Question Answer System using BiDaF

AKSHAY NAVALAKHA

About Me

IIT Gandhinagar
Indian Institute of
Technology Gandhinagar

CORNELL

NVIDIA®

Stanford
ARTIFICIAL
INTELLIGENCE

Agenda

- Motivation for Question Answering System
- Reading Comprehension using Bi-Directional Attention Flow (BiDaf)
- Detail Explanation of Each Layer
- Training Loss and Hyperparameters
- Next Steps

Motivation for Question Answering(QA) System

- There is massive amounts of information in full text documents i.e. the web
- QA systems help to retrieve useful information from the web
- Voice assistants Alexa, Google Assistant, etc. are QA systems
- Other examples of QA systems are chatbots used in customer service

2 Step Process

- Finding the documents that (might) contain the answer
  - Which can be handled by traditional information retrieval/ web search
- Finding the answer in a paragraph or a document
  - This problem is often termed as reading comprehension system

2 Step Process

- Finding the documents that (might) contain the answer
  - Which can be handled by traditional information retrieval/ web search
- Finding the answer in a paragraph or a document
  - This problem is often termed as reading comprehension system

Problem Statement

- The goal is to find an answer in a context given a question
- For example,

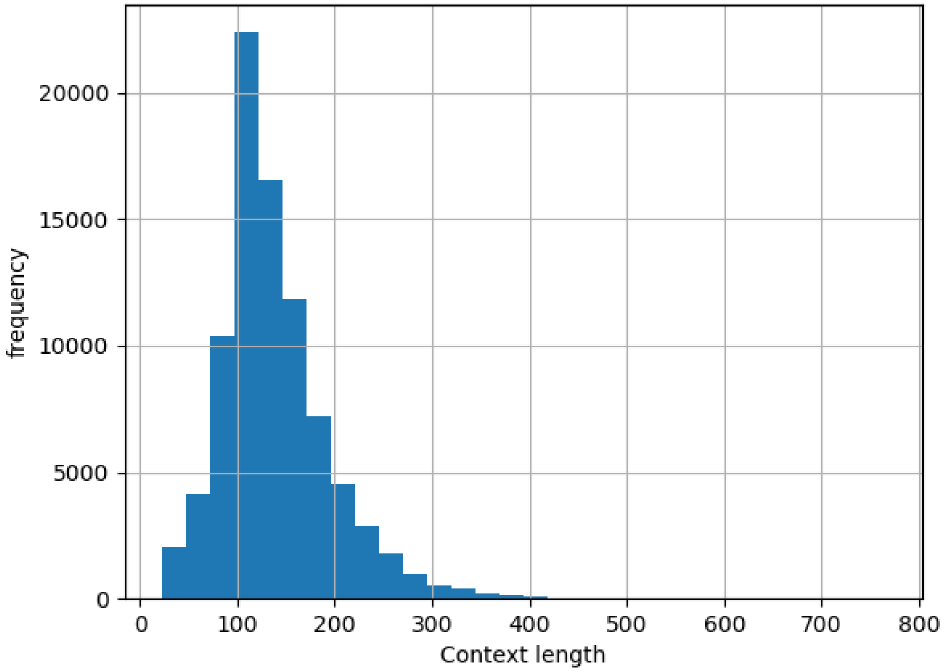
  Context: I am Akshay. I would be presenting at Pune Developer Community today. The presentation starts at 10 am.

