Shanghai Jiao Tong University

X033525

Machine Learning

Quora Question Pairs @ Kaggle

Author:	*Student Number:*
Yuliang Xiu	*116033910045*
Xiaoting Jiang	*116033910035*
Weiyu Cheng	*016033910007*
Bowen Zhang	*016033910035*

June 7, 2017

Contents

1 Problem Description

1.1 Background

Where else but Quora can a physicist help a chef with a math problem and get cooking tips in return? Quora is a place to gain and share knowledge—about anything. It's a platform to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.

Over 100 million people visit Quora every month, so it's no surprise that many people ask similarly worded questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both of these groups in the long term.

Currently, Quora uses a Random Forest model to identify duplicate questions. In this competition, Kagglers are challenged to tackle this natural language processing problem by applying advanced techniques to classify whether question pairs are duplicates or not. Doing so will make it easier to find high quality answers to questions resulting in an improved experience for Quora writers, seekers, and readers.

1.2 Definition

More formally, the duplicate detection problem can be defined as follows: given a pair of questions q1 and q2, train a model that learns the function, where 1 represents that q1 and q2 have the same intent and 0 otherwise:

$$f(q_1, q_2) \rightarrow 0 \quad or \quad 1$$

Submission is evaluated by logloss between predicted value and ground truth:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(P_{ij})$$

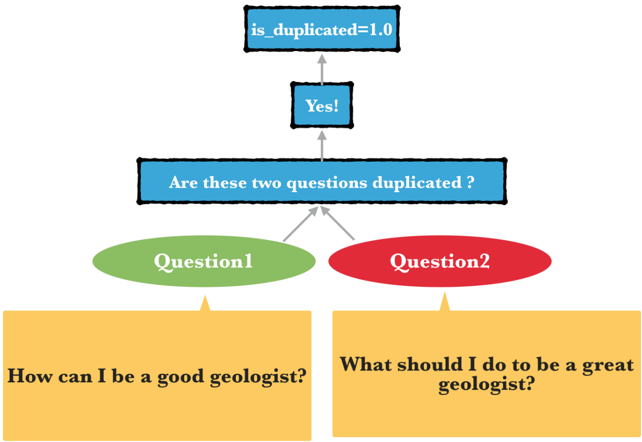


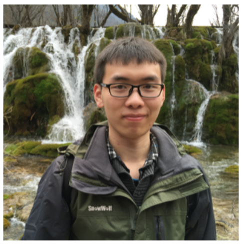
Figure 1: Problem definition diagram

2 Related Work

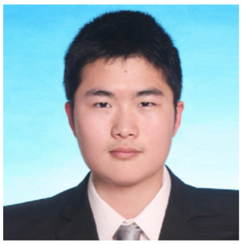
Some paraphrase identification related papers:[6][4][10][4][2][9][1][3]

Model	Source of Word Embeddings	Accuracy
"BiMPM model" [5]	GloVe Common Crawl (840B tokens, 300D)	**0.88**
"LSTM with concatenation" [6]	"Quora's text corpus"	0.87
"LSTM with distance and angle" [6]	"Quora's text corpus"	0.87
"Decomposable attention" [6]	"Quora's text corpus"	0.86
"L.D.C." [5]	GloVe Common Crawl (840B tokens, 300D)	0.86
Max bag-of-embeddings (*this work*)	GloVe Common Crawl (840B tokens, 300D)	0.83
"Multi-Perspective-LSTM" [5]	GloVe Common Crawl (840B tokens, 300D)	0.83
"Siamese-LSTM" [5]	GloVe Common Crawl (840B tokens, 300D)	0.83
"Neural bag-of-words" (max) [7]	GloVe Common Crawl pruned to 1M vocab. (spaCy default)	0.83
"Neural bag-of-words" (max & mean) [7]	GloVe Common Crawl pruned to 1M vocab. (spaCy default)	0.83
"Max-out Window Encoding" with depth 2 [7]	GloVe Common Crawl pruned to 1M vocab. (spaCy default)	0.83
"Neural bag-of-words" (mean) [7]	GloVe Common Crawl pruned to 1M vocab. (spaCy default)	0.81
"Multi-Perspective-CNN" [5]	GloVe Common Crawl (840B tokens, 300D)	0.81
"Siamese-CNN" [5]	GloVe Common Crawl (840B tokens, 300D)	0.80
"Spacy + TD-IDF + Siamese" [8]	GloVe (6B tokens, 300D)	0.79

3 Team Division



Bowen Zhang
DDST Ph.D.



