Detecting Quora Duplicate Questions

Kenan Farmer
Department of Computer Science

Charles Stahl
Department of Physics

Meir Hirsch
Department of Computer Science

PRINCETON UNIVERSITY

Introduction & Motivation

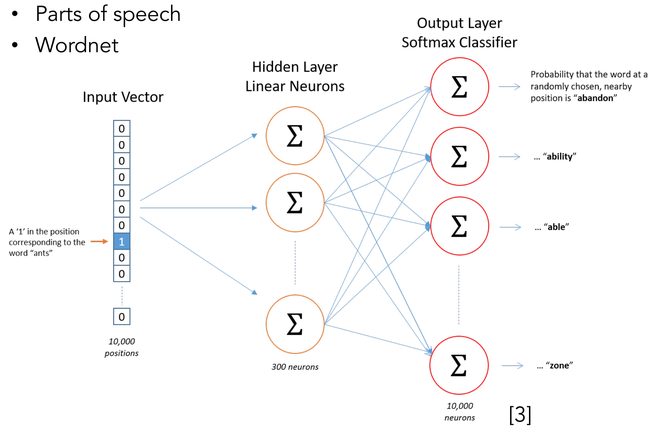
Quora is a website that strives to connect people from different backgrounds to be able to answer questions whose answers are "either locked in people's heads, or only accessible to select groups" [1]. 100 million people visit Quora every month to ask , answer, or view questions. Once these questions are answered, any other person should be able to find the responses. However, sometimes people are unable to find the questions they want and end up asking the same question again. This counteracts Quora's vision of having "only one version of each question\dots [not] a left wing version, a right wing version, a western version, and an eastern version" [1].
The data were made available by Quora on Kaggle, a website that hosts data analysis competitions [2].

Research Questions

- What common features indicate that two questions are duplicates?
- How does word embedding improve prediction capability?
- Is there latent structure indicating the meaning of a question? How do we access it using machine learning methods?
- What advantages does interpretable learning offer that neural networks do not?

Feature Engineering

- Simple similarity score using bag-of-words, tf-idf
- `word2vec:` Embeds words in pre-trained vector space
- Focusing on word differences
- Parts of speech
- Wordnet



[3]

Data

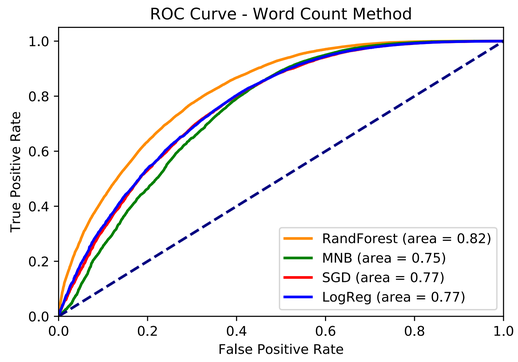
Training data consists of 404,302 question pairs labeled as duplicates or not.
Most non-duplicates are near-duplicates, while

- "What is the step by step guide to invest in share market in india?" and "What is the step by step guide to invest in share market?" are not duplicates
- "How can I be a good geologist?" and "What should I do to be a great geologist?" are duplicates

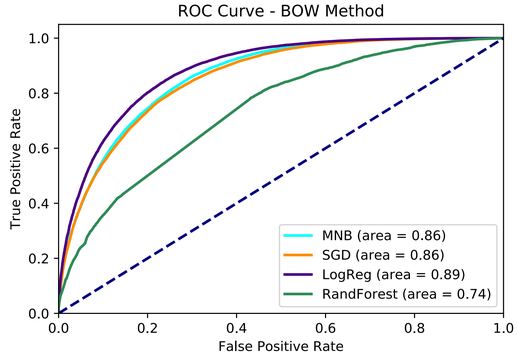
Testing data consists of 2.3 million pairs, made from 4.2 million distinct questions.
The testing set is unlabeled.

Results

Method 1 – Word Counts: Only count the number of differences and the number of similarities.



Method 2 – Bag of Words: Allow weights to be aware of word identities, but not how different words are related.



Log Loss scores on training data:

	MNB	SGD	LogReg	RandForest
Method 1	0.96641	0.55215	0.54342	**0.49019**
Method 2	0.51773	0.46103	**0.41490**	0.61665
M 2, tf-idf	0.45464	0.52297	**0.42771**	0.60837

Examples

"How do iPads get viruses?", vs.
"How do you get rid of a virus on an iPhone?"
4 differences, 8 similarities, 4 stop words.
These are not duplicates.

"Who is your favourite female movie director and why?", vs. "Who is the best female movie director?"
4 differences, 10 words in common, 2 stop words.
These are duplicates.

Methodology

- Feature vectors from all
- Linear regression – baseline
- Logistic regression
- Stochastic gradient descent
- Random forest of decision trees

Assessment

All methods were cross validated using the training data. Quora also provides an unlabeled test set. The website accepts submissions of probabilities for each pair in the test set.
Submissions to Kaggle were scored with a log-loss score, penalizing incorrect guesses with higher certainty.
Our best method so far, word counts, achieved a log-loss of 0.3825, putting us at 1540 out of 2750 on the leaderboard for the challenge.

Looking Forward

Next Steps: finish training a word2vec model and getting a submission from that. Hopefully this will improve our standing.

Long Term: More thoroughly compare our interpretable results with the results from neural networks. On which pairs does the NN perform better? On which does our perform better?

References

1. https://www.quora.com/about
2. https://www.kaggle.com/c/quora-question-pairs
3. http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Generous Thanks to the University of Texas, Austin for the base template design