Project Report

On

QUORA QUESTION PAIRS SIMILARITY DETECTION USING SIAMESE MANHATTAN LSTM

Submitted by

Rameez Raza(182IT018)
Suraj Meshram(182IT018)

Under the Guidance of
Prof. G Ram Mohana Reddy
Dr. Sowmya Kamath S
Dr. Anand Kumar
Dept. of Information Technology, NITK, Surathkal

Date of Submission: 03-04-2019

Department of Information Technology
National Institute of Technology Karnataka, Surathkal.
2018-2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Detecting questions pairs similarity all along is a challengeable task in the field of natural language processing (NLP), since uncertainty and inconstancy of phonetic articulation. Specifically, in the field of community question answering (CQA), the homologous hotspot is focusing on question retrieval.it is essentially a classification problem: given a pair of questions, label it similar, relevant, or irrelevant. To get the most similar question compared with user's query, we proposed a question model building with Siamese Manhattan Long Short-Term Memory (MaLSTM) neural networks, which as well can be used in other fields, such as sentence similarity computation, paraphrase detection, question answering and so on. We evaluated our model in labeled Quora Answers data.

Chapter 1

Introduction

1.1 Motivation

With the explosive growth of the internet information, Question answering (QA) portals, such as Yahoo! Answers, Quora and Baidu Knows, are developing dramatically. Acting as a platform for people to share their knowledge and experience, QA portals have accumulated a lot various forms of data from multiple fields organized as questions and a list of candidate answers. To get the intended answer in QA, there are two steps. Firstly, retrieving similar posted questions and then gathering the candidate answers through those similar questions. Secondly, based on a quality assessment on the candidate answers, picking up the most relevant answer. During the interactive process between users and the QA website, sentence similarity computation, question similarity computation specifically, plays a key role.

In the field of normal language preparing (NLP), a center issue is to decide whether two sentences have around a similar significance. That is, we have to know "Thou workmanship mine" and "You are mine" share a similar importance, and "How old are you?" and "What is your age?" express a similar inquiry. A standout amongst the most common utilizations of inquiry likeness is the network Question Answering (QA) framework.

In QA system after receiving a query from the user, a QA system first retrieves a list of possible candidate questions from a large database (typically via an indexing service), resulting in only a few hundred ones. Then, it selects the most similar question set from the candidate list with a sentence similarity algorithm. The answers to these existing questions are most likely the correct responses to the original query. In such a system, the key issue is to determine the query-question similarity, i.e., find a question that is most similar to the user's query. More formally, the problem is defined as follows:

Given a query Q and a set of relevant question candidates $\{C_1, C_2,..., C_n\}$retrieved from an indexing service, determine whether or not each candidate C_i is similar to Q, and rank them by their similarities to the original query Q.

Determining question similarity is difficult because of the complication of semantics in human languages. The same word may have multiple meanings, and the same meaning can be expressed in many different ways. At the sentence level, the syntactical structures that contribute to paraphrases are even more complicated. Particularly, in community question answering, similarities over the surface text do not always imply that the questions share the same answer. For instance, for the query “What must not I feed a dog?”, similar questions would be “What can’t dogs eat?”, and “What food may make dogs sick?”. On the other

hand, the seemingly similar question “What can I feed a dog?” has just the opposite meanings. Simply taking its answer may lead to tragic consequences.

To cope with this challenge, predecessors proposed a variety of methods, from logic-based inference methods to vector space semantic models, from traditional NLP syntactic parsing or lexical semantic networks based WordNet similarity measurement to recently fashionable deep neural networks, and these methods are gradually getting better performance.we implemented a question similarity computation model adopting deep learning method. We model question using a siamese adaptation of the Long Short-Term Memory (MaLSTM) network for labeled data comprised of pairs of variable-length sequences. we provide word- embedding vectors supplemented with synonymic information to the MaLSTMs, which use a fixed size vector to encode the underlying meaning expressed in a sentence (irrespective of the particular wording/syntax).On the Quora pair dataset, we tested several deep learning methods based on RNNs to compute question similarity. And about the question modeling, we adopted two kinds of architectures, connected and separated. Additionally, from the result MaLSTM is better than RNN. So deep neural networks do not necessarily have to be highly deep. In fact, deep learning is a new thing. During the tuning of model parameters, we tried several sets of parameters, some are important to the experimental results, but some are not. So we need more in-depth investigation for better understanding and using the deep neural networks appropriately In summary, we developed a basic model for question similarity based on MaLSTM. It could be used for sentence similarity as well. And this model will be applied to another dataset in the near future. Using the new model, we can prospect for better results this year.

