

Req2Lib: A Semantic Neural Model for Software Library Recommendation

Zhensu Sun, Yan Liu[§], Ziming Cheng, Chen Yang, Pengyu Che
School of Software Engineering
Tongji University
Shanghai, China
{87su, yanliu.sse, 1750114, 1610833, chepengyu}@tongji.edu.cn

Abstract—Third-party libraries are crucial to the development of software projects. To get suitable libraries, developers need to search through millions of libraries by filtering, evaluating, and comparing. The vast number of libraries places a barrier for programmers to locate appropriate ones. To help developers, researchers have proposed automated approaches to recommend libraries based on library usage pattern. However, these prior studies can not sufficiently match user requirements and suffer from cold-start problem. In this work, we would like to make recommendations based on requirement descriptions to avoid these problems. To this end, we propose a novel neural approach called Req2Lib which recommends libraries given descriptions of the project requirement. We use a Sequence-to-Sequence model to learn the library linked-usage information and semantic information of requirement descriptions in natural language. Besides, we apply a domain-specific pre-trained word2vec model for word embedding, which is trained over textual corpus from Stack Overflow posts. In the experiment, we train and evaluate the model with data from 5,625 java projects. Our preliminary evaluation demonstrates that Req2Lib can recommend libraries accurately.

Index Terms—Library Recommendation, Deep Learning, GitHub

I. INTRODUCTION

Today, third-party libraries play an important role in modern software projects. With them, developers can improve the quality of software project and do not need to "re-implement the wheel". However, there are over 4.4m unique open-source packages on Libraries.io [1] who collects data from package managers and source code repositories. The availability of such a huge amount of reusable libraries facilitates software development and evolution. It also places a barrier for developers to locate useful ones.

Usually, developers search on the Internet based on traditional keyword matching to retrieve libraries. The names of suitable library may be shown in the websites like blogs or forums whose content are not well-organized. Thus, developers have to manually examine many web pages to pick the most appropriate libraries during the process of retrieval. On the one hand, it requires a certain level of experience of related domain. On the other hand, the retrieval consumes considerable energy and time of developers [2]. Thus, it is

difficult and time-consuming to find the appropriate ones without additional guides or experiences.

To save developers from filtering among search results, researchers have proposed models that can recommend relevant libraries for software projects using project content (e.g., libraries, source codes) [3] [4] [5] [6]. Although their models can make recommendations correctly in their working contexts, they have a set of essential defects that do harm to its practicability:

- **Requirement insufficiently matching:** Previous studies are based on library usage pattern [3] [4] [5] or library similarity [6] to make recommendations. All of them are content-based recommendation systems. Their results have no relation with requirements. In a supermarket, customers may pay for high-quality products recommended by sellers even if the products are not on their shopping lists. However, in the domain of software engineering, projects have clear requirements, and libraries that are irrelevant to the requirement will not be "paid". Therefore, recommendations without requirements may not help the development of software projects.
- **An awkward situation caused by cold-start problem:** Previous study suffers from the cold-start problem: it fails when no project content are available. The content of a project need to accumulate for some time before it is enough for an effective recommendation. Thus, their model mainly works during software maintenance and evolution. However, research indicates that projects do not appear to change their dependencies over time [7]. Because of the fear of system errors, developers are not willing to change libraries ("if it ain't broke, don't fix it"). In other words, it is not especially necessary to recommend libraries for projects that are already implemented.

Considering the above issues, a potential way is to recommend relevant libraries directly from requirements. In the process of software development, project requirements are often identified before technology selection. In this way, developers can get recommendations while technology selection. On the one hand, technology selection is the phase that suggestions on libraries are urgently needed. On the other hand, using requirement descriptions, developers do not need to accumulate enough project content. Thus, by directly utilizing requirement

[§]Yan Liu is corresponding author.

This work is supported by Intelligent City Knowledge Service System (<http://icity.ikcest.org/>).

information, recommendations can match the requirements of projects without cold-start problem.

