Are two questions the same?

Dataset: Quora question pairs dataset

	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

404,290 question pairs. 37% of the question pairs are duplicated.

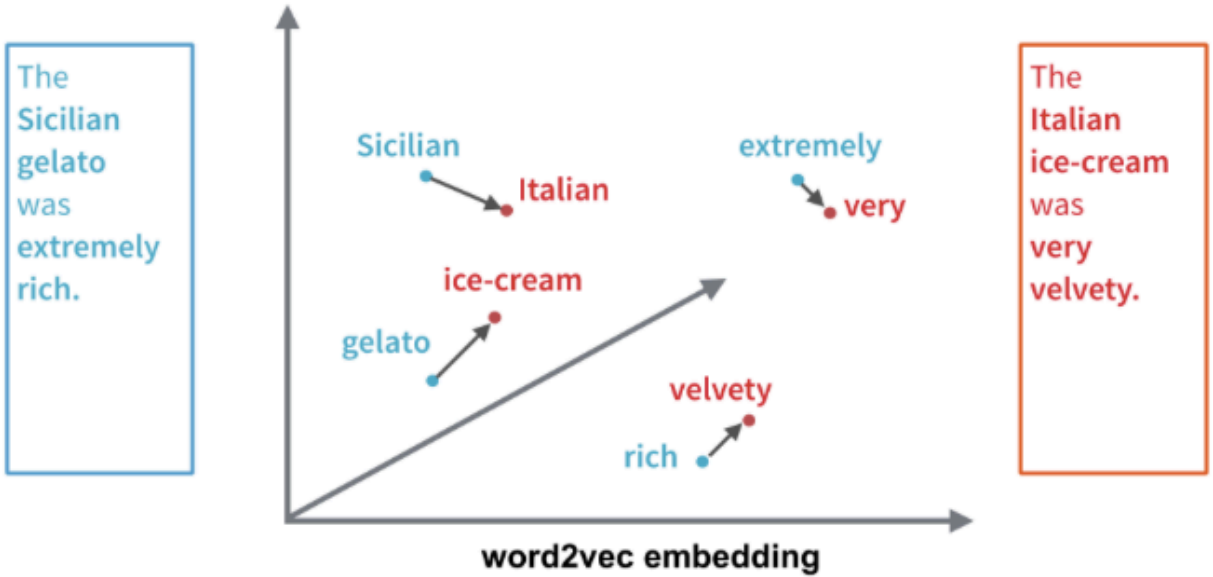
E.g. Not a duplicate question pair

```
('How can I increase the speed of my internet connection while using a VPN?'
 'How can Internet speed be increased by hacking through DNS?')
```

E.g. Duplicate question pair

```
('How can I be a good geologist?', 'What should I do to be a great geologist?')
```

Finding Duplicate question using Word Mover Distance (WMD)



Modeling steps

Preprocessing steps
Lowercase the questions -> Tokenize -> Select only words the English characters -> Remove stop words

