Question Answering through Transfer Learning from Large Fine-grained Supervision Data

Sewon Min[*]
Seoul National University
shmsw25@snu.ac.kr

Minjoon Seo
University of Washington
minjoon@uw.edu

Hannaneh Hajishirzi
University of Washington
hannaneh@uw.edu

Abstract

We show that the task of question answering (QA) can significantly benefit from the transfer learning of models trained on a different large, fine-grained QA dataset. We achieve the state of the art in two well-studied QA datasets, WikiQA and SemEval-2016 (Task 3A), through a basic transfer learning technique from SQuAD. For WikiQA, our model outperforms the previous best model by more than 8%. We demonstrate that finer supervision provides better guidance for learning lexical and syntactic information than coarser supervision, through quantitative results and visual analysis. We also show that a similar transfer learning procedure achieves the state of the art on an entailment task.

1 Introduction

Question answering (QA) is a long-standing challenge in NLP, and the community has introduced several paradigms and datasets for the task over the past few years. These paradigms differ from each other in the type of questions and answers and the size of the training data, from a few hundreds to millions of examples.

We are particularly interested in the context-aware QA paradigm, where the answer to each question can be obtained by referring to its accompanying context (paragraph or a list of sentences). Under this setting, the two most notable types of supervisions are coarse *sentence-level* and fine-grained *span-level*. In sentence-level QA, the task is to pick sentences that are most relevant to the question among a list of candidates (Yang et al., 2015). In span-level QA, the task is to locate the smallest span in the given paragraph that answers the question (Rajpurkar et al., 2016).

In this paper, we address coarser, sentence-level QA through a standard transfer learning[1] technique of a model trained on a large, span-supervised QA dataset. We demonstrate that the target task not only benefits from the scale of the source dataset but also the capability of the fine-grained span supervision to better learn syntactic and lexical information.

For the source dataset, we pretrain on SQuAD (Rajpurkar et al., 2016), a recently-released, span-supervised QA dataset. For the source and target models, we adopt BiDAF (Seo et al., 2017), one of the top-performing models in the dataset's leaderboard. For the target datasets, we evaluate on two recent QA datasets, WikiQA (Yang et al., 2015) and SemEval 2016 (Task 3A) (Nakov et al., 2016), which possess sufficiently different characteristics from that of SQuAD. Our results show 8% improvement in WikiQA and 1% improevement in SemEval. In addition, we report state-of-the-art results on recognizing textual entailment (RTE) in SICK (Marelli et al., 2014) with a similar transfer learning procedure.

2 Background and Data

Modern machine learning models, especially deep neural networks, often significantly benefit from transfer learning. In computer vision, deep convolutional neural networks trained on a large image classification dataset such as ImageNet (Deng et al., 2009) have proved to be useful for initializing models on other vision tasks, such as object detection (Zeiler and Fergus, 2014). In nat-

---

[*] All work was done while the author was an exchange student at University of Washington.

[1] The borderline between transfer learning and domain adaptation is often ambiguous (Mou et al., 2016). We choose the term "transfer learning" because we also adapt the pretrained QA model to an entirely different task, RTE.

Span-level QA		Sentence-level QA				RTE	
	SQuAD		*WikiQA*	*SemEval-2016 Task 3A*			*SICK*
Q	Which company made Spectre?	**Q**	Who made airbus	I saw an ad, data entry jobs online. It required we give a fee and they promise fixed amount every month. Is this a scam?		**P**	Four kids are doing backbends in the park.
C	Spectre (2015) is the 24th James Bond film produced by Eon Productions. It features (...)	**C1**	Airbus SAS is an aircraft manufacturing subsidiary of EADS, a European aerospace company.	well probably is so i be more careful if i were u. Why you looking for online jobs		**H**	Four girls are doing backbends and playing outdoors.
		C2	Airbus began as an union of aircraft companies.	SCAM!!!!!!!!!!!!!!!!!!!!			
		C3	Aerospace companies allowed the establishment of a joint-stock company, owned by EADS.	Bcoz i got a baby and iam nt intrested to sent him in a day care. thats y iam (...)			
A	"Eon Productions"	**A**	C1(Yes), C2(No), C3(No)	C1(Good), C2(Good), C3(Bad)		**A**	Entailment

