Tackling the generalization problem in Reading Comprehension NLP models

Aparna Dhurjati
aparna.dhurjati@utexas.edu

Abstract

In this work we analyze an NLP model's strength, robustness and its generalizability. Our goal is to improve accuracy of a reading comprehension model in two ways: (1) for in-domain data (2) for generalized or advsersarial data. We explore techniques that can provide an insight into a model's strengths and weaknesses. We discuss approaches that can fix the prediction errors and improve the model's robustness. For doing this analysis, we train ELECTRA-small model with the SQuAD(Rajpurkar et al., 2016) dataset for a reading comprehension system for Extractive Question and Answering task. We gather the baseline metrics from this model. We then probe the model with checklist(Ribeiro et al., 2020) sets and adversarial(Bartolo et al., 2020)(Jia and Liang, 2017) sets. We gather the model's mistakes from these evaluations and then experiment with various techniques to fix them. We then reevaluate the model in the same manner as the baseline version and observe the results. Finally, we analyze the results and make conclusions on what techniques can be applied to reading comprehension or similar NLP systems.

1 Introduction

Most current reading comprehension(RC) NLP models today achieve remarkably close to human performance when tested with in-domain data sets. However when faced with adversarial settings, their performance is sub optimal. This naturally raises the question whether these models have actually learned the task of solving reading comprehension. This inferior performance is attributed to the models, exploiting spurious correlations from the examples they are trained on. That is, they tend to capture dataset-specific patterns that hold for the in-domain examples which do not hold in generalized settings. The study of this behavior and its mitigation has seen huge surge of interest from many NLP researchers today. Many researches have shown a model's vulnerabilities to adversarial attacks(Jia and Liang, 2017)(Bartolo et al., 2020) and how it lacks robustness(Ribeiro et al., 2020).

In this work, we construct an RC system using ELECTRA-small by training on SQuAD v1.1(Rajpurkar et al., 2016) dataset and probe the model with the out-of-domain datasets and analyze its behavior. Due to its simplicity and effectiveness, we focus our work on "Extractive QA" or EQA task where the model finds the start and end span of the answer from the context and gives back the prediction. Examples of this task are SQuAD v1.1 and SQuAD 2((Rajpurkar et al., 2018) and NewsQA(Trischler et al., 2017).

We propose an approach where we can improve the model on generalized or adversarial settings while still being strong on the in-domain datasets.

1.1 Proposal

Our proposed approach to improve the model consists of two-phased training as shown in **Figure 1**

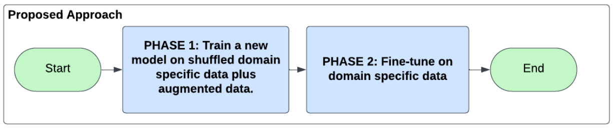
Proposed Approach
Start
PHASE 1: Train a new model on shuffled domain specific data plus augmented data.
PHASE 2: Fine-tune on domain specific data
End

