Building a Robust QA system using an Adversarially Trained Ensemble

Stanford CS224N Default Project (Robust QA track)

Kevin Lee
Department of Electrical Engineering
Stanford University
kelelee@stanford.edu

Abstract

Despite monumental progress in natural language understanding, QA systems trained on giant datasets are still vulnerable to domain transfer. Evidence shows that language models pick up on domain-specific features which hinders it from generalizing to other domains. In this project, we implore the use of adversarial networks to regularize the finetuning process which encourages the generator model to learn more meaningful representations of context and questions. We then construct an ensemble of these models based on each model's performance on specific subgroups of questions.

1 Introduction

In the task of Question Answering (QA), the benchmark on popular datasets such as SQuAD [1] have been beaten numerous times to a point where the state of the art (SOTA) models are now outperforming humans. However, as shown by [2], these models do not generalize well to domains different than the one they are trained on. In other words, the models have overfitted to domain-specific features in the dataset which prevents it from generalizing.

In order to train a model that performs well across domains, a mechanism that either encourages the learning of domain-invariant features or discourages the learning of domain-specific features need to be introduced. To that end, we implore the use of adversarial networks, whereby a discriminator is used to predict the originating domain of the input given a hidden representation from the QA model. To prevent the discriminator from accurately predicting the domain, the QA model would need to capture domain-invariant features from the input, while simultaneously extracting important information to perform accurately on the QA task. Since the QA model architecture remains unchanged, this approach can be applied to practically any QA model.

We then create an ensemble of adversarially trained models in order to maximize performance. Each model in the ensemble is weighted according to their F1 scores achieved on each question class on the in-domain validation set. We elaborate more on question classes in Section 3.3.

2 Related Work

2.1 Pre-trained Language Models

BERT [3] is one of many language models that are pre-trained on a large corpus. BERT randomly masks some input tokens and predicts the masked tokens based on its context. In modern NLP, it has become a common practice to select one of these pre-trained models as a starting point and finetune it on a downstream task such as sentence classification, POS tagging or question answering.

2.2 Adversarial Networks

Adversarial methods have been made popular by the application of Generative Adversarial Networks [4] across many Deep Learning fields, including NLP and computer vision. It involves a generator network and a discriminator network taking turns outwitting each other in a minmax optimization procedure where the generator tries to generate as realistic of a synthesized output as possible while the discriminator learns to distinguished a synthesized output from real ones.

While GANs are typically associated with the generation of realistic images or text, GANs have also been proven to help with domain adaptation/generalization. In Domain-Adversarial Neural Network (DANN) [5], the authors implore the use of an adversarial classifier to predict the domain of the data, which encourages the feature extractor, G_f to focus on domain-invariant features. This approach is then applied to the task of QA by [6], which is what this project is heavily inspired by.

3 Approach

3.1 Baseline:

The baseline system finetunes DistilBERT [7] (a smaller, distilled version of the original BERT model) on a specified domain of training data. We train two baselines: the first on all in-domain training data, and the second on both in-domain and out-of-domain training data. The loss function is the sum of the negative log-likelihood (cross-entropy) loss for the start and end locations. That is, if the gold start and end locations are $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, N\}$ respectively, then the loss for a single example is:

$$\mathcal{L}_{QA} = -\log \boldsymbol{p}_{start}(i) - \log \boldsymbol{p}_{end}(j) \tag{1}$$

The code to train the baseline model has been provided by the course staff at:

`https://github.com/MurtyShikhar/robustqa`

3.2 Adversarial Training

Inspired by [6], this project attempts to use a discriminator network, D to predict the domain of the input context and question. D will be trained simultaneously during the finetuning of the DistilBERT model, G and will be given the [CLS] token representation from G as input. This is illustrated in Figure 1.

For K domains, and N_k data for each domain $k \in [1, K]$, the adversarial loss that is the Kullback-Leibler (KL) divergence between a uniform distribution over K classes denoted as $\mathcal{U}(l)$ and the discriminator's prediction is given by:

$$\mathcal{L}_{adv} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} KL\left(\mathcal{U}(l) || P_\phi(l_i^{(k)} | \mathbf{h}_i^{(k)})\right) \tag{2}$$

The QA model's loss function is $\mathcal{L}_{QA} + \lambda_{adv}\mathcal{L}_{adv}$, where λ_{adv} is a hyperparameter that controls the weight of the adversarial loss $\mathcal{L}_{adv}$ relative to $\mathcal{L}_{QA}$.

As for the discriminator's loss function, $\mathcal{L}_D$, we simply take the negative log-likelihood of the D's prediction. However, since there is a major imbalance in the amount of data between the in-domain and out-of-domain train sets, we provide weights to the discriminator's negative log likelihood loss based on the effective number of samples [8]. Thus, we get the classed-balanced discriminator loss:

$$\mathcal{L}_D = -\frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \frac{1-\beta_k}{1-\beta_k^{N_k}} P_\phi(l_i^{(k)} | \mathbf{h}_i^{(k)}), \beta_k = \frac{N_k - 1}{N_k} \tag{3}$$

To clarify, we set $K = 3$ when training a model with only the in-domain datasets and $K = 6$ when training with both in-domain and out-of-domain datasets. Refer to 4.1 for more information on the datasets.

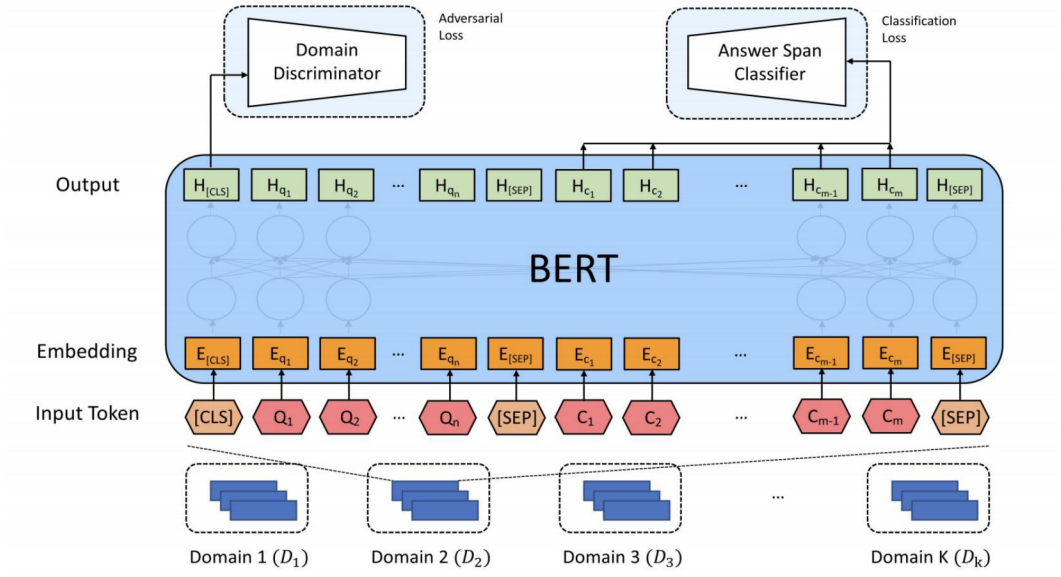

Figure 1: Model architecture for adversarial training

In this project, we adapted the official implementation of [6] from `https://github.com/seanie12/mrqa` into the starter code.

3.3 QA Analysis by Question Class

In order to perform the ensemble step detailed in Section 3.4, we'd have to first split the questions up by class as done in [9]. We start off with the same questions classes from their paper, including combining the "which" and "what" classes, and we've added two of our own classes, namely "instruction" and "yes/no". The distribution of questions classes as well as examples are shown in Table 1. While it is unclear how the question classifier is implemented in [9], we used a simple rule where suppose a class c has a set of phrases $P^c = \{p_0^c, ..., p_n^c\}$, question q falls under class c if q contains any phrase $p_i^c \in P^c$. Refer to Appendix A.1 for more information. This, however, means that questions may be categorized to multiple classes instead of being uniquely categorized as in [9], which is why the percentages in each column of Table 1 does not necessarily add up to 100%. One interesting observation is that the distribution of questions remain consistent across domains. Because of this, we conjecture that this method of separation is robust to domain shifts.