Table 1: Examples of question-context pairs from QA datasets and premise-hypothesis pair from RTE dataset. **Q** indicates question, **C** indicates context, **A** indicates answer, **P** indicates premise and **H** indicates hypothesis.

ural language processing, domain adaptation has traditionally been an important topic for syntactic parsing (McClosky et al., 2010) and named entity recognition (Chiticariu et al., 2010), among others. With the popularity of distributed representation, pre-trained word embedding models such as word2vec (Mikolov et al., 2013b,a) and glove (Pennington et al., 2014) are also widely used for natural language tasks (Karpathy and Fei-Fei, 2015; Kumar et al., 2016). Instead of these, we initialize our models from a QA dataset and show how standard transfer learning can achieve state-of-the-art in target QA datasets.

There have been several QA paradigms in NLP, which can be categorized by the context and supervision used to answer questions. This context can range from structured and confined knowledge bases (Berant et al., 2013) to unstructured and unbounded natural language form (e.g., documents on the web (Voorhees and Tice, 2000)) and unstructured, but restricted in size (e.g., a paragraph or multiple sentences (Hermann et al., 2015)). The recent advances in neural question answering lead to numerous datasets and successful models in these paradigms (Rajpurkar et al., 2016; Yang et al., 2015; Nguyen et al., 2016; Trischler et al., 2016). The answer types in these datasets are largely divided into three categories: sentence-level, in-context span, and generation. In this paper, we specifically focus on the former two and show that span-supervised models can better learn syntactic and lexical features. Among these datasets, we briefly describe three QA datasets to be used for the experiments in this paper. We also give the description of an RTE dataset for an example of a non-QA task. Refer to Table 1 to see the examples of the datasets.

SQuAD (Rajpurkar et al., 2016) is a recent span-based QA dataset, containing 100k/10k train/dev examples. Each example is a pair of context para-

graph from Wikipedia and a question created by a human, and the answer is a span in the context.

SQUAD-T is our modification of SQuAD dataset to allow for sentence selection QA. ('T' for sen*T*ence). We split the context paragraph into sentences and formulate the task as classifying whether each sentence contains the answer. This enables us to make a fair comparison between pretraining with span-supervised and sentence-supervised QA datasets.

WikiQA (Yang et al., 2015) is a sentence-level QA dataset, containing 1.9k/0.3k train/dev answerable examples. Each example consists of a real user's Bing query and a snippet of a Wikipedia article retrieved by Bing, containing 18.6 sentences on average. The task is to classify whether each sentence provides the answer to the query.

SemEval 2016 (Task 3A) (Nakov et al., 2016) is a sentence-level QA dataset, containing 1.8k/0.2k/0.3k train/dev/test examples. Each example consists of a community question by a user and 10 comments. The task is to classify whether each comment is relevant to the question.

SICK (Marelli et al., 2014) is a dataset for recognizing textual entailment (RTE), containing 4.5K/0.5K/5.0K train/dev/test examples. Each example consists of a hypothesis and a premise, and the goal is to determine if the premise is entailed by, contradicts, or is neutral to the hypothesis (hence classification problem). We also report results on SICK to show that span-supervised QA dataset can be also useful for non-QA datasets.

3 Model

Among numerous models proposed for span-level QA tasks (Xiong et al., 2017; Wang and Jiang, 2017b), we adopt an open-sourced model, BiDAF[2] (Seo et al., 2017).

---

[2] https://allenai.github.io/bi-att-flow

BiDAF. The inputs to the model are a question q, and a context paragraph x. Then the model selects the best answer span, which is $\arg\max_{(i,j)} \mathbf{y}_i^{\text{start}} \mathbf{y}_j^{\text{end}}$, where $i <= j$. Here, $\mathbf{y}_i^{\text{start}}$ and $\mathbf{y}_i^{\text{end}}$ are start and end position probabilities of i-th element, respectively.

