Research Statement and a Brief Introduction

EBRAHIM ANSARI

INSTITUTE FOR ADVANCED STUDIES IN BASIC SCIENCES, ZANJAN, IRAN

PRAGUE, SEPTEMBER 18

Outline

Ebrahim Ansari
September 2018

- Who am I?

- My interests

- My last projects

- My current focus

Who am I?

Ebrahim Ansari
September 2018

▶ Born in Mashhad, Iran. September, 06, 1983

▶ BSc. In Computer Science from Yazd University (2005)

▶ MSc. In Computer Eng. from Shiraz University (2008)

▶ PhD. In Computer Eng. from Shiraz University (2014)

   ▶ Sabbatical at University of Pisa (2012-2013)

▶ Assistant Professor at IASBS from 2015.

Main interests

Ebrahim Ansari
September 2018

- ▶ Natural Language Processing

- ▶ Machine Learning

- ▶ Low level programming in C and C++

- ▶ Parallel Programming and HPC

Main Projects – Distributed Frequent Itemset Mining (2009)

Ebrahim Ansari
September 2018

▶ A part of my MSc. Project

▶ Developing an algorithm to increase the efficiency of Apriori (Sequential Algorithm)

▶ Developing a parallel version of Trie-based Apriori

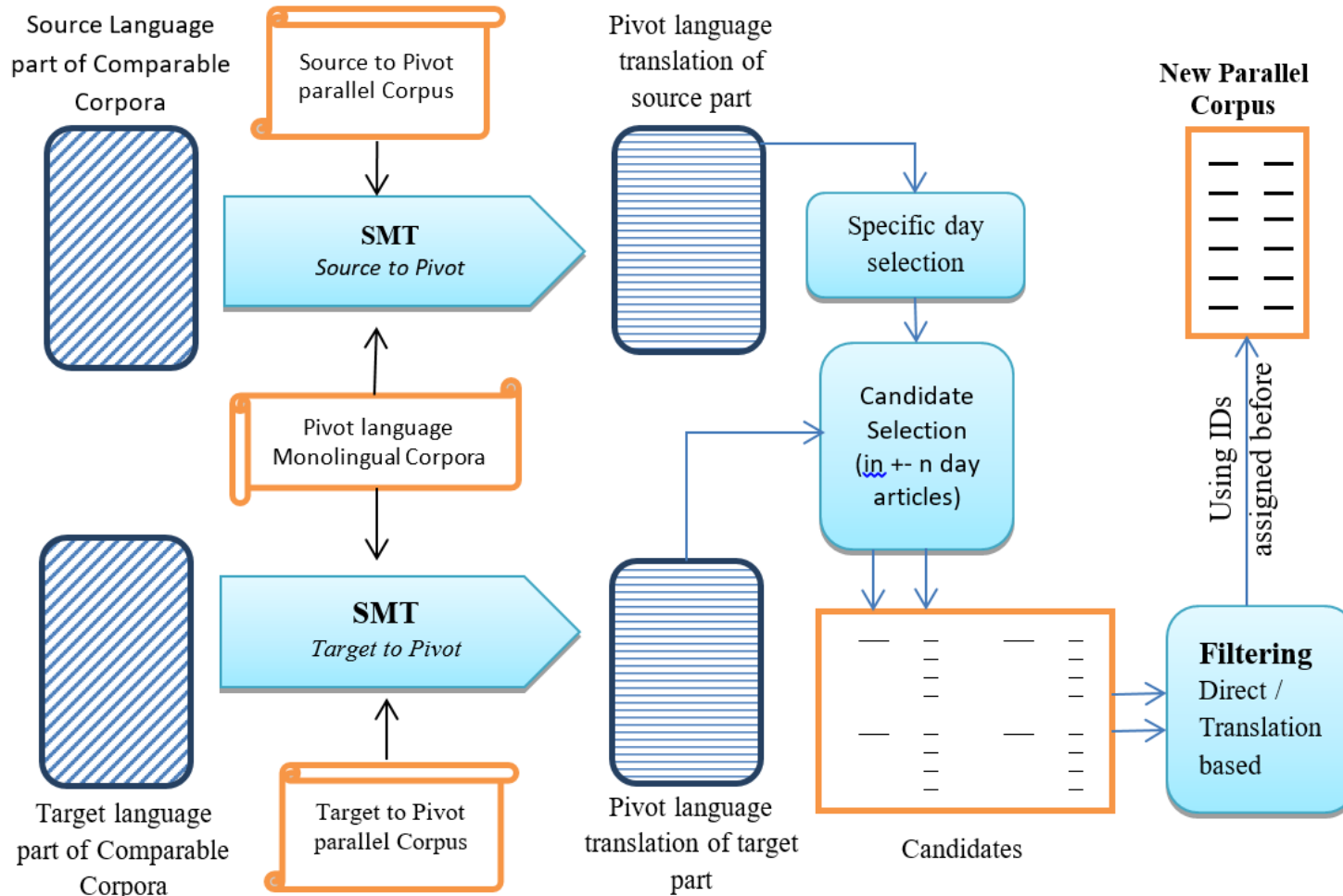▶ Developing a framework for distributed Trie-based Apriori

