Learning-based Translation Performance Prediction

Shujun Wang, Jie Jiao, Mingyu Yang, Xiaowang Zhang*, and Zhiyong Feng

College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China
* Corresponding author:{xiaowangzhang}@tju.edu.cn

Abstract. RDF question/answering (Q/A) can translate questions into SPARQL queries by employing question translation.One of the challenges of RDF Q/A is predicting the performance of questions before they are translated. Performance characteristics, such as the translation time, can help data consumers identify unexpected long-running questions before they start and estimate the system workload for scheduling. In this paper, we adopt machine learning techniques to predict the performance of question translation in RDF Q/A.Our work focuses on modeling features of a question to a vector representation. Our feature modeling method does not depend on the knowledge of underlying systems and the structure of the underlying data, but only on the nature of questions. Then we use these features to train prediction models.Finally, based on this model, we designed a single parallel-batching RDF Q/A application.Evaluations are performed on real-world questions, whose translation time ranges from milliseconds to minutes. The results demonstrate that our approach can effectively predict question translation performance.

Keywords: RDF · Question Answering · Performance Prediction

1 Introduction

RDF Q/A allows users to ask questions in natural languages over a knowledge base represented by RDF. Hence, it has received extensive attention in both natural language processing and database areas. The core task of RDF Q/A is to translate natural language questions into SPARQLs. Prediction of question translation can benefit many system management decisions. The challenge in our work centers on capturing characteristics of questions and representing the characteristics as features for the application of machine learning techniques.

The main contributions of this work are summarized as follows:

– We adopt machine learning techniques to predict the question performance before their execution effectively.

– We propose four ways to model features of a question. The lexical features, part of speech features, and dependency relation features can be acquired from the question's dependency tree. The hybrid feature can be derived from part of speech features and dependency features. All features can be easily obtained without the information provided by the underlying systems.
– The RDF Q/A system we used is one of the most used systems in the community of Semantic Web. Thus our work will benefit a large population of users.

With the decline of computer hardware costs, the parallelism of computers increases gradually. Based on the prediction algorithm proposed above, we designed a single-machine high-parallel RDF Q / A application to implement the specific query transformation process.

2 Feature Modeling

We formulate the problem as follows: Let $N = (W, P, T)$ denote a question, where W is a set of words, which contained in N, P is a set of posTags and T is a dependency tree of N. Feature modeling is the transformer that maps $N \to \mathcal{N}$, where $\mathcal{N} \in R^m$ and m is the number of features.

2.1 Lexical Features

We first focus on each word's characteristics in the question, such as their lengths, and the number of special words. More specifically,

– *Word Length*: the number of words whose length belongs to $[1, 15]$, and the number of words whose length is ≥ 16.
– *Special Words*: We detect the number of three kinds of special words{all upper-case, contains a hyphen and stop word}

Most importantly, we use information entropy $I(N)$ to measure the uncertainty of a question.

$$I(N) = -\sum_{i=1}^{n} p(w_i) \log_2 p(w_i) \tag{1}$$

$w_i \in N$, $P(w_i)$ refers to the probability of w_i appearing in the corpus.

2.2 Part of Speech Features

In the process of translating natural language questions into SPARQL queries in the RDF Q/A system, the part of speech of a word can determine whether the word participates in the construction of the SPARQL query graph. For example, nouns, verbs, and adjectives in questions are important components of the SPARQL query graph. Therefore, in our work, we apply Standford pos tagger to obtain the part of speech of each word contained in N. We collect the number of different parts of speech as part of speech features of a given question. Besides, we further insert the number of words at the beginning of the vector.

2.3 Dependency Relation Features

The above two kinds of features mainly express the characteristics of the words in the questions. In this subsection, we emphasize the relationships between different words.

In our work, we collect the number of different dependencies as dependency relation features. Note that we further insert the height of the dependency tree at the beginning of the vector.

2.4 Hybrid Features

We build hybrid features by selecting the most predictive features based on the part of speech features and dependency relation features.

Definition 1 (Triple). *Let $T = \langle p_i, d, p_j \rangle$, where p_i and p_j are part of speech features, and d is a dependency relation feature between p_i and p_j. For example, there is a triple $\langle WP, nsubj, VBD \rangle$ in Figure 1. T describes the structural characteristics of questions.*

We use T as our hybrid features, which represent the structural characteristics of questions. A synthetic feature vector example is shown below.

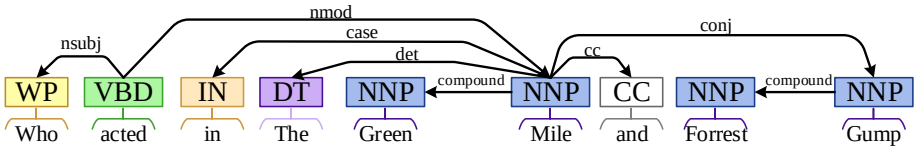


Fig. 1. A Dependency Tree of the Question

Lexcial Features								Part of Speech Features						Dependency Relation Features					
I(N)	1	2	3	4	...	15	SW	Num	WP	VBD	NNP	JJ	...	Height	nsubj	nmod	det	conj	...
0.64	0	1	3	2	...	0	0	9	1	1	4	0	...	3	1	1	1	1	...

