Effect of the Size of Datasets

TOSEM-2021-0216

The size of datasets may have an impact on PLAN. We investigate the effectiveness of the following three datasets: (1) the size of irrelevant API-Q&A pairs, (2) the size of the corpus for word embeddings, and (3) the size of AK datasets.

1 Effect of the size of irrelevant API-Q&A pairs

ARQ1: How effective is PLAN under different sizes of irrelevant API-Q&A pairs?

The size of irrelevant API-Q&A pairs may have impacts on PLAN. In this RQ, we would like to investigate the performance of PLAN under different sizes of irrelevant API-Q&A pairs. We originally collected 9,787 McGill and 6,304 Android irrelevant API-Q&A pairs. However, to mitigate the imbalanced problem, we randomly selected an equal amount of irrelevant API-Q&A pairs based on the number of relevant API-Q&A pairs (i.e., 3,561 McGill and 993 Android irrelevant API-Q&A pairs). We compare the P@5, P@10, P@15, and MRR of PLAN on different sizes of irrelevant API-Q&A pairs, respectively. SO and tutorial datasets are the same as we used in our paper, except that we vary the size of irrelevant API-Q&A pairs in SO datasets. For SO-McGill dataset, the size of irrelevant API-Q&A pairs is set as [10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 200%, All]. For SO-Android dataset, we set the size of irrelevant API-Q&A pairs as [10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 200%, 300%, 400%, 500%, All]. Note that the size of irrelevant API-Q&A pairs refers to the percent of 3,561 McGill and 993 Android irrelevant API-Q&A pairs. "All" means that we use all the collected McGill and Android irrelevant API-Q&A pairs (i.e., 9,787 McGill and 6,304 Android irrelevant API-Q&A pairs). Since the total number of the collected McGill and Android irrelevant API-Q&A pairs are different, the size setting of these two datasets is different.

Figures R1 and R2 present P@5, P@10, P@15, and MRR of PLAN under different sizes of McGill and Android irrelevant API-Q&A pairs, respectively. As shown in Figures R1 and R2, the performance of PLAN increases first until the size is 100%; thereafter it decreases. That is, PLAN can achieve better performance when using 100% McGill and Android irrelevant API-Q&A pairs. Therefore, using 100% McGill and Android irrelevant API-Q&A pairs (i.e., 3,561 McGill and 993 Android irrelevant API-Q&A pairs) is helpful in improving the PLAN's performance and we use this size setting in our work.

2 Effect of the size of the corpus for word embeddings

ARQ2: How effective is PLAN under different sizes of the corpus for word embedddings?

The size of the corpus for word2vec may have impacts on PLAN. We would like to investigate the performance of PLAN under the size of the corpus for word2vec with different

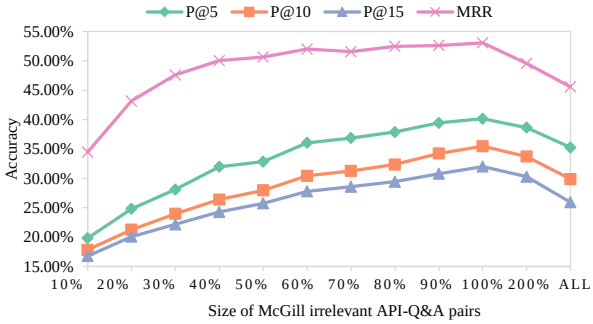

Figure R1: Results for different sizes of Mcgill irrelevant API-Q&A pairs

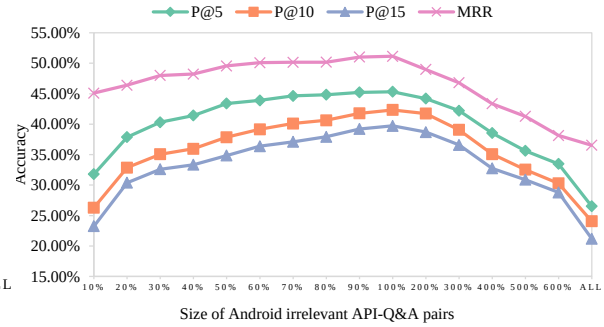

Figure R2: Results for different sizes of Android irrelevant API-Q&A pairs

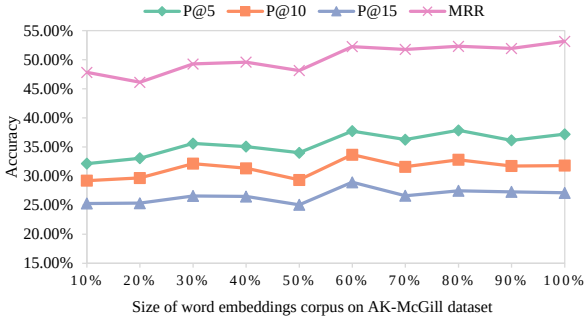

Figure R3: Results for different sizes of the corpus for word embeddings on AK-McGill dataset

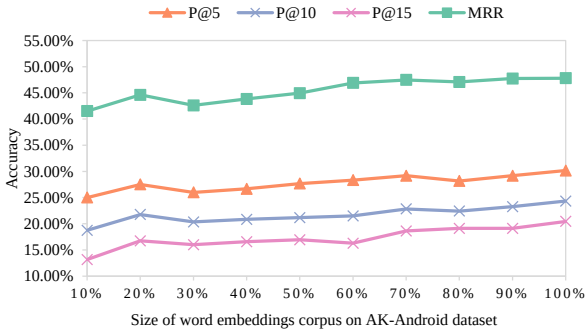

Figure R4: Results for different sizes of the corpus for word embeddings on AK-Android dataset

values. We compare the P@5, P@10, P@15, and MRR of PLAN with different sizes of the corpus for word2vec (i.e., [10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%]), respectively.

Figures R3 and R4 show the results (i.e., P@5, P@10, P@15, and MRR) of different sizes of the corpus for word embedding on AK-McGill and AK-Android datasets, respectively. We can observe that when using 60% or more corpus of word2vec, PLAN can achieve better performance to P@5, P@10, P@15 and MRR on AK-McGill and AK-Android datasets, respectively.

3 Effect of the size of AK datasets

ARQ3: How effective is PLAN under different sizes of AK datasets?

The size of AK (i.e., API-KI) datasets may have impacts on PLAN. We would like to investigate the performance of PLAN under different sizes of AK datasets. We compare the P@5, P@10, P@15 and, MRR of PLAN with different sizes of AK datasets (i.e., [10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%]), respectively.

Figure R5 and Figure R6 present P@5, P@10, P@15 and MRR of different sizes of two AK

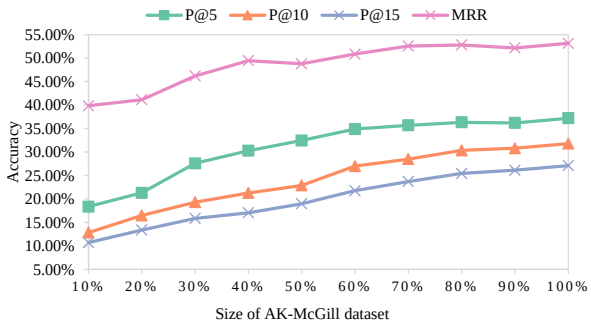

Figure R5: Results for different sizes of AK-McGill dataset

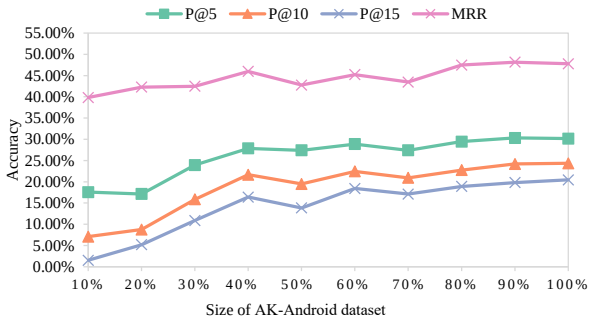

Figure R6: Results for different sizes of AK-Android dataset

datasets, respectively. We can observe that as more API-KI pairs are added, PLAN can achieve better performance on AK-McGill and AK-Android datasets, respectively. The reason may be that the performance of TDML used in PLAN is improved by more training data. That is, TDML can find more relevant API-KI pairs as more training data is added.

References