GMR-165

An Empirical Study of Prompt-based Non-functional Requirements Classification

Allen Kim

Abstract

In modern software development, Non-Functional Requirements (NFR) are essential to satisfy users' needs. Distinguishing different categories of NFR is tedious, error-prone, and time consuming due to the complexity of software systems. In our project, we conducted a comprehensive study to evaluate the performance of prompt-based NFR classification by designing various handcraft templates and soft templates on the pre-trained language model (i.e., BERT). Our experimental results show that handcraft templates can achieve best effectiveness (e.g., 83.52% in terms of F1 score) but with unstable performance for different templates.

Introduction

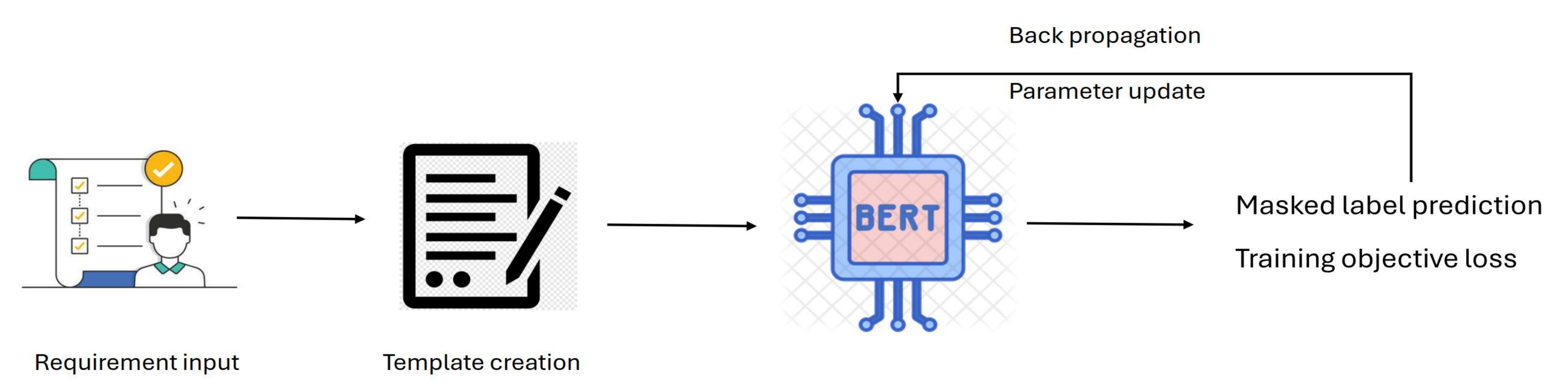
In modern software development, Non-Functional Requirements (NFR) are essential to satisfy users' needs, which define various constraints and qualities that the system must adhere (e.g., quality, usability, security). However, developers always overlook the importance of NFR since they tend to be across various requirement specification documents, making it difficult to locate and consolidate them effectively. Thus, the task of NFR classification is crucial for the whole software development process. Recently, pre-trained foundation models (e.g., BERT, GPT) have been widely used in various AI fields such as natural language processing (NLP). In other way, current survey paper on prompt engineering demonstrates that different prompts can affect the performance of the pre-trained model so that it is necessary to evaluate the impact of different prompt templates on NFR classification. Furthermore, one study also indicates that a learnable tensor can be concatenated with the input embeddings to become a series of soft templates for natural language understanding. In this project, we conduct a comprehensive study by designing various prompt templates (including handcraft templates and soft templates) for NFR classification based on pre-trained BERT model.

Research Question(s)

(1) How do handcraft templates affect the performance of NFR classification?
(2) How do learnable soft templates affect the performance of NFR classification?

Materials and Methods

The overall structure of our project is as follows. Based on the original requirement text, we design various templates (including a masked target label) that can be as the input of pre-trained models (We use pre-trained BERT mode). During the training process, the pre-trained model can predict the masked target label and the training loss is calculated for back propagation to finetune the pre-trained model by updating the parameters. We use cross-entropy loss function in our study since NFR classification is the classic multi-class classification problem.



We design following handcraft templates (P1-P4) and learnable soft templates (P5-P10).

P1: [CLS] Only authorized personnel can access customer records in the database. [SEP] This requirement is related to [M]. [SEP]

P2: [CLS] Following text is [M] requirement. [SEP] Only authorized personnel can access customer records in the database.[SEP]

P3: [CLS] "Only authorized personnel can access customer records in the database. " is a requirement related to [M]. [SEP]

P4: [CLS] Given the following statement: "Only authorized personnel can access customer records in the database. "[SEP] Question: what type of requirement is it? [SEP] Answer: [M]

P5: [CLS] Only authorized personnel can access customer records in the database. [SEP] [P] [P] [M]. [SEP]

P6: [CLS] [P] [P] [M]. [SEP] Only authorized personnel can access customer records in the database.[SEP]

P7: [CLS] Only authorized personnel can access customer records in the database. [SEP] [P] [P] [P] [M]. [SEP]

P8: [CLS] [P] [P] [P] [M]. [SEP] Only authorized personnel can access customer records in the database.[SEP]

P9: [CLS] Only authorized personnel can access customer records in the database. [SEP] [P] [P] [P] [P] [M]. [SEP]

P10: [CLS] [P] [P] [P] [P] [M]. [SEP] Only authorized personnel can access customer records in the database.[SEP]

[CLS] in the template represents a special token in BERT model in the front of the original input text and [SEP] is a separator token to represent the segment of each sentence. [M] is the masked token to represent the requirement category (e.g., performance, security, usability) that can be predicted by BERT model. [P] represents the learnable token that replaces the concrete templates.

In our project, we use the widely used pre-labeled dataset PROMISE with 914 nonfunctional requirements consisting of the following five categories: maintainability, operability, performance, security, and usability.

Results

The shows the results of NFR classification based on the 4 handcraft templates and 6 learnable soft templates in terms of the evaluation metrics precision, recall and F1 score. Please note that all results are calculated as the average values of 10-fold cross-validation based on each template. From the results, we can find the overall performance of learnable soft templates are worse than handcraft templates for all metrics. For example, in terms of F1 score, the best result of learnable templates is 78.79% while the best result of handcraft templates is 83.52%. The possible reason is that there are no meaningful context for the special tokens [P] in the learnable soft templates so that it is not easy to predict the target label accurately. Also, even the handcraft template can achieve better results, the standard deviation of the four templates (1.00) is larger than learnable templates (0.84), showing unstable results for random handcraft templates.

Template	Precision	Recall	F1 score
P1	**83.59%**	**83.46%**	**83.52%**
P2	82.37%	82.50%	82.43%
P3	81.27%	81.97%	81.61%
P4	80.35%	81.27%	80.81%
P5	77.26%	76.64%	76.95%
P6	78.43%	79.17%	78.79%
P7	76.40%	78.35%	77.36%
P8	78.38%	78.12%	78.25%
P9	76.54%	76.53%	76.53%
P10	78.51%	77.60%	78.05%

Conclusions

In this project, we conducted a comprehensive study to evaluate the performance of prompt-based non-functional requirements classification by designing various handcraft templates and soft templates on pre-trained model. Our experimental results show that handcraft templates can achieve best effectiveness (e.g., 83.52% in terms of F1 score) but with unstable performance for different templates.

Contact Information

Allen Kim akim72@students.kennesaw.edu

References

Devlin, Jacob. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Schick, Timo and Sch¨utze, Hinrich, Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676. 2020

Liu, Xiao and Zheng, Yanan and Du, Zhengxiao and Ding, Ming and Qian, Yujie and Yang, Zhilin and Tang, Jie. GPT understands, too. AI Open, 2023

Schick, Timo and Sch¨utze, Hinrich, Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676. 2020

Sayyad, SJ. PROMISE software engineering repository. 2005

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING