View Reviews

Paper ID
4524

Paper Title
Knowledge-Driven Distractor Generation for Cloze-Style Multiple Choice Questions

Track Name
AAAI2021

Reviewer #1

Questions

1. {Summary} Please summarize the main claims/contributions of the paper in your own words. (Do not provide any review in this box)
The paper proposes a new framework for generating distractors for multiple-choice cloze questions and instantiates several new models from its framework.

They compile a “new” dataset that is a concatenation of several existing MCQ datasets, which they use for training and evaluation. Experimental results show that their new models outperform some baselines on human and automatic evaluation.

2. {Novelty} How novel is the paper?
Paper contributes some new ideas

3. {Soundness} Is the paper technically sound?
I have not checked all details, but the paper appears to be technically sound

4. {Impact} How important is the paper likely to be, considering both methodological contributions and impact on application areas?
The paper will have low overall impact

5. {Clarity} Is the paper well-organized and clearly written?
Fair: paper is somewhat clear, but important details are missing or confusing, which hurts readability

6. {Evaluation} Are claims well supported by experimental results?
Moderate: Experimental results are weak: important baselines are missing, or improvements are not significant

7. {Resources} How impactful will this work be via sharing datasets, code and/or other resources? (It may help to consult the paper’s reproducibility checklist.)
Fair: some may find shared resources useful in future work

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper’s reproducibility checklist.)
Good: e.g., code/data available, but some details of experimental settings are missing/unclear

9. {Reasons to Accept} Please describe the paper’s key strengths.

The paper provides a general framework for distractor generation in the multiple choice cloze question-answering setting.

The paper compares their methods against several baselines, demonstrating some slight improvements in automatic and human evaluation.

10. {Reasons to Reject} Please describe the paper’s key weaknesses.
It’s not entirely clear to me that this framework is _that_ novel…the approach of generating distractors and ranking them has been used by several past works. The authors argue that past work falls short, “Since identifying the concrete domain of each question and preparing large-scale domain-specific vocabulary require

substantial human labor, such corpus-based methods cannot be easily applied in real-world scenarios.”.

It’s not entirely clear to me that such labor is substantial, and it seems likely that any “real-world scenarios” this task has would not be completely open-domain; the open-domain assumption seems somewhat artificial unto itself.

Also, the authors claim that their dataset is “new”, which is a bit misleading---the new dataset is just a concatenation of existing datasets used in past work.

11. {Detailed Comments} Please provide other detailed comments and constructive feedback.

I found the paper to be quite dense, especially with respect to the experimental results. Some more prose about what the multitude of numbers in Table 3 would have been appreciated---it’s difficult to draw conclusions from them as-is, especially since the gaps are so small.

When citing BERT as a method, don’t cite HuggingFace transformers---cite the Devlin paper, and then note that it was implemented in Transformers (Wolf et al)

12. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)

1. The details about the human evaluation are very sparse--how many annotators did you use, how many samples were rated, etc? From just looking at the information in the paper, it’s unclear whether the human evaluation has enough statistical power for me to trust it.

14. (OVERALL SCORE)

5 - Below threshold of acceptance

Reviewer #3

Questions

1. {Summary} Please summarize the main claims/contributions of the paper in your own words. (Do not provide any review in this box)

This paper introduces a new method to create and select distractors for cloze-style multiple-choice questions. The method consists of two parts.
First component is candidate set generator which selects a small set of candidate distractors with knowledge base. The second component is a learning-to-rank model to rank the selected candidates.

The experiments show that the method outperforms baseline methods with respect to the plausibility (99.13->99.33) and reliability (1.26->1.34 in 0-2 scale).

2. {Novelty} How novel is the paper?

Paper contributes some new ideas

3. {Soundness} Is the paper technically sound?

I have not checked all details, but the paper appears to be technically sound

4. {Impact} How important is the paper likely to be, considering both methodological contributions and impact on application areas?

The paper will have low overall impact

5. {Clarity} Is the paper well-organized and clearly written?

Good: paper is well organized but language can be improved

6. {Evaluation} Are claims well supported by experimental results?

Moderate: Experimental results are weak: important baselines are missing, or improvements are not significant

7. {Resources} How impactful will this work be via sharing datasets, code and/or other resources? (It may help to consult the paper’s reproducibility checklist.)

