ARTIFICIAL INTELLIGENCE: METHODS & APPLICATIONS
FINAL PRESENTATION

SRI PHANI GORTI - SGORTI3@UIC.EDU

SAIKRISHNA KALAHASTI KARTHIK - SKALAH2@UIC.EDU

MALAVIKA RAMPRASAD - MRAMPR2@UIC.EDU

MAY 3, 2018

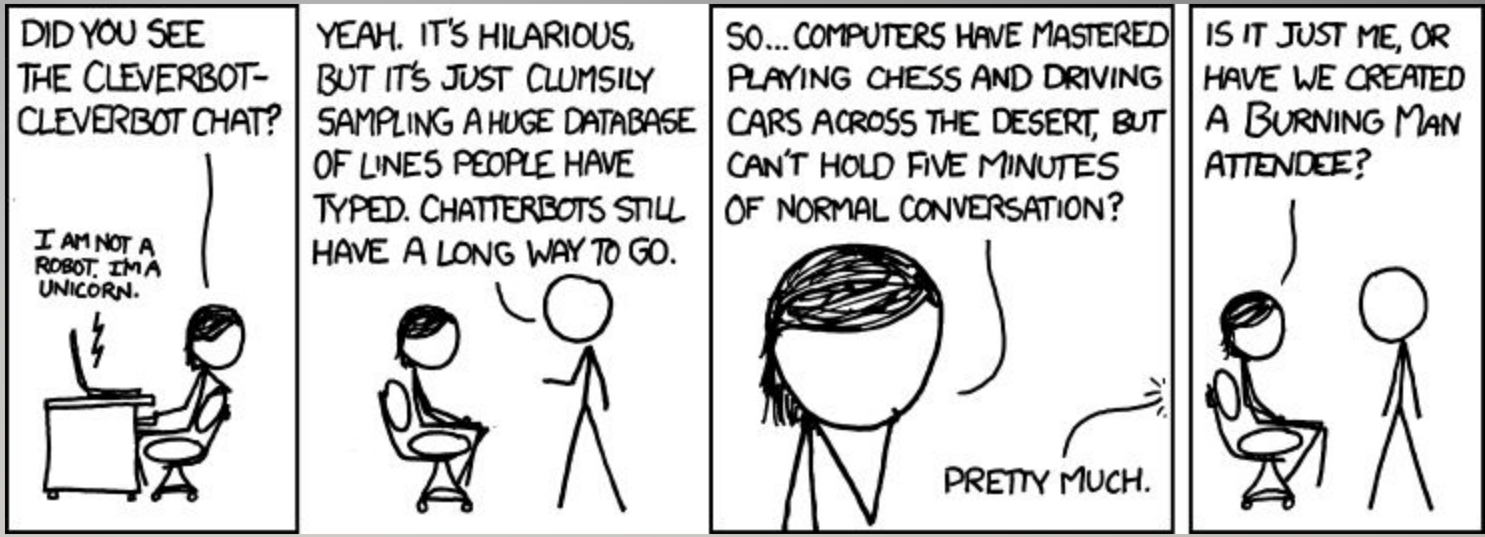
QUORA DUPLICATE QUESTION DETECTION

PROBLEM & DATASET

- Identifying actual duplication or semantic coincidence of two questions on a question answer platform - Quora
- Recognize if two questions are semantically related or have same intent (Y =1) or if they are different altogether(Y=0)
- Dataset: https://www.kaggle.com/c/quora-question-pairs/data
- *id*: unique identifier for the question pair (unused)
- *qid1*: unique identifier for the first question (unused)
- *qid2*: unique identifier for the second question (unused)
- *question1*: full unicode text of the first question
- *question2*: full unicode text of the second question
- *is_duplicate*: label 1 if questions are duplicates, 0 otherwise

Why Duplicate Question Detection?

- Understanding Natural Language using Deep Learning
- Problem can be reformulated into several other active research forms such as:
  - Answering Selection
  - Textual Entailment(Positive)
  - Sentence Similarity


DID YOU SEE THE CLEVERBOT-CLEVERBOT CHAT?
I AM NOT A ROBOT. I'M A UNICORN.
YEAH. IT'S HILARIOUS, BUT IT'S JUST CLUMSILY SAMPLING A HUGE DATABASE OF LINES PEOPLE HAVE TYPED. CHATTERBOTS STILL HAVE A LONG WAY TO GO.
SO... COMPUTERS HAVE MASTERED PLAYING CHESS AND DRIVING CARS ACROSS THE DESERT, BUT CAN'T HOLD FIVE MINUTES OF NORMAL CONVERSATION?
PRETTY MUCH.
IS IT JUST ME, OR HAVE WE CREATED A BURNING MAN ATTENDEE?