Chapter 2

Literature Survey

2.1 Outcome of Literature Survey

References describe the details of the approaches of engagement detection. Following are the outcome of Literature survey given in Table 2.1.

TABLE 2.1: SUMMARY OF EXISTING WORKS

Author	Methodology	Advantages	Disadvantages	Year
Borui Ye, Guangyu Feng, Anqi Cui, and Ming Li	RNN Encoder-Decoder	It works well if the input is small, The proposed model is capable of both classification and candidate ranking.	Hidden vector supposed to contain all the information of input sentence, It does not work well on long sentence since it will start losing out information	2017
Chao An, Jiuming Huang, Shoufeng Chang, and Zhijie Huang	Bidirectional Long Short-Term Memory (BLSTM) neural networks	It works well on long sentence since it uses memory to store long dependencies, bidirectional LSTMs have information about past and future to detect similarity.	Adopted fully connected neural networks uniformly as similarity measurement module, It can not be used where you cannot wait for the whole sequence before starting inference.	2016
Jonas Mueller, and Aditya Thyagarajan	Siamese Manhattan LSTM	It is capable of modeling complex semantics if the representations are explicitly guided.	Relies on pre-trained word-vectors as the LSTM inputs, it will benefit from improvements in the word-embedding method.	2016

2.2 Problem Statements

2.2.1 Objectives

In this project, we aim to develop machine learning and natural language processing system to classify whether question pairs are duplicates or not.

In other words, this semantic question matching problem can be defined as follows: for question pair q1 and q2, train a deep learning model to predict the function:

$$f(q1, q2) \rightarrow 0 \text{ or } 1$$

0 represents that q1 and q2 are not duplicate.

Chapter 3

Methodology

3.1 Manhattan LSTM Model

The proposed Manhattan LSTM (MaLSTM) model is outlined in Figure 3.1 and Figure 3.2. There are two networks $LSTM_a$ and $LSTM_b$ which each process one of the sentences in a given pair, but we solely focus on siamese architectures with tied weights such that $LSTM_a = LSTM_b$ in this work. Nevertheless, the general untied version of this model may be more useful for applications with asymmetric domains such as information retrieval (where search queries are stylistically distinct from stored documents).

