SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Quora Question Pairs
Identify question pairs that have same intent

Arlene Fu

ENSC895 Course Project
Professor: Ivan Bajic

School of Engineering Science

Simon Fraser University

Burnaby, BC, Canada

Project Description

- Kaggle competition hold by Quora
- Finished 6 months ago
- Goal: Develop machine learning and natural language processing system to classify whether question pairs are duplicates or no

Semantic Question Matching

What are the best ways to lose weight?

VS

What are effective weight loss plans?

Provided Data

Train.csv: 64MB with >400,000 pairs

id	qid1	qid2	question1	question2	Is_duplicate
0	1	2	What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
…	…	…	…	…	…

Test.csv: 314MB with >2.3 million pairs

test_id	question1	question2
…	…	…

Evaluation: log loss = $-\frac{1}{N}\sum_{i=1}^{N}[y_i \log p_i + (1-y_i)\log(1-p_i)]$.

PreProcessing

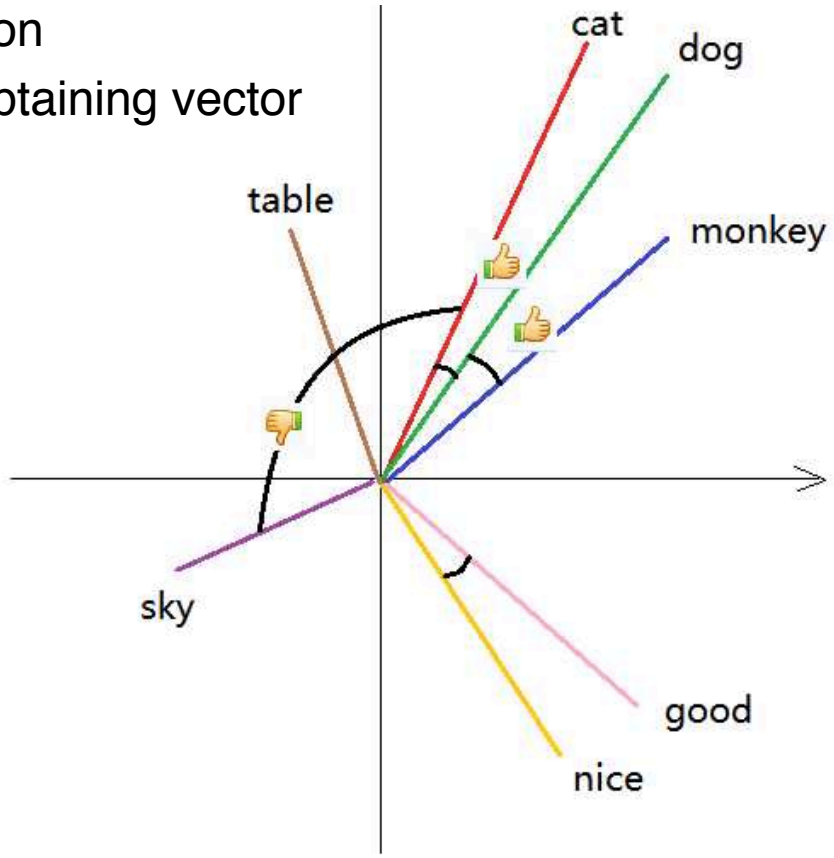
- Problems of input data
  - Questions in training set : genuine examples from Quora (with typo)
  - Questions in test set: computer-generated (Not make sense)

What food fibre?

- Correct typo
- Compare to DefaultDict
- Replace unreasonable words by a common word
- e.g. don't ➔ do not
- Tokenize
- Remove stopwords like and, also, to…
- Lemmatization: e.g. do, did, done → do

Features

- Word Embedding
- GloVe:
  - Global Vectors for Word Representation
  - Unsupervised learning algorithm for obtaining vector representations for words
  - Pre-trained word vector available