To this end, we conduct an exploratory attempt to identify suitable libraries from a description of requirement. Recommendations based on natural language (NL) is beyond the capability of traditional recommendation algorithms like collaborative filtering. To exploit NL, previous researchers treat sentences as bags-of-words (BOW) and try to match the textual similarity [4] [8]. BOW lack an deep understanding of the semantics of NL. A BOW model can not tell the difference between "before long" and "long before".

In this work, we propose an approach based on deep learning, called Req2Lib, which can recommend libraries given a NL description of the requirements. Req2Lib adapt a neural model called Sequence-to-Sequence (seq2seq) [9]. It is trained with a dataset of historical repository content, which contains two parts: requirement descriptions and libraries listed in the configuration files. In addition, we apply a domain-specific embedding model [10] trained over a corpus of Stack Overflow posts to represent words with high-dimensional vectors. With domain-specific embedding, Req2Lib can get more domain knowledge from the embedded input.

In the experiment, we train and evaluate Req2Lib using the descriptions and libraries of 5,625 unique projects. Finally, the experiment results show that Req2Lib can recommend libraries accurately with recall rate@10 of 0.904, precision@10 of 0.316, respectively.

The contributions of our work are as follows:

- We identify a new perspective on the domain of library recommendation which helps to solve the defects above: Recommend useful libraries given requirement descriptions of a project.
- We propose a deep learning approach to accurately recommend available third-party libraries from requirements, which proves the practicality of our perspective.

II. RELATED WORK

Among the previous studies of the recommendation system at the library-level, it is common to use the content of projects (i.e. current library usage, source code) to make recommendations. LibRec [3], proposed by Thung *et al.*, use association rule mining and collaborative filtering to recommend libraries. LibFinder [4], proposed by Ouni *et al.*, improves the performance of LibRec by recommending libraries based on linked-usage of libraries and semantic similarity of identifiers of source code. DSCRec [5] introduces domain-specific categories to the LibRec. Instead of using historical data of projects, Chen *et al.* [6] constructed an analogical-library knowledge base by mining Q&A data from Stack Overflow to calculate the similarity between libraries. However, studies above can not provide solutions to the their defects (i.e., requirement insufficiently matching and cold-start problem).

Our approach differs from all the previous works in the fact that we take the requirement descriptions instead of project libraries as input. To the best of our knowledge, this paper is

the first to investigate the use of requirement descriptions to address the problems of existing studies.

III. REQ2LIB

A. Overview

Req2Lib takes natural language as input and generates a sequence of related libraries. When the model recommends a library, it considers not only the semantic information from input but also the correlation information from the libraries recommended before. In addition, it contains a deep understanding of natural language combined with domain knowledge from the embedding model. Fig. 1 shows the overall framework of our approach. It comprises two steps: a) data preprocessing; b) model training.

B. Preprocessing

1) *Data Collection*: In this step, we collect project requirement descriptions and the corresponding libraries. Particularly, our model utilizes semantic information in NL so that we can use various forms of texts in natural language as long as it contains requirement information (e.g., requirement documents, project description, readme document). On the one hand, the proprietary of requirement documents limits its acquisition. On the other hand, readme documents are usually not well-organized which contains many unrelated information. Thus, we choose the project descriptions as the substitute of requirement documents. We get the data required by Req2Lib from Libraries.io.

2) *Data Processing*: The data we collected have poor quality. Thus, we apply several processing methods to the dataset. An example of the procedure for processing NL is shown in Fig. 2.

Concatenating: The name of a project contains semantic information. To get more information, we splice the project names and stitch them before the input texts.

Filtering: For natural language, we turn all the words into lowercase and remove the stop words¹ and special symbols. At the same time, we use a vocabulary [10] based on Stack Overflow Q&A posts to filter out words that are not related to the software domain. For libraries, we remove low-usage libraries because libraries adopted by few people may have quality defects.