Main Projects – Creating 3D models from High resolution photos (2015)

Ebrahim Ansari
September 2018

- Collaboration with PhotoCore Company in Zurich

- Using sift features collected from high resolution photos

- Developing fast version for feature matching using MPI

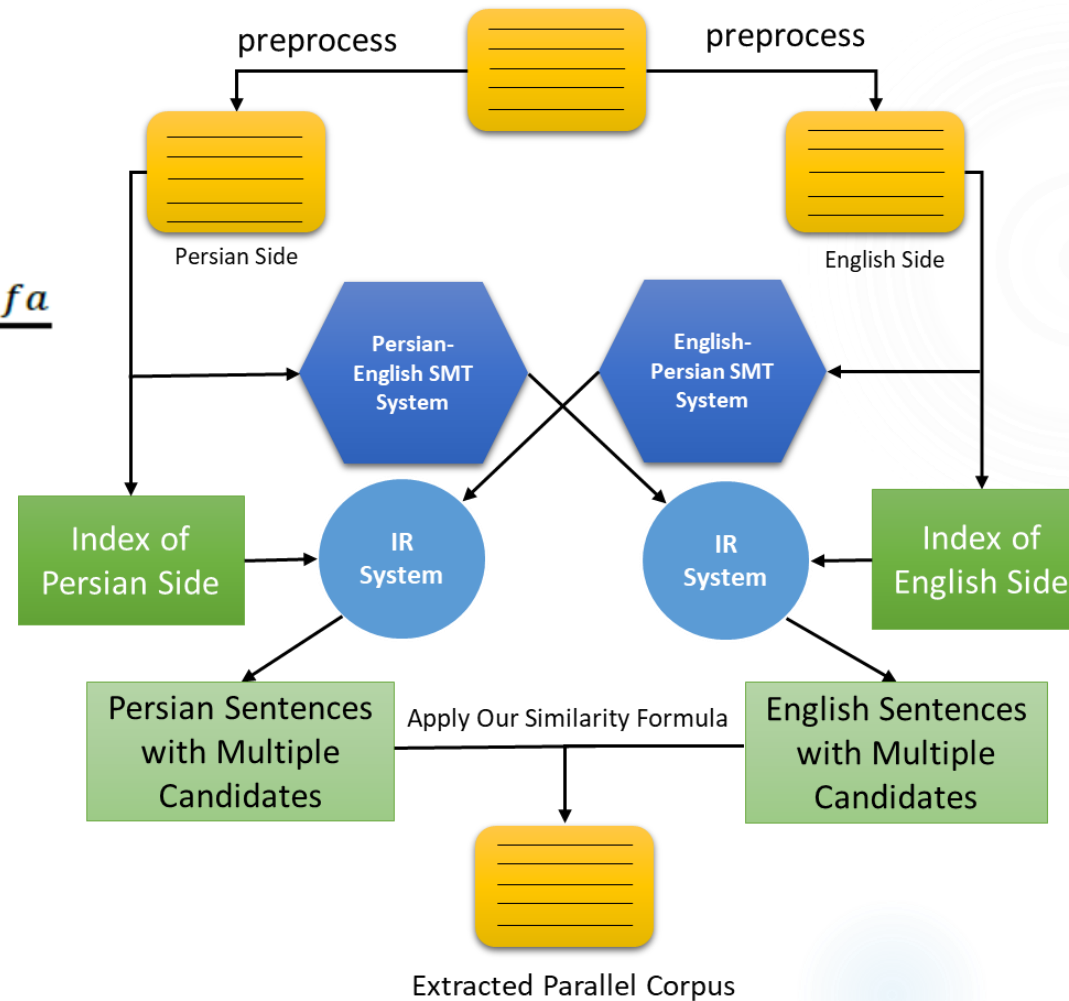- Developing fast version of PatchMatch algorithm

