International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 01 | Jan 2020 www.irjet.net p-ISSN: 2395-0072

Semantic Question Matching

Ruchit Mody[1], Yesha Sanghavi[2], Sejal D'Mello[3]

[1]B.E. student, Dept. of Information and Technology, Atharva college of Engineering, Maharashtra, India
[2]B.E. student, Dept. of Information and Technology, Atharva college of Engineering, Maharashtra, India
[3]Assistant Professor, Dept. of Information and Technology, Atharva college of Engineering, Maharashtra, India

---

Abstract – *Quora is a very popular and known tool used by developers all over the world. Owing to this popularity millions of questions of are asked by users everyday in the hope of getting expert solution. Hence there is a high possibility that numerous users pose questions that might having same intent or purpose. The methods used in this paper will help people discover questions that have already been discussed on the forum previously and keep individuals from replying to a similar question time and time again thereby improving improving user experience for all. The main aim of this project is to allow a user to see all questions that are semantically similar to his/her query so that the most optimal solution is made easily available and thus limiting the amount of duplicate questions being generated on the platform. This paper contributes to build up a proficient Semantic Question Detection system by experimenting with significant kinds of Natural Language Processing techniques used for intent identification*

Key Words: **Natural Language Processing, Semantic Question Matching, Word Movers Distance, Intent Classification, Google News Vector.**

1. INTRODUCTION

Quora is a very popular social media platform where people ask questions and can connect to the actual experts who contribute unique insights and quality answers. But since it has such an enormous user base so it's no surprise that many people ask similarly worded questions. Among the numerous questions posted in forums, two or more of them may express the same point and thus are duplicates of one another. Duplicate questions make site maintenance harder, waste resources that could have been used to answer other questions, and cause developers to unnecessarily wait for answers that are already available. Question duplication is the primary trouble encountered by using Q&A boards like Quora, Stack-overflow, Reddit, and so on. Answers get fragmented throughout distinct versions of the identical query due to the redundancy of questions in those forums. Eventually, this consequences in loss of a practical search, answer fatigue and segregation of information. This paper contributes to develop an efficient question duplication detection system using the power of Machine Learning and Natural Language Processing. The meaning of a sentence does not only depend on the words in it, but also in the order with which they are combined. The semantic similarity can have several dimensions, and sentences may be similar in one aspect but opposite in the other. For example if a person asks, "Did Linked-in acquire Microsoft?" while in the same platform if another user asks, "Did Microsoft acquire Linked-in?" the words used in both cases are same but the order completely changes the meaning of the sentence. In this paper we propose to identify questions with similar intent and then classify them as duplicate or original and to do so we first automatically generate tags, using NLP, for all available questions and after comparing the tags using various methodologies we calculate similarity between them.

2. LITERATURE SURVEY

Early work to distinguish the similitude between sentences utilized physically designed features like word overlap alongside conventional Artificial intelligence and Machine learning algorithms like Support Vector Machines. Semantic analysis of sentences has traditionally concentrated on logical inference based on the Inference Corpus of Stanford Natural Language. The paper by Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kociskya and Phil Blunsom used LSTMs to concentrate on word-by-word recognition methods [1].Neural Network methodologies have been the in a broader selection of NLP tasks. Siamese neural network which contains two sub-networks joined at their outputs was proposed. In spite of the fact that the Siamese architecture is lightweight and easily trainable, there is clearly a lesser impact of correlation between the parameters due to which information loss may occur [2]. The Compare-Aggregate model [3] which indeed captures the interaction between two sentences was proposed in order to ensure that the limitations of the Siamese framework were overcomed.

