Similar Question Detection

Monday August 21st, 2017


Geoffrey Link


James Peng


Brad Putman

Berkeley
UNIVERSITY OF CALIFORNIA

Background Research and Modeling Choices

- Challenges with similar question detection
  - Lexical chasm: “Where can I watch movies online?” vs “Are there any websites for streaming films?”
  - Polysemy: words with multiple meanings depending on context
  - Noisy information: misspelled words, poor grammar, and short questions
- Quora’s model of choice: Random Forest
  - XGBoost with distance metrics
  - Fast, linear model
  - Our best loss score: 0.41
- Convolutional Neural Net (with “GloVe” embeddings)
  - Originally designed for Computer Vision
  - Shown to be effective for NLP (semantic parsing and search query retrieval)
  - Our best loss score: 0.35
- LSTM Neural Net (with “Google News” corpus)
  - Originally designed to encapsulate long term dependencies
  - Captures sequential and syntactic patterns of text (helps with word sense disambiguation).
  - Our best loss score: 0.33

Exploratory Data Analysis

404,290 and 2,345,796 question pairs in training/testing

Unique Questions

66% in training
93% in testing
87% in training+testing

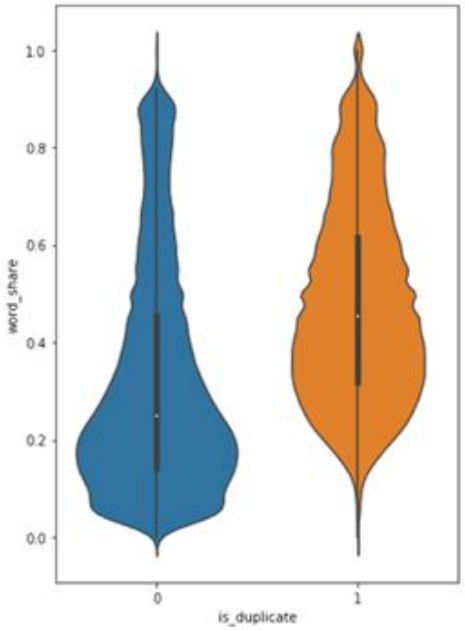
2 and 6 in training/testing

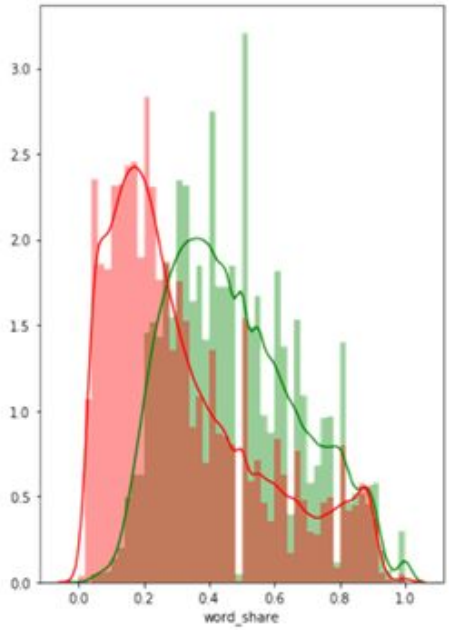
- Clean multi-spaces, newlines, tabs, carriage returns, and .. so on
- Casefolding, e.g. **india** vs **India**, **u.s.** vs **U.S**

Exploratory Data Analysis *(cont.)*

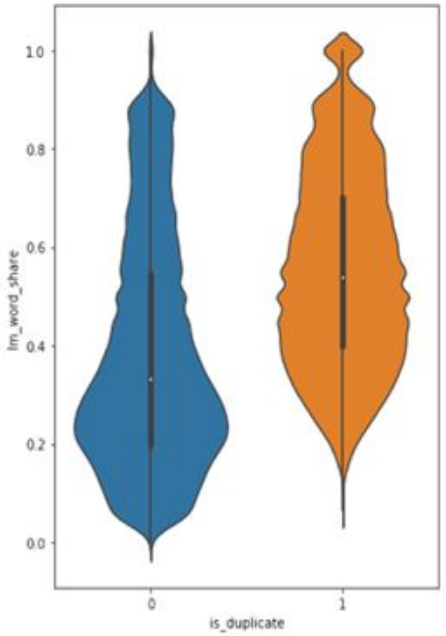
Word Share(%): tokens appeared in both in Q1 and Q2 in a pair