- Main focus was on efficiency.

Main Projects – Parallel Corpora collection (2014)

Source Language part of Comparable Corpora

Source to Pivot parallel Corpus

Pivot language translation of source part

New Parallel Corpus

SMT *Source to Pivot*

Specific day selection

Pivot language Monolingual Corpora

Candidate Selection (in +- n day articles)

Using IDs assigned before

SMT *Target to Pivot*

Target language part of Comparable Corpora

Target to Pivot parallel Corpus

Pivot language translation of target part

Candidates

Filtering Direct / Translation based

Main Projects – Parallel Corpora collection (2017)

Ebrahim Ansari
September 2018

Bi-Directional Translation

$$\frac{\alpha}{\alpha + Penalty} \times \frac{\beta \times Sim_{fa-en} + Sim_{en-fa}}{\beta + 1}$$

preprocess   preprocess

Persian Side   English Side

Persian-English SMT System   English-Persian SMT System

Index of Persian Side   IR System   IR System   Index of English Side

Persian Sentences with Multiple Candidates   Apply Our Similarity Formula   English Sentences with Multiple Candidates

Extracted Parallel Corpus

Main Projects –Lexicon Creation from comparable Corpora (2013)

Ebrahim Ansari
September 2018

- Extracting Bilingual Persian Italian Lexicon from Comparable Corpora Using Different Types of Seed Dictionaries

- Based on Reinhard Rapp's work

- to find a way to combine different dictionaries together in order to produce a better and more accurate lexicon

- New Weighting Schema

# Main Projects – WSD using word embeddings (2016)

Ebrahim Ansari
September 2018

▶ A new unsupervised Word Sense Disambiguation method for Persian using word embeddings

▶ The main idea is to use information provided by surrounding words.

▶ These words are translated into English using a word by word Persian-English dictionary. The translated words are clues for disambiguation.

Main Projects – WSD using word embeddings (2017)

Ebrahim Ansari
September 2018

- Supervised model based on Iacobacci, et al. , 2016

- Four new ideas:

  - Using New Coeffs. in Exponential Decay Strategy

  - Applying dimensionality reduction using PCA

  - Considering new weighting scheme (for imbalanced data)

  - Voting system

Main Projects – WSD using word embeddings (2017) – cont.