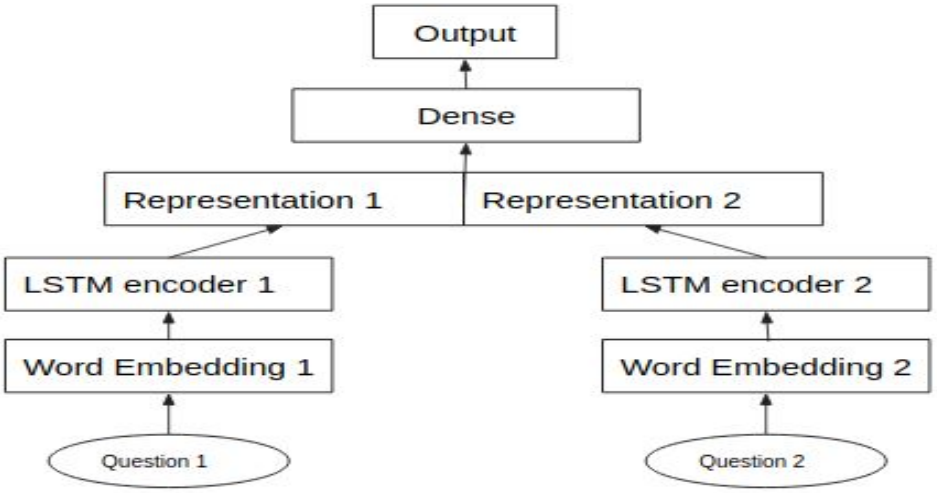

Fig. 3.1 Flowchart of the proposed model

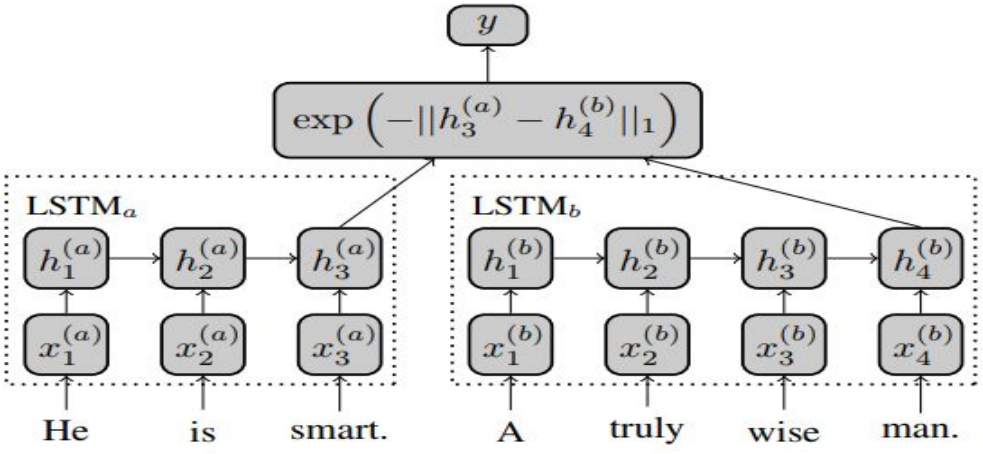

Fig. 3.2 Our model uses an LSTM to read in word-vectors representing each input sentence and employs its final hidden state as a vector representation for each sentence.

More concretely, each sentence (represented as a sequence of word vectors) x1,..., xT, is passed to the LSTM, which updates its hidden state at each sequence-index via equations.

Note that unlike typical language modeling RNNs, which are used to predict the next word given the previous text, our LSTMs simply function as the encoder.

We restrict ourselves to the simple similarity function as given below

$$g(h_{T_a}^{(a)}, h_{T_b}^{(b)}) = \exp(-||h_{T_a}^{(a)} - h_{T_b}^{(b)}||_1) \in [0, 1].$$

Contributions to the system are zero-cushioned successions of word records. These sources of info are vectors of fixed length, where the initial zeros are being disregarded and the nonzeros are lists that remarkably recognize words. Those vectors are then sustained into the installing layer. This layer looks into the relating inserting for each word and embodies all them into a lattice. This lattice speaks to the given content as a progression of embeddings. We utilized Google's word2vec implanting, same as in the first paper. The procedure is portrayed in Fig. 3.4.