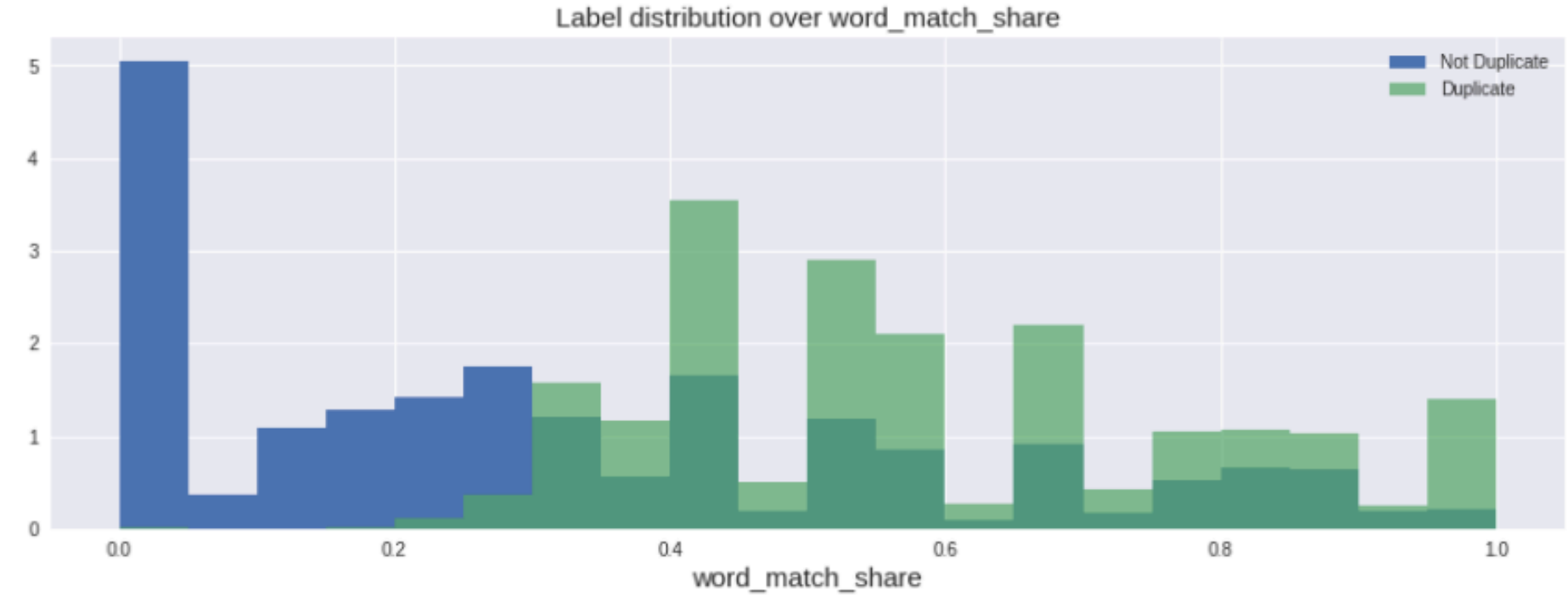


Projection of the embedding vectors to 2-D

Reference: https://www.zhihu.com/question/32275069

Features

- Word Embedding
- Word match & share



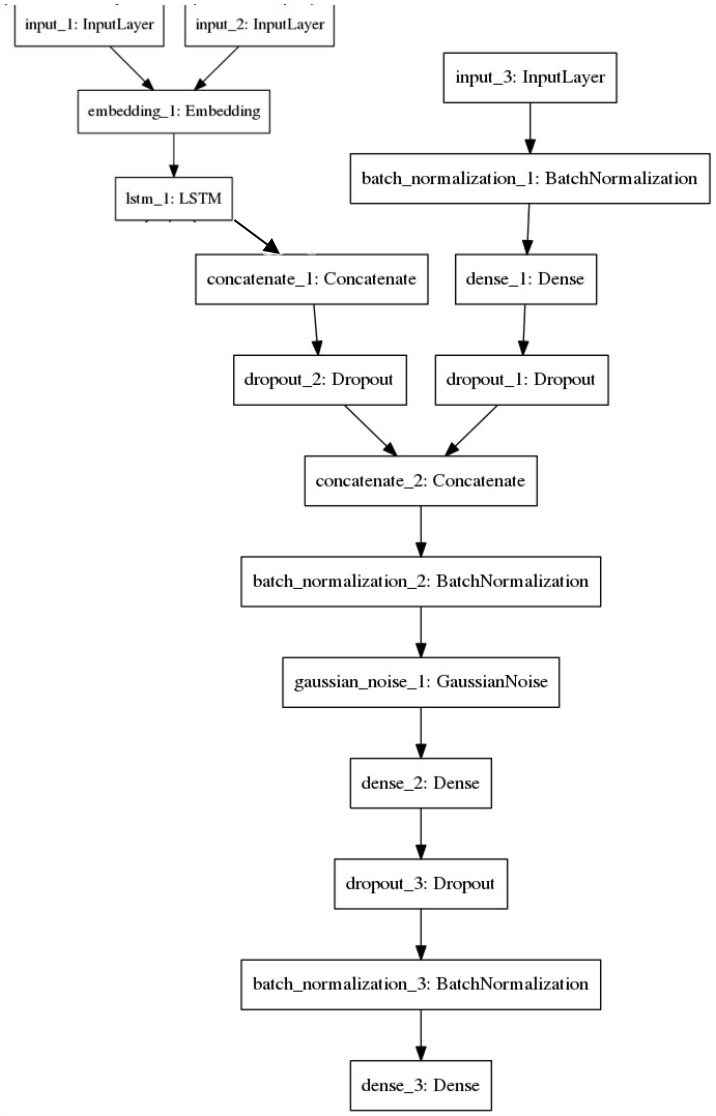
Features

- Word Embedding
- Word match & share
- Magic feature provided by kagglers
  - More frequent questions are more likely to be duplicates
  - Count the neighbors of the question neighbors
  - More common neighbors the question pair have, more likely to be duplicate
- …

Models

- StratifiedKFold: (5 Fold)
  - variation of k-fold which returns stratified folds
  - preserving the percentage of samples for each class
- LSTM, currently logloss=0.13264 on public LB

LSTM model



Post Processing

- Rescaling: Convert training predictions to test predictions
  - 37% positive class in Train
  - Without rescaling:
    - Logloss for Train in LB = 0.6585
    - Logloss for Test in LB = 0.554

$$r = \frac{logloss + log(1-p)}{log\left(\frac{1-p}{p}\right)}$$

    - r is positive class in Test: 16.5%
  - Distribution of Train/Test set is different
  - To avoid oversampling, use method provided by one kaggler:
    - Let $a = \frac{0.165}{0.37}, b = \frac{1-0.165}{1-0.37}$
    - $f(x) = a * x/(a * x + b * (1-x))$

Future work

- XGBoost
  - optimized distributed gradient boosting library designed to be highly efficient, flexible and portable
- LightGBM
- Sentence embedding
- More features

Reference

[1]. https://www.kaggle.com/sudalairajkumar/keras-starter-script-with-word-embeddings/notebook

[2]. https://github.com/aerdem4/kaggle-quora-dup

[3]. https://www.kaggle.com/c/quora-question-pairs/discussion/32819

[4]. https://www.kaggle.com/dasolmar/xgb-with-whq-jaccard/code

[5]. https://www.kaggle.com/c/quora-question-pairs/discussion/31179

[6]. http://blog.csdn.net/lanxu_yy/article/details/29002543

[7]. https://nlp.stanford.edu/projects/glove/

Q & A