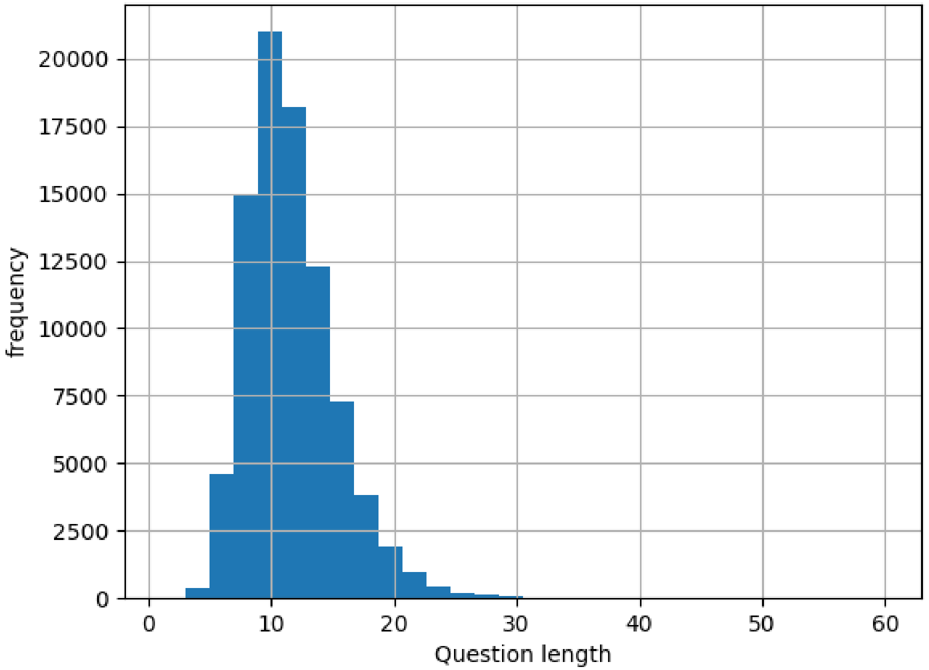
  Question: When does the presentation start?