3 Prediction Model

Support vector regression(SVR) is to find the best regression function by selecting the particular hyperplane that maximizes the margin. The problem is formulated as an optimization problem:

$$min\ \mathbf{w}^T\mathbf{w}, \quad s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi, \xi \geq 0 \tag{2}$$

An advantage of SVR is its insensitivity to outliers.

4 Parallel RDF Q/A

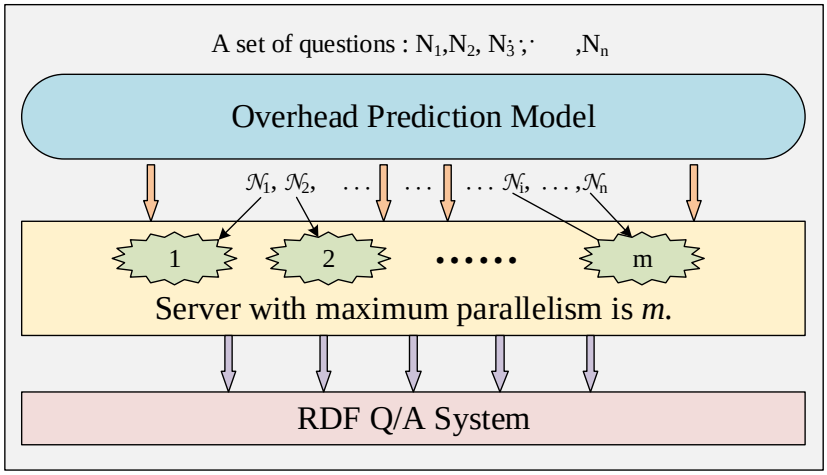


Fig. 2. Framework of Parallel-Batching

Our system's task is to predict the overhead of translating N questions into N SPARQL queries, and then divide the overhead of N questions into M processors, in order to achieve this goal, we design the algorithm 1 to minimize the loss function in Formula 3.

$$Loss = min(max(M_1, M_2, \ldots, M_n)) \tag{3}$$

where M_i is the sum overhead of all questions in the i-th processor.

5 Experiments

We use QALD to verify the effectiveness of our parallel-batching RDF Q/A system. Four experiments with a parallelism of 2,4,6,8 are shown in the following four figures. Each experiment includes ten groups (10 questions in each group) of question translation tests.

In each experiment, we compare our approach with the other three methods in performance, i.e., divide ten questions into M processors according to the number of questions, the number of words, and serial execution.

Experiments show that our prediction model is accurate, and our parallel RDF Q/A system can achieve a single server high parallel question translation.

http://qald.aksw.org/

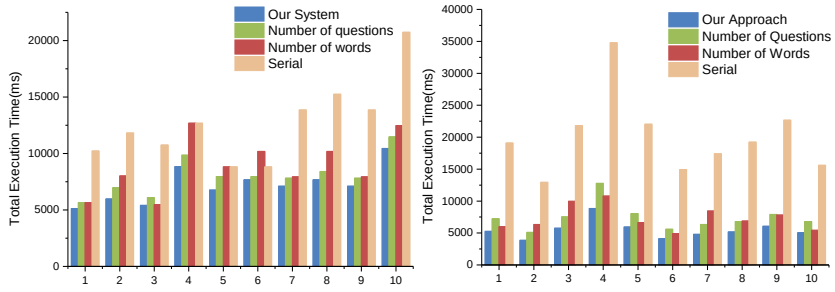
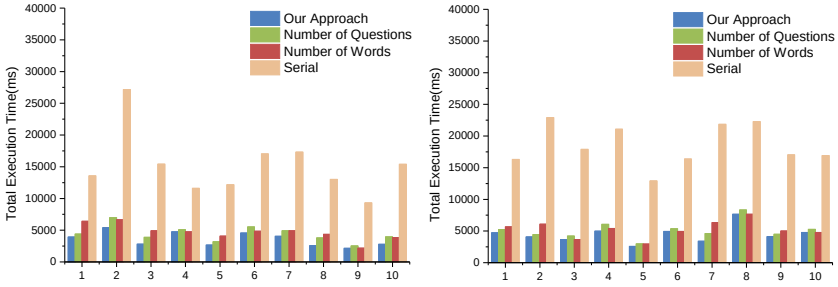


Fig. 3. Efficiency Evaluation



Acknowledgments

This work is supported by the National Key Research and Development Program of China (2017YFC0908401) and the National Natural Science Foundation of China (61972455,61672377). Xiaowang Zhang is supported by the Peiyang Young Scholars in Tianjin University (2019XRX-0032).

References

1. Zhang W., Sheng Q., Qin Y., et al: Learning-based SPARQL query performance modeling and prediction. In *Proc. of WWW 2018*, pp.1015-1035.
2. Chifu A., Laporte L., Mothe J., et al: Query Performance Prediction Focused on Summarized Letor Features. In *Proc. of SIGIR 2018*, pp.1177–1180.
3. Zou L., Huang R., Wang H., et al: Natural language question answering over RDF: a graph data driven approach. In *Proc. of SIGMOD 2014*, pp.313–324.
4. Hu S., Zou L., Yu J., et al: Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs. In *Proc. of ICDE 2018*, pp.1815-1816.
5. Jiao J., Wang S., Zhang X., et al: Multi-Query Optimization in RDF Q/A System. In *Proc. of ISWC 2019*, pp.77–80.