Here, we briefly describe the answer module which is important for transfer learning to sentence-level QA. The input to the answer module is a sequence of vectors $\{\mathbf{h}_i\}$ each of which encodes enough information about the i-th context word and its relationship with its surrounding words and the question words. Then the role of the answer module is to map each vector $\mathbf{h}_i$ to its start and end position probabilities, $\mathbf{y}_i^{\text{start}}$ and $\mathbf{y}_i^{\text{end}}$.

BiDAF-T refers to the modified version of BiDAF to make it compatible with sentence-level QA. ('T' for sen*T*ence). In this task, the inputs are a question q and a list of sentences, $x_1, \ldots, x_T$, where T is the number of the sentences. Note that, unlike BiDAF, which outputs single answer per example, Here we need to output a C-way classification for each k-th sentence.

Since BiDAF is a span-selection model, it cannot be directly used for sentence-level classification. Hence we replace the original answer module of BiDAF with a different answer module, and keep the other modules identical to those of BiDAF. Given the input to the new answer module, $\{\mathbf{h}_1^k, \ldots, \mathbf{h}_N^k\}$, where the superscript is the sentence index ($1 \leq k \leq T$), we obtain the C–way classification scores for the k-th sentence, $\tilde{\mathbf{y}}^k \in [0,1]^C$ via max-pooling method:

$$\tilde{\mathbf{y}}^k = \text{softmax}(\mathbf{W} \max(\mathbf{h}_1^k, \ldots, \mathbf{h}_N^k) + \mathbf{b}) \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{C \times d}, \mathbf{b} \in \mathbb{R}^C$ are trainable weight matrix and bias, respectively, and $\max()$ function is applied elementwise.

For WikiQA and SemEval 2016, the number of classes (C) is 2, i.e. each sentence (or comment) is either relevant or not relevant. Since some of the metrics used for these datasets require full ranking, we use the predicted probability for "relevant" label to rank the sentences.

Note that BiDAF-T can be also used for the RTE dataset, where we can consider the hypothesis as a question and the premise as a context sentence ($T = 1$), and classify each example into 'entailment', 'neutral', or 'contradiction' ($C = 3$).

Transfer Learning. Transfer learning between the same model architectures[3] is straightforward: we first initialize the weights of the target model with the weights of the source model pretrained on the source dataset, and then we further train (fine-tune) on the target model with the target dataset. To transfer from BiDAF (on SQuAD) to BiDAF-T, we transfer all the weights of the identical modules, and initialize the new answer module in BiDAF-T with random values. For more training details, refer to Appendix A.

4 Experiments

Pretrained dataset	Fine-tuned	WikiQA			SemEval-2016		
		MAP	MRR	P@1	MAP	MRR	AvgR
-	-	62.96	64.47	49.38	76.40	82.20	86.51
SQuAD-T	No	75.22	76.40	62.96	47.23	49.31	60.01
SQuAD	No	75.19	76.31	62.55	57.80	66.10	71.13
SQuAD-T	Yes	76.44	77.85	64.61	76.30	82.51	86.64
SQuAD	Yes	79.90	82.01	70.37	78.37	85.58	87.68
SQuAD*	Yes	**83.20**	**84.58**	**75.31**	**80.20**	**86.44**	**89.14**
Rank 1		74.33	75,45	-	79.19	86.42	88.82
Rank 2		74.17	75.88	64.61	77.66	84.93	88.05
Rank 3		70.69	72.65	-	77.58	85.21	88.14

Table 2: Results on WikiQA and SemEval-2016 (Task 3A). The first row is a result from non-pretrained model, and * indicates ensemble method. Metrics used are Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Precision at rank 1 (P@1), and Average Recall (AvgR). Rank 1,2,3 indicate the results by previous works, ordered by MAP. For WikiQA, they are from Wang and Jiang (2017a); Tymoshenko et al. (2016); Miller et al. (2016), respectively. For SemEval-2016, they are from Filice et al. (2016); Joty et al. (2016); Mihaylov and Nakov (2016). SQuAD*&Yes sets the new state of the art on both datasets.