  Answer : 10 am

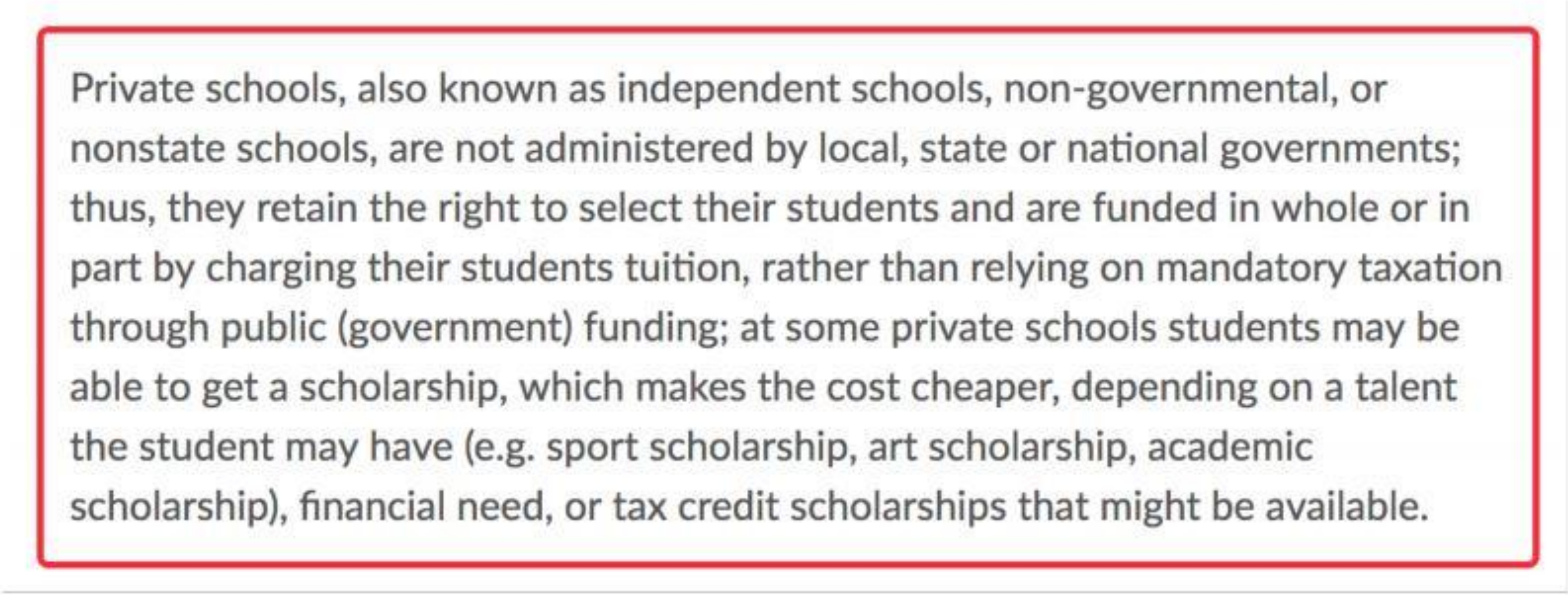
Dataset

- Stanford Question Answering Dataset is used for the problem
- Consist of 100k + questions on different set of Wikipedia articles
- It’s a closed dataset i.e. answer to the question is a part of the context and is a continuous span
- The distribution of the context length and question length





Stanford Question Answering Dataset



Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?
Gold answers: 1. independent 2. independent schools 3. independent schools

Along with sport and art, what is a type of talent scholarship?
Gold answers: 1. academic 2. academic 3. academic

