Benefits of Intermediate Annotations in Reading Comprehension

Dheeru Dua
University of California,
Irvine, CA, USA
ddua@uci.edu

Sameer Singh
University of California,
Irvine, CA, USA
sameer@uci.edu

Matt Gardner
Allen Institute for Artificial
Intelligence, Irvine, CA, USA
mattg@allenai.org

Abstract

Complex, compositional reading comprehension datasets require performing latent sequential decisions that are learned via supervision from the final answer. A large combinatorial space of possible decision paths that result in the same answer, compounded by the lack of intermediate supervision to help choose the right path, makes the learning particularly hard for this task. In this work, we study the benefits of collecting intermediate reasoning supervision along with the answer during data collection. We find that these intermediate annotations can provide two-fold benefits. First, we observe that for any collection budget, spending a fraction of it on intermediate annotations results in improved model performance, for two complex compositional datasets: DROP and Quoref. Second, these annotations encourage the model to learn the correct latent reasoning steps, helping combat some of the biases introduced during the data collection process.

1 Introduction

Recently many reading comprehension datasets requiring complex and compositional reasoning over text have been introduced, including HotpotQA (Yang et al., 2018), DROP (Dua et al., 2019), Quoref (Dasigi et al., 2019), and ROPES (Lin et al., 2019). However, models trained on these datasets (Hu et al., 2019; Andor et al., 2019) only have the final answer as supervision, leaving the model guessing at the correct latent reasoning. Figure 1 shows an example from DROP, which requires first locating various operands (i.e. relevant spans) in the text and then performing filter and count operations over them to get the final answer "3". However, the correct answer can also be obtained by extracting the span "3" from the passage, or by adding or subtracting various numbers in the passage. The lack of intermediate hints makes learning challenging and can lead the model to rely on data biases, limiting its ability to perform complex reasoning.

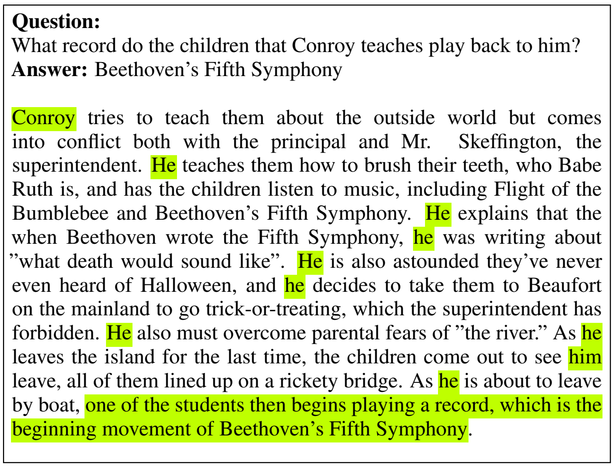
Question:
How many touchdown passes did Cutler throw in the second half?
Answer: 3

.....In the third quarter, the Vikes started to rally with running back Adrian Peterson's 1-yard touchdown run (with the extra point attempt blocked). The Bears increased their lead over the Vikings with Cutler's 3-yard TD pass to tight end Desmond Clark. The Vikings then closed out the quarter with quarterback Brett Favre firing a 6-yard TD pass to tight end Visanthe Shiancoe. An exciting .... with kicker Ryan Longwell's 41-yard field goal, along with Adrian Peterson's second 1-yard TD run. The Bears then responded with Cutler firing a 20-yard TD pass to wide receiver Earl Bennett. The Vikings then completed the remarkable comeback with Favre finding wide receiver Sidney Rice on a 6-yard TD pass on 4th-and-goal with 15 seconds left in regulation. The Bears then took a knee to force overtime.... The Bears then won on Jay Cutler's game-winning 39-yard TD pass to wide receiver Devin Aromashodu. With the loss, not only did the Vikings fall to 11-4, they also surrendered homefield advantage to the Saints.

Figure 1: Example from DROP, showing the intermediate annotations that we collected via crowd-sourcing.