Ebrahim Ansari
September 2018
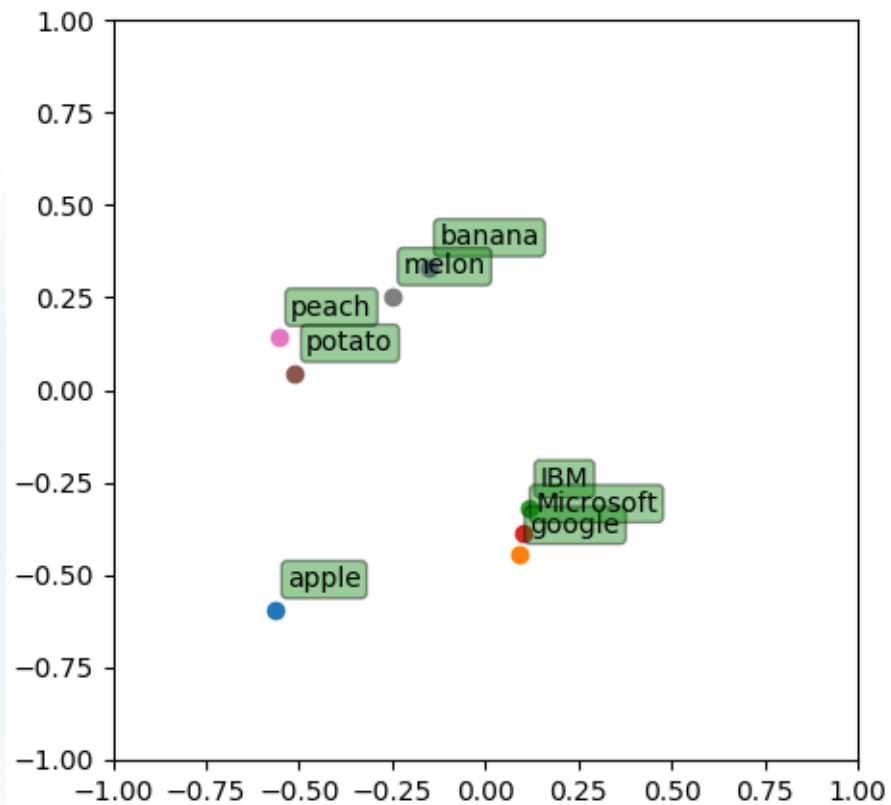
Google News embeddings
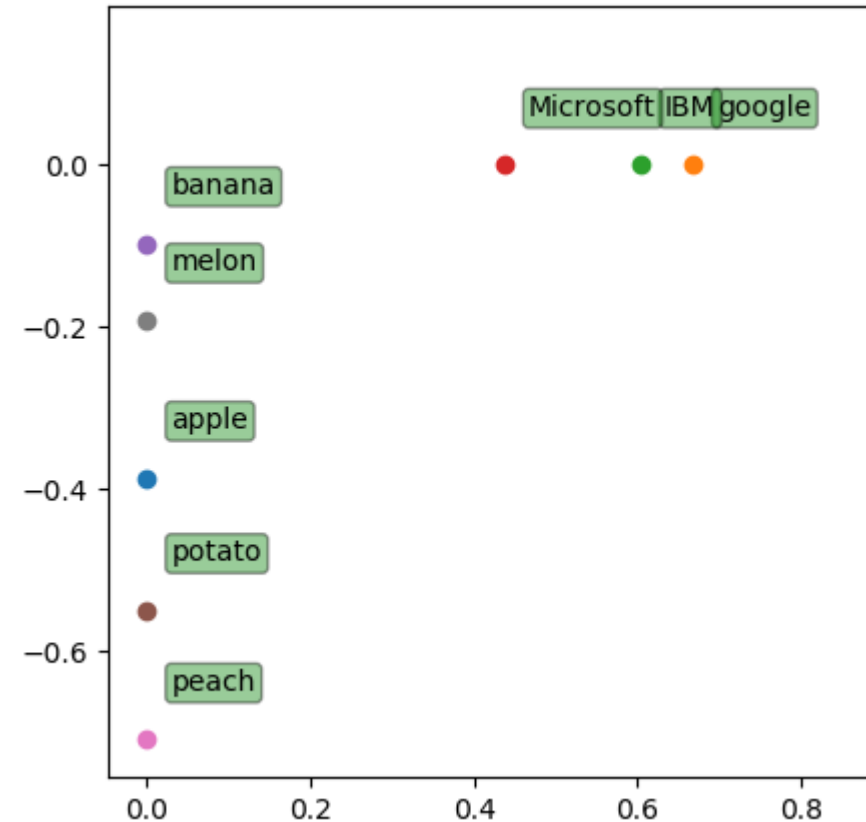
English Wikipedia embeddings