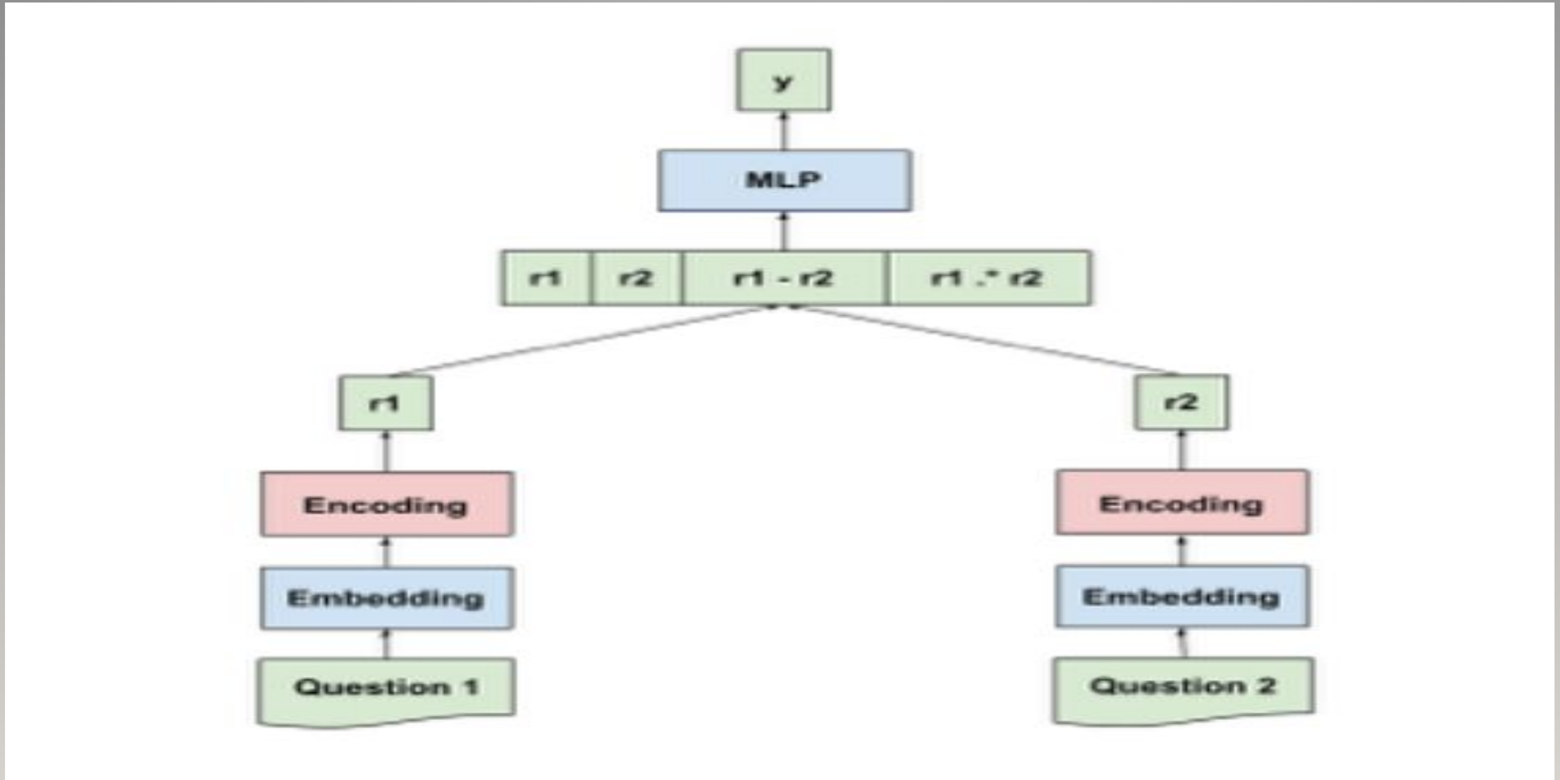
Related Work

- Siamese Manhattan LSTM(Manhattan distance between two questions)
- Bilateral Multi Perspective Matching for Natural Language Sentences
- Hybrid LSTM(LSTM + CNN)

References:

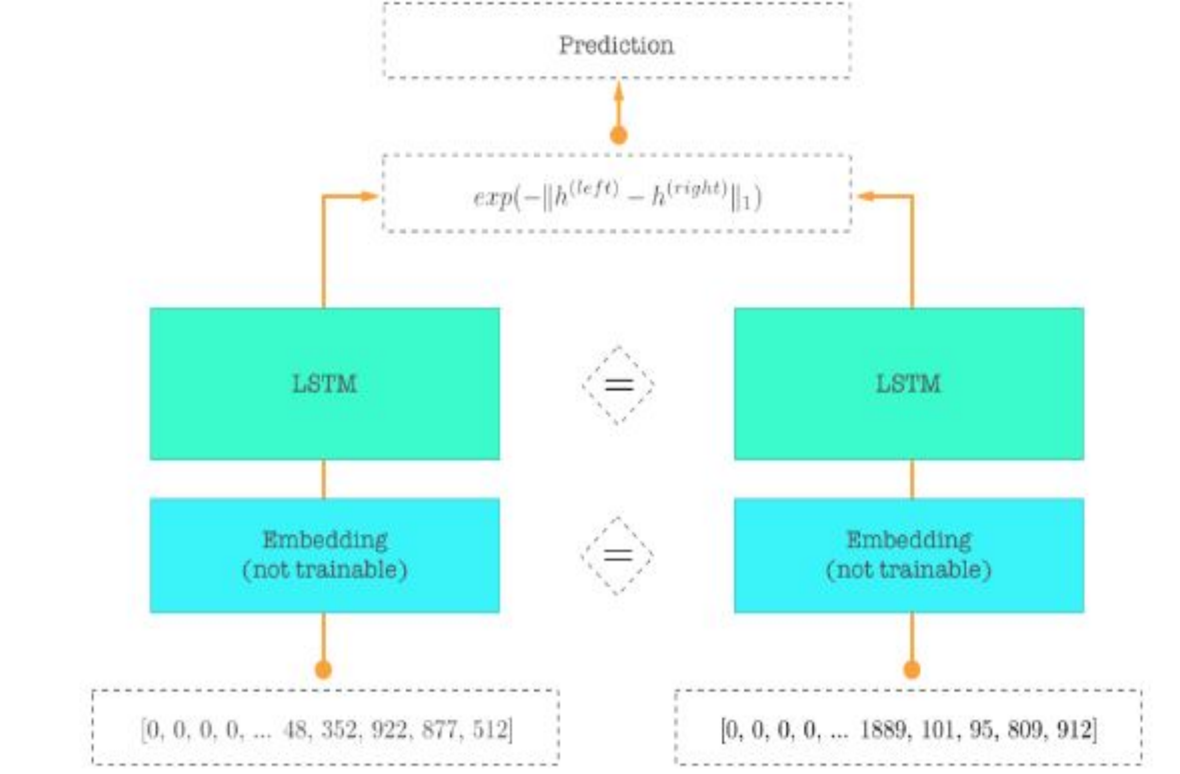
- https://arxiv.org/pdf/1702.03814.pdf
- http://www.mit.edu/~jonasm/info/MuellerThyagarajan_AAAI16.pdf
- https://web.stanford.edu/class/cs224n/reports/2759336.pdf

