Distractor Generation with Generative Adversarial Nets for Automatically Creating Fill-in-the-blank Questions

Chen Liang¹, Xiao Yang², Drew Wham³, Bart Purse³, Rebecca Passonneau², C. Lee Giles¹

¹Information Sciences and Technology  ²Computer Science and Engineering  ³Teaching and Learning with Technology

Background

- Fill-in-the-blank (FITB) question: Branch roots of the primary root of a flowering plant are initiated in the ________.
  (a) cortex  (b) pericycle  (c) epidermis  (d) endodermis

  Stem

  Key
  Distractor

- Distractor generation (DG): generate distractor answers given the question sentence (stem) and the key to the question.

Motivation

- DG is a crucial step for fill-in-the-blank (FITB) question generation.
- Previous DG methods were mostly based on semantic similarities between the key and the candidate distractor:
  - WordNet synonyms [1]
  - Embedding-based similarity [2, 3]
  - Co-occurrence likelihoods [3, 4]
- Existing DG methods have NOT fully explored how to utilize context information (stem) regarding the question.

Our Contribution

- Propose a machine learning-based approach for DG, which is fundamentally different from previous unsupervised similarity-based methods.
- Proposed method only uses stem information and it can be used in combination with existing key-based methods.
- A new dataset collected from college-level biology exams for evaluating DG or FITB generation.

Method – GAN

- Learn distractor distribution conditioned on the stem
- Adapt generative adversarial nets (GANs) [5] to tackle DG
- A generative model G that captures real data (key to the FITB question) distribution given a context (stem).
- A discriminative model D estimates the probability that a sample comes from the real training data rather than G.

Data Preparation

- Utilize the Wikipedia corpus for creating the training set of (stem, key) pairs.
- Substitute the link in a Wikipedia sentence with a blank to get the stem and use the linked Wiki concept as the key.
- Training set: 1.62 million (stem, key) pairs, with a vocabulary of 8879 biology-related concepts
- Test sets:
  - Wiki-FITB: 30 FITB questions based on sentences in Wikipedia, selected by a domain expert
  - Course-FITB: 92 FITB questions from actual exams for two college-level biology courses and biology GRE 2016

Experiment Settings

- For each (stem, key) pair, we apply the proposed DG method to generate a list of distractors.
- Three domain experts were asked collaboratively to label each of the top-4 predictions as a Good, Fair, or Bad distractor.
- GAN: the proposed GAN-based DG model
- W2V: a frequently used similarity-based method, which generates distractors based on the word2vec similarity between the candidate and the key
- GAN+W2V: Use the prediction score and the ranking of GAN and W2V as four features and train a logistic regression classifier to predict the probability of a distractor being good, fair, or bad.

Experiment Results

Table 1: Distractor generation results. Numbers are 95% confidence intervals of percentages of generated distractors being good, fair, or bad, calculated in a leave-one-out manner.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Good</th>
<th>Fair</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN</td>
<td>28.4(10.8)</td>
<td>10.8(5.5)</td>
<td>60.8(10.5)</td>
</tr>
<tr>
<td>W2V</td>
<td>35.8(6.8)</td>
<td>10.0(5.6)</td>
<td>54.2(8.0)</td>
</tr>
<tr>
<td>GAN + W2V</td>
<td>40.0(7.6)</td>
<td>11.7(5.6)</td>
<td>48.3(8.6)</td>
</tr>
</tbody>
</table>

(a) Wiki-FITB

<table>
<thead>
<tr>
<th>Methods</th>
<th>Good</th>
<th>Fair</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN</td>
<td>17.7(5.9)</td>
<td>9.2(3.5)</td>
<td>73.1(5.9)</td>
</tr>
<tr>
<td>W2V</td>
<td>32.9(5.3)</td>
<td>11.9(3.5)</td>
<td>55.2(5.7)</td>
</tr>
<tr>
<td>GAN + W2V</td>
<td>34.3(5.7)</td>
<td>14.1(3.9)</td>
<td>51.6(6.0)</td>
</tr>
</tbody>
</table>

(b) Course-FITB

Table 2: Distractor generation examples for question “Changes in gene frequency over time describes the process of ________,” whose key is Evolution. (Legend: Good, Fair, Bad)

Challenge

- GAN requires that the composition of the generator and the discriminator are fully differentiable.
- Sampling a discrete token is not differentiable.
- Use Gumbel softmax trick (Jang et al., ICLR’17) to enable back-propagating training signals from the discriminator to the generator.

References


Acknowledgements

We gratefully acknowledge partial support from the Pennsylvania State University Center for Online Innovation in Learning.