Question Answering Results. Table 2 reports the state-of-the-art results of our transfer learning on WikiQA and SemEval-2016 and the performance of previous models as well as several ablations that use no pretraining or no finetuning. There are multiple interesting observations from Table 2 as follows:

(a) If we only train the BiDAF-T model on the target datasets with no pretraining (first row of Table 2), the results are poor. This shows the importance of both pretraining and finetuning.

(b) Pretraining on SQuAD and SQuAD-T with no finetuning (second and third row) achieves results close to the state-of-the-art in the WikiQA dataset, but not in SemEval-2016. Interestingly, our result on SemEval-2016 is not better than only training without transfer learning. We conjecture that this is due to the significant difference between the domain of SemEval-2016 and

---

[3] Strictly speaking, this is a domain adaptation scenario.

that of SQuAD, which are from community and Wikipedia, respectively.

(c) Pretraining on SQuAD and SQuAD-T with finetuning (fourth and fifth row) significantly outperforms (by more than 5%) the highest-rank systems on WikiQA. It also outperforms the second ranking system in SemEval-2016 and is only 1% behind the first ranking system.

(d) Transfer learning models achieve better results with pretraining on span-level supervision (SQuAD) than coarser sentence-level supervision (SQuAD-T).[4]

Finally, we also use the ensemble of 12 different training runs on the same BiDAF architecture, which obtains the state of the art in both datasets. This system outperforms the highest-ranking system in WikiQA by more than 8% and the best system in SemEval-2016 by 1% in every metric. It is important to note that, while we definitely benefit from the scale of SQuAD for transfer learning to smaller WikiQA, given the gap between SQuAD-T and SQuAD ($> 3\%$), we see a clear sign that span-supervision plays a significant role well.

Varying the size of pretraining dataset. We vary the size of SQuAD dataset used during pretraining, and test on WikiQA with finetuning. Results are shown in Table 3. As expected, MAP on WikiQA drops as the size of SQuAD decreases. It is worth noting that pretraining on SQuAD-T (Table 2) yields 0.5 point lower MAP than pretraining on 50% of SQuAD. In other words, roughly speaking, span-level supervision data is worth more than twice the size of sentence-level supervision data for the purpose of pretraining. Also, even a small size of fine-grained supervision data helps; pretraining with 12.5% of SQuAD gives an advantage of more than 7 points than no pretraining.

Analysis. Figure 1 shows the latently-learned attention maps between the question and one of the context sentences from a WikiQA example in Table 1. The top map is pretrained on SQuAD-T (corresponding to SQuAD-T&Yes in Table 2) and the bottom map is pretrained on SQuAD (SQuAD&Yes). The more red the color, the higher

---

Percentage of used SQuAD dataset	MAP
100%	79.90
50%	76.94
25%	74.39
12.5%	70.76

Table 3: Results with varying sizes of SQuAD dataset used during pretraining. All of them are finetuned and tested on WikiQA.
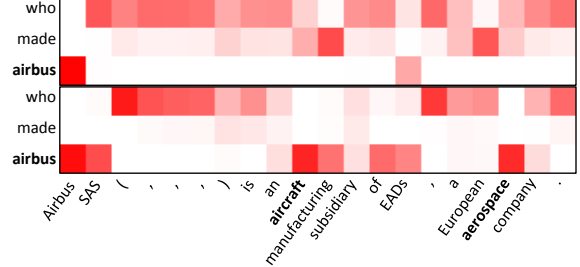

Figure 1: Attention maps showing correspondence between the words of a question (vertical) and the subset of its context (horizontal) in WikiQA for (top) SQuAD-T-pretrained model and (bottom) SQuAD-pretrained model. The more red, the higher the correspondence.

the relevance between the words. There are two interesting observations here.

First, in SQuAD-pretrained model (bottom), we see a high correspondence between question's `airbus` and context's `aircraft` and `aerospace`, but the SQuAD-T-pretrained model fails to learn such correspondence.