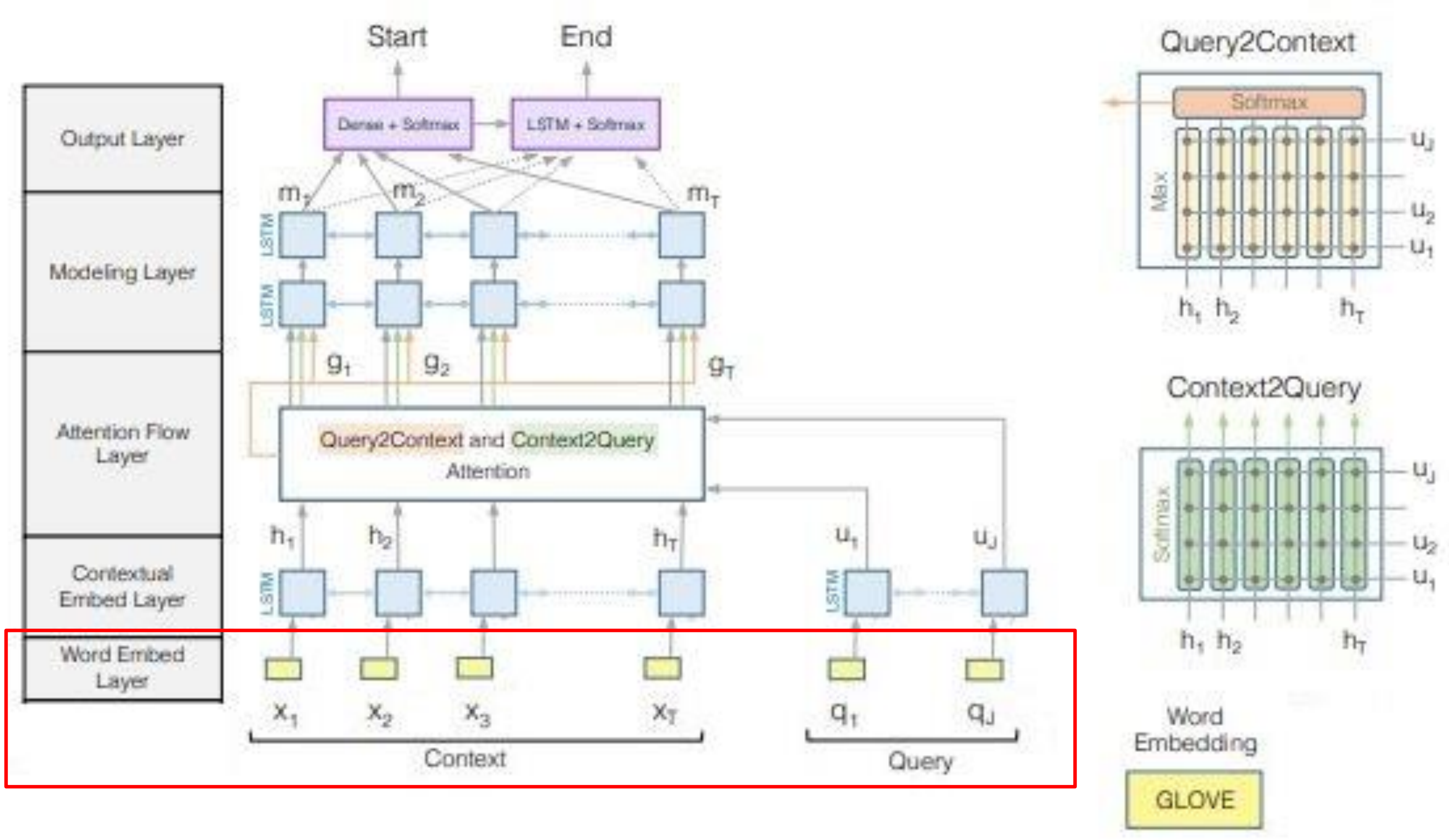
Approach

- Need to develop an end to end system
- Transform the words of context and query into vectors to be read by machine (word embeddings)
- Need each word to be aware of the words coming before and after it (Recurrent Neural Network)
- Find the co-relation between the context and query (Attention Mechanism)
- Find the pair of start and end word indices having the maximum probability
- Start with a baseline and then build further