Current Focus

Ebrahim Ansari
September 2018

▶ Image Captioning and Visual Question Answering using deep learning models

▶ Fake News Detection / Sentiment Analysis on Tweeter

  ▶ English and Persian

▶ Word Sense Disambiguation using DL

Current Focus – Image Captioning

Ebrahim Ansari
September 2018

a brown bear standing on a rock in forest area.

Ebrahim Ansari
September 20

Which one is better?!

a man in a <u>red jacket</u> walks down a walking trail next to a park.

a man with a <u>hat</u> on walking near a park and a <u>bench</u>.

a man in red is walking down the <u>long street</u>.

view down a city walkway and street, with <u>grass</u>, <u>pedestrians</u>, <u>trees</u>, cars on street and parked on side of street, a bench, and some <u>buildings in distance</u>.

Current Focus – Image Captioning Unseen words

Ebrahim Ansari
September 2018

▶ In general, current approaches cannot describe previously unseen compositions of objects, even though the individual objects might have been observed in the training data.

▶ Our main focus is to describe the objects in the test set, which were unseen in the training set.

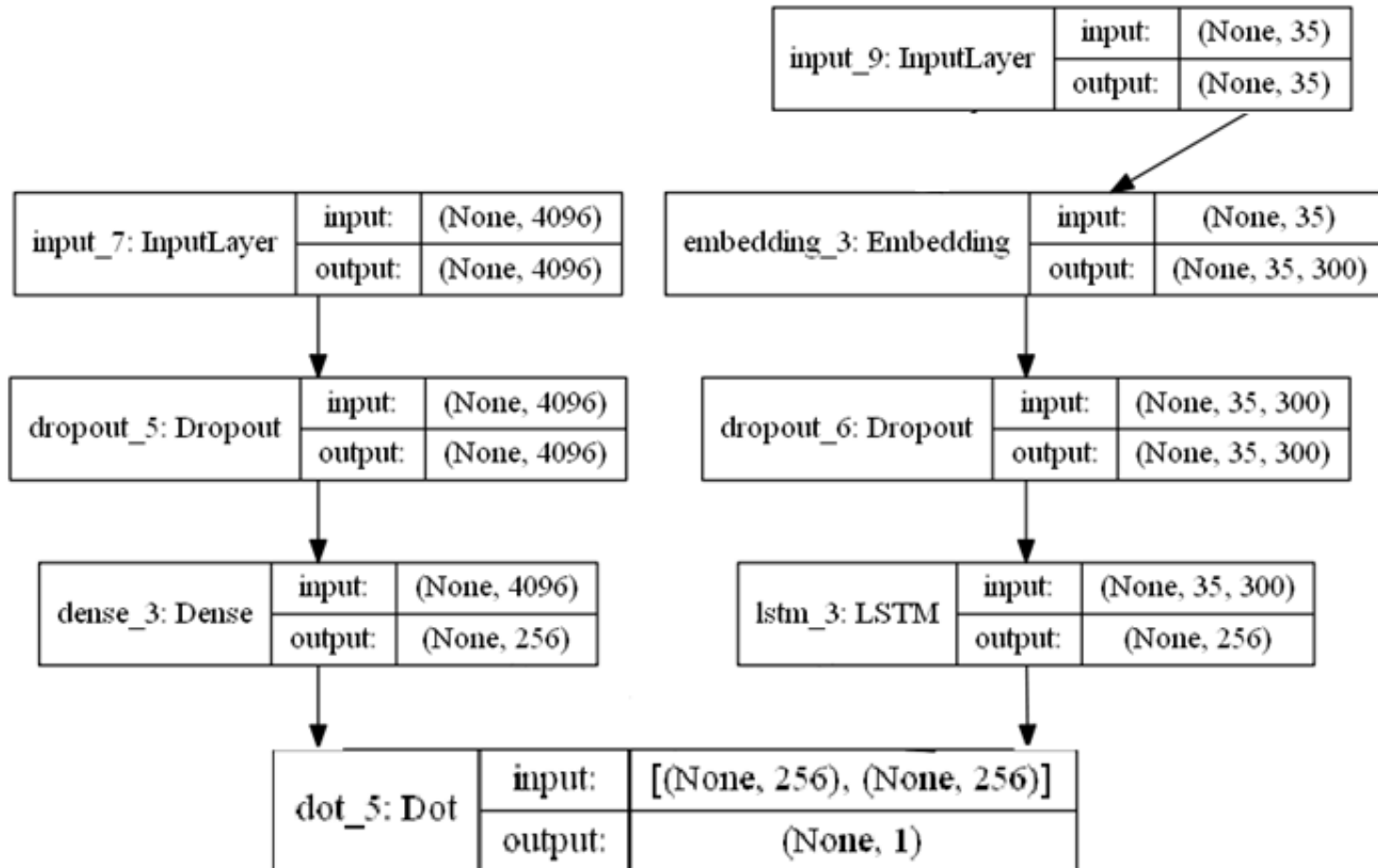# Current Focus – Image Captioning For Persian