Second, we see that the attention map of the SQuAD-pretrained model is more sparse, indicating that it is able to more precisely localize correspondence between question and context words. In fact, we compare the sparsity of WikiQA test examples in SQuAD&Y and SQuAD-T&Y. Following Hurley and Rickard (2009), the sparsity of an attention map is defined by

$$\text{sparsity} = \frac{|\{x \in \mathbb{V} | x \leq \epsilon\}|}{|\mathbb{V}|} \quad (2)$$

where $\mathbb{V}$ is a set of values between 0 and 1 in attention map, and ϵ is a small value which we define 0.01 for here. A histogram of the sparsity is shown in Figure 2. There is a large gap in the average sparsity of WikiQA test examples between SQuAD&Yes and SQuAD-T&Yes, which are 0.84 and 0.56, respectively.

More analyses including error analysis and more visualizations are shown in Appendix B.

Entailment Results. In addition to QA experiments, we also show that the models trained on span-supervised QA can be useful for textual entailment task (RTE). Table 4 shows the trans-
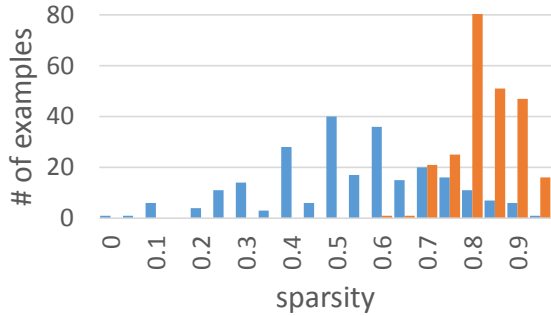
Figure 2: Histogram of the sparsity (Equation 2) of the attention maps of SQuAD-T-pretrained model (SQuAD-T&Yes, blue) and SQuAD-pretrained model (SQuAD&Yes, red). Mean sparsity of SQuAD-pretrained model (0.84) is clearly higher than that of SQuAD-T-pretrained model (0.56).

Pretrained dataset / Previous work	Accuracy
-	77.96
SQuAD-T	81.49
SQuAD	82.86
SQuAD*	84.38
SNLI	83.20
SQuAD-T + SNLI	85.00
SQuAD + SNLI	86.63
SQuAD + SNLI*	**88.22**
Yin et al. (2016)	86.2
Lai and Hockenmaier (2014)	84.57
Zhao et al. (2014)	83.64
Jimenez et al. (2014)	83.05
Mou et al. (2016)	70.9
Mou et al. (2016) (pretrained on SNLI)	77.6

Table 4: Results on SICK after finetuning. The first row is only trained on SICK. * indicates ensemble method.

fer learning results of BiDAF-T on SICK dataset (Marelli et al., 2014), with various pretraining routines. Note that SNLI (Bowman et al., 2015) is a similar task to SICK and is significantly larger (150K/10K/10K train/dev/test examples). Here we highlight three observations:

(a) BiDAF-T pretrained on SQuAD outperforms that without any pretraining by 6% and that pretrained on SQuAD-T by 2%, which demonstrates that the transfer learning from large span-based QA gives a clear improvement.

(b) Pretraining on SQuAD+SNLI outperforms pretraining on SNLI only. Given that SNLI is larger than SQuAD, the difference in their performance is a strong indicator that we are benefiting from not only the scale of SQuAD, but also the fine-grained supervision that it provides.

(c) We outperform the previous state of the art by 2% with the ensemble of SQuAD+SNLI pretraining routine.

It is worth noting that Mou et al. (2016) also shows improvement on SICK by pretraining on SNLI.

5   Conclusion

In this paper, we show state-of-the-art results on WikiQA and SemEval-2016 (Task 3A) as well as an entailment task, SICK, outperforming previous results by 8%, 1%, and 2%, respectively. We show that question answering with sentence-level supervision can greatly benefit from standard transfer learning of a question answering model trained on a large, span-level supervision. We additionally show that such transfer learning can be applicable in other NLP tasks such as textual entailment.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *EMNLP*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. *SemEval* 16:1116–1123.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Niall Hurley and Scott Rickard. 2009. Comparing measures of sparsity. *IEEE Transactions on Information Theory* 55(10):4723–4741.