In this paper, we present three main contributions. First, we show that annotating relevant context spans, given a question, can provide an easy and low-cost way to learn better latent reasoning. To be precise, we show that under low budget constraints, collecting these annotations for up to 10% of the training data (2-5% of the total budget) can improve the performance by 4-5% in F1. We supervise the current state-of-the-art models for DROP and Quoref, by jointly predict the relevant spans and the final answer. Even though these models were not designed with these annotations in mind, we show that they can still be successfully used to improve model performance. Models that explicitly incorporate these annotations might see greater benefits. Our results suggest that future dataset collection efforts should set aside a fraction of budget for intermediate annotations, particularly as the reasoning required becomes more complex.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5627–5634
July 5 - 10, 2020. ©2020 Association for Computational Linguistics



Question:
What record do the children that Conroy teaches play back to him?
Answer: Beethoven's Fifth Symphony

Conroy tries to teach them about the outside world but comes into conflict both with the principal and Mr. Skeffington, the superintendent. He teaches them how to brush their teeth, who Babe Ruth is, and has the children listen to music, including Flight of the Bumblebee and Beethoven's Fifth Symphony. He explains that the when Beethoven wrote the Fifth Symphony, he was writing about "what death would sound like". He is also astounded they've never even heard of Halloween, and he decides to take them to Beaufort on the mainland to go trick-or-treating, which the superintendent has forbidden. He also must overcome parental fears of "the river." As he leaves the island for the last time, the children come out to see him leave, all of them lined up on a rickety bridge. As he is about to leave by boat, one of the students then begins playing a record, which is the beginning movement of Beethoven's Fifth Symphony.

Figure 2: Example collected annotation from Quoref, showing the intermediate steps.

Second, these annotations can help combat biases that are often introduced while collecting data (Gururangan et al., 2018; Geva et al., 2019). This can take the form of label bias—in DROP, 18% of questions have answers 1, 2, or 3—or annotator bias, where a small group of crowd workers creates a large dataset with common patterns. By providing intermediate reasoning steps explicitly, the annotations we collect help the model overcome some of these biases in the training data.

Finally, the intermediate annotations collected in this work, including 8,500 annotations for DROP and 2,000 annotations for Quoref, will be useful for training further models on these tasks. We have made them available at https://github.com/dDua/Intermediate_Annotations.

2 Intermediate Annotations

Intermediate annotations describe the right set of context spans that should be aggregated to answer a question. We demonstrate their impact on two datasets: DROP and Quoref. DROP often requires aggregating information from various events in the context (Figure 1). It can be challenging to identify the right set of events directly from an answer when the same answer can be derived from many possible event combinations. We annotate the entire event span including all the attributes associated with the specific event. Quoref requires understanding long chains of coreferential reasoning, as shown in Figure 2, which are often hard to disentangle, especially when the context refers to multiple entities. We specifically annotate the coreference chains which lead to the entity being queried.

Collection process: We used Amazon Mechanical Turk to crowd-source the data collection. We randomly sample 8,500 and 2,000 QA pairs from the training set for DROP and Quoref respectively. We showed a QA pair and its context to the workers and asked them to highlight "essential spans" in the context. In case of DROP, crowd workers were asked to highlight complete events with all their corresponding arguments in each span. For Quoref, they were asked to highlight the coreference chains associated with the answer entity in the context.

Cost of gathering intermediate annotations: Each HIT, containing ten questions, paid $1, and took approximately five minutes to complete. Overall, we spent $850 to collect 8,500 annotations for DROP and $200 to collect 2,000 annotations for Quoref. If these annotations are collected simultaneously with dataset creation, it may be feasible to collect them at a lower cost, as the time taken to read the context again will be avoided.

3 Experiments and Results

In this section, we train multiple models for the DROP and Quoref datasets, and evaluate the benefits of intermediate annotations as compared to traditional QA pairs. In particular, we will focus on the cost vs benefit tradeoff of intermediate annotations, along with evaluating their ability to mitigate bias in the training data.