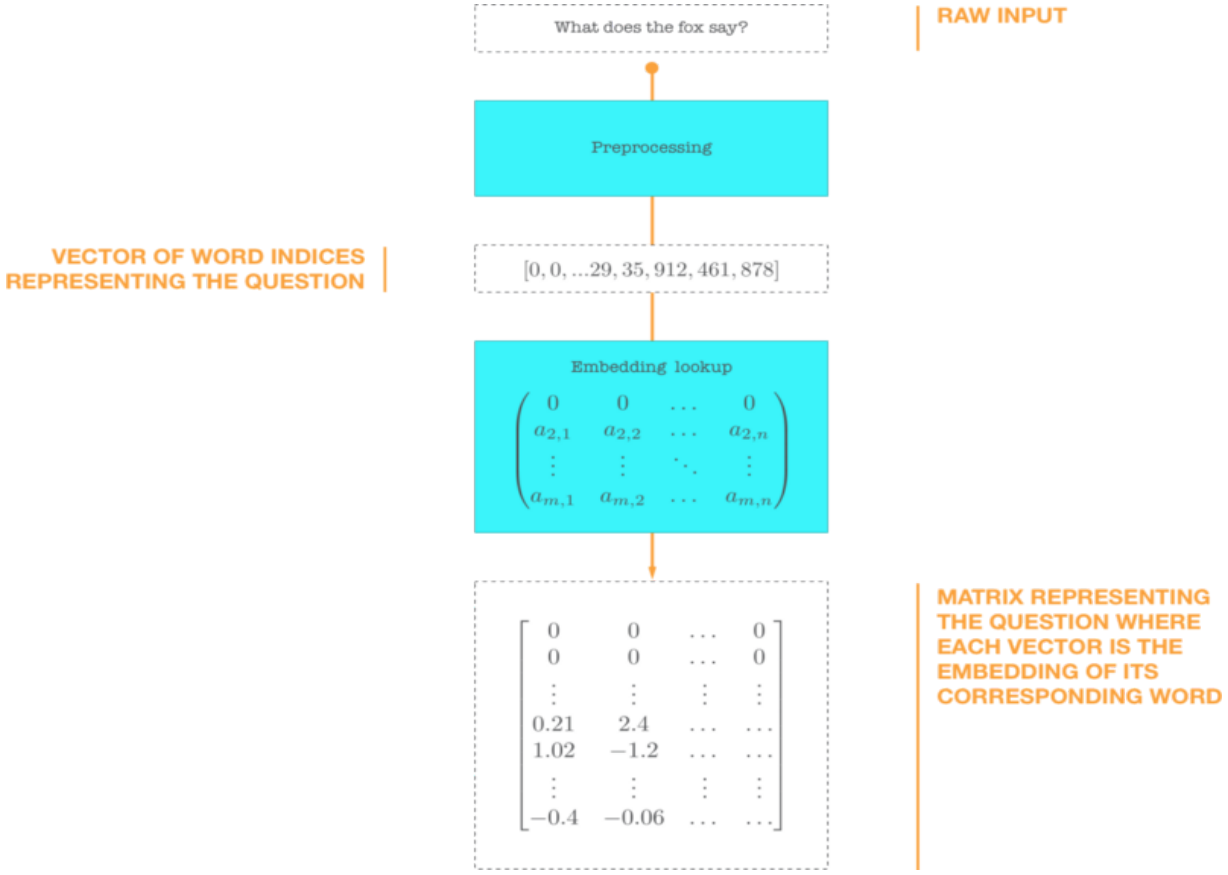

Fig 3.3 Embedding process

3.2 Dataset

There are over 400,000 lines of potential question duplicate pairs. Each line contains IDs for each question in the pair, the full text for each question, and a binary value that indicates whether the line truly contains a duplicate pair.

1. Id the id for each question pair
2. Qid_1 the id for question 1 in the pair
3. Qid_2 the id for question 2 in the pair
4. Question1 the full text for question1
5. Question2 the full text for question2
6. Is_duplicate yes(1) or no(0)

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?	0
2	2	5	6	How can I increase the speed of my internet connection while using a VPN?	How can Internet speed be increased by hacking through DNS?	0
3	3	7	8	Why am I mentally very lonely? How can I solve it?	Find the remainder when [math]23^{24}[/math] is divided by 24,23?	0
4	4	9	10	Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?	Which fish would survive in salt water?	0
5	5	11	12	Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1

Fig 3.4 Screenshot of the dataset

Chapter 4

Results and Discussion

To properly evaluate the model performance, lets plot training data vs validation data accuracy and loss as shown in Fig. 4.1 and Fig 4.2.

