Quora Duplicate Question Pair Detection

Sakya Basak , Gururaj K , Rahul Bansal

Abstract

Determining whether two questions are asking the same thing can be challenging, as word choice and sentence structure can vary significantly. Traditional natural language processing techniques such as shingling[1] have been found to have limited success in separating related question from duplicate questions. Using a dataset of 400,000 labeled question pairs provided by question-and-answer forum Quora, we explore a series of deep learning methodologies for detecting duplicate question pairs. We explore 3 approaches based on LSTM networks on this data set. The first model uses a Siamese architecture with the learned representations from a single LSTM running on both sentences. The second method uses two LSTMs with the two sentences in sequence, and the second attending on the first (word-by-word attention) and the third one is a hybrid model. It has been seen that these deep learning techniques outperform traditional NLP methods and simple neural network baseline.

Introduction

Question-and-answer websites such as Quora provide users with a platform to ask questions that other users on the site may answer. However, many of the questions being asked at any given time have already been asked by other users, usually with different wording or phrasing. Ideally, these duplicate questions would be merged together into a single canonical question. A key element for question and answer website efficiency such as Quora relies on properly categorizing questions such no questions with identical intent are duplicated. In order to detect duplication we need to understand semantic relatedness of sentences. Detecting duplicate questions and merging them into one saves time, makes searches efficient and prevents redundancy in users' feeds.

Problem Statement

In this project, we address the problem of detecting duplicate and semantically equivalent questions in a supervised learning setting . This problem makes it necessary for us to recognize semantic relations between pairs of questions using neural networks architecture. LSTM[2] and Attention based RNNs or LSTM with siamese architecture are successful to identify entailment relationship between sentences[3].

In this project, we will try to leverage LSTM based recurrent architecture to identify semantically equivalent questions. Some of the challenges include (i) Question pairs may share limited common words but might be semantically similar (ii) Question pairs may share a lot of words in common differing only by few but might be semantically quite different . Our Model tries to learn these patterns.

Related Work

In traditional natural language processing (NLP), w-shingling[1] has been successfully used to quantify the similarity between two text documents. However, duplicate questions can be rephrased in many ways, and so techniques such as this that rely on word overlap fall short for this task. Over the years, CNNs have shown significant improvement in this task over traditional NLP techniques[4]

However in recent times[5], Wang applied bidirectional LSTMs to the problem of duplicate question pair detection and achieved state of the art results. This work motivates us to apply LSTM encoding to this task. Furthermore, latest research on RNN models[6] have shown that attention mechanism focus only on specific words during training[6]. This facilitates our problem by ensuring that only relevant words are taken into account.

Solution Sketch

We plan to implement multiple architectures and compare their performances. However the higher level idea for performing the task is as follows:
The raw questions are converted into pre trained word embedding using Glove in the embedding layer.The embedding matrix is then passed into the encoding layer which consists of a Siamese network of LSTMs. The output of the hidden network of these LSTMs are combined through appropriate linear transformation,which may vary in the various architectures that we plan to implement. Finally we plan to use softmax output and cross-entropy loss for training.
In case of sequence to sequence based LSTMs, we plan to implement global attention layer using LSTM cell of question 1 which acts as a context vector for LSTM cell of question 2. The final hidden layer output represents the concatenation of output from LSTM cell of question 2 with corresponding context vector. We also plan to experiment with an additional LSTM layer on top of the previous final layer. Here also we use cross-entropy loss for training.

Datasets and Metrics

We plan to use the dataset released by Quora [1] that consists of over 400,000 lines of potential duplicate pairs. Each pair of questions we have a sample ID, an individual question ID, the questions and their corresponding labels (duplicate = 1, not duplicate = 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

Figure 1: Sample lines of the dataset

As for metrics we have decided to list the precision,recall,F1 score and accuracy of the model both wrt training set and test set.

Baselines

As a baseline we plan to use a multilayer perceptron whose input is a concatenated feature vector consisting of 2 vectors of dimension d where d is the word embedding dimension and each vector is the average of the word embeddings in the respective question. This produces a bag of words input vector. The concatenated vectors are then fed through a fully connected layer that outputs the confidence values for the two classes as the final output.In order for our model to be considered non-trivial, we should expect an accuracy score above 63 percent,which is the number of non duplicate examples[7] to account for the bias in the labels, otherwise our models could simply label all samples as non-duplicate and achieve this performance.
The accuracy score of the baseline model is 0.7263 and the F1 score is 0.6490.[8]
Among the best implementations to date is an LSTM RNN architecture constructed by Quora that has an accuracy of 86 percent[7]

Experiment Hypothesis

We preprocess the data ,doing a sanity check on the question pairs present and plan to divide into a 90/10 split between training and test. We plan to use GloVe pre-trained word vectors to initialize our word embeddings[9]. In assessing the performance of our models, we consider primarily accuracy metric, but also report precision ,recall, and F1 score.

References

[1] A. Broder. On the resemblance and containment of documents. pages 21–, 1997.

[2] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *CoRR*, abs/1503.00075, 2015.

[3] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. Reasoning about entailment with neural attention. *CoRR*, abs/1509.06664, 2015.

[4] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics, 2014.

[5] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *CoRR*, abs/1702.03814, 2017.

[6] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[7] Natural language inference for quora dataset.

[8] Baseline models.

[9] Glove vector embeddings: https://nlp.stanford.edu/projects/glove.