

Predicting Quora Question Similarity with a Siamese deep MaLSTM network



Data Processing

“Can you identify question pairs that have the same intent?”

Number of question pairs for training: 323164
Total number of questions in the training data: 646328
Number of questions that appear multiple times: 119193
Total percentage of Duplicate pairs: 36.88%
Number of question pairs for testing: 81126
Total number of questions in the testing data: 162252

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

		question1	question2	is_duplicate
0		['step', 'step', 'guide', 'invest', 'share', 'market', ...]	['step', 'step', 'guide', 'invest', 'share', 'market', ...]	0
1		['story', 'kohinoor', 'koh', 'i', 'noon', 'diamond']	['would', 'happen', 'indian', 'government', 'stole', ...]	0
2		['increase', 'speed', 'internet', 'connection', 'using', 'vpn']	['internet', 'speed', 'increased', 'hacking', 'dns']	0
3		['mentally', 'lonely', 'solve', 'it']	['find', 'remainder', 'math', '23', '24', 'math', ...]	0
4		['one', 'dissolve', 'water', 'quickly', 'sugar', 'salt', ...]	['fish', 'would', 'survive', 'salt', 'water']	0

Exploratory Data Analysis

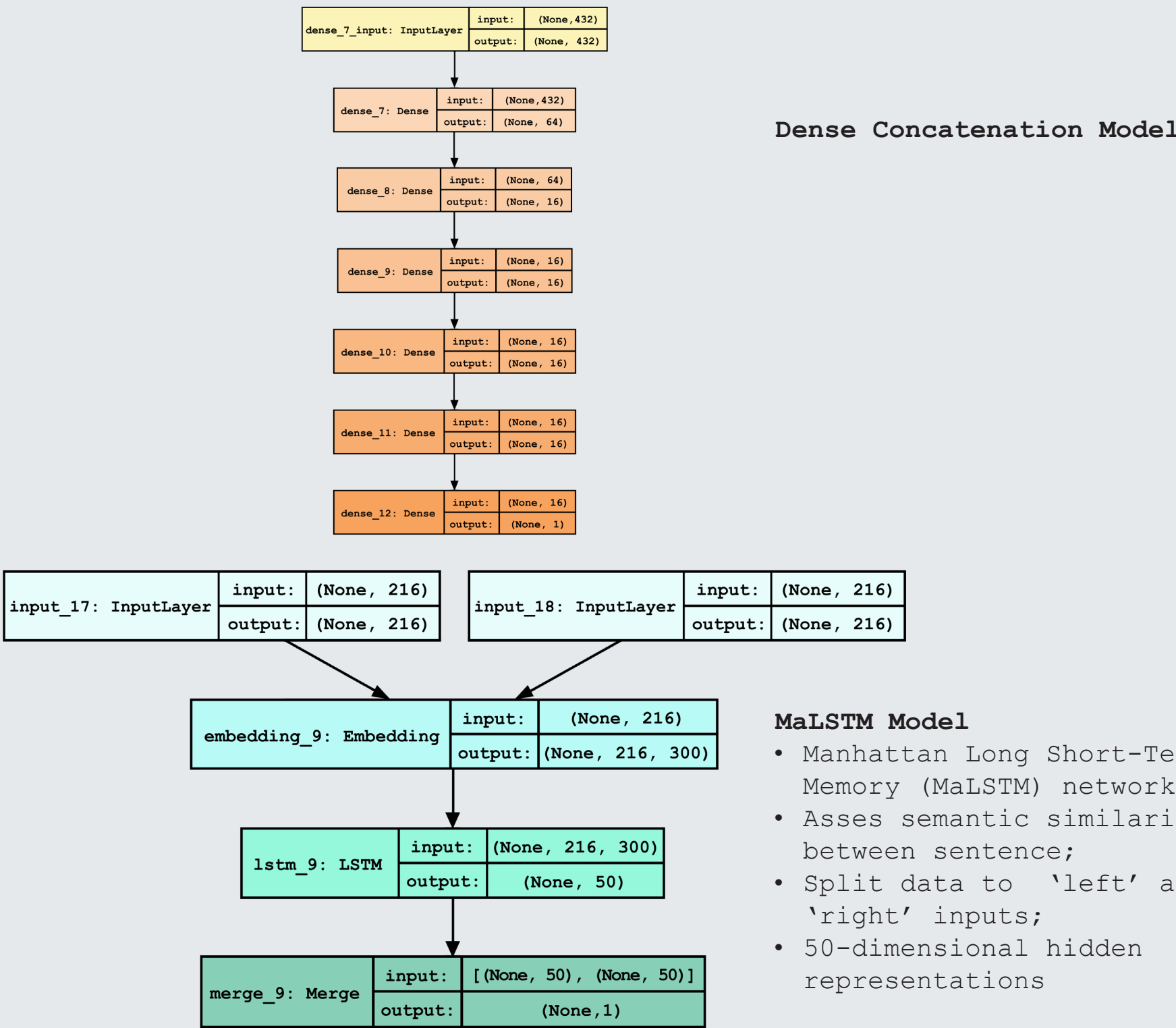
Pre-Processing

- Load the train, train labels and test data sets;
- Join the train labels with the train data set;
- Remove id rows from both train and test data sets.

Text Cleaning

- Convert words to lower case;
- Remove punctuation;
- Normalization (What's to What is)
- Remove stop words;
- Split sentence into words separated by white space.

Model Selection



Model Evaluation

Feature Extraction

		question1	question2	is_duplicate
0		[1, 2, 3, 4, 5, 4, 6, 7, 8, 9, 10, 8, 11]	[1, 2, 3, 4, 5, 4, 6, 7, 8, 9, 10]	0
1		[1, 2, 3, 12, 13, 14, 15, 16, 17]	[1, 18, 19, 20, 3, 21, 22, 23, 3, 13, 14, 15, ...]	0
2		[25, 26, 15, 27, 3, 28, 29, 30, 31, 32, 33, 34]	[25, 26, 30, 28, 35, 36, 5, 37, 38, 39]	0
3		[48, 41, 15, 42, 43, 44, 25, 26, 15, 45, 46]	[47, 3, 48, 49, 50, 51, 52, 50, 2, 53, 5, 52, 51]	0
4		[54, 55, 56, 8, 57, 58, 59, 60, 61, 62, 63, 64]	[54, 65, 18, 66, 8, 60, 57]	0

Words to Indices

- Convert words to indices;
- Start at index 1 to reserve 0 for zero padding.

Embedding

- Use Google's Word2Vec embedding to turn words into their embedding;
- Model with 300 dimensional vectors for 3 million words and phrases (pre-trained over about 100 billion words);
-

Model Training

Technical Information