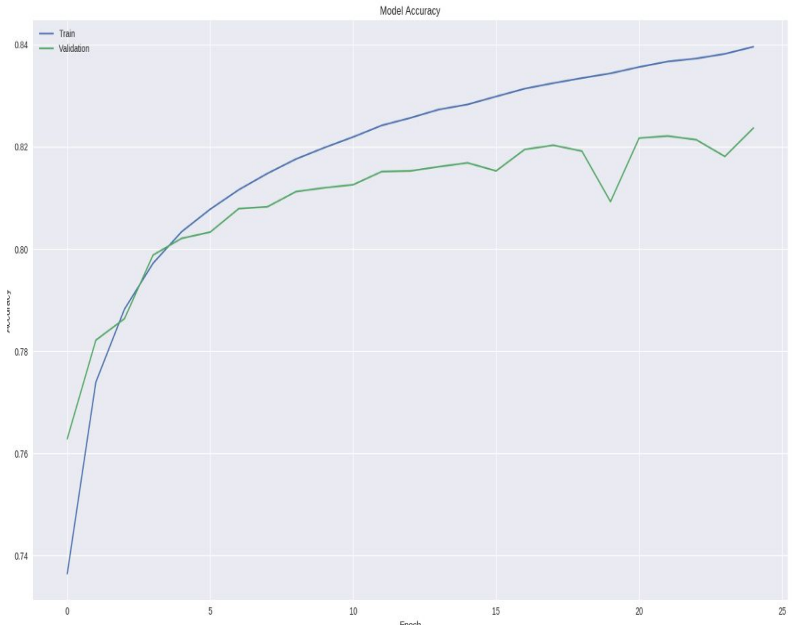

Fig. 4.1 Accuracy Graph

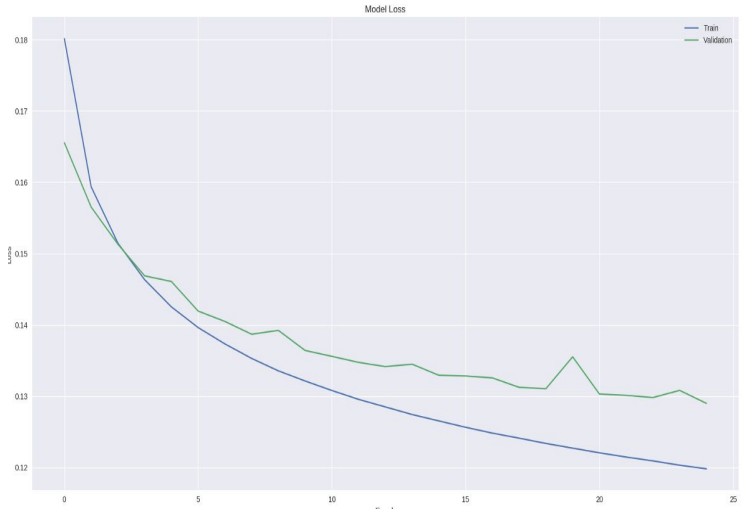

Fig. 4.2 Loss Graph

Table 4.1 Accuracy Result for a different model

Model	Accuracy	Recall	F1
RNN	0.678	0.581	0.593
BLSTM	0.726	0.683	0.668
MaLSTM	0.793	0.671	0.658

The above table 4.1 shows the comparison of different model based on accuracy, recall, and F1 value. From the table, it can be considered that the MaLSTM model is more accurate as compared to the RNN and BLSTM model.

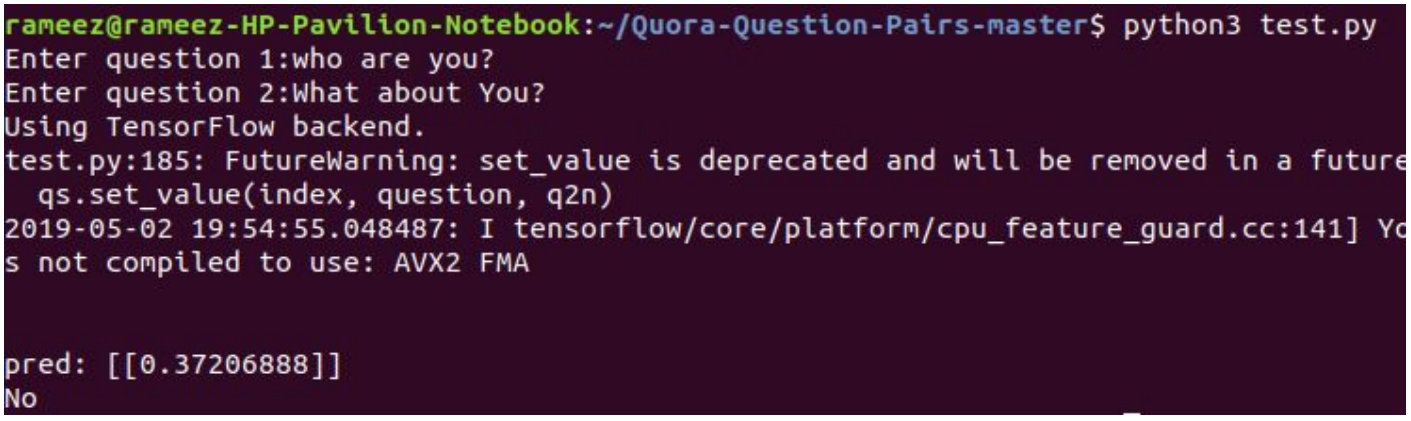

Fig 4.3 The Output of the proposed model

The above output in Fig 4.3 shows a similar value of two questions.

Chapter 5

Conclusion and Future Work

On the Quora pair dataset, we tested several deep learning methods based on RNNs to compute question similarity. And about the question modeling, we adopted two kinds of architectures, connected and separated. Additionally, from the result LSTM is better than RNN. So deep neural networks do not necessarily have to be highly deep. In fact, deep learning is a new thing. During the tuning of model parameters, we tried several sets of parameters, some are important to the experimental results, but some are not. So we need more in-depth investigation for better understanding and using the deep neural networks appropriately In summary, we developed a basic model for question similarity based on MaLSTM. It could be used for sentence similarity as well. And this model will be applied to another dataset in the near future. Using the new model, we can prospect for better results this year. Furthermore, another dataset in future work could extend this model to related tasks including question answering, paraphrase detection, paraphrase generation and so on.

References

[1] Borui Ye, Guangyu Feng, Anqi Cui, Ming Li.Learning Question Similarity with Recurrent Neural Networks, IEEE International Conference on Big Knowledge (ICBK), 2017.

[2] Chao An, Jiuming Huang, Shoufeng Chang, Zhijie Huang. Question Similarity Modeling with Bidirectional Long Short-Term Memory Neural Network, IEEE First International Conference on Data Science in Cyberspace (DSC), 2016.

[3] Jonas Mueller; Aditya Thyagarajan; Siamese Recurrent Architectures for Learning Sentence Similarity, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2016.