Figure 1: Two-phased training proposal

With our proposed approach, we were able to improve our baseline model by 0.25

Error Category	**Example Failure**	**Details**
n-gram spurious correlation	**C:**: ... The Broncos' seven sacks tied a Super Bowl record set by the Chicago Bears in Super Bowl XX. Kony Ealy tied a Super Bowl record with three sacks. Jordan Norwood's 61-yard punt return set a new record, surpassing the old record of 45 yards set by John Taylor in Super Bowl XXIII. Denver was just 1-of-14 on third down, while Carolina was barely better at 3-of-15. The two teams' combined third down conversion percentage of 13.8 was a Super Bowl low. Manning and Newton had quarterback passer ratings of 56.6 and 55.4, respectively, and their added total of 112 is a record lowest aggregate passer rating for a Super Bowl. Manning became the oldest quarterback ever to win a Super Bowl at age 39, and the first quarterback ever to win a Super Bowl with two different teams, while Gary Kubiak became the first head coach to win a Super Bowl with the same franchise he went to the Super Bowl with as a player. **Q:** Who did the Broncos tie with the most sacks in a Super Bowl? **A:** Chicago Bears **P:** Kony Ealy	When the highlighted word was changed to "Bowl", then the model was able to predict correctly as Chicago Bears.
Mistaken stop words	**C:**: The two most prominent Norman families to arrive in the Mediterranean were descendants of Tancred of Hauteville and the Drengot family, of whom Rainulf Drengot received the county of Aversa, the first Norman toehold in the south, from Duke Sergius IV of Naples in 1030. The Hauteville family achieved princely rank by proclaiming prince Guaimar IV of Salerno D̈uke of Apulia and Calabria.̈ He promptly awarded their elected leader, William **Iron Arm**, with the title of count in his capital of Melfi. The Drengot family thereafter attained the principality of Capua, and emperor Henry III legally ennobled the Hauteville leader, Drogo, as d̈ux et magister Italiae comesque Normannorum totius Apuliae et Calabriae(D̈uke and Master of Italy and Count of the Normans of all Apulia and Calabria)̈ in 1047. **Q:** Who was Count of Melfi? **A:** William Iron Arm **P:** William	Model treated "Iron Arm" as stop words. When the space (e.g Iron-Man) was removed, the model could predict that word in addition to William as the answer.
Punctuation sensitivity	**C:**: Fresno has three large public parks, two in the city limits and one in county land to the southwest. Woodward Park, which features the Shinzen Japanese Gardens, numerous picnic areas and several miles of trails, is in North Fresno and is adjacent to the San Joaquin River Parkway. Roeding Park, near Downtown Fresno, is home to the Fresno Chaffee Zoo, and Rotary Storyland and Playland. Kearney Park is the largest of the Fresno **region's** park system and is home to historic Kearney Mansion and plays host to the annual Civil War Revisited, the largest reenactment of the Civil War in the west coast of the U.S. **Q:** Which park is home to the Kearney Mansion? **A:** Kearney Park **P:** Roeding Park	Model predicted correctly when the apostrophe was removed.
Attention to words occurring after the question-words	**C:**: Luther and his wife moved into a former monastery, "The Black Cloister", **a wedding present** from the new elector John the Steadfast (1525). They embarked on what appeared to have been a happy and successful marriage, though money was often short... **Q:** What was the Black Cloister? **A:** former monastery **P:** a wedding present	A noun after the words "The Black Cloister" was picked up rather than looking before the words.
Lexical overlap	**C:**According to Tesla, Edison remarked, T̈here's fifty thousand dollars in it for you2̆014if you can do it.̈542̆01357 :64 This has been noted as an odd statement from an Edison whose company was stingy with pay and who did not have that sort of cash on hand. After months of work, Tesla fulfilled the task and inquired about payment. Edison, saying that he was only joking, replied, T̈esla, you don't understand our American humor.̈64 Instead, Edison **offered** a US$10 a week raise over Tesla's US$18 per week salary; Tesla refused the offer and immediately resigned **Q:** How much did Tesla say Edison offered him to redesign his motor and generators? **A:** fifty thousand dollars **P:** US$10 a week	The model matched the word "offered". When this was changed to "gave", the model predicted the correct answer.

Table 1: **Before**: Sample error categories and examples from SQuAD v1.1 and Adversarial QA validation datasets with the baseline model.

F1 points for the in-domain dataset and 7 F1 points for the out-of-domain Adversarial QA(Bartolo et al., 2020) dataset. We describe our baseline model and the implementation details of our proposal in the following sections.

2 Implementation

2.1 Baseline model

We first trained a model as our baseline model. We trained ELECTRA-small on SQuAD v1.1(Rajpurkar et al., 2016) training data set for 3 epochs. The training time was approximately 2 hours on Google Colab GPU. This baseline model achieved an EM score of 78.46 and F1 score 86.35 on the SQuAD v1.1 validation set.