Ebrahim Ansari
September 2018

- This project defined in our University

- Using current models

- First results are acceptable

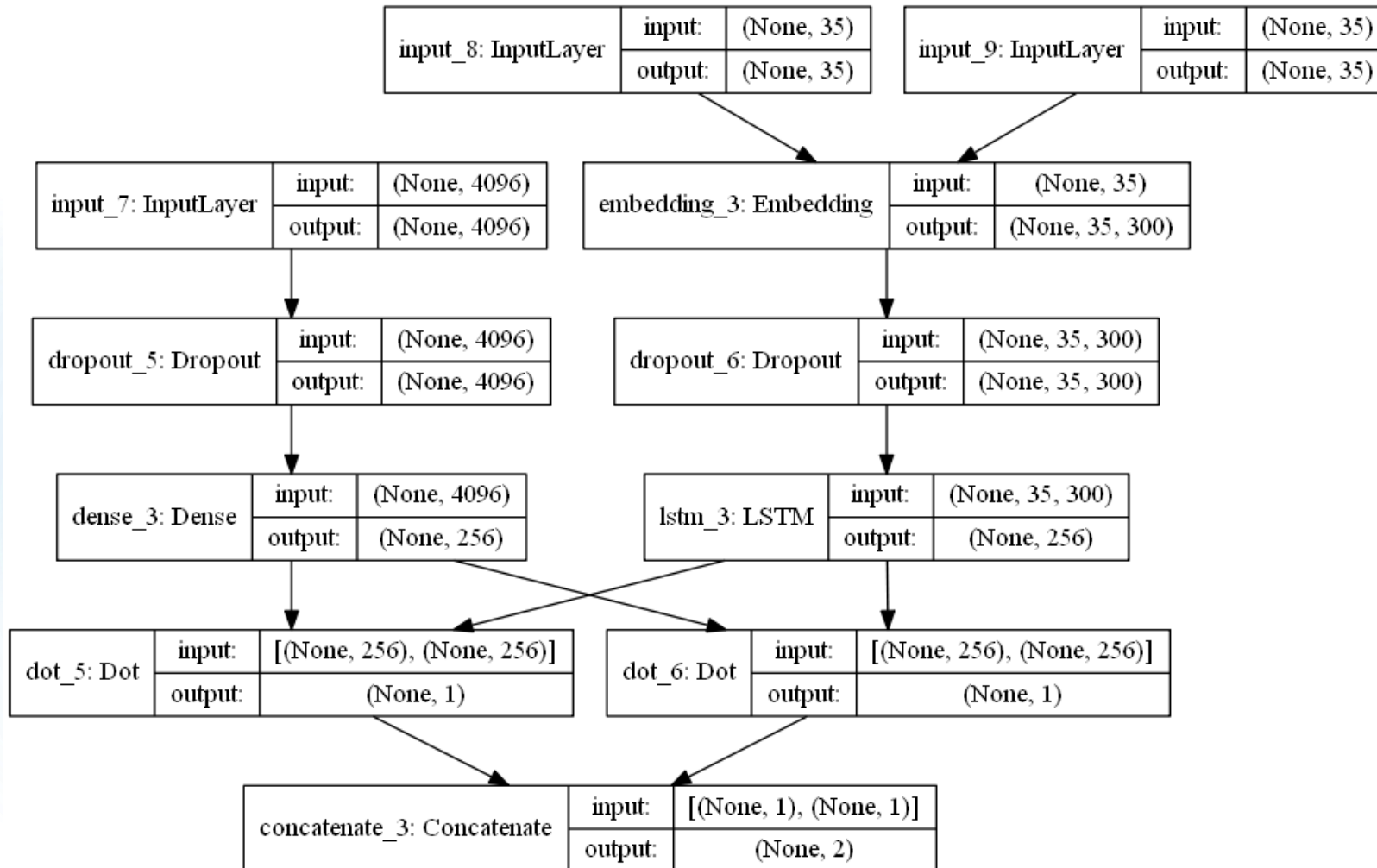# Current Focus – Image Captioning Corpus creation from twitter

