View Reviews

Paper ID
6729

Paper Title
Joint Training of Classification and Similarity Models for Intention Detection in Task-specific Chatbot

Track Name
AAAI2021

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)
In this paper, the authors provide an algorithm for intent detection for task-oriented chatbot systems. The proposed algorithm has two subcomponents i.e. a classification model and a similarity model. These subcomponents are trained using multi-task learning. The classification model is used to classify intents while the similarity model is used to help with distinguishing between positive and negative intent classes. The experiments in the paper show that the proposed model outperforms the baseline significantly, especially in the low resource settings.

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.)
Not applicable: no shared resources

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper's reproducibility checklist.)
Meets Minimum Standard: e.g., code/data unavailable, but paper is clear enough that an expert could confidently reproduce

9. {Reasons to Accept} Please describe the paper's key strengths.
- The proposed approach is novel and well described in the paper.
- Authors conducted an extensive set of experiments and ablations to show effectiveness of the approach.

11. {Reasons to Reject} Please describe the paper's key weaknesses.
- The baselines used for the comparison could be stronger, for example, it could be based on the previous works of joint-training of NER and intent detection using BERT type of model.
- It is not clear from the description if the pre-trained models used in baseline were fine-tuned to the data in hand, which is important to know to understand the results.
- There is not much insight shared in the paper about what exactly (examples or analysis) was improved by introducing the similarity model.

12. {Detailed Comments} Please provide other detailed comments and constructive feedback.
- Baseline used for comparisons could be from the previous work specifically done for the SLU tasks. For example, there are papers which use BERT type of model and train NER and intent prediction model jointly has shown to perform much better.
- Some analysis of what/where the proposed idea of using similarity model is helping should have been useful.

13. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
- Was the pre-trained model used in baseline were fine-tuned using the dataset and task? Could you provide more details?

15. (OVERALL SCORE)
6 - Above threshold of acceptance

Reviewer #4

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 addresses the problem of training an sample-efficient short-text classification model. It proposes a new model that is composed of a classification model and a similarity model.

2. {Novelty} How novel is the paper?
Main ideas of the paper are known or incremental 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 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?
Good: Experimental results are sufficient, though more analysis would significantly add support to the claims

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.)
Not applicable: no shared resources

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the

paper's reproducibility checklist.)
Meets Minimum Standard: e.g., code/data unavailable, but paper is clear enough that an expert could confidently reproduce

9. {Reasons to Accept} Please describe the paper's key strengths.
The paper combines two traditional methods in one model, and it uses detailed abalation experiments to show the improvements of each part.

11. {Reasons to Reject} Please describe the paper's key weaknesses.
The motivation of this paper is not clear. The paper says in line 96 on page 2 "classification model cannot use data augmentation method" and "similarity model doesn't perform well on a new task-specific dataset". I cannot agree with these conclusions, and it is difficult to understand why putting a classification model and a similarity model together can solve these problems.
The authors seem to be simply putting a classification model and a matching model together, and show this is better than a single classificaiton model and a matching model. The matching models and the classification models used in this paper are all well-known models, and the authors haven't made any improvement. The combined solution is complicated and the conclusion is too straight forward, and I don't think the paper is telling us something new.

12. {Detailed Comments} Please provide other detailed comments and constructive feedback.
Please improve the motivation part. The paper says in line 96 on page 2 "classification model cannot use data augmentation method" and "similarity model doesn't perform well on a new task-specific dataset", and it is difficult to understand why putting a classification model and a similarity model together can solve these problems.

The author can summarize an insight that can lead to the performance improvement. An insight is something general, important, and can can be applied in other models/domains. For example, careful parameter tuning can lead to performance improvement, but there is no insight and thus have little academic value.

13. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
1.The paper says in line 96 on page 2 "classification model cannot use data augmentation method" and "similarity model doesn't perform well on a new task-specific dataset". Why putting a classification model and a similarity model together can solve these problems?
2.How to solve the error propogation problem? If the stage one model has made a wrong top-k prediction, however good the stage two model is, the result will always be wrong. Please analyse the error cases and categorize them.

15. (OVERALL SCORE)
4 - Reject

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 proposes a novel model, called similarity model fused with classification model (SFC), which combines a classification model and a similarity model, and also borrows the power of multi-task training as well. Extensive experiments were conducted on 4 public and 1 private datasets, and show that the proposed model outperforms very strong baselines (i.e., BERT, RoBERTa and ALBERT based pretrained models) by over 2 percentages in average F1

score.

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?
Good: paper is well organized but language can be improved

6. {Evaluation} Are claims well supported by experimental results?
Good: Experimental results are sufficient, though more analysis would significantly add support to the claims

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.)
Not applicable: no shared resources

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.
Overall, this paper is well organized and clearly written.
This paper proposes a novel model, called similarity model fused with classification model (SFC), which combines a classification model and a similarity model, and also borrows the power of multi-task training as well.
Extensive experiments on several datasets demonstrate the effectiveness of the proposed SFC model and its advantages over state-of-the-art strong baselines.

11. {Reasons to Reject} Please describe the paper's key weaknesses.
None

12. {Detailed Comments} Please provide other detailed comments and constructive feedback.
(1) Some details of experimental settings are missing/unclear. For instance, before starting to fine-tune the sentence-pair model on task-specific dataset, the authors first fine-tune RoBERTa on Quora dataset (Iyer, Dandekar, and Csernai 2017), which contains 404,290 potential duplicate question pairs, for transfer learn-ing. However, in the following experiments, I haven't find experiments on the impact of this fine-tuning process, and all the baseline didn't invovled in this fine-tuning process. Maybe this is unfair.

13. {QUESTIONS FOR THE AUTHORS} Please provide questions for authors to address during the author feedback period. (Please number them)
(1) Can you explain the impact of the fine-tuning process on Quora dataset?
(2) Will you release you code if your paper were accepted?

15. (OVERALL SCORE)
7 - Accept

Reviewer #6

Questions

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

In this paper, the author proposes a method for intent detection for task-specific chatbot. It makes use of joint training of a classification and a similarity model. In the first stage, a classification model is employed based on RoBERTa to produce top K candidates. Then, in the second stage, sentence-pair similarity model is formulated as multi-task learning to pick the the class label with highest similarity.

2. {Novelty} How novel is the paper?

Main ideas of the paper are known or incremental 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 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.)

Not applicable: no shared resources

8. (Reproducibility) Would the experiments in the paper be easy to reproduce? (It may help to consult the paper's reproducibility checklist.)

Meets Minimum Standard: e.g., code/data unavailable, but paper is clear enough that an expert could confidently reproduce

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

1. The model shows that it can substantially help with limited amount of data.
2. The method proves to improve over strong baseline.

11. {Reasons to Reject} Please describe the paper's key weaknesses.

1. The method seems time-consuming, especially the second stage. It is not sure that it scales well with large data.

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

In this paper, the author proposes a method for intent detection for task-specific chatbot. It makes use of joint training of a classification and a similarity model. In the first stage, a classification model is employed based on RoBERTa to produce top K candidates. Then, in the second stage, sentence-pair similarity model is formulated as multi-task learning to pick the the class label with highest similarity.
The method is novel. Experimental results show that it can substantially help with limited amount of data. Even with full amount of data, it still improves over the baseline.

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

1. 1. The method seems time-consuming, especially the second stage. Can you provide some experimental figures wrt. training/ response time?

15. (OVERALL SCORE)

6 - Above threshold of acceptance