3.4 Ensemble

Following the approach of [9], we train a few models and evaluate their in-domain validation F1 scores on the classes of questions as detailed in Section 3.3. We then construct an ensemble of these models where the output of each model is weighted by the respective F1 scores. Models that output the same answer will have their weights added. Finally, we select the prediction that carries the highest weight as the output of the ensemble.

4 Experiments

4.1 Data:

The datasets are provided by the course. Table 2 specifies the data splits:

Class	in-domain train	in-domain dev	oo-domain train	oo-domain dev	oo-domain test	Sample
during	0.89%	0.92%	0.52%	0.26%	0.18%	What was the win/loss ratio in 2015 for the Carolina Panthers during their regular season?
how is / are	0.87%	1.14%	1.31%	0.79%	0.87%	If a detention requires a pupil to just sit there, how are they required to sit?
how big / size	8.36%	8.28%	6.04%	3.93%	2.80%	How large was the audience BSkyB said they could reach?
how many / much	7.30%	7.04%	4.99%	2.09%	1.38%	How many miners died in the typhoid outbreak of 1854?
how old	0.97%	0.81%	0.79%	0.79%	0.30%	How old are the fossils found that represent ctenophhores?
what	46.53%	43.15%	63.25%	65.18%	66.22%	What were the reasons why residents moved to the town of Fresno Station?
when	12.77%	13.99%	7.61%	9.42%	9.20%	When was the Tower Theatre built?
where	8.24%	8.94%	7.87%	10.21%	4.82%	Where is Audra McDonald from?
who / whom	24.39%	26.42%	25.72%	25.13%	25.94%	Who is Kearney Boulevard named after?
why	0.75%	0.86%	0.79%	0.52%	0.34%	Why does Fresno only have UHF television stations?
instruc-tion	0.26%	0.25%	0.00%	0.00%	0.00%	Name a text that might be used by a religious teacher to teach
yes/no	1.18%	0.68%	0.26%	0.00%	0.89%	Are there any regions where the Treaty of European Union excludes from jurisdiction?
undefined	2.09%	2.77%	0.26%	0.00%	0.23%	More in the present prevalence of civil disobedience has turned and said to be?

Table 1: Distributions and sample questions per class.

Dataset	Question Source	Passage Source	Train	dev	Test
in-domain datasets					
SQuAD [1]	Crowdsourced	Wikipedia	50000	10,507	-
NewsQA [10]	Crowdsourced	News articles	50000	4,212	-
Natural Questions [11]	Search logs	Wikipedia	50000	12,836	-
oo-domain datasets					
DuoRC [12]	Crowdsourced	Movie reviews	127	126	1248
RACE [13]	Teachers	Examinations	127	128	419
RelationExtraction [14]	Synthetic	Wikipedia	127	128	2693

Table 2: Statistics for datasets used for building the QA system for this project. **Question Source** and **Passage Source** refer to data sources from which the questions and passages were obtained. Table borrowed from [15]

4.2 Evaluation Method:

Performance is measured via two metrics: Exact Match (EM) score and F1 score. EM is a binary measure (i.e. true/false) of whether the system output matches the ground truth answer exactly, while F1 is the harmonic mean of precision and recall. The EM and F1 scores are averaged across the entire evaluation dataset to get the final reported scores.

4.3 Experimental details:

We first train the baseline models as outlined in Section 3.1. We used the AdamW [16] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 3×10^{-5}.

We then include the adversarial components as described in 3.2. We follow [6] in that D is a fully-connected network (FCN) with 3 hidden layers with ReLU activations after each layer followed by a dropout layer with $p_{dropout} = 0.1$. Due to limitations in resources, we only managed to slightly tune λ_{adv} by attempting two values, 0.01 and 0.1. We also attempt to train models with and without the out-of-domain train set, but we made sure to always validate on the in-domain dev set for all training so that we are not "cheating" by optimizing directly on the new domain.

All training runs are done with a batch size of 16 and max epochs of 3. For all runs, F1 scores converge within 3 epochs. Table 2 shows the loss curves over each training step for the Adversarial, $\lambda_{adv} = 0.1$ + Finetune model (See section 4.4), while Table 3 shows the Validation EM and F1 scores.

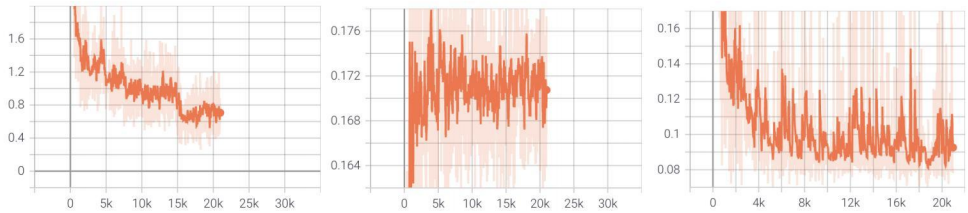

Figure 2: Loss curves versus number of steps for adversarial training with $\lambda_{adv} = 0.1$. Figures are obtained from TensorBoard with a smoothing of 0.8. From left to right: $\mathcal{L}_{QA}$, $\lambda_{adv}\mathcal{L}_{adv}$, $\mathcal{L}_D$.

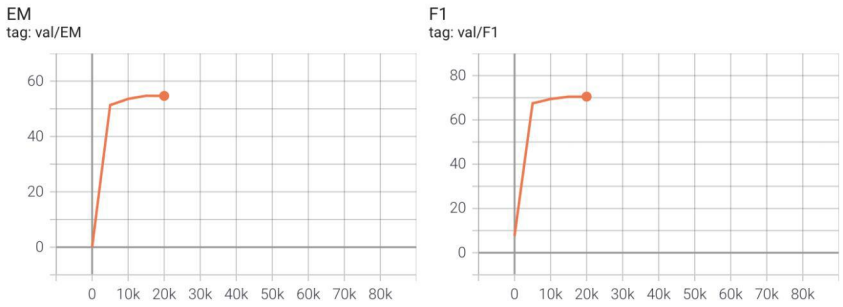

Figure 3: Validation EM and F1 scores versus number of steps for adversarial training with $\lambda_{adv} = 0.1$. Figures are obtained from TensorBoard.

4.4 Results:

Table 3 shows the EM and F1 scores of models with varying values λ_{adv} and training dataset. Note that all models are validated against the in-domain dev set during training and evaluated against the out-of-domain dev and test sets. The results for the out-of-domain test set are obtained by submitting to the Default Final Project (RobustQA Track) test leaderboard on Gradescope. See that the F1 scores of models trained with $\lambda_{adv} = 0.1$ outperform the baselines.

Table 4 shows the breakdown of F1 scores for the same models and evaluation set, but per question class. Due to limited access on the test leaderboard, we are unable to provide these scores for the ensemble on the oo-domain test set.

Model	Train Dataset	Validation Dataset	Evaluation Dataset	EM	F1
Baseline					
Baseline	in-domain train	in-domain dev	oo-domain dev	33.25	48.43
Baseline + Finetune	in-domain train + oo-domain train	in-domain dev	oo-domain dev	34.29	48.65
Adversarial					
Adversarial, $\lambda_{adv} = 0.01$	in-domain train	in-domain dev	oo-domain dev	32.72	47.06
Adversarial, $\lambda_{adv} = 0.1$	in-domain train	in-domain dev	oo-domain dev	31.94	49.18
Adversarial, $\lambda_{adv} = 0.1$ + Finetune	in-domain train + oo-domain train	in-domain dev	oo-domain dev	34.55	50.46
			oo-domain test	41.10	59.75
Ensemble					
Ensemble	N/A	N/A	oo-domain dev	**35.34**	**50.53**
			oo-domain test	**42.84**	**61.29**