3.1 Setup

We study the impact of annotations on DROP on two models at the top of the leaderboard: NABERT[1] and MTMSN (Hu et al., 2019). Both the models employ a similar arithmetic block introduced in the baseline model (Dua et al., 2019) on top of contextual representations from BERT (Devlin et al., 2019). For Quoref, we use the baseline XLNet (Yang et al., 2019) model released with the dataset. We supervise these models with the annotations in a simple way, by jointly predicting intermediate annotation and the final answer. We add two auxiliary loss terms to the marginal log-likelihood loss function. The first is a cross-entropy loss between the gold annotations (g) and predicted annotations, which are obtained by passing the final BERT representations through a linear layer to get a score per token p, then normalizing each token's score of being selected as an annotation

[1] https://github.com/raylin1000/drop_bert

with a sigmoid function.

$$\mathcal{L}_1(\theta) = \alpha_1 CE(g, \sigma(p)) \tag{1}$$

The second is an L_1 loss on the sum of predicted annotations, encouraging the model to only select a subset of the passage.

$$\mathcal{L}_2(\theta) = \alpha_2 \sum_{\ell=0}^{|tokens|} \sigma(p_l)$$

The hyper-parameters α_1 and α_2 were used to balance the scale of both auxiliary loss terms with the marginal log-likelihood.

3.2 Cost vs Benefit

To evaluate the cost-benefit trade-off, we fix the total collection budget and then vary the percentage of budget that should go into collecting intermediate annotations. As shown in Figure **??**, the model achieves better performance (+1.7% F1) when spending \$7k where 2% budget is used for collecting intermediate reasoning annotations as compared to model performance when spending \$10k for collecting only QA pairs. Overall, from Figure 3 we can see that allocating even 1% of the budget to intermediate annotations provides performance gains. However, we observe that allocating a large percentage of the budget to intermediate annotations at the expense of QA pairs reduces performance. In our experiments, we find that the sweet-spot percentage of the budget and training-set that should be allocated to intermediate annotations is 2% and ∼10% respectively.

3.3 Bias Evaluation

Unanticipated biases (Min et al., 2019; Manjunatha et al., 2019) are often introduced during dataset collection due to many reasons (eg., domain-specific contexts, crowd-workers distributions, etc.). These "dataset artifacts" can be picked up by the model to achieve better performance without learning the right way to reason. We explore two examples of such dataset artifacts in DROP and Quoref.