Lemmatization: Using the Stanford Core NLP tool², we reduce the words to its base form. For instance, the word 'libraries' will be replaced by the word 'library'. In this way, a word in different forms can share the same weights in our model.

Sorting: The deep learning model that we use is a sequential model. To improve the performance, libraries should be listed in a fixed order. Thus, we sort the libraries of each project descending by library frequency in the total data.

Tokenization: We tokenize the descriptions and obtain a sequence of tokens. As for libraries, each library is treated as a

¹<https://www.textfixer.com/tutorials/common-english-words.txt>

²<https://stanfordnlp.github.io/CoreNLP/>

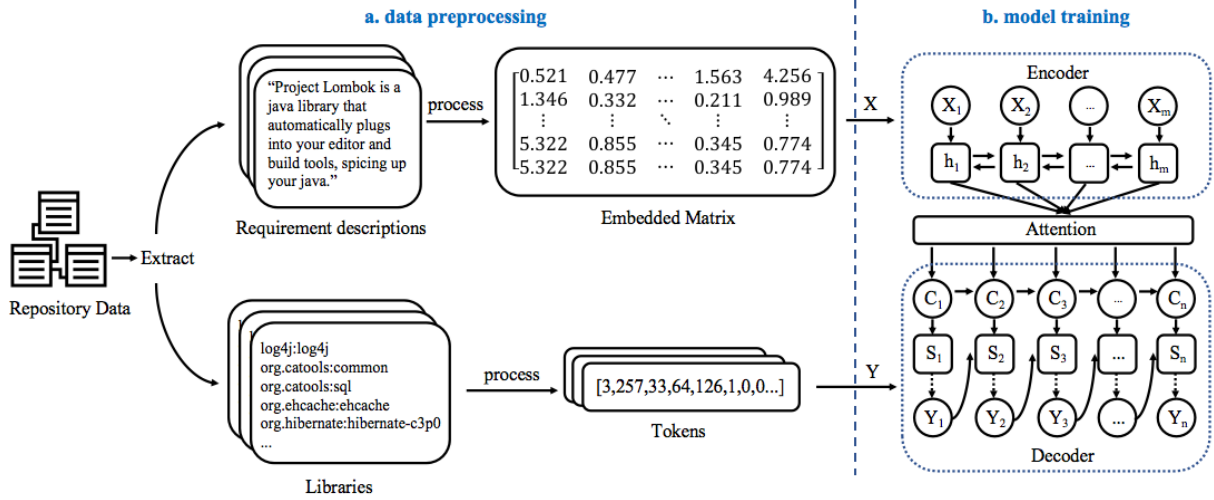


Fig. 1. Overview of Req2Lib. The dotted lines denote the masked softmax layer

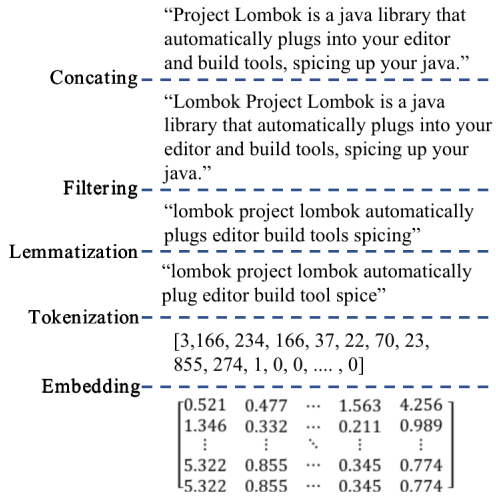


Fig. 2. An example of description processing.

token. With randomly generated vocabularies, we represent the tokens with numerical IDs. In this way, we turn the natural languages and libraries into sequences of numerical tokens with fixed length.