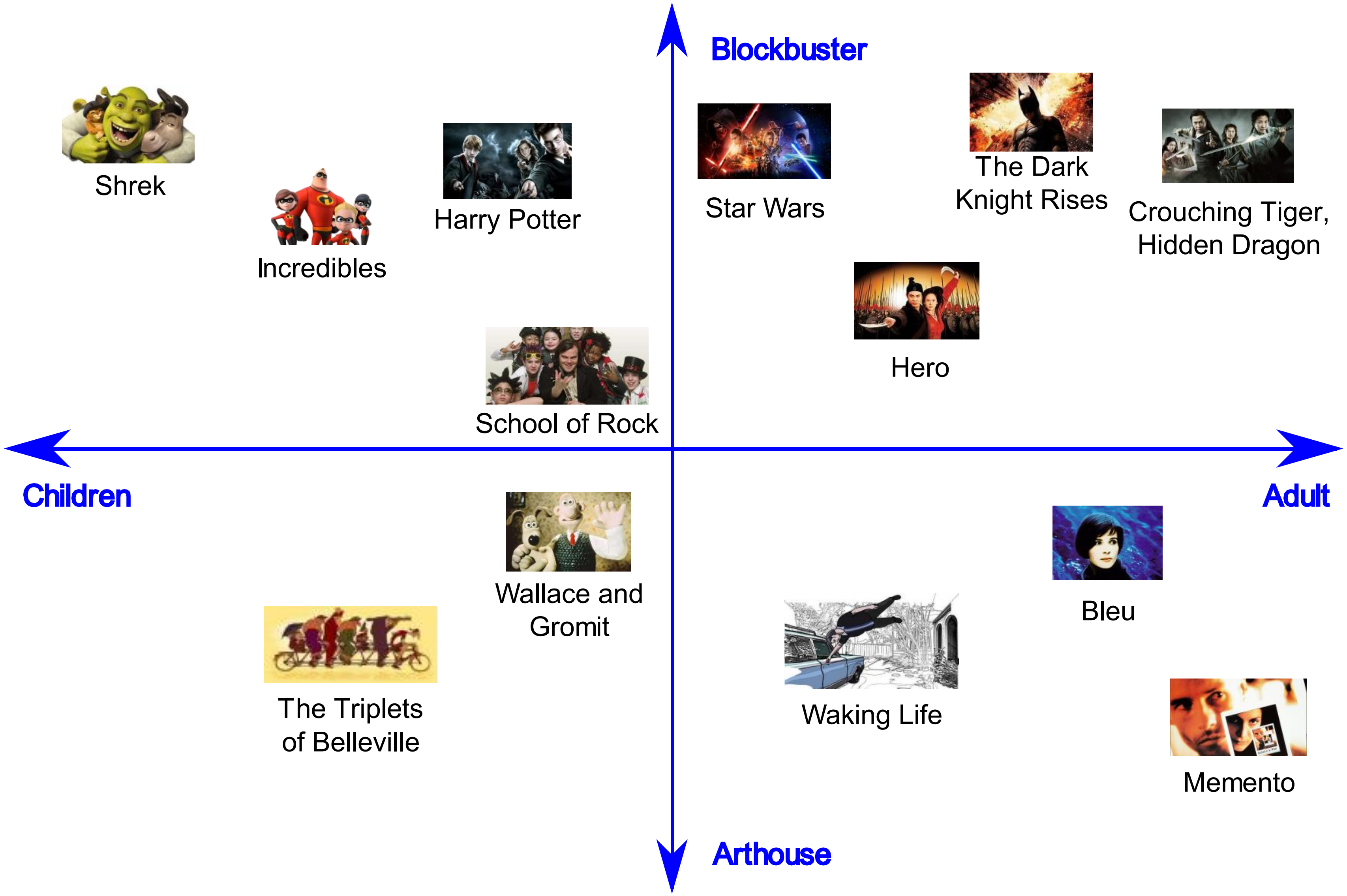
Agenda

- Motivation for Question Answering System
- Reading Comprehension using Bi-Directional Attention Flow (BiDaf)
- Detail Explanation of Each Layer
- Training Loss and Hyperparameters
- Next Steps

Bidaf Model For QA System



Example with Movie Embeddings

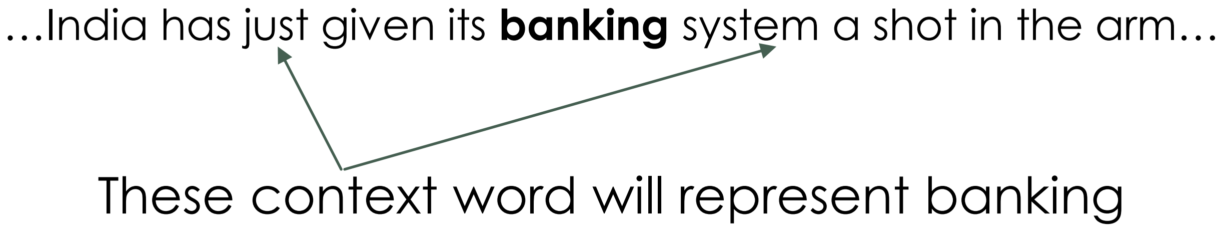


Word Embeddings: Representing Words

- A word’s meaning is given by the words that frequently occur close-by
- When a word w appears in its text, its context is given by the set of words that appear nearby
- Use many context of w to build a representation of w
- For example

…government debt problems turning into **banking** crises as happened in 2009…

…saying that Europe needs unified **banking** regulation to replace the hodgepodge…