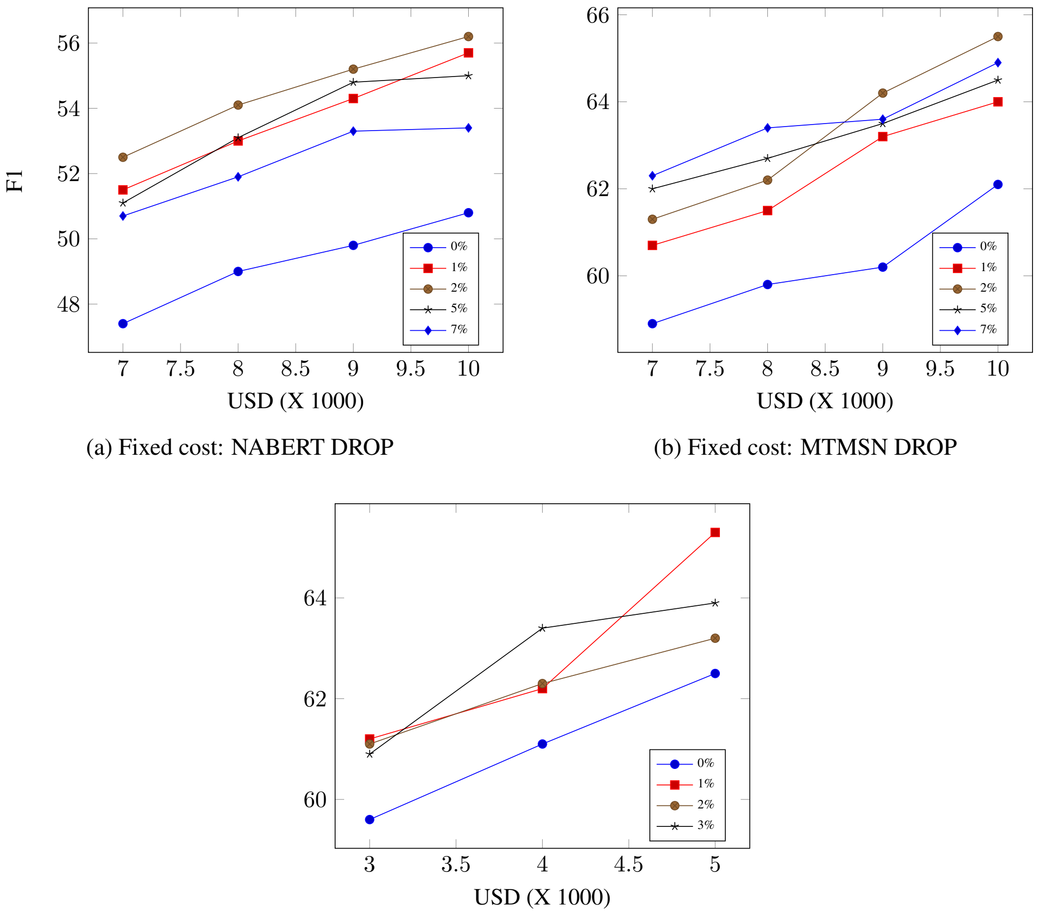
In DROP, around 40% of the passages are from NFL game summaries. The frequency of counting and arithmetic questions from this portion of the data resulted in the answers 1, 2, and 3 making up 18% of the entire training set. To study the effect of biased answer distribution on model performance, we sample 10k QA pairs with answers ∈ [0,9] from the training set *randomly* as a biased training set. We also sample QA pairs from the validation set *uniformly* for each answer ∈ [0,9] thus ensuring that each answer has equal representation in the unbiased validation set.

Dataset	Baseline		More QA pairs		Annotations	
	F1 (%)	Conf. loss	F1 (%)	Conf. loss	F1 (%)	Conf. loss
DROP	24.6	101.5	25.5	107.5	28.1	94.5
Quoref	61.8	103.0	62.7	109.0	64.3	97.0

Table 1: F1 performance and confusion loss (lower is better) of models in three settings: baseline with 10k(DROP) and 5k(Quoref) QA pairs, additional QA pairs worth \$250 and \$100 for DROP and Quoref respectively, and additional annotations worth \$250 and \$100 for DROP and Quoref respectively. To put confusion loss in perspective, the *best* confusion loss, i.e. perfect diffusion, is 90.1 for DROP and 87.0 for Quoref.

In Quoref, we found that around 65% of the answers are entity names present in the first sentence of the context. Similar to DROP, we create a biased training set with 5k QA pairs from the original training data, and an unbiased validation set with equal representation of answers from the first sentence and the rest of the context.

We investigate the effects of spending a small additional budget, either by adding more QA pairs (from the biased data distribution) or by collecting intermediate annotations, on this bias. We use two metrics to measure the extent to which bias has been mitigated. The first is the original metric for the task, i.e. F_1, that measures how accurate the model is on the unbiased evaluation. Further, we also want to evaluate the extent to which the errors made by the model are unbiased; in other words, how much is the error *diffused* over all possible answers, rather than only over the biased labels. We compute *confusion loss* (Machart and Ralaivola, 2012) as the metric for this, which measures error diffusion by computing the highest singular value of the unnormalized confusion matrix after setting the diagonal elements (i.e. true positives), to zero (Koço and Capponi, 2013) (lower confusion loss implies more diffusion). In an ideal scenario, all labels should have an equally likely probability of being a mis-prediction. Higher confusion loss implies that if we consider mis-classifications of a model we see that it has a tendency of over-predicting a specific label, making it biased towards that specific class.



(a) Fixed cost: NABERT DROP

(b) Fixed cost: MTMSN DROP

(c) Fixed cost: Quoref XLNet

Figure 3: Performance of model for varying percentage of budget invested in collecting intermediate annotation. The calculation were done with cost as \$0.4 and \$0.7 for a QA pair in DROP and Quoref respectively.

Table 1 shows that along with higher improvements in F_1 on providing annotations as compared to more QA pairs, we also see a reduction in the confusion loss with annotations indicating bias mitigation.

Further, we also find that for DROP, the false positive rate for top-3 common labels fell down from 47.7% (baseline) to 39.6% (with annotations), while the false positive rate for the bottom-7 increased from 30.4%(baseline) to 36.3%(with annotations), further demonstrating mitigation of bias. The confusion matrices are included in Appendix.

3.4 Qualitative Result

Figure 4 shows a DROP example where the model trained without annotations is not able to determine the right set of events being queried, returning an incorrect response. The model trained with annotations can understand the semantics behind the query terms “first half” and “Cowboys”, to arrive at the correct answer. The curves depicting quanti-

How many times did the Cowboys score in the first half?

Still searching for their first win, the Bengals flew to Texas Stadium for a Week 5 interconference duel with the Dallas Cowboys. In the first quarter, Cincinnati trailed early as Cowboys kicker Nick Folk got a 30-yard field goal, along with RB Felix Jones getting a 33-yard TD run. In the second quarter, Dallas increased its lead as QB Tony Romo completed a 4-yard TD pass to TE Jason Witten. The Bengals would end the half with kicker Shayne Graham getting a 41-yard and a 31-yard field goal. In the third quarter, Cincinnati tried to rally as QB Carson Palmer completed an 18-yard TD pass to WR T. J. Houshmandzadeh. In the fourth quarter, the Bengals got closer as Graham got a 40-yard field goal, yet the Cowboys answered with Romo completing a 57-yard TD pass to WR Terrell Owens. Cincinnati tried to come back as Palmer completed a 10-yard TD pass to Houshmandzadeh (with a failed 2-point conversion), but Dallas pulled away with Romo completing a 15-yard TD pass to WR Patrick Crayton.

Figure 4: Predicted relevant spans for question answered correctly with annotation (prediction: “3”) and incorrectly without annotations (prediction: “2”) by MTMSN model trained on DROP

tative performance gains with varying amounts of annotations and QA pairs are in the appendix.

4 Related Work

Similar to our work, Zaidan et al. (2007) studied the impact of providing explicit supervision via *rationales*, rather than generating them, for varying fractions of training set in text classification. However, we study the benefits of such supervision for complex compositional reading comprehension datasets. In the field of computer vision, Donahue and Grauman (2011) collected similar annotations, for visual recognition, where crowd-workers highlighted relevant regions in images.

Within reading comprehension, various works like HotpotQA (Yang et al., 2018) and CoQA (Reddy et al., 2019) have collected similar reasoning steps for entire dataset. Our work shows that collecting intermediate annotations for a fraction of dataset is cost-effective and helps alleviate dataset collection biases to a degree. Another line of work (Ning et al., 2019) explores the cost vs. benefit of collecting full vs. partial annotations for various structured predictions tasks. However, they do not focus on intermediate reasoning required to learn the task.

Our auxiliary training with intermediate annotations is inspired by extensive related work on training models using *side information* or *domain knowledge* beyond labels (Mann and McCallum, 2008; Chang et al., 2007; Ganchev et al., 2010; Rocktaschel et al., 2015). Especially relevant is work on supervising models using explanations (Ross et al., 2017), which, similar to our annotations, identify parts of the input that are important for prediction (Lei et al., 2016; Ribeiro et al., 2016).