Embedding: Using neural models, we can present the words by real-valued vectors. Our model utilizes a domain-specific pre-trained word2vec model [10] that is trained over textual corpus from Stack Overflow posts. The word2vec model [11] can captures detailed relational structure, and a broad range of topics in the field. With the pre-trained model, we convert the tokens into 200-length high-dimensional vectors that contains related software domain knowledge.

After the above processing steps, we convert the descriptions and libraries into a sequence of dense embedding vectors and a sequence of tokens separately.

C. Sequence-to-Sequence

The library recommendation task can be treated as a multi-label classification problem by taking libraries as labels. Thus, we implement our model based on the work of Yang et al [12]. They proposed a variant of the seq2seq model with a novelty decoder for multi-label classification. The overview structure is shown in the model training part of Fig. 1.

Encoder: We adopt a deep neural network called Long Short-Term Memory (LSTM) [13]. To fully utilize the context of input, we use a bi-direction LSTM as the encoder to extract semantic information from the input in both forward and reverse directions. At each step t , it reads the t -th embedded token x_t from the input sequence, then computes the hidden states h_t , namely,

$$\vec{h}_t = \overrightarrow{LSTM}(x_t, \vec{h}_{t-1}); \overleftarrow{h}_t = \overleftarrow{LSTM}(x_t, \overleftarrow{h}_{t+1}) \quad (1)$$

By concatenating \vec{h}_t and \overleftarrow{h}_t , we get the final output of the Encoder, a sequence of hidden state h_1, h_2, \dots, h_n called context vector. It represents the semantic information and other latent features extracted from the embedded input.

Attention: For the contributions of each keyword to the recommended result is different, we use an attention component [14] to score each hidden state of encoder and compute the weighted average to generate new context vectors. The weights α_{ti} to the i -th word of time step t is computed as follow:

$$e_{ti} = v_a^T \tanh(W_a s_t + U_a h_i); \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^m \exp(e_{tj})} \quad (2)$$

where W_a, U_a, v_a are weights parameters of linear layers and s_t is the current hidden state of decoder. With weights, the new context vectors are computed as follow:

$$c_t = \sum_{i=1}^m \alpha_{ti} h_i \quad (3)$$

Decoder: To predict the current results from the context vector c_t , we use a LSTM model as the decoder. It takes the

context vector, previous prediction, and previous hidden state as input. In other words, it take use of semantic information and library linked-usage information. Its hidden state s_t is computed by:

$$s_t = LSTM([emb(y_{t-1}); c_{t-1}], s_{t-1}) \quad (4)$$

where $[emb(y_{t-1}); c_{t-1}]$ is the concatenation of $emb(y_{t-1})$ and c_{t-1} . $emb(y_{t-1})$ is the embedding of a library whose id has the biggest value in y_{t-1} , y_{t-1} is the probability distribution of the model output in the previous step, which is computed as follow:

$$o_t = W_o \text{relu}(W_d s_t + V_d c_t); y_t = \text{softmax}(o_t + I_t) \quad (5)$$

where W_o, W_d, V_d are the weight parameters of the linear layers. I_t is the mask at time step t to prevent the decoder predicting a repeated library i , which is calculated by:

$$(I_t)_i = \begin{cases} -\infty & \text{if } i \text{ has repeated at previous steps} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Loss: The goal of the model is to minimize the value of loss function. On the basis of cross-entropy loss, we apply a popularity-based coefficient to the weight of each item. Thus, we use the weighted cross-entropy loss as the loss function of our model. The coefficients are computed by:

$$L(y) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^n w_j \log p(y_j^i); w_j = 1 - \frac{f_j}{\sum_{k=1}^n f_k} \quad (7)$$

where N is the total number of a training batch, n is the number of target libraries of each project, f_j is the frequency of a library whose ID is j .

Cross-entropy loss is widely used in classification models. The reason why we use weighted one is to set barriers to popular libraries. Among the libraries used by developers, there are many popular ones that are acceptable for almost all projects (e.g., Junit). The model can always recommend these common libraries to reduce the loss value. However, from the user's point of view, predictable recommendations are not particularly useful. The predictable recommendation is a new problem, to which previous studies do not provide any solutions.

IV. EXPERIMENTS

A. Dataset

Our study uses the dataset extracted from Libraries.io [1] on 22 Dec 2018. To improve the data quality, we filter the project with the following criteria:

- The project should use Java as the programming language with over ten stars. Projects with low stars may have problems in quality.
- Following the work of Thung et al [3], the project should use at least ten libraries.
- To filter out the descriptions that contains little valid information, the project should have a piece of descriptive text which contains more than three words.

- The project should be unique.

Finally, we obtained 5,625 pieces of data. The collected projects contain 1468 libraries and .80% of them are randomly picked as training sets, and the rest are used as test sets.

B. Evaluation Metrics

Recall rate@k: Recall rate@k is widely used by previous studies to evaluate their models [3] [8]. It is calculated as follows:

$$\text{Recallrate@k} = \frac{\sum_{i=1}^N \text{isFound}(S_i)}{N} \quad (8)$$

and

$$\text{isFound}(x) = \begin{cases} 1 & x \text{ is in ground truth} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where S_i is the i _th model result.

Precision@k: Precision@k is defined as the proportion of ground truth hits among top-k model results. It is calculated by:

$$\text{Precision@k} = \frac{TP}{TP + FP} \quad (10)$$

where TP is the number of correct ones among top k model results, FP is the number of incorrect ones among top k model results.

Popularity-Stratified Recall@k Metrics of previous studies mainly focus on the accuracy of models. To assess predictable recommendations, we adopt popularity-stratified recall (PSR)@k [15]. On the basis of classical recall, PSR introduces the inverse response probability as the weights of each library. In this way, popular libraries acquire a lower weight. It is calculated by:

$$\text{PSR@k} = \frac{\sum_{i \in S_u^{+,k}} s_i}{\sum_{i \in S_u^+} s_i}; s_i \propto 1/(N_{obs,i}^+)^{\beta} \quad (11)$$

where $S_u^{+,k}$ is the sequence of recommended results for project i , S_u^+ is the sequence of ground truth for project i , $N_{obs,i}^+$ is the frequency of a library i among all the projects. We set β to 0.2 to prevent the models who only recommend unpopular items from winning.

C. Details

According to the above methods, we obtain the dataset and train the model. During training, we use the Adam optimization algorithm [16] to minimize the weighted cross-entropy loss. To improve the accuracy, we adopt several tricks on seq2seq:

Dropout regularization [17]: The key idea of dropout to randomly drop units from the neural network during training. It can prevent neural network from over-fitting.

Gradient clipping [18]: This strategy limit the gradient to a maximum number, which helps to deal with exploding gradients and vanishing gradients problem

Beam search [19]: It is applied for measuring sub-optimal recommendation. In this way, we can get the global optimal result.

D. Results

With the experimental results obtained from the model on the test set, we compare the recommended libraries with actually used libraries and calculate the evaluation metrics with different recommendation size k . The results is shown in Table. I and Fig. 3.

For the recall rate@ k , with the increase of recommendation list sizes k , the performance of our method for recall rate@ k has improved. When the size reaches 10, our model can recommend at least one useful library for more than 90% of projects. For the precision@ k , a more significant k results in a lower precision, because confident results have a higher priority in our model. Req2Lib do not perform well on PSR@ k for the reason that PSR is a fairly strict criterion.