2.2 Analysis on the baseline model

We evaluated the performance of the baseline model on three datasets: (1) SQuAD v1.1 (2) Adversarial QA(Bartolo et al., 2020) whose dataset is constructed with model-in-the-loop methodology (3) checklists(Ribeiro et al., 2020), which contains test suite for behavior testing("black-box" testing, as per the authors). **Table 4** presents the scores from SQuAD 1.1 and Adversarial QA data. We conducted an error analysis on a random sample of 20 erroneous predictions from our model

Error Category	Example	FIXED/NOTFIXED/NEW
n-gram spurious correlation	**C:**: ... The Broncos' seven sacks tied a Super Bowl record set by the Chicago Bears in Super Bowl XX. Kony Ealy tied a Super Bowl record with three sacks. Jordan Norwood's 61-yard punt return set a new record, surpassing the old record of 45 yards set by John Taylor in Super Bowl XXIII... **Q:** Who did the Broncos tie with the most sacks in a Super Bowl? **A:** Chicago Bears **P:** Chicago Bears	FIXED
Punctuation sensitivity	**C:**: ...Roeding Park, near Downtown Fresno, is home to the Fresno Chaffee Zoo, and Rotary Storyland and Playland. Kearney Park is the largest of the Fresno region's park system and is home to historic Kearney Mansion and plays host... **Q:** Which park is home to the Kearney Mansion? **A:** Kearney Park **P:** Kearney Park	FIXED
Lexical overlap	**C:**According to Tesla, Edison remarked, T̈here's fifty thousand dollars in it for you2̆014if you can do it.̈:542̆01357 :64 This has been noted as an odd statement from an Edison whose company was stingy with pay and who did not have that sort of cash on hand. After months of work,... **Q:** How much did Tesla say Edison offered him to redesign his motor and generators? **A:** fifty thousand dollars **P:** fifty thousand dollars	FIXED
Mistaken stop words	**C:**: ...D̈uke of Apulia and Calabria.̈ He promptly awarded their elected leader, William **Iron Arm**, with the title of count in his capital of Melfi. The Drengot family thereafter attained the principality of Capua, and emperor Henry III legally ... **Q:** Who was Count of Melfi? **A:** William Iron Arm **P:** William	NOT FIXED
Attention to words occurring after the question-words	**C:** Luther and his wife moved into a former monastery, "The Black Cloister", **a wedding present** from the new elector John the Steadfast (1525). They embarked on what appeared to... **Q:** What was the Black Cloister? **A:** former monastery **P:** a wedding present	NOT FIXED
Attention to words occurring after the question-words (picked up "NFC" as "NFL")	**C:** Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 ... **Q:**Which NFL team won Super Bowl 50? **A:** Denver Broncos **P:** Carolina Panthers	NEW!
Correlation "When" and "Time"	**C:** ...The showś premise is that this is a life process of **Time** Lords through which the character of the Doctor takes on a new body and, to some extent, new personality, which occurs after sustaining an injury which would be fatal to most other species... **Q:****When** does Doctor Who transition to a new body? **A:** after sustaining an injury **P:** Time Lords	NEW!

Table 2: **After**: Fixed and new errors from SQuAD v1.1 evaluation with the final model(model v4).

and analyzed their potential sources of errors and categorized them. The most common errors were on questions which required complex type of reasoning on domain knowledge. The most common mistake we observed is that the model simply picked up words from the context that had high lexical overlap with the questions. See **Table 1** for examples of some of the common phenomena. See **Table 3** for the results from running the checklists. As highlighted in the table, it exhibited strengths only in these categories: **Synonyms**, **Question Typos** and **Random perturbations to the context**. We also noticed that the failure rate was 100% in almost all other categories, which implied this RC model was not suited for any real world tasks. While our baseline model performed reasonably well on SQuAD v1.1. dataset, it is evident that it has not generalized well to out-of-domain datasets and that it is not robust enough. Our goal therefore became to improve our model in two ways: (1) improve the accuracy of the model for SQuAD domain (2) generalize the model for adversarial settings.

Dataset	F1 score	EM score
SQuAD v1.1	86.35	78.46
Adversarial QA	28.39	18.73

Table 4: Baseline model scores.

3 Experiments

For our experiments we used Adversarial QA(Bartolo et al., 2020). This is a pre-generated dataset and is created with model-in-the-loop methodology. As the authors claim, models trained on data collected with a model-in-the-loop generalize well to non-adversarially collected data as well. Since this idea matches what we intend to do, we chose this dataset for our experiments. This dataset contained 30K training examples.