The recent study by Zhang et al shows the effect of word embedding. Their methodology, PCQADup, relies on extracted features of question pairs. They announced a noteworthy improvement in results contrasted with DupPredictor and Dupe [4],[5].In the paper "Mining Duplicate Questions in Stack Overflow" by Ahasanuzzaman Muhammad, Chanchal K. Roy and Kevin A. Schneider , for detecting duplicate questions, they used a discriminative model classifier together with BM25 scoring function. With a wide number of questions concerning multiple programming languages, they tested their strategy. As per the results of

their proposed strategy Dupe, it was found that Dupe outperforms Dup-Predictor, Stack Overflow's only usable duplicate query tracker [6]. To test the similarities between two interrogative segments in a set, a Duplicate question detection approach with a Jaccard coefficient in the paper by Y. Wu, Q. Zhang, and X. Huang , "Efficient near- duplicate detection for Q&A forum," was utilized. A Duplicate question detection dataset resorting to the Baidu Zhidao, a Chinese question and answer platform supported by the search engine Baidu, was developed. The program achieved a f-score of 60.29 by practicing with 3M pairs and checking on 3k pairs [7],[8].

3. PROPOSED SYSTEM AND ARCHITECTURE

3.1 Architecture Description

The proposed approach which we will develop is to match an input question with all its possible forms of duplicates. We first perform dataset preprocessing by removing all the stopwords, question marks and other unwanted characters. We need our model to understand that the words "ask" and "asked" are simply various tenses of a similar action word and Stemming and Lemmatization are concepts of NLP that's helps us to do just that. It helps us to convert various types of a word to a center root. After preprocessing we generate tags or keywords to every question. For every question, if 50% tags of that question match with the tags of another we run it through a Natural Language Processing algorithm used for classification.

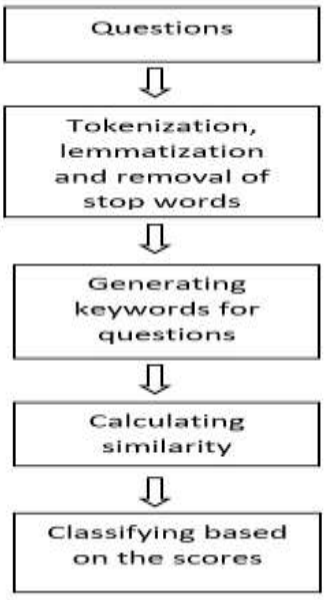

Figure 3.1: Architecture

3.2 IMPLEMENTATION

Our Research is carried out in the following stages:

A. Data Preprocessing

All the available questions are first cleaned by converting all to lowercase, removing stopwords, question marks and other unwanted characters. After which we use stemming to convert the words into its equivalent core words.

B. Generation of Tags

For all the available questions we automatically generate keywords/tags using TFIDF(Term Frequency Inverse-Document Frequency).This is a method used to evaluate the importance that a particular word has as compared to all the other words in the document. Words having the highest TFIDF score are chosen as the tags for that question. We developed a python code that takes Question as an input and finds out the number of words in that question. Using this it generates a particular number of tags which is equal to 30% of the length of question. For Example if a question of 11 words is asked such that "What are the different steps in becoming a freelance datascientist?" The 3 tags generated by the model are "steps","freelance","datascientist"

C. Calculation of Similarity

Since the machine cannot understand human language in this step we first convert all the words of the question into vectors. To do so we use Google News Vectors. Google News Vector includes word vectors for a vocabulary of 3 million words and phrases that are trained on around 100 billion words from a Google News dataset [9]. For every question if 50% tags of that question match with the tags of another we run it through a NLP algorithm used for classification. In that model we use Word Movers Distance to calculate similarity between the questions. WMD calculates the dissimilarity between two texts as the amount of distance that the embedded words of one text needs to travel to reach the embedded words of another text

D. Developing a Front End

To make our model more user friendly we developed a webpage using HTML5 and CSS. Using this web page any layman can work with our Python Model. We have used a flask framework that helps us to connect our python code to an HTML web page. Using Flask we can now create a web app that will trigger our python model on a button click. HTML5 and CSS will make our web page responsive and mobile friendly and accessible from devices of all sizes

