SIMILAR QUESTION IDENTIFICATION USING DEEP LEARNING

[1]Pragati P. Mahale, [2]Pragya Kumari, [3]Mohini J. Rana, [4]Ashish H. Patil, [5]Rushikesh A. Daga

[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student
[1]Information Technology,
[1]AISSMS Institute of Information Technology, Pune, India

Abstract: Similar Question Identification makes the use of Natural Language Processing (NLP) which is a subject under dynamic research. The task of identifying similar questions falls under Semantic Analysis (understanding the meaning of text). It is mainly used in the identification of similar questions on Q/A platforms. Sometimes the user submits a question that describes the same problem as the one which had already been asked, leading to generation of duplicate questions. In this project, we make use of Natural Language Processing to aid and automate the process of identification of similar questions on online user forums. Two questions are similar if they can agreeably be answered by same answer. Similar Question Identification is associated with Semantic Text Analysis, which evaluates the extent to which two text chunks are semantically equivalent. We take the words in the question in simple English, perform some processing on the words, also termed as tokens, and then use the result of processing on frequency of words to identify similar questions.

IndexTerms **– Deep Learning, Natural Language Processing, Semantic Analysis**

I. INTRODUCTION

Question-answering (Q&A) community sites, such as Yahoo! Answers, Quora and Stack Exchange, have attained a lot of consideration in the present-day. Most Q&A community sites propose users to look for an equivalent answer before posting a new question. Nonetheless, this is not consistently an effortless job because distinct users could explain the same question in quite various ways. A few online question-answering communities, such as the Stack Exchange, have a similarity scheme. Exact duplicates, such as copy-and-paste questions, and nearly exact duplicates are generally quickly detected, closed and removed from the community. However, some similar questions are kept. The primary reason for that is that a same question can be posted in different words, and a user might not be able to look up the answer to it if they are asking it in a different way. In this study, we describe two questions as semantically equivalent or similar if they can be satisfactorily answered by the exact same answer. Detecting semantically similar questions is a strenuous job due to two main factors: (1) one question can be defined in many different ways; and (2) two questions might be defined differently but are looking for same answer. Therefore, conventional similarity measures based on word overlap such as Shingling and Jaccard coefficient (Broder, 1997) and its variations (Wu et al., 2011) are unable to capture many instances ofsemantic equivalence.

II. LITERATURE SURVEY

This paper[1] presents the results of experimentation with widely used algorithms for automatic similar question detection. When these methodologies are applied to progressively larger datasets, thus enabling us to study and evaluate the learning profiles of this task under different approaches. The study was made possible by utilizing a new dataset with over 400,000 pairs labeled with respect to their elements that consist of duplicate interrogative segments, released by Quora question answering engine for research purposes. A major observation is that the more knowledgeably complex type of approach used, deep learning, has clearly a higher degree of performance than other methodologies, given it can be trained over a considerably large dataset.

In this paper[2], DupPredictor, an automated duplicate question detector is used. This approach gets a question as input and identifies all the questions similar to this question by considering various factors. The title, description of a question and tags associated with the question are extracted. User needs to input the title, description and tags of a question. By using a topic model latent topics are processed. Then, four similarity scores are calculated for every question by analyzing their titles, descriptions, latent topics, and tags. A new similarity score is formed by combining the previous four similarity scores. DupPredictor measures the similarity of two questions by contrasting their evident factors, which are titles, descriptions, and tags of the questions and their dormant factors correspondingto the topic dispersions that are learned from the naturallanguage descriptions of thequestions. Four factors of DupPredictor have been described as: title similarity component, description similarity component, topic similarity component and tag similarity component. All the four components are combined on the basis of weights assigned to them that are learnt automatically from training data. DupPredictor was evaluated on questions of Stack Overflow site containing more than two million questions and measured on the basis of recall rate.

This paper[3] intents to identify semantically similar questions from online user forums. A number of experiments were carried out using data from two different Stack Exchange online user forums. Analysis of standard machine learning algorithms such as Support Vector Machines (SVM) and a convolutional neural network (CNN) was done. The CNN method creates distributed vector representations for pairs of questions and scores them using a similarity measure. First,

they calculated in-domain word embedding vs. the ones trained on Wikipedia, then anticipated the effect of the training set size, and then calculated some features of domain transformation. Experimental results showed that the convolutional neural network (CNN) with in-domain word embedding can achieve high performance.

This paper[3] proposes a method for classifying questions that are semantically similar on the basis of a convolutional neural network(CNN). Experiments showed that the CNN can accomplish high accuracy exclusively when the word embedding is trained before on in-domain data. The results of an SVM-based approach to this task depend largely on the size of the training data. On the other hand, the CNN with in-domain word embedding grants high performance.

This paper[4] presents a report on the observations of an experimental study on the application of character-level embedding and basic convolutional neural network (CNN) to the shared task of duplicate statement detection in Russian. The credibility of this algorithm was tested in the basic run of Task 2 of that shared task and gave out competitive results. The result of this algorithm is tested against a word-level NN and various other approaches such as rule-based and classical machine learning.

