عنوان مقاله:

ParSQuAD: Persian Question Answering Dataset based on Machine Translation of SQuAD ۲.۰

محل انتشار:

فصلنامه بین المللی وب پژوهی, دوره 4, شماره 1 (سال: 1400)

تعداد صفحات اصل مقاله: 13

نویسندگان:

Negin Abadani - *Department of Software Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan*

Jamshid Mozafari - *Department of Software Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan*

Afsaneh Fatemi - *Department of Software Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan*

Mohamadali Nematbakhsh - *Department of Software Engineering, Faculty of Computer Engineering, University of Isfahan, Isfahan*

Arefeh Kazemi - *Department of Linguistics, University of Isfahan, Isfahan, Iran*

خلاصه مقاله:

Recent developments in Question Answering (QA) have improved state-of-the-art results, and various datasets have been released for this task. Since substantial English training datasets are available for this task, the majority of works published are for English Question Answering. However, due to the lack of Persian datasets, less research has been done on the latter language, making comparisons difficult. This paper introduces the Persian Question Answering Dataset (ParSQuAD) based on the machine translation of the SQuAD ۲.۰ dataset. Many errors have been discovered within the process of translating the dataset; therefore, two versions of ParSQuAD have been generated depending on whether these errors have been corrected manually or automatically. As a result, the first large-scale QA training resource for Persian has been generated. In addition, we trained three baseline models, i.e., BERT, ALBERT, and Multilingual-BERT (mBERT), on both versions of ParSQuAD. mBERT achieves scores of ۵۶.۶۶% and ۵۲.۸۶% for F۱ score and exact match ratio respectively on the test set with the first version and scores of ۷۰.۸۴% and ۶۷.۷۳% respectively with the second version. This model obtained the best results out of the three on each version of .ParSQuAD

کلمات کلیدی:

Question Answering, Persian Machine Reading Comprehension, Persian Question Answering Dataset, SQUAD

لینک ثابت مقاله در پایگاه سیویلیکا: