

Problem statement

"Identify if a pair of questions are similar in content"

Why?



Re-use content



Simpler data storage

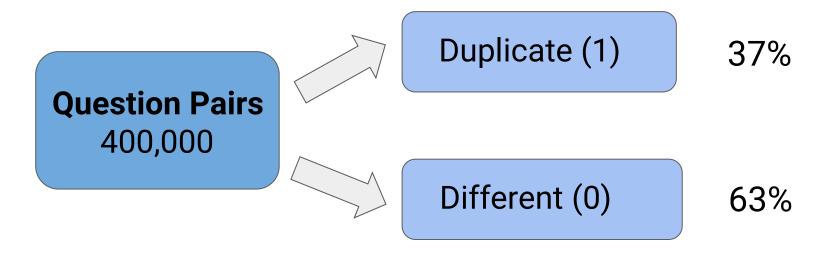




Customer Experience

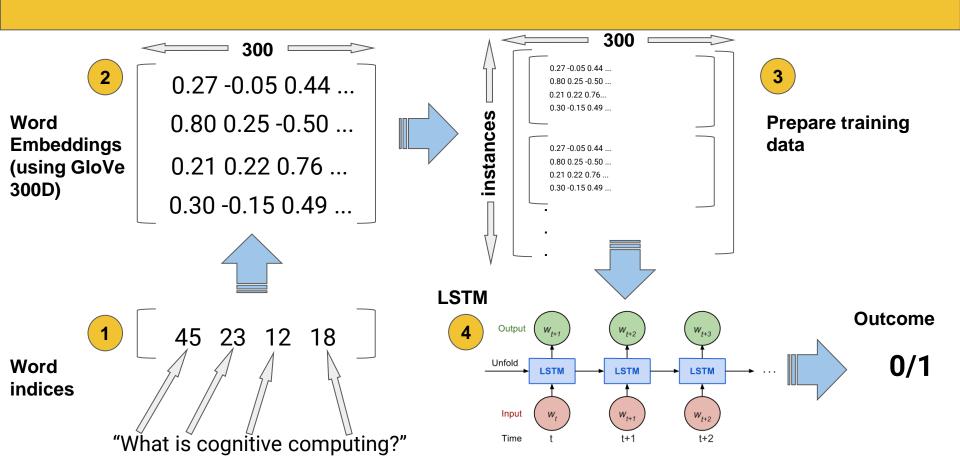
Data storage costs

Data

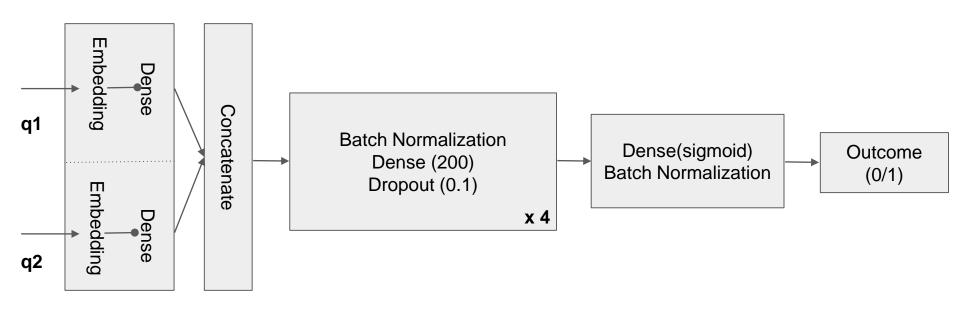


	Question 1	Question 2	Is_duplicate (target)
	Which one dissolves in water quickly sugar, salt, methane or carbon dioxide?	Which fish would survive in salt water?	0
	Astrology: I am a Capricorn Sun Cap moon and cap risingwhat does that say about me?	I'm a triple capricorn (Sun, moon and ascendant in Capricorn). What does this say about me?	1

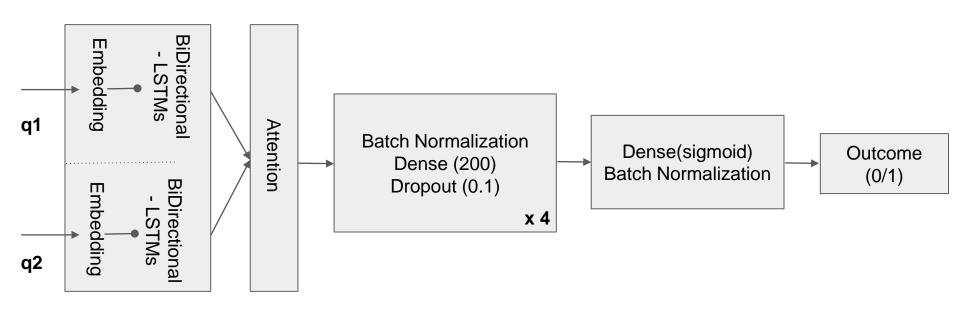
Modelling Approach



Modelling Approach: Bag of Embeddings



Modelling Approach: Bi-LSTM with Attention



Model Evaluation

Loss function: Binary cross entropy (Log Loss)

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

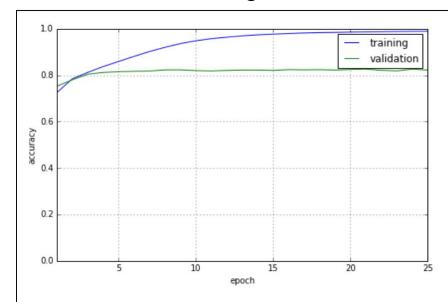
Metric Scores -

Baseline accuracy: 63%

Validation accuracy: 82.63%

Test set accuracy: 82.43%

Learning Curve



Model Performance Comparison

Baseline Accuracy: 63%

(If all predictions are 0)

79.24%

Max, SGD, Sigmoid 80.83%

Max, Adam, Sigmoid 80.95%

Mean, Adam, Sigmoid



82.63%

Attention, Adam, Sigmoid

Next steps

- Use different pre-trained embeddings for the model.
 e.g. Word2Vec, fasttext
- Try different similarity measure in embedding concatenation. e.g. Manhattan distance
- Extract and combine more other NLP features.
 e.g. number/proportion of common words
- Another interesting problem that utilizes the same concept is that of question answering using a context passage. We can attempt that.



Appendix

Approach

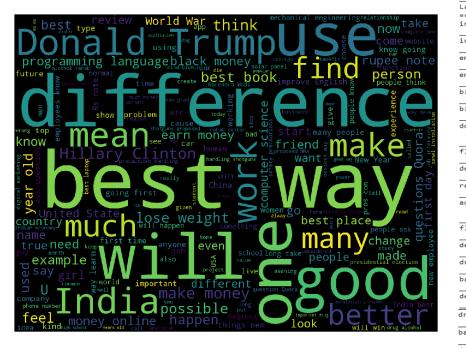
Data preprocessing

Generated tokenized word indices

Pre-trained GloVe word embeddings

Create word embedding matrices

Prepare training data and label tensors



Layer (type)	Output	Shape	Param #	Connected to
input_3 (InputLayer)	(None,	25)	0	
input_4 (InputLayer)	(None,	25)	0	
embedding_3 (Embedding)	(None,	25, 300)	28679100	input_3[0][0]
embedding_4 (Embedding)	(None,	25, 300)	28679100	input_4[0][0]
bidirectional_3 (Bidirectional)	(None,	25, 128)	439296	embedding_3[0][0]
bidirectional_4 (Bidirectional)	(None,	25, 128)	439296	embedding_4[0][0]
dot_2 (Dot)	(None,	128, 128)	0	bidirectional_3[0][0] bidirectional_4[0][0]
flatten_3 (Flatten)	(None,	16384)	0	dot_2[0][0]
dense_7 (Dense)	(None,	3200)	52432000	flatten_3[0][0]
reshape_2 (Reshape)	(None,	25, 128)	0	dense_7[0][0]
add_2 (Add)	(None,	25, 128)	0	bidirectional_3[0][0] reshape_2[0][0]
flatten_4 (Flatten)	(None,	3200)	0	add_2[0][0]
dense_8 (Dense)	(None,	200)	640200	flatten_4[0][0]
dropout_5 (Dropout)	(None,	200)	0	dense_8[0][0]
batch_normalization_5 (BatchNorm	(None,	200)	800	dropout_5[0][0]
dense_9 (Dense)	(None,	200)	40200	batch_normalization_5[0][0]
dropout_6 (Dropout)	(None,	200)	0	dense_9[0][0]
patch_normalization_6 (BatchNorm	(None,	200)	800	dropout_6[0][0]
dense_10 (Dense)	(None,	200)	40200	batch_normalization_6[0][0]
dropout_7 (Dropout)	(None,	200)	0	dense_10[0][0]
patch_normalization_7 (BatchNorm	(None,	200)	800	dropout_7[0][0]
dense_11 (Dense)	(None,	200)	40200	batch_normalization_7[0][0]
dropout_8 (Dropout)	(None,	200)	0	dense_11[0][0]
patch_normalization_8 (BatchNorm	(None,	200)	800	dropout_8[0][0]
dense_12 (Dense)	(None,	1)	201	batch_normalization_8[0][0]

Applications

Improve efficiency of question answer websites (Quora, Stack Overflow, Reddit)

ChatBots

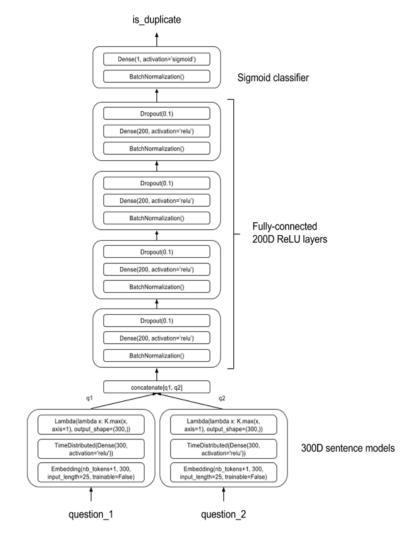
Enhance customer services and experience

Question generation systems for competitive exams

Examples

Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?	Which fish would survive in salt water?	0
Astrology: I am a Capricorn Sun Cap moon and cap risingwhat does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1

Approach: Model building



Approach: Model with attention

