SEMANTIC QUESTION PAIR MATCHING WITH DEEP LEARNING

***Abstract*—** Q&A forums like Quora, Stack-overflow, Reddit, etc are highly susceptible to question pair duplication. Two questions asking the same thing could be too different in terms of vocabulary and syntactic structure, which makes identifying their semantic equivalence challenging. In this paper, we've explored deep learning methodology of determining semantic equivalence between pairs of questions using a dataset released by Quora of more than 400,000 questions pairs through Machine Learning with Natural Language Processing. Even if a model cannot describe exactly the reality, it could be very helpful if it is close enough, with our model we have tried to achieve an accurate prediction of semantic relatedness between common queries Our machine learning approach is based upon Levenshtein distance between two sentences and the sentence-vector encoding using our own unique Word2Vec model to experiment with a variety of distance metrics and predict their semantic equivalence. We compare the standard supervised classification methods such as Logistic Regression, KNN and Random Forest with our artificial neural network (ANN). Our experimental results show that the artificial neural network with word embeddings achieves high performance, achieving an F1-score of 0.6529 with 0.7236 accuracy on the test set.

Keywords— semantic analysis, duplicate questions, natural language processing, machine learning, word embeddings