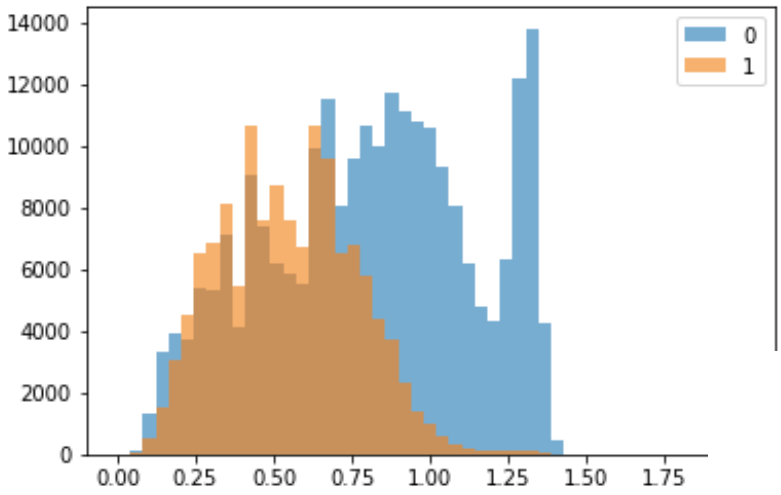
Load Google news vectors

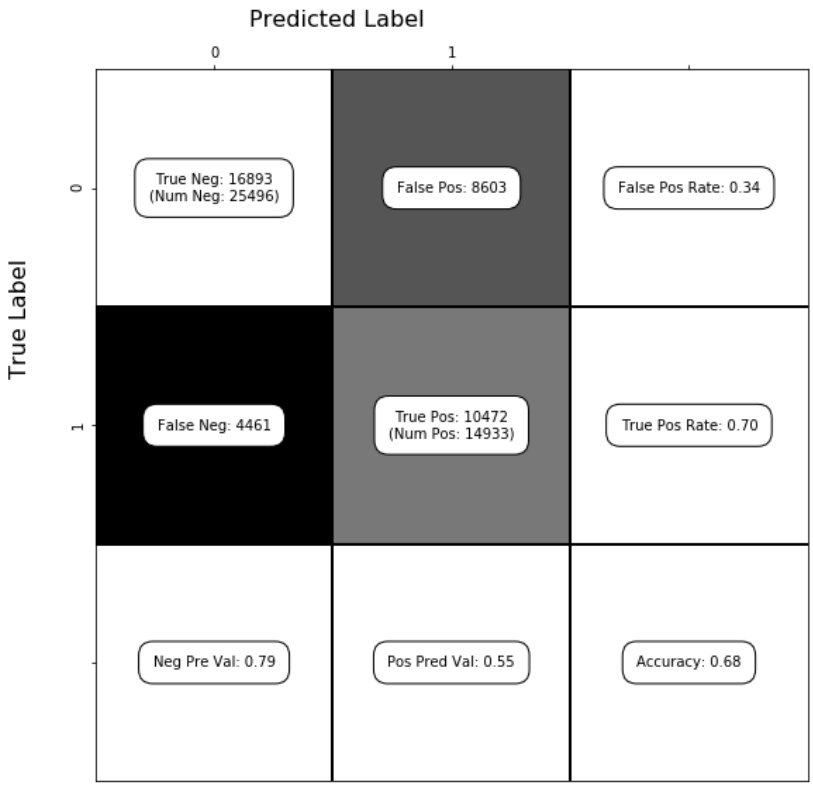
Calculate WMD

Split dataset in to train (90%) and validation (10%)

Use decision tree to predict duplicated question pairs using WMD

Distribution of WMD – with google vectors





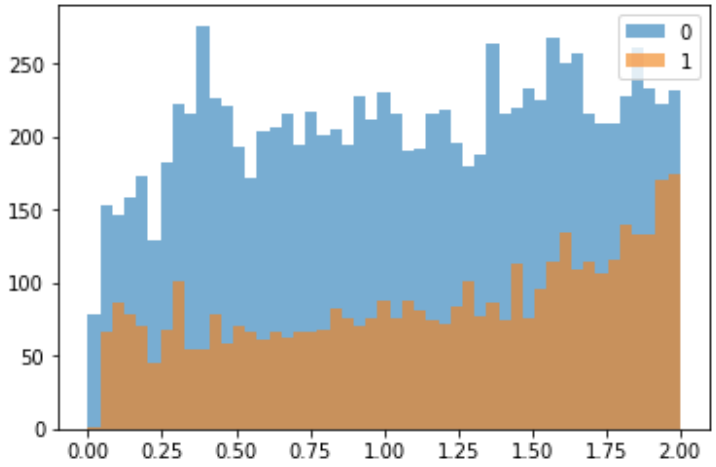
Results

Accuracy – 67.7%
(Baseline accuracy 63.1%)
Logloss – 0.5445

Alternative approach to WMD calculation

Train word embeddings using question pair corpus

Distribution of WMD – with new embeddings



No separation of distributions! Not useful!