This paper[4] showed the performances of a range of experiments to take on the task of paraphrase detection for Russian under the circumstances and with the datasets of the corresponding shared task organized in 2016. Specifically, the character- based convolutional neural network model attains competitive results for the task of detecting similar sentences.

In this paper[5], they showed an experimental study to examine duplicate question detection in Stack Overflow. For this objective, they carried out an individual reproduction of two prior processes, DupPredictor[2] and Dupe. An experimental study was performed on DupPredictor[2] and Dupe. The outcomes, not persuasive when strived with distinct set of tools and data sets, show that the constraints to reproduce these processes are high. Furthermore, when tested with more recent data, a performance degradation of both reproductions in terms of recall-rate over time was perceived as the number of questions increased. Findings infer that the subsequent works concerning detection of duplicated questions in Online User Forums necessitate intensive research to assert their findings.

This paper[6] proposes a method of priority judgment emanating from the sentiment words inside these question articles in cases of enormous questions to resources or products. This method utilizes LDA which is renowned as an “unsupervised learning technique” to excerpt question articles as the “question target”, and then the question target part stated in questions, known as the “review points” are extracted by KeyGraph based on question targets. For those sentiments of question articles’ authors from drawn out review points, evaluation expression dictionary is used to provide “sentiment words” to question articles. For the configuration of question targets and review points, question articles with more negative emotions are contemplated as higher priority questions. The aforementioned method is rehearsed on the question articles posted on the online support forums to implement a priority judgment analysis on them. After collating the ranking of priorities provisioned by users with support center experience, it is entrenched that the proposed method could correctly categorize articles with higher priority and the emotion words given to articles with medium priority are also conclusive.

In this paper[7], two methodologies are used to enhance the detection accuracy. In the first method, the posted questions in Stack Overflow consist of not only natural languages but also artificial languages such as source codes and system logs. Every type of data is confined with a different HTML markdown. The posted data or question is split into the distinctive types of data and analyze them to improve the efficiency. In the second method, it has been marked that Wordnet is not very effectual in detecting similar questions. This is because Wordnet is manually supported and that being the case; it barely catches up new technical terms. Hence, the method of Word2vec is introduced. This method is based on the Distributional Hypothesis and maps word co-occurrence relations of each word to a low dimensional vector space by utilising a two-layer neural network. It is computationally economical to create a Word2vec model, so it facilitates us to catch up new words by periodically refreshing the model. This method is expected to reimburse for the weakness of the BOW (Bag of words) based way.

III. Future Scope and Research Area

The current systems use different approaches for addressing the problem. These methods have different learning approaches. There are many different models related with the task. One way to take advantage of the fact is to ensemble the methods together to obtain a better predictive performance which could not be obtained from any of the constituent learning algorithms alone. In addition to CNN, other profiles can ensemble to give better accuracy.

IV. Conclusion

Natural language processing is one of the most challenging problems. To overcome this problem, more accurate and efficient systems should be implemented. The current systems do not provide accuracy as much as we expect it to be. A more principled and computationally intensive model can be created. We discuss limitation and future scope for each of these systems. We hope our study paper help others understand the current systems for semantic analysis between two questions.

V. ACKNOWLEDGMENT

Thanks to DigitalMain Inc. and Vizerto product development team for inspiring us to pursue research in this field of study. We also thank them for active feedback and input on the significance of this work and its uses in different business and non-commercial use-cases. Vizerto learns about the business language of a specific organization and creates reusable knowledge through curated question/answer pairs that can be utilized by all the team members to improve the speed of decision making, responding to a customer in a timely manner while creating scale for experts to support a wider audience.

REFERENCES

[1] "Learning Profiles in Duplicate Question Detection"
Chakaveh Saedi, Jo˜ao Rodrigues, Jo˜ao Silva, Ant´onio Branco, Vladislav Maraev , IEEE 2017

[2] "Multi-factor duplicate question detection in stack overflow"
Yun Zhang, David Lo, Xin Xia, Jian-Ling Sun, Springer 2015

[3] "Detecting semantically equivalent questions in online user forums"
Dasha Bogdanova, C´ıcero dos Santos, Luciano Barbosa and Bianca Zadrozny, ACM 2015

[4] "Character-Level Convolutional Neural Network for Paraphrase Detection and Other Experiments"
Vladislav Maraev(B), Chakaveh Saedi, Jo˜ao Rodrigues, Ant´onio Branco, and Jo˜ao Silva, Springer 2018

[5] "Duplicate Question Detection in Stack Overflow: A Reproducibility Study"
Rodrigo F. G. Silva, Klérisson Paixão, Marcelo de Almeida Maia, IEEE 2018

[6] "Urgent Question Detection based on the Review Points and Sentiment Words" Koji Wajima, Tetsuji Satoh, ACM 2015.

[7] "Two Improvements to Detect Duplicates in StackOverflow", Yuji Mizobuchi and Kuniharu Takayama, IEEE 2017