Our initial experiment was to fine-tune the baseline model on the Adversarial QA dataset. We fine-tuned the baseline model with 30K Adversarial training examples for 3 epochs. The training loss was 1.806. This fine-tuned

Test type	**Failure %**	**Failure Example**
MFT: Comparisons	99.2%	**C:** Joe is younger than Marie. **Q:**Who is less young?**A:** Marie **P:** Joe
MFT: Properties to categories	100%	**C:** There is a figure in the room. The figure is round and old. **Q:** What shape is the figure? **A:** round **P:** round and old
MFT: Intensifiers	100%	**C:** Ashley is somewhat hopeful about the project. Jennifer is hopeful about the project. **Q:** Who is most hopeful about the project? **A:** Jennifer **P:** Ashley
MFT: Profession vs nationality	66.8%	**C:** William is a Chinese organizer. **Q:** What is William's job? **A:** organizer **P:** Chinese organizer
MFT: Animal vs Vehicle	25.8%	**C:** Karen has an iguana and a tractor. **Q:** What vehicle does Karen have? **A:** tractor **P:** iguana and a tractor
MFT: Synonyms	2.8%	**C:** Carol is very vocal. Jeff is very angry. **Q:** Who is outspoken? **A:** Carol **P:** Jeff
MFT: Comparison to antonym	100%	**C:** Gary is poorer than Annie. **Q:** Who is richer? **A:** Annie **P:** Gary
MFT: Change in profession	0%	
INV: Negation in context	93.8%	**C:** Jack is not an actress. Jean is. **Q:** Who is an actress? **A:** Jean **P:** Jack
INV: Negation in Q only	100%	**C:** Linda is a nurse. Robert is a producer. **Q:** Who is not a nurse? **A:** Robert **P:** Linda
MFT: Basic coref, he / she	100%	**C:** Steven and Jill are friends. He is an investor, and she is a historian. **Q:** Who is an investor? **A:** Steven **P:** Steven and Jill
MFT: Basic coref, his / her	100%	**C:** Ben and Emily are friends. Her mom is an analyst. **Q:** Whose mom is an analyst? **A:** Emily **P:** Emily and Ben
MFT: former/latter	100%	**C:** Ray and Larry are friends. The former is an advisor. **Q:** Who is an advisor? **A:** Ray **P:** Ray and Larry
MFT: subj/obj distinction	95.8%	**C:** Angela understands Diana. **Q:** Who understands? **A:** Angela **P:** Diana
MFT: subj/obj distinction with 3 agents	100%	**C:** Lawrence loves George. George loves Albert. **Q:** Who loves George? **A:** Lawrence **P:** Albert
INV: Question typo	17.1%	**C:** On May 21, 2013, NFL owners at their spring meetings in Boston voted and awarded the game to Levi's Stadium. The $1.2 billion stadium opened in 2014. It is the first Super Bowl held in the San Francisco Bay Area since Super Bowl XIX in 1985, and the first in California since Super Bowl XXXVII took place in San Diego in 2003. **Orig Q:** When was San Francisco voted to be the location for Super Bowl 50? **Orig P:** May 21, 2013 **Q:** When was San Francisco voted to be **thel ocation** for Super Bowl 50? **P:** 1985
INV: Add random sentence to context	8.6%	**C:** On May 21, 2013, NFL owners at their spring meetings in Boston voted and awarded the game to Levi's Stadium. The $1.2 billion stadium opened in 2014. It is the first Super Bowl held in the San Francisco Bay Area since Super Bowl XIX in 1985, and the first in California since Super Bowl XXXVII took place in San Diego in 2003. **Q:** When was San Francisco voted to be the location for Super Bowl 50? **P:** May 21, 2013 **C: The league announced on October 16, 2012, that the.** On May 21, 2013, NFL owners at their spring meetings in Boston voted and awarded the game to Levi's Stadium. The $1.2 billion stadium opened in 2014. It is the first Super Bowl held in the San Francisco Bay Area since Super Bowl XIX in 1985, and the first in California since Super Bowl XXXVII took place in San Diego in 2003. **Q:** When was San Francisco voted to be the location for Super Bowl 50? **P:** 2003

Table 3: **Before:** Checklist results from the **baseline** model. C=Context Q=Question A=Answer P=Prediction MFT= Mininum Functional Test. INV=Invariance to perturbations. Low failure rates where highlighted in yellow.

model performed well on the Adversarial QA dataset. However it had brought down the scores for the in-domain dataset. We called this version of the model generated from this experiment as v1.

