Duplicate Question Detection

Neda Zolaktaf and Vaishnavi Malhotra

SFU Simon Fraser University

1. Problem Statement

Duplicate question detection in forums

- Duplicates pairs are labeled 1, non-duplicates are labeled 0.

Motivation

- Users often ask same question
- Discovering duplicates is valuable for users to get an answer immediately

3. Traditional Models

Two stage approach

- **Feature extraction** needs vocabulary, use the frequency of words, or use TFIDF term weighting [1]
- **Classifier** Logistic Regression, Random Forest, etc

Example: TFIDF vector + Logistic Regression

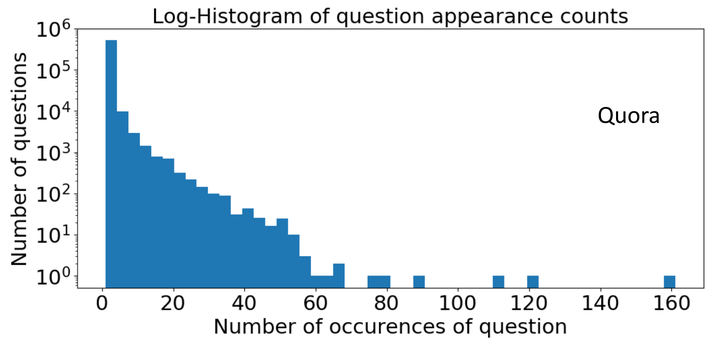
Limitations

- Extracted features may not be good, word order can be lost, used different n-grams sizes, linear classifier, no natural notion of similarity between words

2. Datasets and Analysis

	Quora	Apple	Android	Sprint	Superuser	AskUbuntu
# of examples in train set	323,479	183,174	191,577	735,280	897,365	1,722,980
# of examples in test set	80,871	457,95	478,96	183,822	224,343	430,746
% of duplicates	36.93%	0.99%	0.99%	0.99%	0.99%	0.99%

Question Pairs	Duplicate
Question 1: Restore Windows file extension to default **Question 2:** File transfer tools for linux with support of zero-copy transfer	0
Question 1: How can I change Excel 2007 shortcuts from foreign to English ? **Question 2:** Color coded Windows 7 Preview Pane ?	0
Question 1: What are the best indoor exercises to remove belly fat completely and lose weight in minimum time ? **Question 2:** How can I lose belly fat if I 'm skinny ?	1



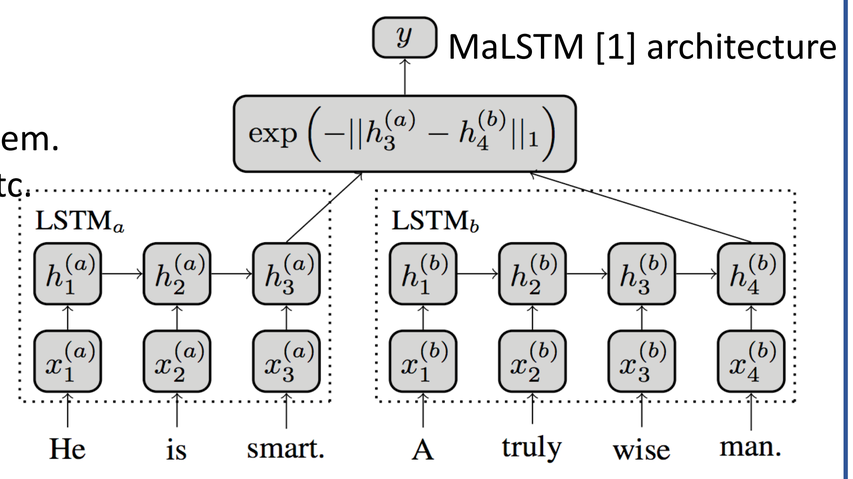
4. Siamese Neural Network

Siamese Network

- networks that have two or more identical sub-networks in them.
- used for sentence similarity, recognizing forged signatures, etc.

Manhattan LSTM (MaLSTM)[1]

- Long-Term Short-Term Memory (LSTM)
- Use output as semantic representation of question
- Computes sentence similarity using Manhattan distance
- Implemented in Keras
  - Input layer uses word2vec (word embeddings dimension = 300)
  - One-layer LSTM (hidden dimension =50 and 100 for quora)
  - epochs = 5 (30 for quora)
  - batchsize = 1024
  - Adam Optimizer

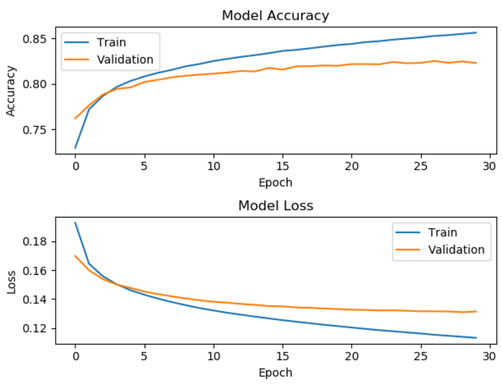


Objective function is Mean Squared Error.
We tried changing the architecture for cross entropy loss but did not obtain improvements.

