QADiver: Interactive Framework for Diagnosing QA Models

Gyeongbok Lee[1,2]* alias_n@yonsei.ac.kr
Sungdong Kim[2] sungdong.kim@navercorp.com
Seung-won Hwang[1] seungwonh@yonsei.ac.kr
[1] Yonsei University
[2] Clova AI Research, NAVER Corp.

Abstract

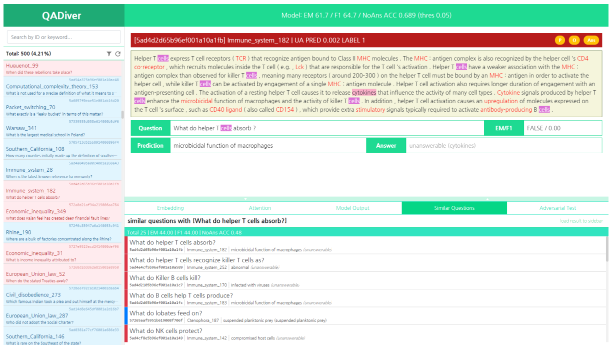
Question answering (QA) extracting answers from text to the given question in natural language, has been actively studied and existing models have shown a promise of outperforming human performance when trained and evaluated with SQuAD dataset. However, such performance may not be replicated in the actual setting, for which we need to diagnose the cause, which is non-trivial due to the complexity of model. We thus propose a web-based UI that provides how each model contributes to QA performances, by integrating visualization and analysis tools for model explanation. We expect this framework can help QA model researchers to refine and improve their models.

Introduction

With the help of large-scale question answering datasets and deep learning frameworks released to the public, question answering (QA) models have been improved rapidly as community efforts. Recently, the reported accuracy of state-of-the-art models has exceeded human performance in SQuAD task, where plausible answers are extracted for the given question and context document pair. However, it is still non-trivial to reproduce such accuracy reported in the paper in production settings.

This paper proposes a diagnosing tool, for troubleshooting the performance gap. For example, is the performance lower than expected because of the biased training to certain types of questions or texts? Is the model attending to wrong words for question or text understanding? Is the embedding suitable for the given task? At the same time, there is a concerning observation that these models can be easily perturbed by a simple adversarial example added (Jia and Liang 2017). A desirable tool may support developers to easily perturb training to identify such vulnerability.

We will demonstrate **QADiver**, a data-centric diagnosing framework for QA model, with diverse interactive visualization and analysis tools for a full pipeline of the attention-based QA model. The framework is connected with the target model to retrieve answer span prediction, inner-model values (such as attention), and no answer probability for given context and question from the model.



Model: EM 61.7 / F1 64.7 / NoAns ACC 0.689 (thres 0.05)

Total: 500 (4.21%)

Huguenot_99
When did these rebellions take place?

5ad54a375b96ef001a10ac48
Computational_complexity_theory_153
What is not used for a precise definition of what it means to s⋯

5a6057f0eae51e001ab14d20
Packet_switching_70
What exactly is a "leaky bucket" in terms of this matter?

57339555d058e614000b5df6
Warsaw_341
What is the largest medical school in Poland?

5705f13d52bb8914006896f4
Southern_California_108
How many counties initially made up the definition of souther⋯

5ad4a049ba00c4001a268e43
Immune_system_28
When is the latest known reference to immunity?

5ad4d2d65b96ef001a10a1fb
Immune_system_182
What do helper T cells absorb?

572a0d21af94a219006aa784
Economic_inequality_349
What does Rajan feel has created deep financial fault lines?

572f6c85947a6a140053c941
Rhine_190
Where are a bulk of factories concentrated along the Rhine?

5727e9523acd2414000def96
Economic_inequality_31
What is income inequality attributed to?

57268d2ddd62a815002e8950
European_Union_law_52
When do the stated Treaties apply?

5728eef92ca10214002daab4
Civil_disobedience_273
Which famous Indian took a plea and put himself at the mercy⋯

5ad14d8e645df0001a2d16b7
European_Union_law_287
Who did not adopt the Social Charter?

5ad0381a77cf76001a686e33
Southern_California_146
What is rare on the Southeast of the state?

[5ad4d2d65b96ef001a10a1fb] Immune_system_182 | UA PRED 0.002 LABEL 1

Helper T cells express T cell receptors (TCR) that recognize antigen bound to Class II MHC molecules . The MHC : antigen complex is also recognized by the helper cell 's CD4 co-receptor , which recruits molecules inside the T cell (e.g. , Lck) that are responsible for the T cell 's activation . Helper T cells have a weaker association with the MHC : antigen complex than observed for killer T cells , meaning many receptors (around 200-300) on the helper T cell must be bound by an MHC : antigen in order to activate the helper cell , while killer T cells can be activated by engagement of a single MHC : antigen molecule . Helper T cell activation also requires longer duration of engagement with an antigen-presenting cell . The activation of a resting helper T cell causes it to release cytokines that influence the activity of many cell types . Cytokine signals produced by helper T cells enhance the microbicidal function of macrophages and the activity of killer T cells . In addition , helper T cell activation causes an upregulation of molecules expressed on the T cell 's surface , such as CD40 ligand (also called CD154) , which provide extra stimulatory signals typically required to activate antibody-producing B cells .

Question: What do helper T cells absorb ?
EM/F1: FALSE / 0.00
Prediction: microbicidal function of macrophages
Answer: unanswerable (cytokines)

similar questions with [What do helper T cells absorb?]