Our second experiment was to train on Adversarial QA first and then train on in-domain. We trained a new ELECTRA-small model on 30K Adversarial QA training examples for 3 epochs. The training loss was 2.69. We then fine-tuned this model on the SQuAD training dataset for another 3 epochs. The training loss was 1.023. This model evaluation resulted in mixed set of results just as in experiment 1.

Test type	Failure %	Failure Example
MFT: Comparisons	97.2%	**C:** Julia is wiser than Dick. **Q:**Who is less wise?**A:** Dick **P:** Julia
MFT: Properties to categories	95.2%	**C:** There is a thing in the room. The thing is big and pink. **Q:** What size is the thing?**A:** big **P:** big and pink
MFT: Intensifiers	98.4%	**C:** Johnny is pleased about the project. Martha is incredibly pleased about the project. **Q:** Who is most pleased about the project? **A:** Martha **P:** Johnny
MFT: Profession vs nationality	48.4%	**C:** Melissa is a Nigerian administrator. **Q:** hat is Melissa's job? **A:** administrator **P:** Nigerian administrator
MFT: Animal vs Vehicle	94.8%	**C:** Albert has a rabbit and a truck. **Q:** What vehicle does Albert have? **A:** truck **P:** a rabbit and a truck
MFT: Synonyms	5.6%	**C:** Bob is very modest. Kathleen is very outspoken. **Q:** Who is humble? **A:** Bob **P:** Kathleen
MFT: Comparison to antonym	100%	**C:** Peter is smaller than Greg. **Q:** Who is bigger? **A:** Greg **P:** Peter
MFT: Change in profession	0%	
INV: Negation in context	59.4%	**C:** JAndrew is an author. Jeff is not. **Q:** Who is not an author? **A:** Jeff **P:** Andrew
INV: Negation in Q only	100%	**C:** Gary is a nurse. Kathleen is a DJ. **Q:** Who is not a nurse? **A:** Kathleen **P:** Gary
MFT: Basic coref, he / she	100%	**C:** Ron and Amy are friends. He is a writer, and she is an investigator. **Q:** Who is a writer? **A:** Ron **P:** Ron and Amy
MFT: Basic coref, his / her	99%	**C:** Jimmy and Eleanor are friends. Her mom is an auditor. **Q:** Whose mom is an auditor? **A:** Eleanor **P:** Jimmy and Eleanor
MFT: former/latter	100%	**C:** Fiona and Louis are friends. The former is an executive. **Q:** Who is an executive? **A:** Fiona **P:** Fiona and Louis
MFT: subj/obj distinction	71.8%	**C:** Mary remembers Claire. **Q:** Who remembers? **A:** Mary **P:** Claire
MFT: subj/obj distinction with 3 agents	88.1%	**C:** Frances is bothered by Kim. Frances bothers Chris. **Q:** Who is bothered by Frances? **A:** Chris **P:** Kim
INV: Question typo	11.4%	**C:** Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title.... **Orig Q:** Which NFL team represented the AFC at Super Bowl 50? **Orig P:** Denver Broncos **Q:** Which NFL team represented **theA FC** at Super Bowl 50? **P:** Carolina Panthers
INV: Add random sentence to context	8.6%	**Orig C:** On May 21, 2013, NFL owners at their spring meetings in Boston voted and awarded the game to Levi's Stadium. The $1.2 billion stadium opened in 2014. It is the first Super Bowl held in the San Francisco Bay Area since Super Bowl XIX in 1985, and the first in California since Super Bowl XXXVII took place in San Diego in 2003. **Q:** When was San Francisco voted to be the location for Super Bowl 50? **P:** May 21, 2013 **C: The league announced on October 16, 2012, that the.** On May 21, 2013, NFL owners at their spring meetings in Boston voted and awarded the game to Levi's Stadium. The $1.2 billion stadium opened in 2014. It is the first Super Bowl held in the San Francisco Bay Area since Super Bowl XIX in 1985, and the first in California since Super Bowl XXXVII took place in San Diego in 2003. **Q:** When was San Francisco voted to be the location for Super Bowl 50? **P:** 2003