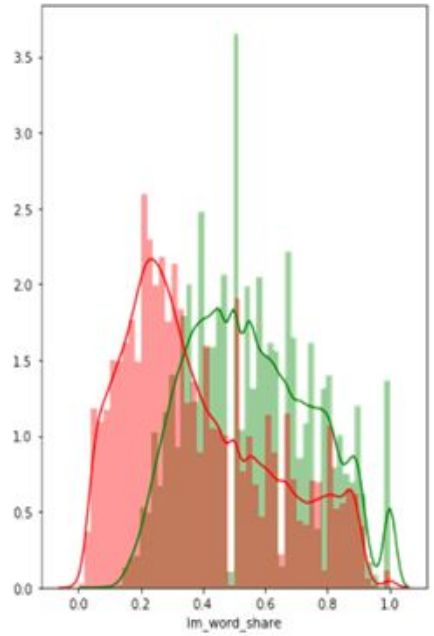
Calculated “word share” in tokens and lemmas





Tokens





Lemma

Berkeley
UNIVERSITY OF CALIFORNIA

Understanding Our Model Results

Models	Classification Similarity
CNN vs. LSTM	92.97%
LSTM vs. XGBoost	74.44%
CNN vs. XGBoost	72.12%
CNN vs. LSTM vs. XGBoost	69.76%

Semantic Emphasis

- Good at matching general topics
- Vulnerable to overgeneralization
- False positives

Syntactic Emphasis

- Vulnerable to lexical chasm
- Better at particularization
- False negatives

Berkeley
UNIVERSITY OF CALIFORNIA

Examples of Labeling Discrepancies

LSTM marked duplicate
CNN marked not duplicate

"What are books do you plan to read in 2015?"

=

"What books do you have on your across 'to-read list'?"

LSTM marked duplicate
XGBoost marked not duplicate

"How did 4Chan respond to the Trump election victory in November?"

≠

"Can Donald Trump realistically defeat Hillary Clinton in 2016?"

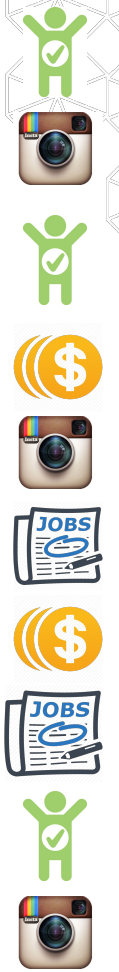
Conclusion

- Our language models confirmed observations noted in earlier papers
- Asking questions is a very noisy process (hard to establish unbiased ground truths)
- In our opinion, a LSTM model is the best choice
  - Highest accuracy, low bias, low maintenance
  - False positives are better than false negatives
  - Syntactic emphasis is less reliable (spelling/grammar issues, lexical chasm)
- Future work:
  - Topic modeling (possibly paired with a syntactic model)
  - Polysemy-induced false positives
  - Incorporate user-provided answers into classification process
  - Incorporate metadata (time, location, etc.) into classification process