Total 25 | EM 44.00 | F1 44.00 | NoAns ACC 0.48

What do helper T cells absorb?
5ad4d2d65b96ef001a10a1fb | Immune_system_182 | microbicidal function of macrophages *(unanswerable)*

What do helper T cells recognize killer T cells as?
5ad4e4cf5b96ef001a10a589 | Immune_system_252 | abnormal *(unanswerable)*

What do Killer B cells kill?
5ad4d2105b96ef001a10a1c7 | Immune_system_170 | infected with viruses *(unanswerable)*

What do B cells help T cells produce?
5ad4d2d65b96ef001a10a1fc | Immune_system_183 | microbicidal function of macrophages *(unanswerable)*

What do lobates feed on?
57265aaf5951b619008f706f | Ctenophora_187 | suspended planktonic prey (suspended planktonic prey)

What do NK cells protect?
5ad4cf8e5b96ef001a10a149 | Immune_system_142 | compromised host cells *(unanswerable)*

Figure 1: Snapshot of QADiver Framework

Our framework targets SQuAD 2.0 dataset (Rajpurkar, Jia, and Liang 2018), a machine reading comprehension benchmark that contains both answerable and unanswerable question for given context. In this task, not only finding plausible answer span in context but predicting question is answerable or not is also crucial. By exploring large question-answer instances, we expect developers can make a better diagnosis and find insights, than those made from qualitative observations of a few instances. The demonstration video for the framework is available in https://youtu.be/V6c8nls6Qcc.

Key Features

Figure 1 shows a snapshot of our system, where the cause of low accuracy can be diagnosed as one or more of the following problems:

Dataset Bias

Left-sidebar in Figure 1 shows data instances form SQuAD 2.0 development set instances. To support a quick exploration of how such instances are distributed for answers predicted right or wrong, we color each to blue and red respectively. This color code is used for the entire system. The top right area in Figure 1 shows more detailed statistics about the instance selected: a question, corresponding context, gold answers and prediction result from the model

*This work was done during an internship at Clova AI Research, NAVER.

including unanswerable probability and EM/F1 scores. Answer span and out-of-vocabulary words are highlighted. The user can also switch between the original text and preprocessed text by the model for both context and question to identify cases where the underlying NLP models, such as tokenizer, are not working as intended.

Embedding Analysis

QA model performance also depends on the effectiveness of word embedding. To visualize such effectiveness, we allow developers to visualize word vector (and nearby words in the space by clicking a word. The user can restrict words in the context, or use the whole vocabulary set for the dataset. For a fast similarity search of dense vectors, we use FAISS (Johnson, Douze, and Jégou 2017) library.

Neural Model Internals

Visual representation for internal states of the neural model shows whether right words are highlighted for understanding questions or finding answers. The most common example is visualizing attention and output layer in the model. For attention layer, we visualize a context-question attention matrix heatmap for the given instance. Similarly, we provide an interpreted version of the model output used for answer span prediction and answerability decision by listing the top-k words with the highest weights. Users can also see the list of answer span candidates (including "unanswerable" case) and its certainty as a colored heatmap.

Question Bias

A model can be biased to answer a certain type of questions particularly well. To diagnose such a case, we identify similar questions to the given instance, labeled with prediction result and evaluation metrics (EM/F1). For a desirable question embedding, projecting similar question types close in the embedding space, we use well-studied features from the question and gold answer such as: answer length, the existence of number and entity, and 2-word question prefix like *What is* and *How many*. Feature values of questions in the same class are mean-aggregated to generate a global statistic vector. To represent the characteristic of each question, local features like word match ratio between context and question and one-hot vector for frequent words are used so that similar types of questions have high similarity. As each question is vectorized, top similar questions for the instance can be retrieved from the whole dataset by the similarity search.

Adversarial Test

As we overviewed, many existing QA models are reportedly lacking the robustness over adversarial examples. Using our tool, the user can easily perform the adversarial test for each instance in two ways: manual modification and rule-based test. First, the user can modify some words in context document and question from the data viewer by double-clicking target word and replacing to another. After this edit, we show an updated prediction and EM/F1 score.

Instead of perturbing each word, which may be costly, the user may create reusable adversarial rules, about word and its part-of-speech (POS) tag. We use NLTK toolkit (Bird, Klein, and Loper 2009) for POS tagging and word tokenization used in rule matching. We also provide pre-defined adversarial rules from SEAR (Ribeiro, Singh, and Guestrin 2018) for those wanting to check the robustness for common cases.

Related Work

Due to the growing demand for model interpretability, many visualization tools for QA models were proposed: AllenNLP machine comprehension demo[1] provides answer span highlights and attention matrix visualization. (Rücklé and Gurevych 2017) shows attention visualization on context and question text, and provides a comparison between the two model. (Shusen Liu and Bremer 2018) proposes bipartite graph attention representation and hierarchical visual for highly asymmetric attention. This tool also supports word- and attention-level perturbation by user edit.

Acknowledgements

This work was supported by the ICT R&D program of MSIT/IITP. [No.2017-0-01778, Development of Explainable Human-level Deep Machine Learning Inference Framework]

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Jia, R., and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on EMNLP*.

Johnson, J.; Douze, M.; and Jégou, H. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don't know: Unanswerable questions for squad. *Proceedings of the 56th Annual Meeting of the ACL.*

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the ACL*, volume 1, 856–865.

Rücklé, A., and Gurevych, I. 2017. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. In *Proceedings of the 55th Annual Meeting of the ACL-System Demonstrations (ACL 2017)*, 19–24. Association for Computational Linguistics.

Shusen Liu, Tao Li, Z. L. V. S. V. P., and Bremer, P.-T. 2018. Visual interrogation of attention-based models for natural language inference and machine comprehension. In *Proceedings of the 2018 Conference on EMNLP: System Demonstrations*.

[1] http://demo.allennlp.org/machine-comprehension