Table 5: **After:** Checklist results from the **final** model v4. Improvements highlighted in green and the dips are highlighted in red. C=Context Q=Question A=Answer P=Prediction MFT= Mininum Functional Test. INV=Invariance to perturbations

This is our model v2.

Our third experiment was to train on shuffled data from both adversarial and in-domain. We used SQuAD 2.0 and Adversarial QA data concatenated together into a file. We then shuffled this data manually and used this file as the training input. SQuAD 2.0 consists of questions that do not have an answer. Since our model training needs an answer, we manually updated the training dataset to use answer span as -1 and answer text as '`@@@`' for those that did not have an answer. The total number of training examples resulting from this concatenated set was 160K. We used this file to train a new ELECTRA-small model. After training on this dataset for 3 epochs, the training loss was 1.462. This model when evaluated gave promising improvements in the scores on both SQuAD v1.1 and Adversarial QA datasets. This is our model v3.

Our fourth and final experiment is our proposal. In this experiment, we fine-tuned the model we obtained in experiment 3, that is v3. We fine-tuned this model with in-domain SQuAD v1.1 data for 1.5 epochs. The training loss was 0.804. This brought in the best results thus far. This is our model v4, which is the final version.

We discuss the results from these experiments in the following section.

4 Results

We evaluated all four versions of the model on SQuAD v1.1 and Adversarial QA. Below **Table 6** shows their evaluation scores. Higher scores are highlighted.

Model version	Dataset	EM score	F1 score
Baseline	SQuAD	78.46	86.35
Baseline	Adv QA	18.73	28.39
v1	SQuAD	78.21	69.38
v1	Adv QA	48.60	61.74
v2	SQuAD	55.78	67.20
v2	Adv QA	23.6	33.81
v3	SQuAD	78.69	86.47
v3	Adv QA	28.66	38.67
v4	**SQuAD**	**78.86**	**86.57**
v4	**Adv QA**	**25.16**	**35.66**

Table 6: Model scores from the four experiments. Scores highlighted are the best. Scores in bold are best for both the datasets

From the experiments, we observed that, if the model has generalized well on adversarial data, it did not do as well as the baseline model on the in-domain data. However the approach taken in our fourth experiment (which is our proposed approach) is where the model demonstrated reasonably higher scores for both, as shown in bold in the table for model v4. Compared to the baseline model, this version of the model achieved a 0.25 point increase in the F1 score on the SQuAD v1.1 dataset and a 7 point increase on the Adversarial QA dataset, thus proving the effectiveness of our two-phased training approach.

To reiterate, the two-phase approach is the approach where we train a model first on mixed and shuffled examples collected from both adversarial and in-domain, followed by fine-tuning on in-domain only examples for relatively short number of epochs. This can bring in the best results. **Figure 2** shows the results in perspective.