Weiyu Cheng
DDST Ph.D.

Xiaoting Jiang
DDST M.Phil.

Yuliang Xiu*
MVIG M.Phil.

* Team leader

Original Dataset Cleaning	Manual Feature Engineering	Sentence to Vector Embedding	Classifier Design and Boosting

4 Dataset Preprocessing

4.1 Dataset Attribute

- Training dataset size : 404290
- Testing dataset size: 2345896
- Dataset format:

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

4.2 Data Cleaning

- **Correct wrong labels:** There are many wrong labeled records in original dataset, which are caused by redundant blank. Before training phrase, we must correct these records to make sure all records are labeled correctly.

  question1: How can I be a good geologist?

  question2: How can I be a good geologist   ?

  i_duplicated: 0 $\rightarrow$ 1

- **Remove special characters:** $\aleph$, $\clubsuit$, © . . .
- **Change abbreviation:** what's $\rightarrow$ what is   can't $\rightarrow$ cannot
- **Use standard format:** e-mail $\rightarrow$ email
- **Replace reference words:** Due to quora is an american QA website, so when users use "us", it always refers to american, so we change all us $\rightarrow$ american.

5 Manual Feature Extraction

We extract 28 kinds of manual features between ques1 and ques2 as follows:

<table>
<tr><th>Manual Features</th><th>Description</th><th>Manual Features</th><th>Description</th></tr>
<tr><td>word_match</td><td>number of common words</td><td>fuzz_qratio</td><td rowspan="7">Fuzz Ratio</td></tr>
<tr><td>tf_idf_word_match</td><td>sum(tf_idf [q1&q2])/sum(tf_idf [q1||q2])</td><td>fuzz_wratio</td></tr>
<tr><td>sentiment</td><td>sentiment (positive or negative)</td><td>fuzz_partial_ratio</td></tr>
<tr><td>diff_len</td><td rowspan="3">Length Difference</td><td>fuzz_partial_token_set_ratio</td></tr>
<tr><td>diff_len_char</td><td>fuzz_partial_token_sort_ratio</td></tr>
<tr><td>diff_len_word</td><td>fuzz_token_set_ratio</td></tr>
<tr><td>q1_freq</td><td>frequence of question1 in all data</td><td>fuzz_token_sort_ratio</td></tr>
<tr><td>q2_freq</td><td>frequence of question2 in all data</td><td>cosine_distance</td><td rowspan="7">Distance</td></tr>
<tr><td>q1_q2_intersect</td><td>intersection of q1&q2 related sents</td><td>cityblock_distance</td></tr>
<tr><td>q1_q2_wm_ratio</td><td>sum(inter [q1&q2])/sum(inter [q1||q2])</td><td>jaccard_distance</td></tr>
<tr><td>wmd</td><td>Word Mover's Distance</td><td>canberra_distance</td></tr>
<tr><td>norm_wmd</td><td>norm of Word Mover's Distance</td><td>euclidean_distance</td></tr>
<tr><td>skew_vec</td><td>skewness of sentence</td><td>minkowski_distance</td></tr>
<tr><td>kur_vec</td><td>kurtosis of sentence</td><td>braycurtis_distance</td></tr>
</table>

Figure 2: Manual feature table

6 Sentence Embedding

During sentence embedding phrase, we firstly remove stop words and punctuations from question sentence, and then do stemming on the rest sentence. Secondly, we use keras *text_to_sequence* function to convert sentence to number sequence, and then do padding operation (30 dimensions) on sequence. Thirdly, use pre-trained embedding matrix such as **Glove**[7] or **FastTest**[5] to do embedding operation, which return a 30×300 matrix. Finally, put this matrix into Bidirection-LSTM[8] to extract sequential information from this sequence.