Sergio Jimenez, George Duenas, Julia Baquero, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. 2014. Unal-nlp: Combining soft cardinality features for semantic textual similarity, relatedness and entailment. In *SemEval Workshop*.

Shafiq Joty, Alessandro Moschitti, Fahad A Al Obaidli, Salvatore Romeo, Kateryna Tymoshenko, and Antonio Uva. 2016. Convkn at semeval-2016 task 3: Answer and question selection for question answering on arabic and english fora. *SemEval* pages 896–903.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*.

Alice Lai and Julia Hockenmaier. 2014. Illinois-lh: A denotational and distributional approach to semantics. *SemEval* .

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *NAACL-HLT*.

Todor Mihaylov and Preslav Nakov. 2016. Semanticz at semeval-2016 task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. *SemEval* pages 879–886.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *EMNLP*.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications? In *EMNLP*.

Preslav Nakov, Llus Mrquez, Alessandro Moschitti, Walid Magdy Mubarak Hamdy Hamdy, abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. *SemEval* pages 525–545.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *NIPS Workshop*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830* .

Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. 2016. Convolutional neural networks vs. convolution kernels: Feature engineering for answer sentence reranking. In *NAACL-HLT*.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *ACM SIGIR*.

Shuohang Wang and Jing Jiang. 2017a. A compare-aggregate model for matching text sequences. In *ICLR*.

Shuohang Wang and Jing Jiang. 2017b. Machine comprehension using match-lstm and answer pointer. In *ICLR*.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814* .

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *ICLR*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *EMNLP*.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *TACL* .

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *ECCV*.

Jiang Zhao, Tian Tian Zhu, and Man Lan. 2014. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. *SemEval* pages 271–277.

A Training details

Parameters. For pretraining BiDAF on SQuAD, we follow the exact same procedure in Seo et al. (2017). For pretraining BiDAF-T on SQuAD-T, we use the same hyperparameters for all modules except the answer module, for which we use the hidden state size of 200. The learning rate is controlled by AdaDelta (Zeiler, 2012) with the initial learning rate of 0.5 and minibatch size of 50. We maintain the moving averages of all weights of the model with the exponential decay rate of 0.999 during training and use them at test. The loss function is the cross entropy between $\tilde{\mathbf{y}}^k$ and the one-hot vector of the correct classification.

Convergence. For all settings, we train models until performance on development set continue to decrease for 5k steps. Table 5 shows the median selected step on each setting.

Dataset	Pretrained	selected step
SQuAD	-	18k
SQuAD-T	-	50k
WikiQA	-	6k
WikiQA	SQuAD-T	6k
WikiQA	SQuAD	3k
SemEval-2016	-	9k
SemEval-2016	SQuAD-T	4k
SemEval-2016	SQuAD	3k
SICK	-	13k
SNLI	-	55k
SICK	SQuAD-T	9k
SNLI	SQuAD-T	31k
SICK	SQuAD	18k
SNLI	SQuAD	49k
SICK	SQuAD-T + SNLI	7k
SICK	SQuAD + SNLI	7k

Table 5: Median global step, which has the best performance on development set.

B More Analysis

Attention maps. We show some more examples of attention maps in Figure 3. (Top) We see high correspondence between same word from question and context such as `senator` and `john`, in SQuAD-pretrained model, but the SQuAD-T-pretrained model fails to learn such correspondence. (Bottom) We see high correspondence between `stems` from question and `stem` from context (left) as well as `plant` from question and `plants` from context (right), in SQuAD-pretrained model, but the SQuAD-T-pretrained model fails to learn such correspondence.