5 Conclusion

We show that intermediate annotations are a cost-effective way to not only boost model performance but also alleviate certain unanticipated biases introduced during the dataset collection. However, it may be unnecessary to collect these for entire dataset and there is a sweet-spot that works best depending on the task. We proposed a simple semi-supervision technique to expose the model to these annotations. We believe that in future they can be used more directly to yield better performance gains. We have also released these annotations for the research community at `https://github.com/dDua/Intermediate_Annotations`.

Acknowledgements

This work was supported in part by Allen Institute of Artificial Intelligence, in part by Amazon, and in part by the National Science Foundation (NSF) grant #CNS-1730158.

References

Daniel Andor, Luheng He, Kenton Lee, and Emily Pitler. 2019. Giving bert a calculator: Finding operations and arguments with reading comprehension. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 280–287.

Pradeep Dasigi, Nelson F Liu, Ana Marasovic, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Jeff Donahue and Kristen Grauman. 2011. Annotator rationales for visual recognition. In *2011 International Conference on Computer Vision*, pages 1395–1402. IEEE.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL*.

Kuzman Ganchev, Joao Graca, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research (JMLR)*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sokol Koço and Cécile Capponi. 2013. On multi-class classification through the minimization of the confusion matrix norm. In *Asian Conference on Machine Learning*, pages 277–292.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *Annual Meeting of the Association for Computational Linguistics (ACL).*

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. *MRQA Workshop*.

Pierre Machart and Liva Ralaivola. 2012. Confusion matrix stability bounds for multiclass classification. *arXiv preprint arXiv:1202.6221*.

Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9562–9571.

Gideon S. Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 870–878.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. *Annual Meeting of the Association for Computational Linguistics (ACL).*

Qiang Ning, Hangfeng He, Chuchu Fan, and Dan Roth. 2019. Partial or complete, that's the question. *Annual Meeting of the Association for Computational Linguistics (ACL).*

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

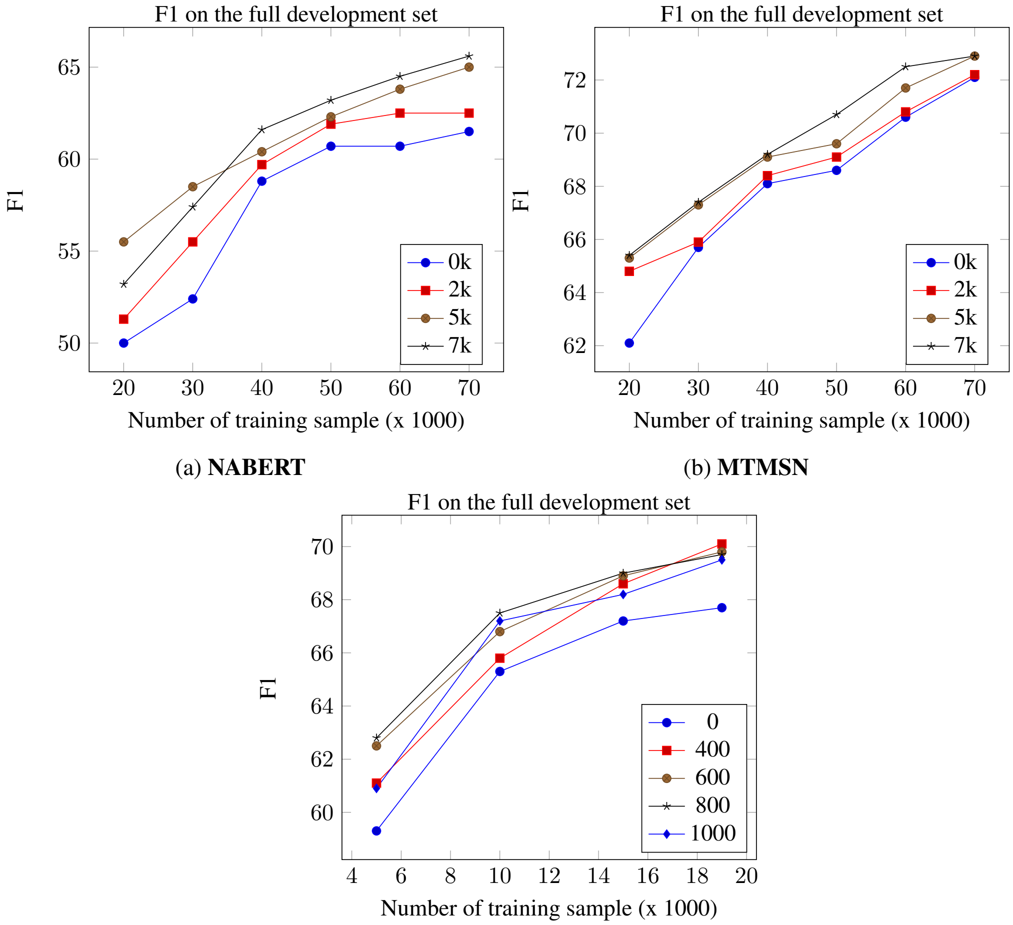
Tim Rocktaschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *IJCAI*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Annual Meeting of the Association for Computational Linguistics (ACL).*