4. OUTPUT

Figure 4.1: Sample result for no match

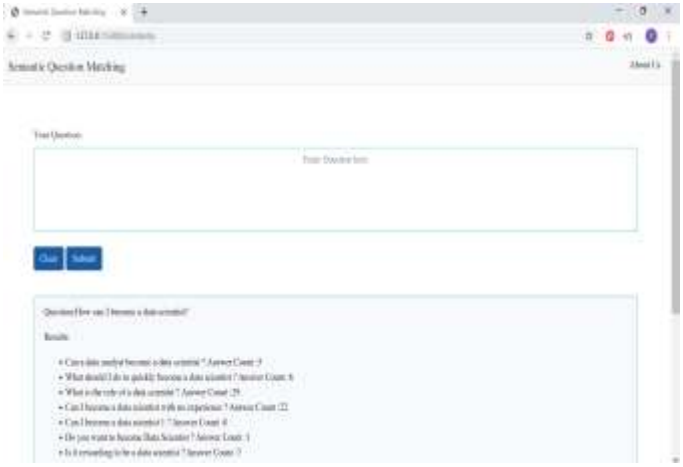

Figure 4.2: Sample result for matching questions

5. CONCLUSION

In this paper we try to identify the problem, of duplicate questions. We propose an unsupervised approach to duplication detection by developing a model that is able to identify the base intent of the question. Our paper is able to reflect the various advantages pertaining to identifying duplicate questions on question and answer forums which will lead to cheaper data storage as less data would be stored and the user experience would be improved as they can expect faster response to the questions. With the help of our project the clients will be able to find great answers without investing more time and energy searching for best answer among its similar questions. Our approach includes cleaning and tokenizing the questions in the initial phase to prepare the data set of questions. In our second stage we develop a model that is able to automatically generate tags/keywords for all the Questions. The third stage is the stage in which our model is prepared based on Natural Language Processing. It uses Google News vectors to covert words into vectors after which we run our similarity calculation model. WMD is a method that allows us to assess the "distance" between two documents in a meaningful way, no matter they have or have no words in common[10]. The final stage includes the implementation of the GUI where in an interactive website is created using Flask, HTML5 and CSS. Hence, this research work adopts Natural Language Processing to address the problem of question duplication in Q&A forums like Quora to classify whether questions are duplicates or non- duplicates.

ACKNOWLEDGEMENT

We would like to thank our project guide Prof. Sejal D'mello for her enormous cooperation and guidance. We have no words to express our gratitude for a person who wholeheartedly supported the project and gave freely of her valuable time while making this project. All the inputs given by her found a place in the project. She has always been a source of inspiration for us.

REFERENCES

[1] Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, Phil Blunsom." Reasoning about entailment with neural attention", In ICLR 2016
[2] Wang, Zhiguo, Wael Hamza, and Radu Florian. "Bilateral Multi-Perspective Matching for Natural Language Sentences.", In arXiv.org ,2017.
[3] Wang, Shuohang, and Jing Jiang. "A Compare-Aggregate Model for Matching Text Sequences.", In arXiv.org ,2016
[4] Rodrigo F. G. Silva, Kl□sson Paix□, Marcelo de Almeida Maia " Duplicate Question Detection in Stack Overflow: A Reproducibility Study " , 978-1-5386-4969-5/18/$31.00 , 2018 IEEE SANER 2018, Campobasso, Italy RENE Track.
[5] W. E. Zhang, Q. Z. Sheng, J. H. Lau, and E. Abebe, "Detecting duplicate posts in programming qa communities via latent semantics and association rules," in 26th International Conference on World Wide Web (WWW), Geneva, Switzerland, 2017, pp. 1221–1229.
[6] Muhammad Asaduzzaman , Chanchal K. Roy , Kevin A. Schneider "Mining Duplicate Questions of Stack Overflow", IEEE/ACM 13th Working Conference on Mining Software Repositories , September 2017.
[7] Chakaveh Saedi, Joao Rodrigues, Joao Silva, Antonio Branco, Vladislav Maraev, "Learning Profiles in Duplicate Question Detection", IEEE International Conference on Information Reuse and Integration (IRI), 2017 .
[8] Y. Wu, Q. Zhang, and X. Huang, "Efficient near- duplicate detection for Q&A forum," in Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), 2011, pp. 1001–1009.
[9] Ashwin Dhakal,Arpan Poudel ,Sagar Pandey " Exploring Deep Learning in Semantic Question Matching", IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu (Nepal) , June 2018
[10] Naveen Saini, Pushpak Bhattacharyya, Sriparna Saha, Dhiraj Chakraborty," Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures" ,In PLoS ONE November 2019