Finding Duplicate Questions using a Siamese Manhattan LSTM

(As discussed by Elior Cohen in a Medium post)

Modeling steps

Preprocessing steps
Lowercase the questions -> Tokenize

Create the vocabulary of the questions and associate each word with a unique index

Convert each word in questions to their respective index

Load Google news vectors

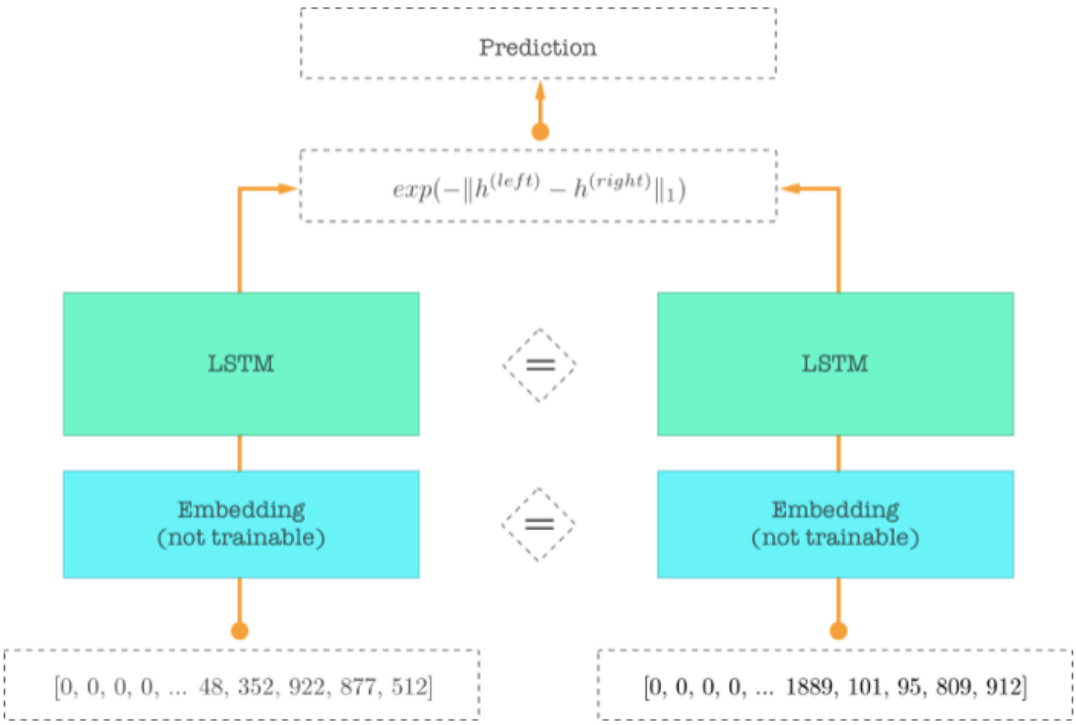
Create embedding matrix

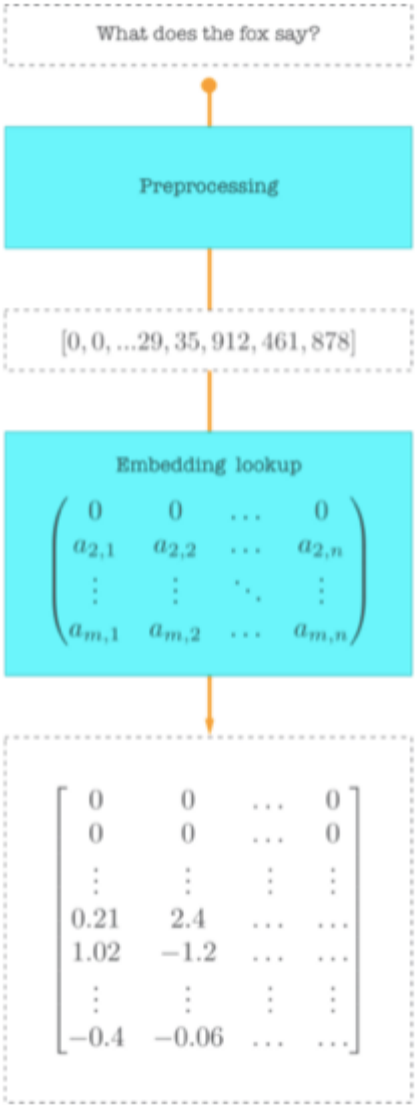
Zero padding of the questions

Split dataset in to train (90%) and validation (10%)

Fit MaLSTM model

Model Architecture

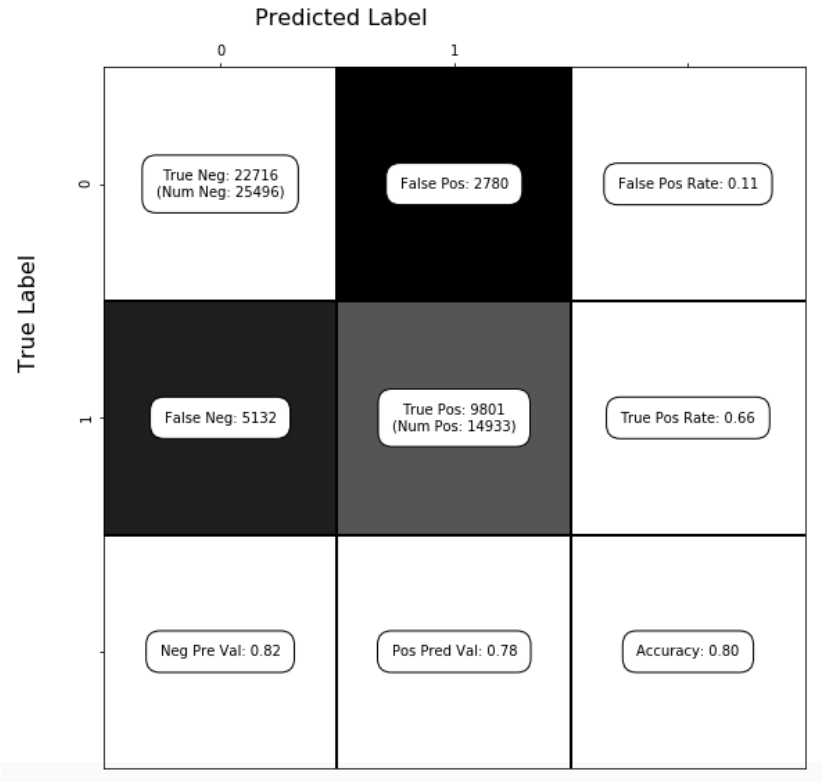




Results

Accuracy – 80.4%

Logloss – 0.4377



Samples question pairs from validation set

TP

```
('What is the best way to get free traffic to my website?', 'How can I get traffic for my web
site?')
('What is the most complex romantic situation you have ever faced?', 'What is the most comple
x romantic situation you ever faced?')
('How do you earn money from internet?', 'How does one earn money online without an investmen
t from home?')
('Which is better: Uber or Ola? Why?', 'What is better Uber or Ola? Why?')
('How do you know if it is time for divorce?', 'How do you know if you should divorce?')
```

TN

```
('How can I integrate my Posterous blog into a Cargo Collective page?', 'How can I stop Poste
rous resizing my images?')
('Magic Tricks: What do I need to do to build a coat rack that I can pull out of a bag as Mar
y Poppins does. I was hoping to find a retractable one but no luck. What do magicians use whe
n pulling out absurdly long objects from small containers?', 'What is the best magic show you
have ever seen? Were you able to figure out the tricks?')
('How do I convince my girlfriend when her parents are scolding her badly?', 'How can I know
that my girlfriend will really try to convince her parents?')
('What do you think about cultural appropriation? Is it really a thing?', "What's the big dea
l about cultural appropriation?")
('How do I become a real estate billionaire?', 'Can you become a real estate billionaire with
out making a company? If so, how?')
```

FP

```
('How do I get freelance work?', 'How do I get my first freelancing work?')
('How do I increase my presence of mind?', 'How can one improve their presence of mind?')
('What are the best language exchange websites?', 'What is the best free website to find a la
nguage exchange?')
('What are the best free Microsoft Office alternatives?', "What's the best free alternative t
o Microsoft Office?")
('How do I  post a trending post on Quora?', 'How do I post in Quora?')
```

click to scroll output; double click to hide

FN

```
('What makes a qualified doctor successful or failed?', 'What makes a successful doctor?')
('How do I choose the best taxi services in Udaipur?', 'How do I choose best taxi service in
Udaipur? How can I find that?')
('What does an actuary do?', 'What do actuaries do?')
('What is the best question asked in your interview?', "What is the most interesting question
you've been asked in an interview?")
('What are the most interesting fields ML is being used in today?', 'What are the most intere
sting applications of ML today?')
```

DEMO

Duplicate question pair identifier