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using annotator rationales to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

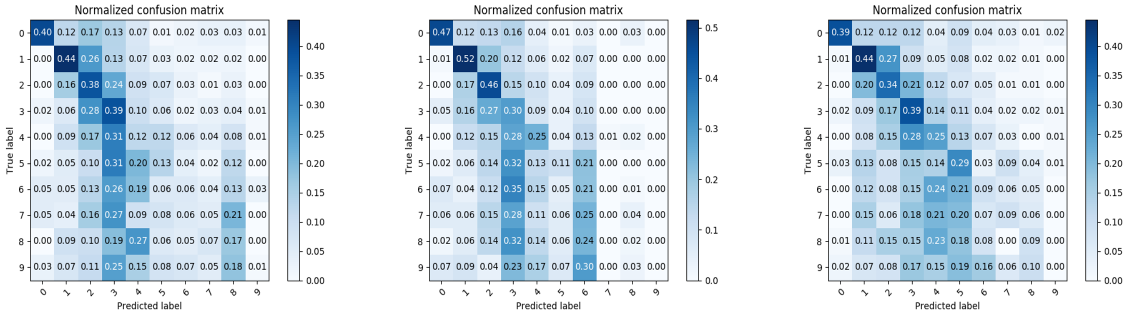


(a) **NABERT**

(b) **MTMSN**

(c) **Quoref-XLNet**

Figure 5: Performance of model trained on varying amount of annotations used in training