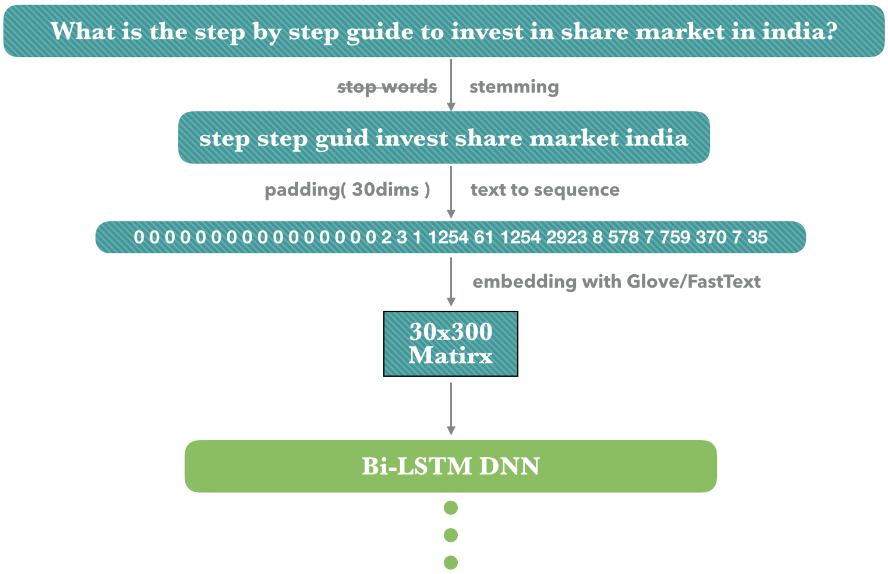


Figure 3: Sentence embedding procedure

```
embedding_layer = Embedding(nb_words,
        EMBEDDING_DIM,
        weights=[embedding_matrix],
        input_length=MAX_SEQUENCE_LENGTH,
        trainable=False)
lstm_layer =  Bidirectional(LSTM(num_lstm, dropout=rate_drop_lstm,
 ↪  recurrent_dropout=rate_drop_lstm))
```

7 Classification Model

Input of network is one question pair. To extract sequential information from questions, we use two separately Bidirection-LSTM on question1 and question2 embedding matrix. Meanwhile, we calculate manual features or traditional features between question1 and question2, and use fully connected layer to align dimension to 200. Finally we concatenate two sequential features and manual features together and put forward through one fully connected layer to get final predicted value.

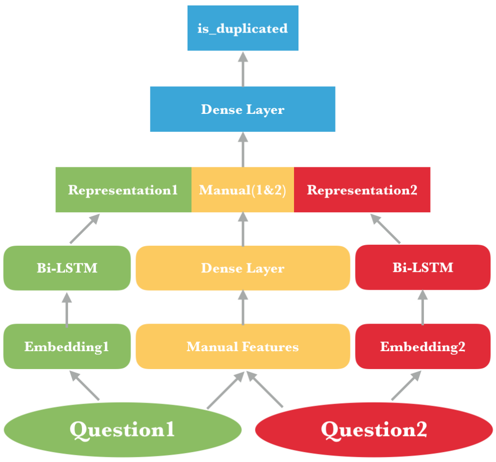
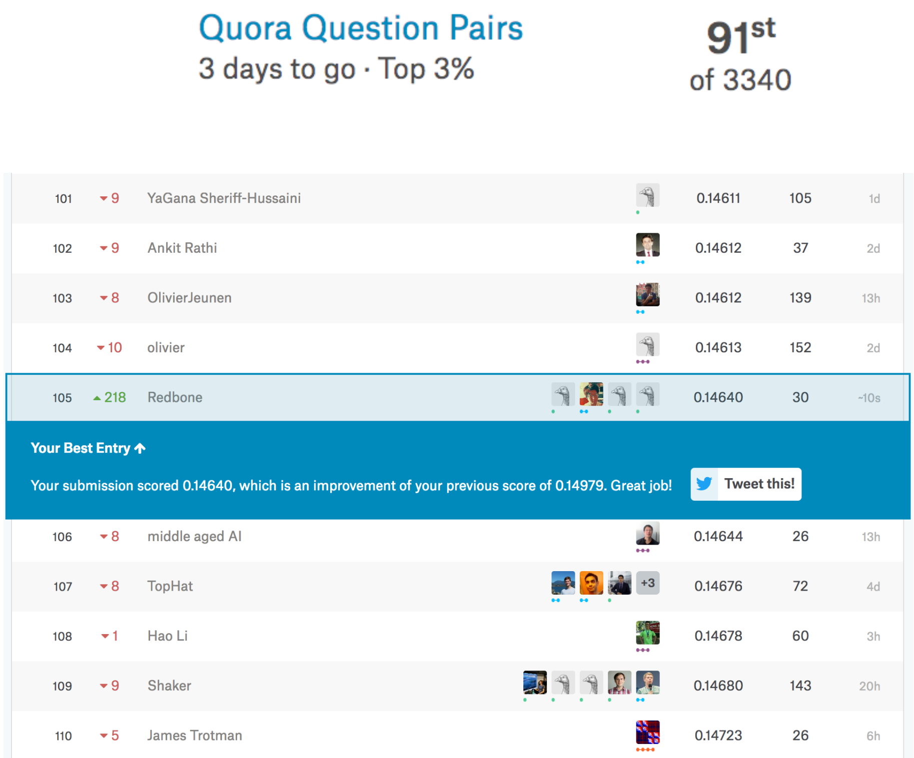