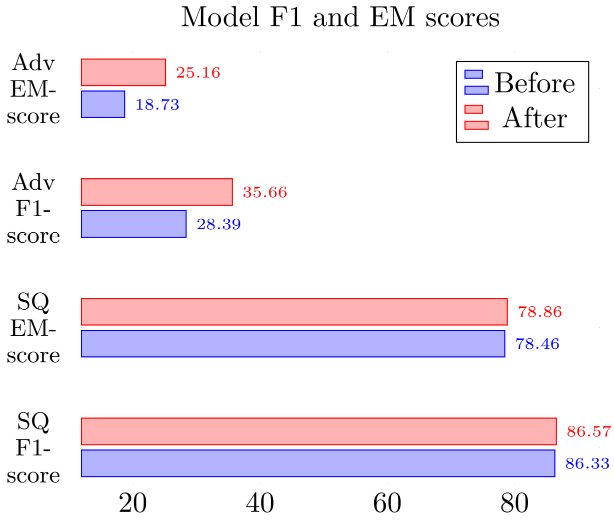

Figure 2: Results from two-phased training

We sampled some of the predictions from this final model. This version of the model has learned to improve on error categories related to **lexical overlap** and **punctuation sensitivity**. **Table 2** lists some of the error categories that the model has improved on. The improvements are as expected because model has now the exposure to adversarial and difficult training examples that do not just require superficial correspondences between words. We also note that there is a side effect to this improvement. It failed on some of the examples where it originally did well. These are captured in the same table as "NEW". We think this behavior is due to model learning to ignore certain patterns due to the mix of adversarial training data. Finally, we ran the checklist suite of tests on this retrained model. The checklist results are available in **Table 5**. We observed that the model has certainly improved in few of the categories as highlighted in that table.

5 Analysis and discussion

Understanding an NLP model's behavior requires a deeper understanding on the data used to train it. There are many data construction factors that directly dictate a model's strength, generalizability and its robustness. These include the methods used to collect the data, quality of the data, whether the data is synthetic or real, the amount of data collected and the annotation artifacts. Among these many factors, annotating the data is crucial, challenging and time-consuming. It can also be error-prone if right techniques are not employed. Some of the approaches currently used today include expert annotation, crowd-sourcing by non-experts and model-in-the-loop. SQuAD v1.1 dataset(Rajpurkar et al., 2016) uses crowd-sourcing by non-experts. Hence we think that this dataset, though diverse, is an easy dataset for an NLP model to learn, and therefore leading to low generalizability. The Adversarial QA that we chose for our experiments closes this gap to some extent by ensuring that the annotations are provided by a model in the loop.(Bartolo et al., 2020). In this methodology, a contemporary model is used either as a filter or directly during annotation, to identify samples wrongly predicted by the model. These examples are then used on the targeted model. The Adversarial QA data that we used for our experiments was constructed such examples with three models in the loop: BiDAF, BERT and RoBERTa.

As stated, we used ELECTRA-small as our base model. This model is similar to BERT which is good at learning surface level correspondences using self-attention. Thus, our baseline model learned to look for an answer in the context that contained similar structure as the question. The SQuAD dataset contained questions and answers that had the same structure. Hence we obtained high scores on SQuAD with our the baseline model to begin with.

SQuAD trained models in general do not have a structured model of facts or semantics that would help identify when the context truly answers the question. They overly fixate on identifying answers of the right type. This is the reason why it did not perform as well when the baseline model was evaluated on generalized datasets.

5.1 Further research

We find many opportunities to do further experimentation to improve a model on reading comprehension(RC) systems with Extractive Question and Answering(EQA) tasks. A lot of interesting and promising research work has been done in this space in recent years. Some notable ones include: (1) training on data that contains adversarial changes(Jia and Liang, 2017) to the original data by adding a sentence at the end of the context or by adding a random sentence into anywhere in the context. (2) using DROP(Dua et al., 2019) Discrete Reasoning Over Paragraphs datasets which requires to search in multiple positions in the context, and the perform discrete operations over them (such as addition, counting, or sorting) (3) Dataset Cartography(Swayamdipta et al., 2020) which creates a "data map" on the training data by identifying hard, medium and easy examples. (4) forgettable examples to tackle spurious correlations(Yaghoobzadeh et al., 2021) (5) training on triviaQA(Joshi et al., 2017) which helps achieve complex and cross-sentence reasoning. We intend to continue our work by combining these techniques with our own.

6 Conclusion

We have presented an intuitive and simple two-phased approach for training reading comprehension(RC) systems for Extractive Question and Answering(EQA) tasks so they generalize well to adversarial settings while maintaining high scores for in-domain questions. We think that this approach can also applied to other NLP tasks such as NLI or MNLI with little or no modifications.

References

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL 2018*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Association for Computational Linguistics (ACL)*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.