Appendix Slides

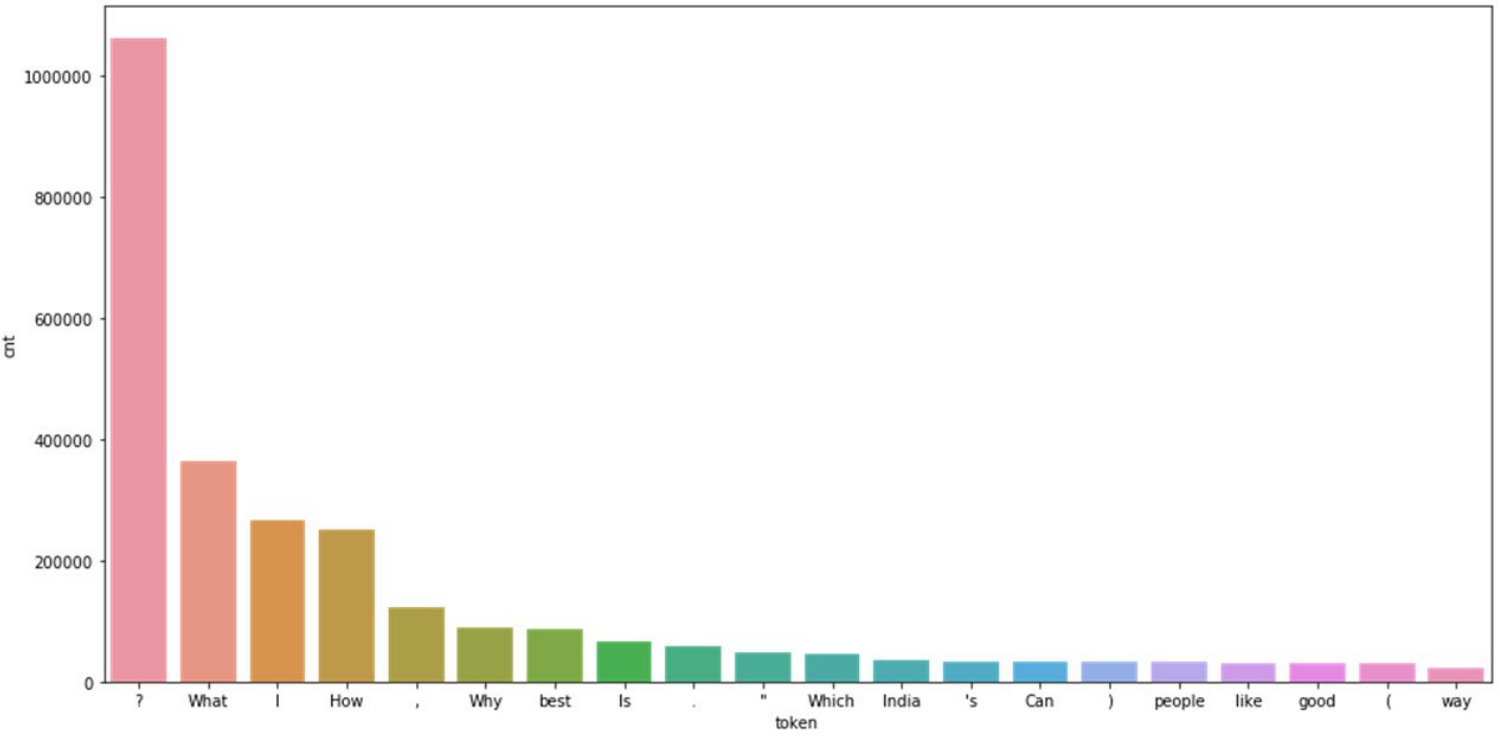
Exploratory Data Analysis *(cont.)*

	0	1
0	What are the best ways to lose weight?	161
1	How can you look at someone's private Instagram account without following them?	120
2	How can I lose weight quickly?	111
3	What's the easiest way to make money online?	88
4	Can you see who views your Instagram?	79
5	What are some things new employees should know going into their first day at AT&T?	77
6	What do you think of the decision by the Indian Government to demonetize 500 and 1000 rupee notes?	68
7	Which is the best digital marketing course?	66
8	How can you increase your height?	63
9	How do I see who viewed my videos on Instagram?	61



Exploratory Data Analysis *(cont.)*

* in training data only
* stop words removed



	cnt	token
0	1062468	?
1	364172	What
2	266739	I
3	251839	How
4	122290	,
5	89778	Why
6	85987	best
7	67188	Is
8	58370	.
9	49283	"
10	45965	Which
11	36599	India
12	33568	's
13	32763	Can
14	32125	)
15	31998	people

Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis *(cont.)*

	0	1
0	What	2686
1	How	1848
2	What is	1212
3	What are	803
4	How do	773
5	What is the	766
6	What is?	680
7	Why	653
8	Is	543
9	How do I	531