Fair: some may find shared resources useful in future work

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper's reproducibility checklist.)
Good: e.g., code/data available, but some details of experimental settings are missing/unclear

9. {Reasons to Accept} Please describe the paper's key strengths.
- The proposed method is unique and interesting.
- The experiments are conducted appropriately.
- The paper is clearly written overall.

10. {Reasons to Reject} Please describe the paper's key weaknesses.
- The amount of improvement does not look impactful.

11. {Detailed Comments} Please provide other detailed comments and constructive feedback.
Thank you for the interesting work. I enjoy reading it.
Here are my comments and questions.

1. line 134. ""These survivors managed to swim to the bank," where bank is the key, we would like to generate candidates like shore":

I was confused here because "shore" can also fit as a correct answer rather than the distractor.

2. Web-search score (line 187~191) :

the algorithm is not clear to me. it will be more helpful if the algorithm is written more formally (or with a figure).

3. human evaluations (line 256~272)

It will be important to report the inter annotator agreement for each metric (plausibility and reliability)

4. Table 6.

I would like to see more examples in other domains too, because the paper mentions that the proposed method works in the open-domain (in the abstract).

5. Typo in table 3:

+DS(lise-wise) --> DS(list-wise)

12. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
see above

14. (OVERALL SCORE)
7 - Accept

Reviewer #4

Not Submitted

Reviewer #5

Questions

1. {Summary} Please summarize the main claims/contributions of the paper in your own words. (Do not provide any review in this box)
This paper addresses the important problem of distractor generation for close-style MCQs. With digital spread of pedagogy increasing, research towards making assessment stronger will be of great value. The paper proposes a technique based on learning-to-rank and showed improvement on a dataset which contains existing as well as new set of questions.

2. {Novelty} How novel is the paper?
Paper makes non-trivial advances over past work

3. {Soundness} Is the paper technically sound?
I have not checked all details, but the paper appears to be technically sound

4. {Impact} How important is the paper likely to be, considering both methodological contributions and impact on application areas?
The paper will impact a moderate number of researchers

5. {Clarity} Is the paper well-organized and clearly written?
Excellent: paper is well organized and clearly written

6. {Evaluation} Are claims well supported by experimental results?
Moderate: Experimental results are weak: important baselines are missing, or improvements are not significant

7. {Resources} How impactful will this work be via sharing datasets, code and/or other resources? (It may help to consult the paper's reproducibility checklist.)
Fair: some may find shared resources useful in future work

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper's reproducibility checklist.)
Good: e.g., code/data available, but some details of experimental settings are missing/unclear

9. {Reasons to Accept} Please describe the paper's key strengths.
Very well motivated and illustrated in the Introduction.
A much larger and heterogenous dataset which could help progressing this line of research.
Novel use of LDA for candidate generation

10. {Reasons to Reject} Please describe the paper's key weaknesses.
Feature based learning to rank has already been used for distractor selection using similar unsupervised/similarity based features (Liang et al. 2018). The contribution there is incremental.
The shortcoming of prior techniques requiring domain knowledge is not well supported. For example, it is not clear why Liang et al. 2018 is domain dependent. Additionally, an MCQ does not have each question from different domain ["identifying the concrete domain of each question"]. Authors have gone overboard with this claim without substantial evidence.
A new dataset is surely welcome but a large fraction of it is a collation of existing datasets.
As the experiment is conducted on the collated dataset, comparison with prior published results on component datasets is unavailable. It is strongly recommended that such comparison is shown e.g. comparison on SciQ and MCQL separately with LR+LM & LR+RF.
Missing Reference:
Automatic distractor generation for multiple-choice English vocabulary questions
Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa & Hiroyuki Obari

11. {Detailed Comments} Please provide other detailed comments and constructive feedback.
The paper claims it to be not domain specific (i.e. general). It appears that the primary source of this strength is a general purpose KB which is able to generate candidates from different domains. Use of such external knowledge base is only an incremental contribution.
A few qualitative examples in experimentation towards supporting the general nature of the technique would have been good.

The human evaluation should be done in a more realistic control-test manner. For example, 50 MCQs are given

to each evaluator of which half of the questions have distractors generated by system-1 (or ground truth) and the other half system generated distractors along with the key. If there is a statistically significant difference the performance of the evaluator between these two sets - then we can conclude one is better than the other.

“to evaluate the proficiency of language learners” - how is it helping assessing language learning skills?

Line 81: “Previous DG methods” - non standard acronym. Should have been used with expansion first.

12. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
None

14. (OVERALL SCORE)
5 - Below threshold of acceptance