QUORA DUPLICATE QUESTION DETECTION

Sri Phani Gorti
sgorti3@uic.edu

Saikrishna Kalahasti Karthik
skalah2@uic.edu

Malavika Ramprasad
mrampr2@uic.edu

ABSTRACT

In this paper, we introduce a siamese model for analysing and evaluating semantic similarity of a couple of sentence. We implemented this model on Quora dataset, where in we try to identify duplicate questions. This model makes use of a simaese network of Bi-Directional LSTMs trained on the vector representation of questions. The resultant vectors from the model were compared using Manhattan distance to analyse the semantic similarity. We achieved a training accuracy of 82.75%.

Keywords

Deep Learning; Semantic Similarity; Duplicate detection; Siamese Network; Bi-LSTM; Machine Learning.

1. INTRODUCTION

Understanding deep NLP techniques has recently become the focus of active research. Duplicate question detection is one such task which involves identifying the semantic meaning between two segments and thus marking it duplicate. Motivation for this research topic is coming from the usefulness of resorting to it to support several other active research such as answering selection, textual entailment, and also conversational interfaces, in general.

For instance, when applied to the answering selection, duplicate question detection can measure the semantic similarity between the question and the pool of answer list to select the best matched answer.

And when using in conversational interface, it can be used to compare the newly entered question with the existing question-answer pair in the database and reply with the corresponding stored answer if similar question is found. Thus, helping to avoid human intervention.

In this paper we address the problem of actual duplication or exact semantic coincidence between questions. Solving this problem, at the very least, will turn out to be useful for any question answering forum such as Quora, StackExchange in order to organize and deduplicate their knowledge base. We propose a Siamese Bi-LSTM framework which uses Manhattan distance to calculate the similarity between the question pair in the Quora data set.

2. RELATED WORK

Due to its importance across diverse applications, semantic similarity evaluation was selected as the first task of SemEval 2014, where numerous researchers applied methods to a labeled dataset containing pairs of sentences involving compositional knowledge (SICK) [3]

Previous work on semantic relatedness of sentences has focused on logical inference and entailment through based on the Stanford Natural Language Inference Corpus [2]. This work [3] proposes an siamese LSTM framework that is applied to assess semantic similarity between sentences using a fixed size vector to encode the underlying meaning expressed in a sentence.

In [4], the work that is proposed is a Siamese based LSTM network for labeled data comprised of pairs of variable-length sequences. This model is applied to assess semantic similarity between sentences from quora dataset pair.

3. APPROACH

Our proposed model, Manhattan Bi-LSTM is depicted in Fig 1. It consists of two Bi-LSTM network which processes each sentence separately. Since it is a Siamese architecture, both the network Bi-LSTM$_a$ and Bi-LSTM$_b$ have the same parameters. Bi-LSTM$_a$= Bi-LSTM$_b$.

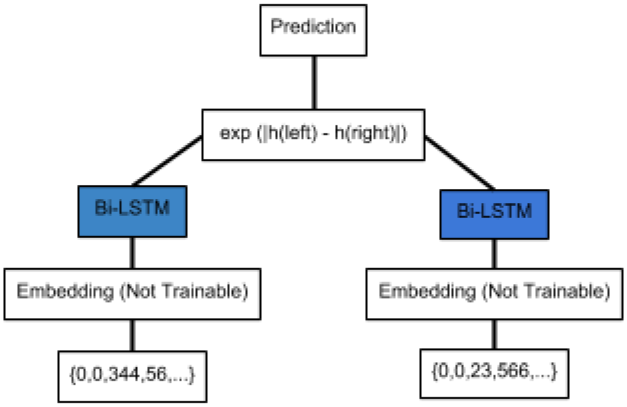
Prediction
exp (||h(left) - h(right)||)
Bi-LSTM
Bi-LSTM
Embedding (Not Trainable)
Embedding (Not Trainable)
{0,0,344,56,...}
{0,0,23,566,...}

Figure 1: Our model uses a Bi-LSTM to read in word-vectors representing each input sentence and employs its final hidden state as a vector representation for each sentence. Subsequently, the similarity is found using Manhattan distance.

DATASET

The dataset used for this analysis was provided by Quora, released as their first public dataset. It consists of 404352 question pairs in a tab-separated format:

- id: unique identifier for the question pair (unused)
- qid1: unique identifier for the first question (unused)
- qid2: unique identifier for the second question (unused)
- question1: full text of the first question
- question2: full text of the second question
- is_duplicate: binary value 1 if questions are duplicates, 0 otherwise

Of the over 400K question pairs, 149302 are duplicates, or roughly 37% of the full dataset.

In our model we split the data into approx 10% for validation and 90% for test.

DATA PREPROCESSING

The dataset approximately has 405,000 question pairs. The words in the questions were first converted to lowercase letters in order to eliminate any word dissimilarity due to irregular capitalization. The work observed some inconsistencies with respect to the usage of words, such as short forms, for instance, what's and what is, or multiple representation for the same word, for instance, eg or e g.. Due to these reasons, the dataset was cleaned using a simple regular expression, and removed multiple representation of the same word/phrase. The punctuations were removed, as it would not significantly affect the meaning/end result, but only disambiguate existing words. Words that do not add any significant meaning, or does not create compelling semantic shifts such as stop words were removed from the dataset.

WORD EMBEDDING

The vector representations for the words were done using Word2Vec. Several approaches of word embeddings, including gloVe, tf-idf, tf-idf with appended synonyms, were attempted. However, Word2Vec seemed to give better results over other vector representations. Each word is represented as a 300-size vector. In order to represent a sentence, question in our case, we represented it in a {(n+1) x 300} vector space, where n is the size of the dictionary created from the entire training dataset.

SIAMESE BI-LSTM

The proposed Manhattan Bi-LSTM model consists of two similar Bi-LSTM network. Each sentence (represented as a sequence of word vectors) $x_1,...,x_T$, is passed to the Bi-LSTM, which updates its hidden state at each sequence. These inputs to the network are zero-padded sequences of word indices. These inputs are vectors of fixed length, where the first zeros are being ignored and the non-zeros are indices that uniquely identify words.

The vectors from the Bi-LSTM that hold the semantic meaning of each question is then passed through the similarity function (Manhattan distance) as defined below.

$$\exp(-\|h1^{(a)} - h2^{(b)} \|_1)$$

LOSS FUNCTION

Our model predicts the similarity for a given pair of questions, and we train the siamese network using backpropagation-through-time under the mean squared-error (MSE) loss function (after rescaling the training-set relatedness labels to lie ∈ [0, 1]).

4. RESULT

TRAINING

We used Tensor Flow's Bi-LSTM to be able to unroll the model by sentence length within batches. We used the AdamOptimizer implemented in tensorflow and tried a different learning rates.

In running our experiments, we used 64 sample batch sizes and trained for a total of 20 epochs. We chose N=50 as our sequence length for all models. Adam optimization was used. A learning rate of 0.01 was chosen, which turned out to be better than the other learning rates that was tried. Softmax cross entropy was chosen for the loss. All of the implementation was written in Tensor Flow, with separate run scripts for training and evaluation.

Hybrid LSTM	Siamese Manhattan LSTM	BiMPM on Quora dataset	Siamese Manhattan with Bi-LSTM
81.05	82.5	88.17	82.75

Table 1: Dev accuracy comparison

Table 1 shows the dev accuracy rate among few models that were implemented for quora data set. Our model out performed the Siamese manhattan LSTM based network.

CONCLUSION

In this project, we produced competitive results on the Quora duplicate dataset problem using Bi-LSTMs. In the given time we reached 82.75% accuracy. The Siamese architecture outperformed the sequence-to-sequence, though the attention methods were also well above the bag of words baseline. It also outperformed the existing Siamese model which uses two LSTM network.

5. REFERENCES

[1] Chakaveh Saedi, Joao Rodrigues, Joao Silva, Antonio Branco, Vladislav Maraev. Learning Profiles in Duplicate Question Detection

[2] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In EMNLP, 2015.

[3] Marelli et. al. - SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment

[4] Jonas Mueller, Aditya Thyagarajan; Siamese Recurrent Architectures for Learning Sentence Similarity

[5] Elkhan Dadashov, Sukolsak Sakshuwong, Katherine Yu. ; Quora Question Duplication

[6] Dasha Bogdanova , Cıicero dos Santos,, Luciano Barbosa and Bianca Zadrozny;Detecting Semantically Equivalent Questions in Online User Forums

[7] Shankar Iyar, Nikhil Dandekar, and Kornél Csernai. "First Quora Dataset Release: Question Pairs," 24 January 2016.

[8] Zhiguo Wang, Wael Hamza and Radu Florian. "Bilateral Multi-Perspective Matching for Natural Language Sentences,"

[9] Lili Jiang, Shuo Chang, and Nikhil Dandekar. "Semantic Question Matching with Deep Learning," 13 February 2017

[10] Eren Golge. "Duplicate Question Detection with Deep Learning on Quora Dataset," 12 February 2017.