Developing a BERT based triple classification model using knowledge graph embedding for question answering system

Phuc Do[1] · Truong H. V. Phan[1,2]

Abstract

The current BERT-based question answering systems use a question and a contextual text to find the answer. This causes the systems to return wrong answers or nothing if the text contains irrelevant contents with the input question. Besides, the systems haven't answered yes-no and aggregate questions yet. Besides that, the systems only concentrate on the contents of text regardless of the relationship between entities in the corpus. This systems cannot validate the answer. In this paper, we presented a solution to solve these issues by using the BERT model and the knowledge graph to enhance a question answering system. We combined content-based and linked-based information for knowledge graph representation learning and classified triples into one of three classes such as base class, derived class, or non-existent class. We then used the BERT model to build two classifiers: BERT-based text classification for content information and BERT-based triple classification for link information. The former was able to make a contextual embedding vector for representing triples that were used to classify into the three above classes. The latter generated all path instances from all meta paths of a large heterogeneous information network by running the Motif Search method of Apache Spark on a distributed environment. After creating the path instances, we produced triples from these path instances. We made content-based information by converting triples into natural language text with labels and considered them as a text classification problem. Our proposed solution outperformed other embedding methods with an average accuracy of 92.34% on benchmark datasets and the Motif Finding algorithm with an average executive time improvement of 37% on the distributed environment.

Keywords BERT based triple classification model · Knowledge graph embedding · Meta-path · Motif finding

1 Introduction

Using the knowledge graph (KG) for enhancing the question answering system is a promising study in recent years and plays an important role in natural language Q&A systems [9]. Plus, the emergence of a pre-trained language model as BERT brings significant improvements in QA systems [12]. However, the KG based QA systems only focus on entities as head, predicate, tail, and links without content. Meanwhile, BERT-based QA systems use the content of the contextual text to find answers without relationships of entities. So, we combined both BERT and knowledge graph to take advantage of their pros. The KG-BERT model that is proposed by Liang Yao also uses BERT and knowledge graph to classify triples [15] for knowledge graph completion. In our model, we use KG-BERT and knowledge graph to improve the question answering system. The aggregate question can be answered by collecting all triples of knowledge graph satisfying the question. Our method generated triples by obtaining all meta-paths from a given HIN and then used Motif Search of Apache Spark to discover all path instances of the meta-paths in distributed environment because of the large knowledge graph. Afterward, our model generated all triples from these path instances. These triples can be based triple for meta-path of length 1 or derived triple for meta-path of length more than 1, then we convert the triple <h, p, r> into text by using the description of head, tail and predicate of triple. The text description of triple is fed into BERT as text classifier. Our model used a softmax classifier to compute the score function of triples. Our model classified triples into three classes such

✉ Phuc Do
    phucdo@uit.edu.vn

    Truong H. V. Phan
    truongphv.ncs@grad.uit.edu.vn; truong.phv@vlu.edu.vn

[1] University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam

[2] Van Lang University, Ho Chi Minh City, Vietnam