Table 3: EM and F1 scores of trained models

After obtaining results from the adversarial models, we construct the ensemble. In practice, we would use several models trained under the best hyperparameter configuration. Due to resource constraints, we chose to construct an ensemble using the following previously-trained models:

- Baseline
- Baseline + Finetune
- Adversarial, $\lambda_{adv} = 0.1$
- Adversarial, $\lambda_{adv} = 0.1$ + Finetune

5 Analysis

We have shown that the adversarial training yields higher F1 and EM scores, which confirms our suspicion that the baseline model had picked up domain-specific features during training, and that training the model to become more domain-agnostic helps it become more robust to domain shifts. Furthermore, we observe no decline in validation F1 scores on the in-domain dev set nor increase in time to convergence, indicating that there is no trade-off in performance across the domains.

Next, we see that constructing an ensemble that weighs models based on their respective performance on each question class delivered promising results. One possible explanation for this is because these question classes remain applicable across the different domains. However, from Table 1 we see that there is an imbalance in questions by class, which might've cause models to focus on answering questions of the more dominant classes. Ideally, we would want question classes to be balanced.

6 Conclusion

We've applied an adversarial approach towards training models for QA tasks which allows them to perform more robustly across unseen domains. We've also used a technique of analyzing performance on question classes to determine the weights for each trained model in an ensemble in order to boost performance.

If we were to instead adopt a Mixture of Experts (MoE) [17] approach where each model's output is gated by a separate DistilBERT model, that gating model might learn to rely on domain-specific

Model	during	how is/ are	how big/ size	how many/ much	how old	what	when	where	who/ whom	why	in-struc-tion	yes/ no	un-de-fined
Baseline													
Baseline	70.67 0.00	64.59 66.67	64.88 **64.44**	64.32 **33.33**	70.78 33.33	70.52 52.34	70.37 46.69	66.67 45.19	**77.70** 37.45	69.38 20.00	**71.99** N/A	49.53 N/A	59.54 N/A
Baseline + Finetune	73.39 0.00	65.22 33.33	**66.08** 43.33	**64.99** 6.25	71.22 33.33	69.98 53.60	68.75 47.14	66.50 60.45	77.14 37.92	67.81 **47.86**	62.74 N/A	**51.03** N/A	**60.12** N/A
Adversarial													
Adversarial, $\lambda_{adv} = 0.01$	72.14 0.00	64.83 66.67	64.20 47.88	63.09 12.50	**71.96** 33.33	70.12 52.01	70.15 46.75	65.99 44.33	77.29 36.32	69.18 42.86	62.50 N/A	49.96 N/A	57.99 N/A
Adversarial, $\lambda_{adv} = 0.1$	72.78 0.00	**67.76** 66.67	65.05 43.21	63.05 6.25	70.78 33.33	**70.70** 53.79	**70.90** 42.63	**67.04** 48.84	77.68 **40.49**	**71.57** 42.86	67.19 N/A	50.29 N/A	58.98 N/A
Adversarial, $\lambda_{adv} = 0.1$ + Finetune	**73.99** 0.00	64.53 66.67	64.23 59.78	63.28 27.08	69.27 33.33	70.45 **55.51**	69.87 **46.94**	65.38 **50.64**	76.90 40.15	68.17 33.33	66.13 N/A	50.61 N/A	58.98 N/A
Ensemble													
Ensemble 1	75.16 0.00	67.56 66.67	66.27 50.00	65.11 6.25	72.47 33.33	72.04 55.22	71.74 39.81	68.07 55.41	78.87 40.25	71.85 42.86	71.68 TBD	52.11 TBD	61.02 TBD

Table 4: F1 scores by question class. Scores on top are evaluated against the in-domain dev set, while the scores on the bottom are against the oo-domain dev set. Highest F1 scores per class per evaluation dataset (excluding the ensemble of models) are bolded.

features in determining the model weights which will cause the overall ensemble to perform worse on the out-of-domain datasets.