5. Pipeline

python · Natural Language Analysis with Python NLTK · pandas · Keras · scikit learn · matplotlib · pandas · Flask · React

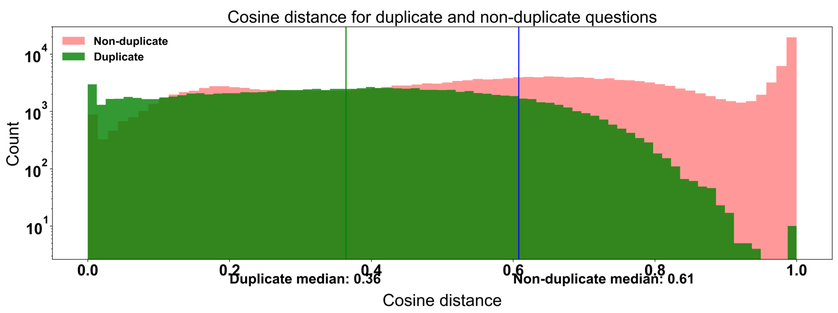
6. Experiments



Evaluating MaLSTM on Quora

Model	Quora		Apple		Android		Sprint		Superuser		AskUbuntu	
	Accuracy	Fscore	Accuracy	Fscore	Accuracy	Fscore	Accuracy	Fscore	Accuracy	Fscore	Accuracy	Fscore
TFIDF + Logistic Regression	0.7554	0.6336	0.9906	0.1161	0.9901	0.1960	0.9901	0.01	0.9902	0.0319	0.9899	0.0091
TFIDF + Random Forest	0.7783	0.6248	0.9911	0.1948	0.9922	0.4335	0.9900	0.0010	0.9909	0.1639	0.9917	0.3370
MaLSTM	0.6756	0.2814	0.991	0.002	0.9899	0.001	0.9917	0.05	0.9904	0.006	0.9900	0.003

Evaluation on individual datasets.

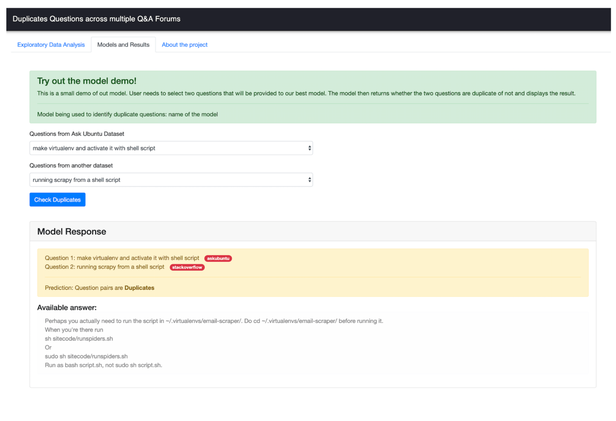


Model	Precision	Recall	Fscore
BOW + Xgboost	0.82	0.61	0.70
Word level TFIDF + Xgboost	0.82	0.62	0.71
n-gram tfidf + Xgboost	0.81	0.41	0.54
Character-level TFIDF + Xgboost	0.84	0.72	0.78

Integrated Quora and AskUbuntu and evaluated on integrated dataset.

7. Interface

1. Exploratory Data analysis
2. Demo
3. Project information



8. Conclusion & Future Work

- Evaluated different methods on 5 datasets
  - On quora TFIDF + Logistic obtains best Fscore on class 1
- Integrated Quora and AskUBuntu datasets into one dataset
  - Evaluated TFIDF at different levels + Xgboost.
  - Character-level TFIDFf + Xgboost obtains best Fscore on class 1
- The datasets are highly imbalanced
  - Consider other objective functions
- Consider alternative neural network models.

9. References

1. Jonas Mueller and Aditya Thyagarajan, ``Siamese Recurrent Architectures for Learning Sentence Similarity'', AAAI, 2016.
2. https://medium.com/mlreview/implementing-malstm-on-kaggles-quora-question-pairs-competition-8b31b0b16a07
3. train.py for [2] based on code from: https://github.com/likejazz/Siamese-LSTM