSVMmodels on the *bAbI* datasets for each subtask. In our experiments, we observed a similar general pattern of improvement of LSSVM over the baselines as well as the improvement due to multi-task learning. Again task classification helped the multi-task learner the most and the QA classification helped more than the QClassification. The results on performance within the subtasks described in the main paper are substantiated by these numbers.

3 Structures Learned

Some more examples of the *text-entailing* structures learned by of model on the MCTest real data are given in Figure 1



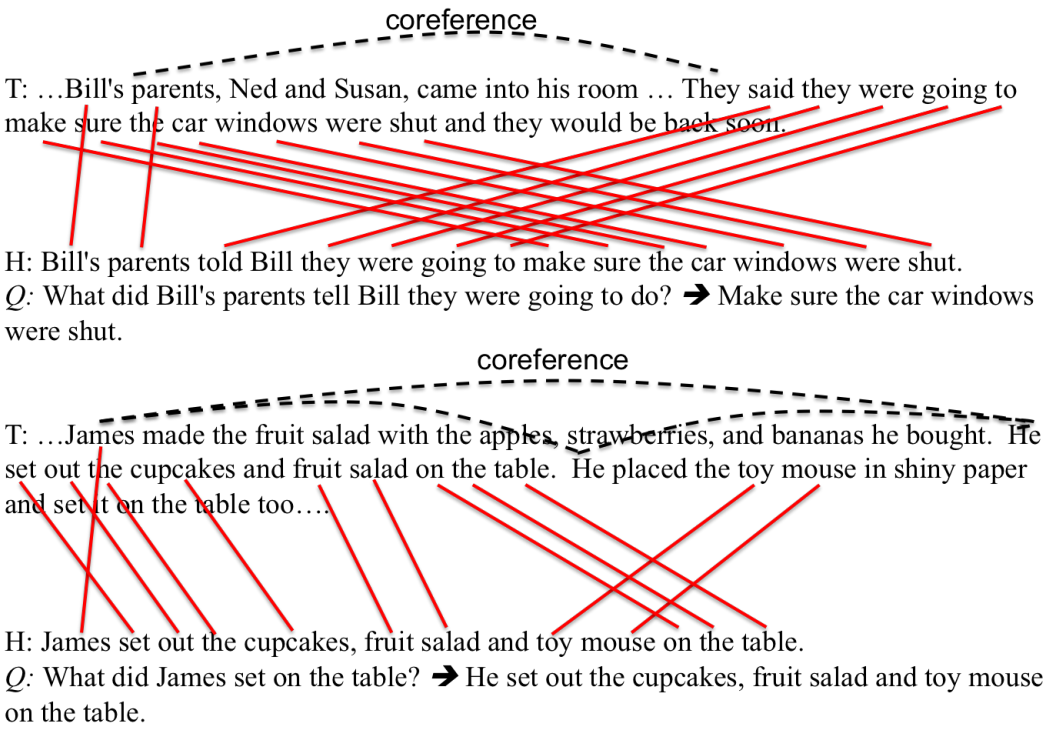


Figure 1: Some more latent *answer-entailing* structures learned by our model.

.

			Single	Multiple	All
LSSVM		**Sentence**	62.16/0.854	60.23/0.825	61.28/0.839
		Subset	61.83/0.841	65.75/0.862	63.97/0.852
		Subset+	61.12/0.835	66.67/0.864	64.15/0.852
		Subset+/Negation	63.24/0.857	66.15/0.863	64.83/0.861
	MultiTask	**Subset+/Negation QClassification**	64.34/0.860	66.46/0.864	65.50/0.863
		Subset+/Negation QAClassification	66.18/0.863	67.37/0.866	66.83/0.865
		Subset+/Negation TaskClassification	67.65/0.867	67.99/0.869	67.83/0.868
Baselines		**SW**	54.56/0.785	54.04/0.784	54.28/0.784
		SW+D	62.99/0.834	58.00/0.805	59.93/0.818
		RTE	69.85/0.869	42.71/0.728	55.01/0.791
		LSTM	62.13/0.833	58.84/0.811	60.33/0.821
		QANTA	63.23/0.842	59.45/0.820	61.00/0.830

Table 1: Comparison of variations of our method against several baselines on the MCTest-500 dataset. The table shows two statistics, accuracy and NDCG_4 (written as accuracy/NDCG_4) on the test set of MCTest-500. All differences between the baselines and LSSVMs, the improvement due to negation and the improvements due to multi-task learning are significant ($p < 0.01$) using the two-tailed paired T-test.

Tasks	Baslines				LSSVM						
									MultiTask		
	SW	RTE	LSTM	QANTA	Sentence	Subset	Subset+	Subset+/Negation	Subset+/Negation QClassification	Subset+/Negation QAClassification	Subset+/Negation TaskClassification
Single Supporting Fact	36	98	50	89	100	100	100	100	100	100	100
Two Supporting Facts	2	79	20	69	60	91	92	91	93	93	94
Three Supporting Facts	7	46	20	42	52	84	86	84	86	87	88
Two Arg. Relations	50	54	61	68	89	91	91	90	92	93	93
Three Arg. Relations	20	31	70	63	84	89	89	88	91	90	91
Yes/No Questions	49	48	48	54	58	58	58	78	81	84	85
Counting	52	11	49	55	61	59	63	61	65	64	64
Lists/Sets	42	34	45	47	55	72	73	71	77	80	82
Simple Negation	62	56	64	72	63	63	64	76	79	80	81
Indefinite Knowledge	45	43	44	68	74	74	78	87	88	91	92
Basic Coreference	25	31	72	80	91	93	96	96	97	97	98
Conjunction	9	59	74	86	94	91	91	90	95	96	97
Compound Coreference	26	72	94	95	86	89	89	88	93	93	94
Time Reasoning	19	68	27	43	65	68	70	68	71	74	76
Basic Deduction	20	49	21	72	76	74	78	76	80	81	82
Basic Induction	43	53	23	55	57	59	61	58	61	63	64
Positional Reasoning	46	66	51	55	81	85	88	88	90	91	90
Size Reasoning	52	77	52	63	78	82	84	83	85	87	89
Path Finding	0	11	8	45	9	9	9	9	11	11	11
Agents Motivations	76	91	91	93	66	69	70	68	69	69	70
Mean Performance	34	54	49	66	70	75	77	78	79	81	82

Table 2: Comparison of accuracies on the variations of our method against several baselines on 20 Tasks of the bAbI dataset. All integer differences are significant ($p < 0.01$) using the two-tailed paired T-test.