With that said, one clear limitation of the question class analysis is in its phrase matching mechanism. It relies on the set of phrases used to capture the majority of question distribution created by language variations and typos. Another limitation is that each question class may only apply to specific languages, and in our case, English.

References

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016.

[2] Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. Learning and evaluating general linguistic intelligence, 2019.

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

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016.

[6] Seanie Lee, Donggyu Kim, and Jangwon Park. Domain-agnostic question-answering with adversarial training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, 2019.

[7] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples, 2019.

[9] Anna Aniol and Marcin Pietron. Ensemble approach for natural language question answering problem, 2019.

[10] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830, 2016.

[11] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

[12] Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension. *CoRR*, abs/1804.07927, 2018.

[13] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. *CoRR*, abs/1704.04683, 2017.

[14] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[15] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. *CoRR*, abs/1910.09753, 2019.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[17] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

A Appendix

A.1 Question Class Phrases

As described in Section 3.3, each question class has a set of phrases that determine if a question falls under its class or not. These phrases are listed in Table 5 in the form of a Regex. For more information on Regex, refer to `https://en.wikipedia.org/wiki/Regular_expression`. If a question does not satisfy any of these classes, it is put under the "undefined" class.

A.2 Sample QA

We list some results from the model trained using both in-domain and out-of-domain training datasets with $\lambda_{adv} = 0.1$.

Question: Where are pyrenoids found?
Context: The chloroplasts of some hornworts and algae contain structures called pyrenoids. They are not found in higher plants. Pyrenoids are roughly spherical and highly refractive bodies which are a site of starch accumulation in plants that contain them. They consist of a matrix opaque to electrons,

Class	Regex representation of phrases
during	during
how is / are	(h[ow]{1,2}|hoe) (does|did|are|can|could|did|do|does|had |has|have|is|might|was|were|will|would)
how big / size	(h[ow]{1,2}|hoe) (big|common|deep|early|far |high|large|long|often|soon|well|wide)?
how many / much	((h[ow]{1,2}|hoe)\s?(many|much|man|may)|amount|number)
how old	((h[ow]{1,2}|hoe) old| age)
what	((\.las|by|is|of)\??$|w[ah]{1,3}[trdsy]|w[hc]+ich?)
when	(date)|(w[ah]{1,3}t (year|month|date|time))|(wh[ae]+n)
where	where?
who / whom	w+hi?om?
why	wh[iy]
instruction	^((define)|(identify)|(list)|(name? (a|some|the|two))
yes/no	^(are|can|could|did|do|does|has|have|is|was?|were|will|would)

Table 5: Regex representation of each question class phrases.

surrounded by two hemispherical starch plates. The starch is accumulated as the pyrenoids mature. In algae with carbon concentrating mechanisms, the enzyme rubisco is found in the pyrenoids. Starch can also accumulate around the pyrenoids when CO2 is scarce. Pyrenoids can divide to form new pyrenoids, or be produced "de novo".
Answer: The chloroplasts of some hornworts and algae
Prediction: chloroplasts

Question: what is the oasis in ready player one
Context: BPB In the 2040s , the world has been gripped by an energy crisis from the depletion of fossil fuels and the consequences of global warming , and overpopulation , causing widespread social problems and economic stagnation . To escape the decline their world is facing , people turn to the OASIS , a virtual reality simulator accessible by players using visors and haptic technology such as gloves . It functions both as an MMORPG and as a virtual society , with its currency being the most stable in the real world . It was created by James Halliday who , when he died , had announced in his will to the public that he had left an Easter egg inside OASIS , and the first person to find it would inherit his entire fortune and the corporation . The story follows the adventures of Wade Watts , starting about five years after the announcement , when he discovers one of the three keys pointing to the treasure . EEPE
Answer: a virtual reality simulator accessible by players using visors and haptic technology such as gloves
Prediction: a virtual reality simulator