Error Analysis. Table 7 shows the comparison between answers by SQuAD-T-pretrained model
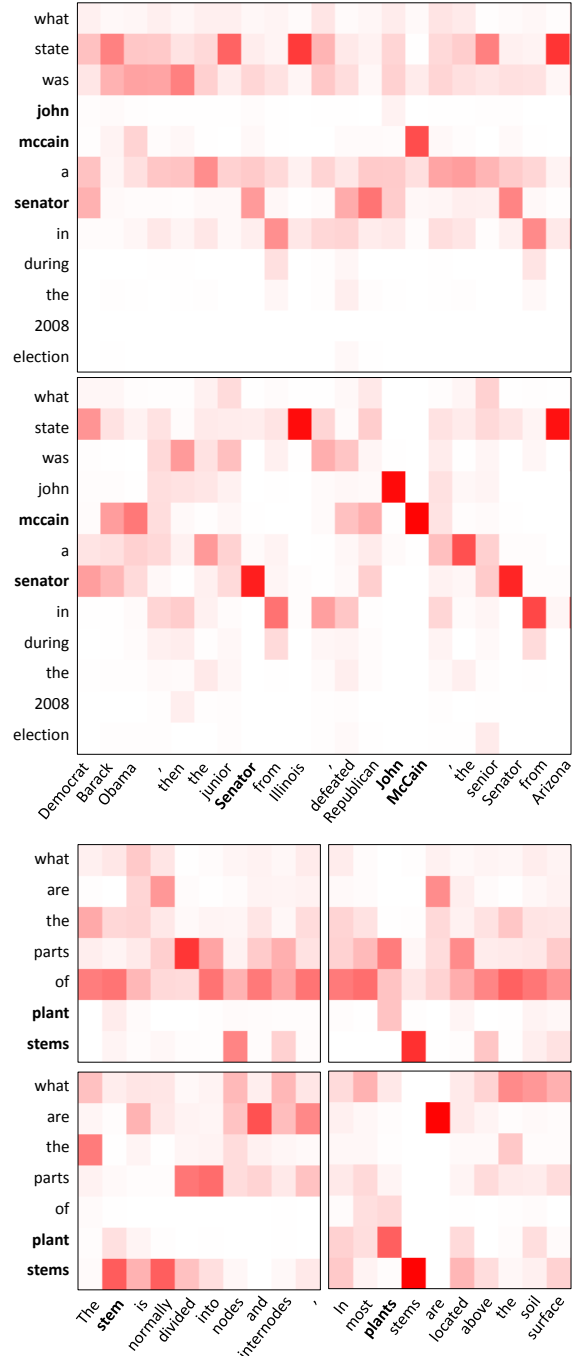

Figure 3: More attention maps showing correspondence between the words of a question (vertical) and one of candidate sentences (horizontal) in WikiQA for (top in each subfigure) SQuAD-MC-pretrained model and (bottom in each subfigure) SQuAD-pretrained model. The more red, the higher the correspondence.

and SQuAD-pretrained model on the example of WikiQA and SemEval-2016 from Table 1. On WikiQA, SQuAD-T-pretrained model selects C2 instead of the groundtruth answer C1. On SemEval-2016, SQuAD-pretrained model ranks C3 (bad comment) higher than C2 (good comment).

Category Id	Category	Example Question	Relevant Sentence
1	Exact Match	When did SpongeBob first air?	The pilot episode of SpongeBob SquarePants first aired in the United States on May 1, 1999, following the ...
2	Paraphrase	When was How the West Was Won filmed?	How the West Was Won is a 1962 American epic Western film.
3	No Clear Clue	When was Mary Anderson born?	Mary Anderson (1866-1953) was a real estate developer, ...
4	Need prior sentence (pronoun)	When did Texas become a state?	In 1845, it joined the United States as the 28th state.
5	Need prior sentence (context)	How do you play spades?	Its major difference as compared to other Whist variants is that, instead of trump being decided by the highest bidder or at random, the Spade suit is always trump, hence the name.
6	Hard to answer	How kimberlite pipes form?	Volcanic pipes are relatively rare.

