Distractor Re-ranking For Automatic Quiz Generation

Girish Kumar (girishk@Stanford.edu)
Andrew Wang (andy2000@Stanford.edu)

Abstract

In this project we seek to tackle the complex and interesting problem of question generation for the purpose of enhancing the educational experience of students. Learning through evaluation has been widely known to be a very effective method of assessing a students knowledge and offering helpful insights in targeted learning methods. We used three methods: a pointwise ranking SVM, a listwise ranking neural net and a GAN combining both frameworks. The listwise neural net yielded the best results.

Data

The data we used to train our models came from the SciQ dataset which contains 13679 crowdsourced science exam questions about Physics, Chemistry and Biology, among others. The initial candidate list is obtained by doing a similarity search over an unsupervised word2vec model with the correct answer as a key.

- Question Sentence: string
- Correct Answer: string
- Human Generated Distractors: list of 3 strings
- Distractor from nearest-neighbor search and not contained in the human-generated list: list of 100 strings

Figure 1. SciQ Data Format

References

1. Manish Agarwal and Prashanth Mannem. Automatic gap-fill question generation from text books. In Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, pages 56–64. Association for Computational Linguistics, 2011.
2. Jinnie Shin, Qi Guo, and Mark J. Gierl. Multiple-choice item distractor development using topic modeling approaches. Frontiers in Psychology, 10:825, 2019.
3.Data.allenai.org. (2019). *Pt. 1 - SciQ Dataset*. [online] Available at: http://data.allenai.org/sciq/pt1/ [Accessed 10 Jun. 2019].
4. Girish Kumar, Matthew Henderson, Shannon Chan, Hoang Nguyen, and Lucas Ngoo. Question-answer selection in user to user marketplace conversations. arXiv preprint arXiv:1802.01766, 2018.

Features

For each question q, answer a, and list of candidate distractors, the model ranks each distractor. In all models we encode the data using Tensorflow-Hub's pre-trained universal sentence encoder. An example data entry is included below.

What type of organism is commonly used in preparation of foods such as cheese and yogurt?

Correct Answer	Distractor 1	Distractor 2	Distractor 2
mesophilic organisms	protozoa	gymnosperms	viruses

Source SciQ Data Part 1 (3)

Models

Ranking SVM

The ranking SVM concatenates the encoded data and predicts 1 if d_i is from a human generated sample, and predicts 0 otherwise. At run time distractors will be ranked according to scored output by SVM

Neural Dot Product Ranker

We use a neural network to rank human-generated distractors favorably across a list of candidates. The network is implemented as a single layer, feed forward transformation. Separating the model into two networks allows the network to run efficiently on varying sizes of D.

$$P_i = P(d_i \mid \mathbf{B},\ q,\ a) \approx \frac{e^{\mathbf{h}(\psi(q);\psi(a))^{\mathrm{T}}\mathbf{g}(d_i,\ \mathbf{D})}}{\sum_{j=0}^{N} e^{\mathbf{h}(\psi(q);\psi(a))^{\mathrm{T}}\mathbf{g}(d_j,\ \mathbf{D})}} \quad (2)$$

Ranking GAN

This model seeks to combine the two approaches taken by the SVM and Neural Dot models. To do so, we looked to use IRGANs. The generative retrieval models $p_\theta(d \mid q,a)$ The discriminator is a binary classifier where $f_\phi(q,a,d)$. The derived optimal parameter for the discriminative retrieval model is given below.

$$\phi^* = \underset{\phi}{\operatorname{argmax}} \sum_{i=1}^{N} \mathbb{E}_{d\sim p_{\text{human}}(d|q_i,a_i)}[\log(\sigma(f_\phi(q_i,a_i,d)))] + \mathbb{E}_{d\sim p_{\theta^*}(d|q_i,a_i)}[1-\log(\sigma(f_\phi(q_i,a_i,d)))] \quad (3)$$

As for the generator since the selection of distractors is discrete, a policy gradient is instead used as

$$\nabla_\theta J^G(q_i,a_i) \simeq \frac{1}{K}\sum_{k=1}^{K} \nabla_\theta \log p_\theta(d_k|q_i,a_i)\ \log(1+\exp(f_\phi(d_k,q_i))) \quad (4)$$

A sampling approximation is performed in which d_k is the k-th distractor sampled from $p_\theta(d \mid q_i,a_i)$. The 1 + exp log term acts as the reward for the policy.

Results

After running all algorithms with a 80-20 split of the dataset we achieved the following results.

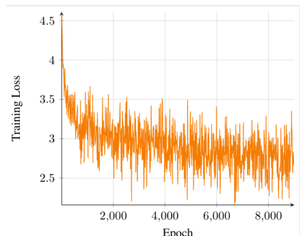
	Baseline Word2Vec	SVM	Neural Ranker	Ranking GAN
Precision@3	0.043	0.314	0.472	0.306

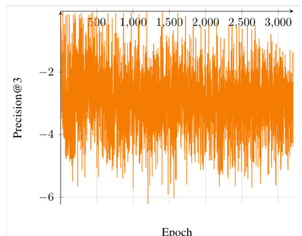
The Neural Dot Product model yielded the best performance with a 47.2% accuracy.

	Question: What type of cartilage contains no collagen? *Answer: lamprey cartilage*
SVM	Distractor: joint cartilage Distractor: vertebrae Distractor: limbs
Neural Ranker	Distractor: joint cartilage Distractor: shark cartilage Distractor: fetal cartilage
GAN	Distractor: joint cartilage Distractor: shark cartilage Distractor: fetal cartilage
	Question: Convex lenses are thicker in the middle than at the edges so they cause rays of light to converge, or meet, at a point called what? *Answer: focus*
SVM	Distractor: the center Distractor: the base Distractor: the apex
Neural Ranker	Distractor: the center Distractor: the base Distractor: the apex
GAN	Distractor: relevance Distractor: 'how' Distractor: debate

Discussion

The reason Neural Dot performs better than the SVM is most likely because the Neural Dot is trained directly to rank. Furthermore, the GAN performed worse compared to the neural dot most likely due to the difficulty of training stability as its loss fluctuates greatly during training.





However, analysis of the qualitative data shows that generally predictions were reasonable. The most common errors were syntactic e.g. Misplaced punctuation. Thus our results show promise for the potential of automatic question generation and should be something continually and increasingly explored as an application of ML for educational purposes in the future.

Future Work

Future work will focus on improving some training properties of GANs for reasons mentioned above. Work on data development will include improving the syntactic form of distractors and even potentially tackling the problem of distractors being actual answers to questions.