…India has just given its **banking** system a shot in the arm…

These context word will represent banking

Word Embeddings: Visualization

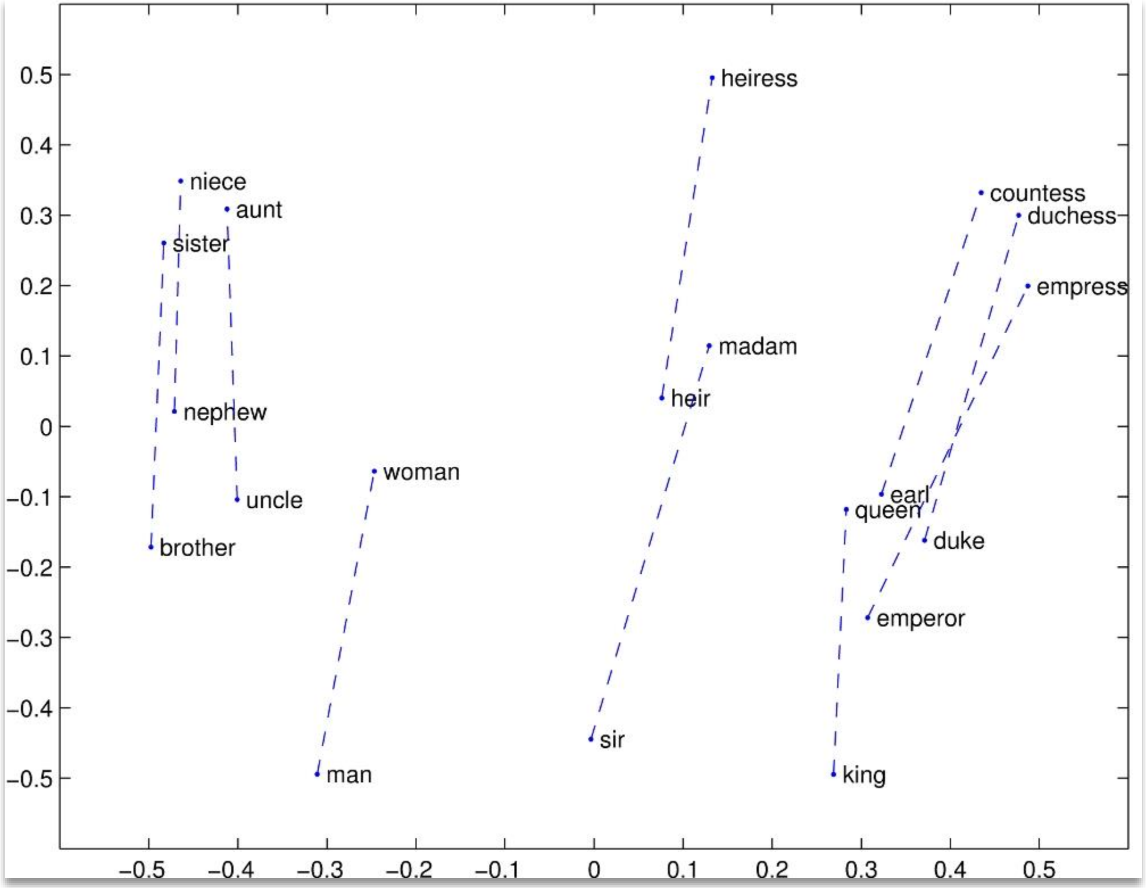


Figure: Glove vector visualizations between opposite gender

- Using word embeddings we get word vector analogies
- Example,

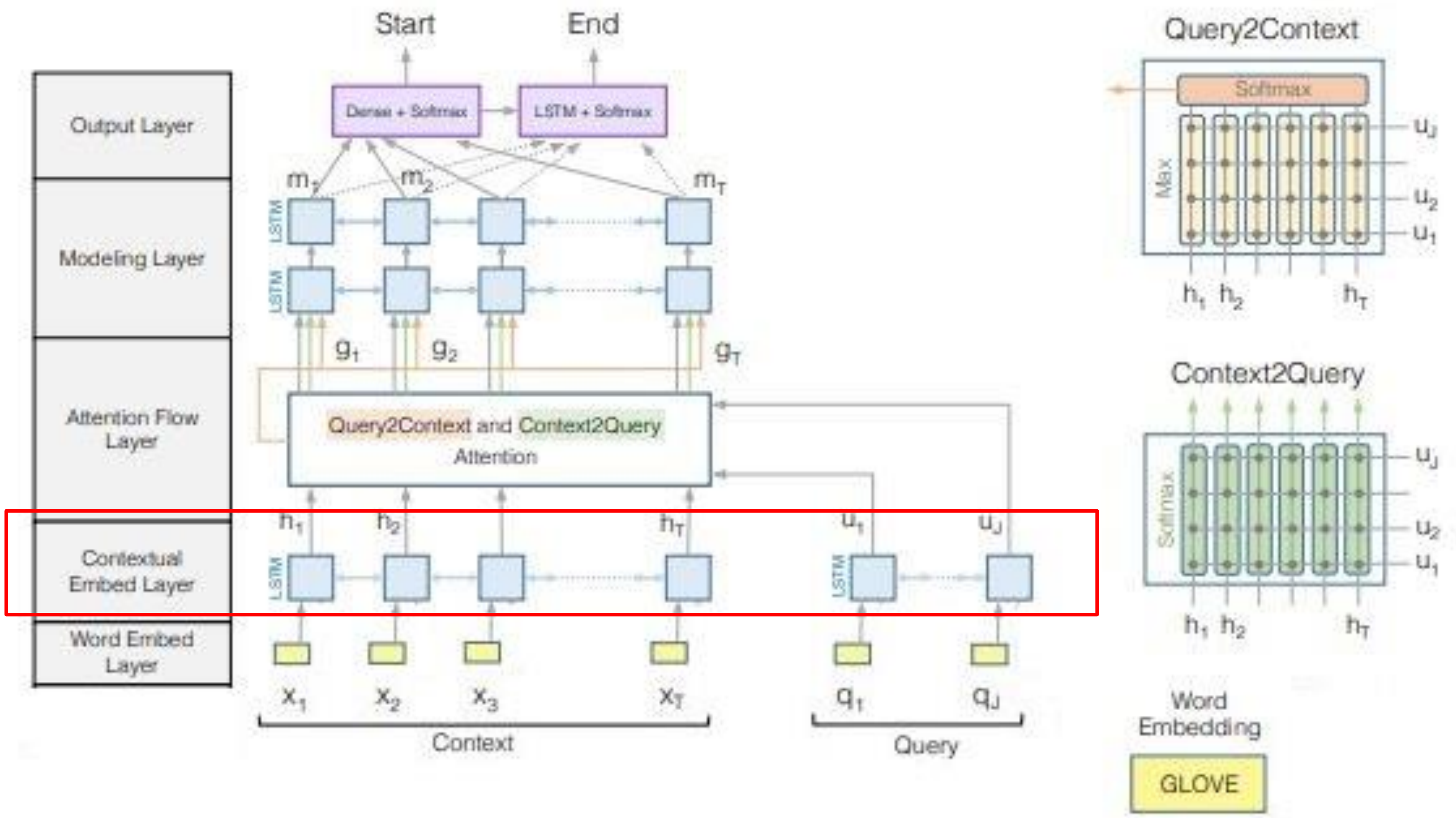
Man : woman :: king : ?

Subtracting the vector of man from woman and adding king would give as queen

Embeddings

- Embedding is a process of mapping discrete – categorical – variable to a vector
- Neural network embeddings have 3 purposes