Category Id	Category	Example Question Id	Example Question
1	Asking information	Q347_R25	hi all is there any IKEA showroom in and around DOHA? Kindly reply thank you
2	Asking opinion or recommendation in specific situation	Q326_R90	Salam I am mechanical Eng. 15 years experience i got a job for Rasgas co. direct hire I am married and i have 2 kids 5 and 3 years old.my life style is average. Is 8.000 QR enough as a basic salary? (...)
3	Asking feelings in specific situation	Q348_R67	oh i wish they will build Disneyland in Qatar :) how do you think guys? It will be perfect for Qatar :)
4	Asking abstract thing	Q341_R11	id like to get to know more about Al Jazeera International from anyone on QATAR LIVING who works at Al Jazeera.
5	Not Asking	Q337_R21	I just stumbled across this news article about the the American university campuses at Education City and thought some of you may also find it interesting.
6	Asking a lot of things at once	Q337_R16	How good are Karwa services? Are they : 1. Courteous/Rude? 2. Taking the correct route/Longer route? (...) 7. A pleasure/displeasure to ride?

Table 6: Examples from each category on (top) WikiQA and (bottom) SemEval-2016 (Task 3A).

	WikiQA	SemEval-2016
SQuAD-T&Yes	C2 > C1 > C3	C1 > C3 > C2
SQuAD&Yes	C1 > C2 > C3	C2 > C1 > C3
Groundtruth	C1(Y), C2(N), C3(N)	C1(Good), C2(Good), C3(Bad)

Table 7: Comparison of ranked answers by SQuAD-T-pretrained model (SQuAD-T&Yes) and SQuAD-pretrained model (SQuAD&Yes) of examples from WikiQA and SemEval-2016 (Task 3A) in Table 1.

Pretrained dataset		total	Category Id					
SQuAD-T-Y	SQuAD-Y		1	2	3	4	5	6
total		100	37	38	6	15	2	2
Correct	Correct	49	28	14	3	4	0	0
Wrong	Correct	26	8	14	1	3	0	0
Correct	Wrong	9	1	4	1	3	0	0
Wrong	Wrong	16	0	6	1	5	2	2

Table 8: Comparison of performance of SQuAD-T-pretrained model (SQuAD-T-Y) and SQuAD-pretrained model (SQuAD-Y) on WikiQA.

Pretrained dataset		total	Category Id					
No Pretrain	SQuAD-Y		1	2	3	4	5	6
total		100	29	38	7	12	9	5
Correct	Correct	30	12	11	2	5	0	0
Wrong	Correct	22	6	10	0	2	2	2
Correct	Wrong	5	0	1	2	1	0	1
Wrong	Wrong	43	11	16	3	4	7	2

Table 9: Comparison of performance of model without pretraining (No Pretrain) and SQuAD-pretrained model (SQuAD-Y) on SemEval-2016 (Task 3A).

In addition, we sampled 100 example randomly from WikiQA and SemEval-2016, and classified them into 6 categories(Table 6). In Table 8, we compare the performance on these WikiQA examples by SQuAD-T-pretrained model and SQuAD-pretrained model. It shows that span supervision clearly helps answering questions on Category 1

and 2, which are easier to answer, with answering correctly on most of the questions in Category 1. Similarly, we show the comparison of the performance on classified examples of the model without pretraining and SQuAD-pretrained model on SemEval-2016. It also shows that span supervision helps answering questions asking information or opinion/recommendation.

C More Results

SQuAD-T. To better understand SQuAD-T dataset, we show the performance BiDAF-T with different training routines. We get MAP 89.46 and accuracy 85.34% with SQuAD-trained BiDAF model, and MAP 90.18 and accuracy 84.69% with SQuAD-T-trained BiDAF-T model. There is no large gap between the two models, as each paragraph of SQuAD-T has 5 sentences on average, which makes the classification problem easier than WikiQA.

SNLI. Other larger RTE datasets such as SNLI also benefit from transfer learning, although the improvement is smaller. We confirm the improvement by showing that the result on SNLI when pretraining on SQuAD with BiDAF is 82.6%, which is slightly higher than that of the model pretrained on SQuAD-T (81.6%). This, however, did not outperform the state of the art (88.8%) by Wang et al. (2017). This is mostly because BiDAF (or BiDAF-T) is a QA model, which is not designed for RTE tasks.