Ebrahim Ansari
September 2018

input_9: InputLayer	input:	(None, 35)
	output:	(None, 35)

input_7: InputLayer	input:	(None, 4096)
	output:	(None, 4096)

embedding_3: Embedding	input:	(None, 35)
	output:	(None, 35, 300)

dropout_5: Dropout	input:	(None, 4096)
	output:	(None, 4096)

dropout_6: Dropout	input:	(None, 35, 300)
	output:	(None, 35, 300)

dense_3: Dense	input:	(None, 4096)
	output:	(None, 256)

lstm_3: LSTM	input:	(None, 35, 300)
	output:	(None, 256)

dot_5: Dot	input:	[(None, 256), (None, 256)]
	output:	(None, 1)

Current Focus – Image Captioning Negative/Positive scoring

Ebrahim Ansari
September 2018

Current Focus – Fake News Detection

Ebrahim Ansari
September 2018

Is this a real picture? Who is the first rider? What was the photographer's point? Explain the use of satire.

Fake News Detection (Buntain and Goldbeck 2017)

Ebrahim Ansari
September 2018

- ▶ Reading and Preprocessing data from pheme dataset

- ▶ Convert all tweets to vectors (300) using trained word2vec model

- ▶ Create our Sequential model using:
  - ▶ Two LSTM layers
  - ▶ Two CNN layers
  - ▶ One dense layer

Fake news detection - What are not considered

Ebrahim Ansari
September 2018

- ▶ Comments of tweets
  - ▶ Using sentiment analysis (in progress)

- ▶ Likes, retweets and mentions
  - ▶ Using a ranking system similar to PageRank but considering positive and negative scores

დიდი მადლობა

merci

谢谢

Хвала

با تشکر از شما

dziękuję

danke

ధన్యవాదాలు

אַ דאַנק

cảm ơn bạn

dankie jy

شكراً

hvala

ありがとう

감사합니다

ju faleminderit

thank you

ขอบคุณ

शुक्रिया

Дзякуй

eskerrik asko

gràcies

gracias

grazie

তোমাকে ধন্যবাদ

நன்றி

děkuji

σας ευχαριστώ

takk

Terima kasih

آپ کا شکریہ

aliquam

спасибо