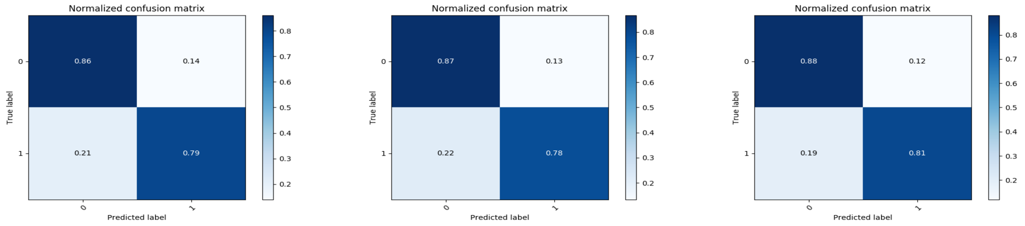


(a) 10k samples

(b) Additional QA pairs worth $250

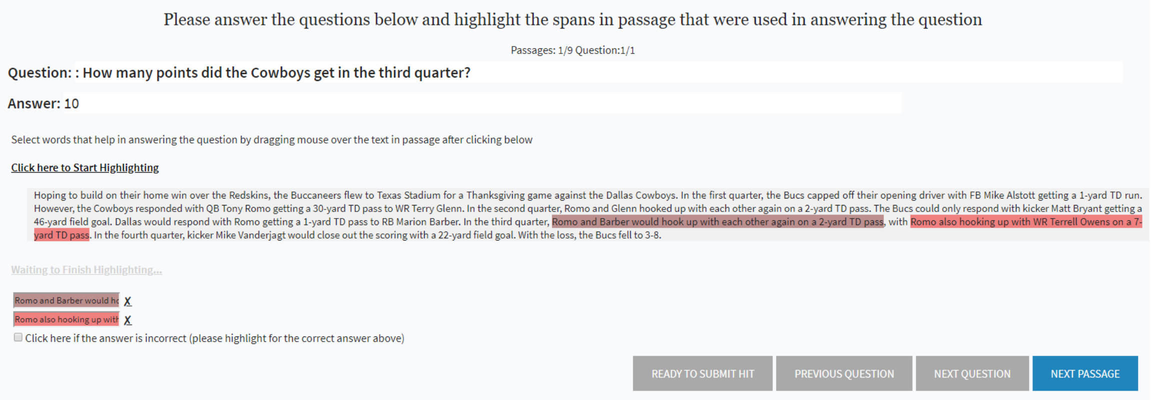
(c) Annotations worth $250

Figure 6: For the same cost intermediate annotations helps diffuse biased over-representation of number 3 as compared to adding more question-answer pairs



(a) 5k training samples (b) Additional QA pairs worth $100 (c) Annotations worth $100

Figure 7: For the same cost intermediate annotations helps diffuse biased over-representation of number 3 as compared to adding more question-answer pairs



Please answer the questions below and highlight the spans in passage that were used in answering the question

Passages: 1/9 Question:1/1

Question: : How many points did the Cowboys get in the third quarter?

Answer: 10

Select words that help in answering the question by dragging mouse over the text in passage after clicking below

Click here to Start Highlighting

Hoping to build on their home win over the Redskins, the Buccaneers flew to Texas Stadium for a Thanksgiving game against the Dallas Cowboys. In the first quarter, the Bucs capped off their opening driver with FB Mike Alstott getting a 1-yard TD run. However, the Cowboys responded with QB Tony Romo getting a 30-yard TD pass to WR Terry Glenn. In the second quarter, Romo and Glenn hooked up with each other again on a 2-yard TD pass. The Bucs could only respond with kicker Matt Bryant getting a 46-yard field goal. Dallas would respond with Romo getting a 1-yard TD pass to RB Marion Barber. In the third quarter, Romo and Barber would hook up with each other again on a 2-yard TD pass, with Romo also hooking up with WR Terrell Owens on a 7-yard TD pass. In the fourth quarter, kicker Mike Vanderjagt would close out the scoring with a 22-yard field goal. With the loss, the Bucs fell to 3-8.

Waiting to Finish Highlighting...

Romo and Barber would hc X
Romo also hooking up witl X

□ Click here if the answer is incorrect (please highlight for the correct answer above)

Figure 8: HIT interface used for collection annotations

Question: What is the full name of Mary Harriette's father?

Motteux was also without heirs and bequeathed Sandringham, together with another Norfolk estate and a property in Surrey, to the third son of his close friend, Emily Lamb, the wife of Lord Palmerston. At the time of his inheritance in 1843, Charles Spencer Cowper was a bachelor diplomat, resident in Paris. On succeeding to Motteux's estates, he sold the other properties and based himself at Sandringham. He undertook extensions to the hall, employing Samuel Sanders Teulon to add an elaborate porch and conservatory. Cowper's style of living was extravagant he and his wife spent much of their time on the Continent and within 10 years the estate was mortgaged for £89,000. The death of their only child, Mary Harriette, from cholera in 1854 led the couple to spend even more time abroad, mainly in Paris, and by the early 1860s Cowper was keen to sell the estate.

Figure 9: Predicted relevant spans for question answered correctly with annotation (prediction:"Charles Spencer Cowper") and incorrectly without annotations (prediction:"Lord Palmerston") by XLNet on Quoref