TABLE I
OBTAINED EVALUATION RESULTS

Metrics	k=1	k=5	k=10	k=20
Recall rate@k	0.684	0.880	0.904	0.931
Precision@k	0.684	0.419	0.316	0.227
PSR@k	0.026	0.096	0.156	0.237

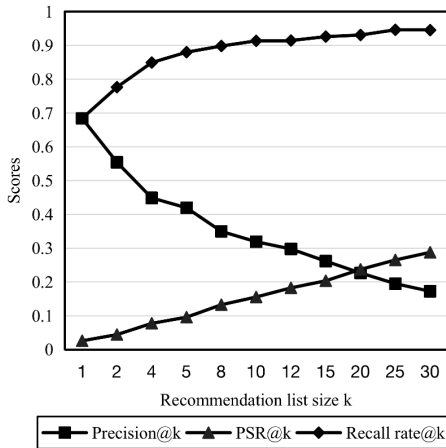


Fig. 3. Different performance on varying recommendation list sizes.

E. Threats to Validity

The results can not present the price for changing the way of making recommendation. The reason is that there are no similar studies on requirement-based library recommendation. Furthermore, it is hard for a sequential model to compare with previous studies [3] [4] directly. In their experiments, for a project, half of its libraries are used as model input. The rest need to be kept for evaluation. Different from them, our model recommend all libraries given requirement descriptions. As a sequential model, the first recommended library will influence the choice of next library. Thus, evaluating half of them can not present the actual accuracy of our approach. A deeper study is needed.

V. CONCLUSION

This paper presents our new perspective on library recommendation to avoid the essential defects of previous studies. We also propose a deep learning approach called Req2Lib to recommend libraries given requirement descriptions. The experimental results of Req2Lib demonstrate that the perspective we identified is practical. Besides, the promising results also indicate that it worth a further study for requirement-based library recommendation using deep learning.

REFERENCES

- [1] J. Katz, "Libraries.io Open Source Repository and Dependency Metadata," Dec. 2018.
- [2] K. Kevic and T. Fritz, "Automatic search term identification for change tasks," in *Companion Proceedings of the 36th International Conference on Software Engineering*, pp. 468–471, ACM, 2014.
- [3] F. Thung, D. Lo, and J. Lawall, "Automated library recommendation," *Proceedings - Working Conference on Reverse Engineering, WCRE*, pp. 182–191, 2013.
- [4] A. Ouni, R. Gaikovina Kula, M. Kessentini, T. Ishio, D. M. German, and K. Inoue, "Search-based software library recommendation using multi-objective optimization," *Information and Software Technology*, vol. 83, pp. 55–75, 2016.
- [5] D. Katsuragawa, A. Ihara, R. G. Kula, and K. Matsumoto, "Maintaining third-party libraries through domain-specific category recommendations," *2018 IEEE/ACM 1st International Workshop on Software Health (SoHeal)*, pp. 2–9, 2018.
- [6] C. Chen, S. Gao, and Z. Xing, "Mining Analogical Libraries in Q&A Discussions – Incorporating Relational and Categorical Knowledge into Word Embedding," *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, vol. 1, pp. 338–348, 2016.
- [7] A. Zerouali, E. Constantinou, T. Mens, G. Robles, and J. González-Barahona, "An empirical analysis of technical lag in npm package dependencies," in *New Opportunities for Software Reuse* (R. Capilla, B. Gallina, and C. Cetina, eds.), (Cham), pp. 95–110, Springer International Publishing, 2018.
- [8] F. Thung, S. Wang, D. Lo, and J. Lawall, "Automatic recommendation of API methods from feature requests," *2013 28th IEEE/ACM International Conference on Automated Software Engineering, ASE 2013 - Proceedings*, pp. 290–300, 2013.
- [9] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.
- [10] V. Efstathiou, C. Chatzilenas, and D. Spinellis, "Word embeddings for the software engineering domain," pp. 38–41, 2018.
- [11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [12] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: Sequence Generation Model for Multi-label Classification," 2018.
- [13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [15] H. Steck, "Item popularity and recommendation accuracy," *RecSys'11 - Proceedings of the 5th ACM Conference on Recommender Systems*, pp. 125–132, 2011.
- [16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [18] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, pp. 1310–1318, 2013.
- [19] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," *arXiv preprint arXiv:1606.02960*, 2016.