Question: dendrites and cell bodies are components of what type of matter found in the brain
Context: BPB Grey matter (or gray matter) is a major component of the central nervous system , consisting of neuronal cell bodies , neuropil (dendrites and myelinated as well as unmyelinated axons) , glial cells (astrocytes and oligodendrocytes) , synapses , and capillaries . Grey matter is distinguished from white matter , in that it contains numerous cell bodies and relatively few myelinated axons , while white matter contains relatively few cell bodies and is composed chiefly of long - range myelinated axon tracts . The colour difference arises mainly from the whiteness of myelin . In living tissue , grey matter actually has a very light grey colour with yellowish or pinkish hues , which come from capillary blood vessels and neuronal cell bodies . EEPE
Answer: Grey matter
Prediction: Grey matter

Question: Why do polar water bodies support a higher amount of life?
Context: Free oxygen also occurs in solution in the world's water bodies. The increased solubility of O 2 at lower temperatures (see Physical properties) has important implications for ocean life, as polar oceans support a much higher density of life due to their higher oxygen content. Water polluted

with plant nutrients such as nitrates or phosphates may stimulate growth of algae by a process called eutrophication and the decay of these organisms and other biomaterials may reduce amounts of O 2 in eutrophic water bodies. Scientists assess this aspect of water quality by measuring the water's biochemical oxygen demand, or the amount of O 2 needed to restore it to a normal concentration.
Answer: higher oxygen content
Prediction: higher oxygen content

Question: What position does Von Miller play for the Denver Broncos?
Context: The Broncos took an early lead in Super Bowl 50 and never trailed. Newton was limited by Denver's defense, which sacked him seven times and forced him into three turnovers, including a fumble which they recovered for a touchdown. Denver linebacker Von Miller was named Super Bowl MVP, recording five solo tackles, 2½ sacks, and two forced fumbles.
Answer: linebacker
Prediction: Denver linebacker

Question: When could VA Dundee?
Context: The VA is in discussion with the University of Dundee, University of Abertay, Dundee City Council and the Scottish Government with a view to opening a new £43 million gallery in Dundee that would use the VA brand although it would be funded through and operated independently. As of 2015, with costs estimated at £76 million, it is the most expensive gallery project ever undertaken in Scotland. The VA Dundee will be on the city's waterfront and is intended to focus on fashion, architecture, product design, graphic arts and photography. It is planned that it could open within five years. Dundee City Council is expected to pay a major part of the running costs. The VA is not contributing financially, but will be providing expertise, loans and exhibitions.
Answer: within five years
Prediction: within five years.

Question: the creation of human beings in the kumulipo happens during which wā or period of creation
Context: BLiB In the ninth wā , Lailai takes her eldest brother Kii as a mate and the first humans are born from her brain . EELiE
Answer: the ninth wā
Prediction: the ninth wā

Question: What is the moniker that is being used to describe the region's diversified technology?
Context: The Tech Coast is a moniker that has gained use as a descriptor for the region's diversified technology and industrial base as well as its multitude of prestigious and world-renowned research universities and other public and private institutions. Amongst these include 5 University of California campuses (Irvine, Los Angeles, Riverside, Santa Barbara, and San Diego); 12 California State University campuses (Bakersfield, Channel Islands, Dominguez Hills, Fullerton, Los Angeles, Long Beach, Northridge, Pomona, San Bernardino, San Diego, San Marcos, and San Luis Obispo); and private institutions such as the California Institute of Technology, Chapman University, the Claremont Colleges (Claremont McKenna College, Harvey Mudd College, Pitzer College, Pomona College, and Scripps College), Loma Linda University, Loyola Marymount University, Occidental College, Pepperdine University, University of Redlands, University of San Diego, and the University of Southern California.
Answer: The Tech Coast
Prediction: Tech Coast