Related Work - Blueprint



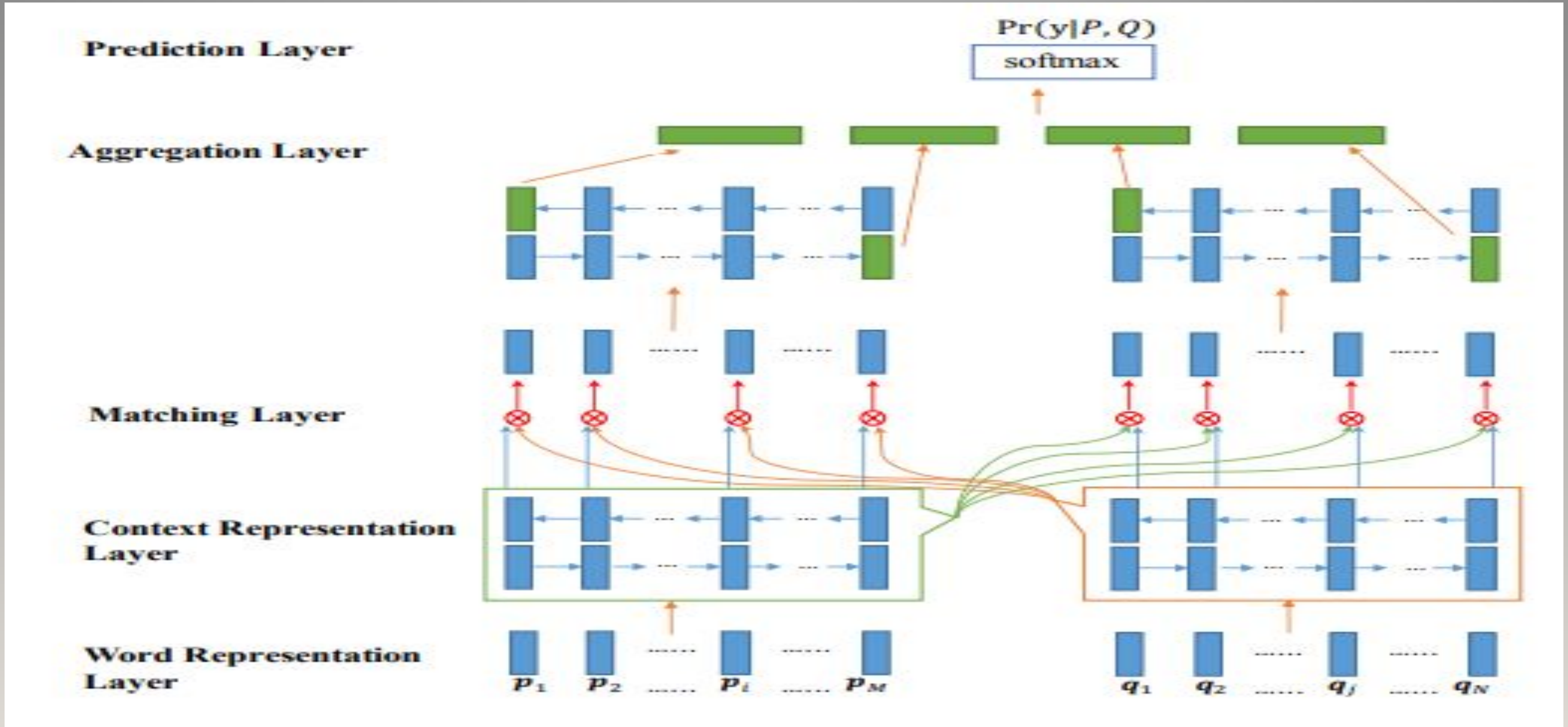
Source: https://web.stanford.edu/class/cs224n/reports/2759336.pdf

Related Work - Ma LSTM



Source: https://medium.com/mlreview/implementing-malstm-on-kaggles-quora-question-pairs-competition-8b31b0b16a07

Related Work - BiMPM



Source: https://arxiv.org/pdf/1702.03814.pdf

Our Model

- Implemented Manhattan Bi- LSTM - Siamese Network
- Data Cleaning - Removing punctuations, short forms, converting to lowercase
- Data Preprocessing :
  - Word Vectorization using Google News Vectors
  - Obtained word level embeddings for each question represented as a matrix
- Trained the model using shared LSTM and Manhattan Distance
- Bidirectional LSTM(for Faster implementation)
- Loss Function: Mean Squared Error
- Optimizer: Adadelta
- Libraries used: nltk - stopwords, pandas, numpy, Keras with Tensorflow backend

Dev Accuracy across models

Hybrid LSTM (LSTM + CNN)	Siamese Manhattan LSTM	BiMPM Quora Dataset	Our Model - Siamese approach (Bi-LSTM + Manhattan distance)
81.05	**82.5**	**88.17**	**82.75**

What’s up with Siamese Architecture?

- Siamese architectures is a class of Neural networks which contain two or more identical subnetworks i.e, shares same parameters and weights
- Popularly used for detecting sentence similarity, answer selection task etc
- *Sharing parameters and weights means less work to do with two models in place and model is less likely to overfit!*
- Easier to train

METRICS & RESULTS

- We used Batch SIze = 64
- No. of Epochs = 25 for (MaLSTM) 15 (MaBiLSTM)
- Adadelta with clipping norm to avoid any gradient explosion
- Planned to use different optimizers like Adam, SGD

	Loss	Accuracy (%)
Training Data(3 lakh question pairs)	0.1260	82.75
Validation Data(40000 question pairs)	0.1337	81.19

DEMO

QUESTIONS?