Figure 4: Classifier network framework

```
sequence_1_input = Input(shape=(MAX_SEQUENCE_LENGTH,), dtype='int32')
embedded_sequences_1 = embedding_layer(sequence_1_input)
x1 = lstm_layer(embedded_sequences_1)
sequence_2_input = Input(shape=(MAX_SEQUENCE_LENGTH,), dtype='int32')
embedded_sequences_2 = embedding_layer(sequence_2_input)
y1 = lstm_layer(embedded_sequences_2)
z1 = Input(shape=(x_train.shape[1],), dtype='float32')
z1_dense = Dense(num_dense/2, activation=act)(z1)
merged = concatenate([x1, y1, z1_dense])
merged = BatchNormalization()(merged)
merged = Dropout(rate_drop_dense)(merged)
merged = Dense(num_dense, activation=act)(merged)
merged = BatchNormalization()(merged)
merged = Dropout(rate_drop_dense)(merged)
preds = Dense(1, activation='sigmoid')(merged)
```

8 Final Result

- Best Rank: 91/3379 (top 3%)
- Prize: Silver Medal (within top 5%)
- Loss value: 0.14128
- Accuracy: 90.13%
- Competition page: Leaderboard of quora question pair
- Github code: kaggle_quora@github



Quora Question Pairs
3 days to go · Top 3%
91st
of 3340

101	▼9	YaGana Sheriff-Hussaini	0.14611	105	1d
102	▼9	Ankit Rathi	0.14612	37	2d
103	▼8	OlivierJeunen	0.14612	139	13h
104	▼10	olivier	0.14613	152	2d
105	▲218	Redbone	0.14640	30	~10s
106	▼8	middle aged AI	0.14644	26	13h
107	▼8	TopHat	0.14676	72	4d
108	▼1	Hao Li	0.14678	60	3h
109	▼9	Shaker	0.14680	143	20h
110	▼5	James Trotman	0.14723	26	6h

Your Best Entry ↑
Your submission scored 0.14640, which is an improvement of your previous score of 0.14979. Great job!
Tweet this!

Figure 5: Final rank

References

[1] *Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks*, 2015.

[2] *A Decomposable Attention Model for Natural Language Inference*, 2016.

[3] William Blacoe and Mirella Lapata. A Comparison of Vector-based Representations for Semantic Composition. *EMNLP-CoNLL*, 2012.

[4] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv.org*, page arXiv:1508.05326, August 2015.

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[6] Tomas Mikolov, Kai Chen 0010, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, cs.CL:arXiv:1301.3781, 2013.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove - Global Vectors for Word Representation. *EMNLP*, 2014.

[8] M Schuster and K K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45:2673–2681, November 1997.

[9] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *CoRR*, 2015.

[10] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral Multi-Perspective Matching for Natural Language Sentences. *CoRR*, cs.AI:arXiv:1702.03814, 2017.