    - Finding the nearest neighbor in the embedding space. This can be used to make recommendations based on user interest or cluster categories
    - As an input to machine learning algorithm for a supervised task
    - For visualization of concepts and relations between categories

Contextual Embed Layer



RNN Encoder Layer

- Would like each word in the context to be aware of the words before and after it
- A bidirectional LSTM will help us do that
- Output of the layer is a series of hidden layers in forward and backward direction
- Similarly we can use the same RNN layer to create question hidden states

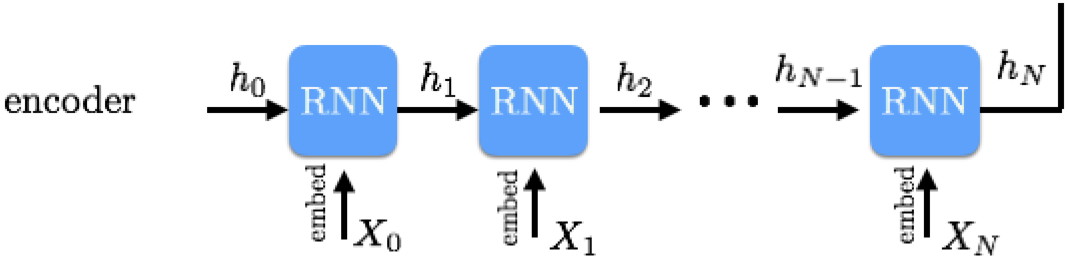
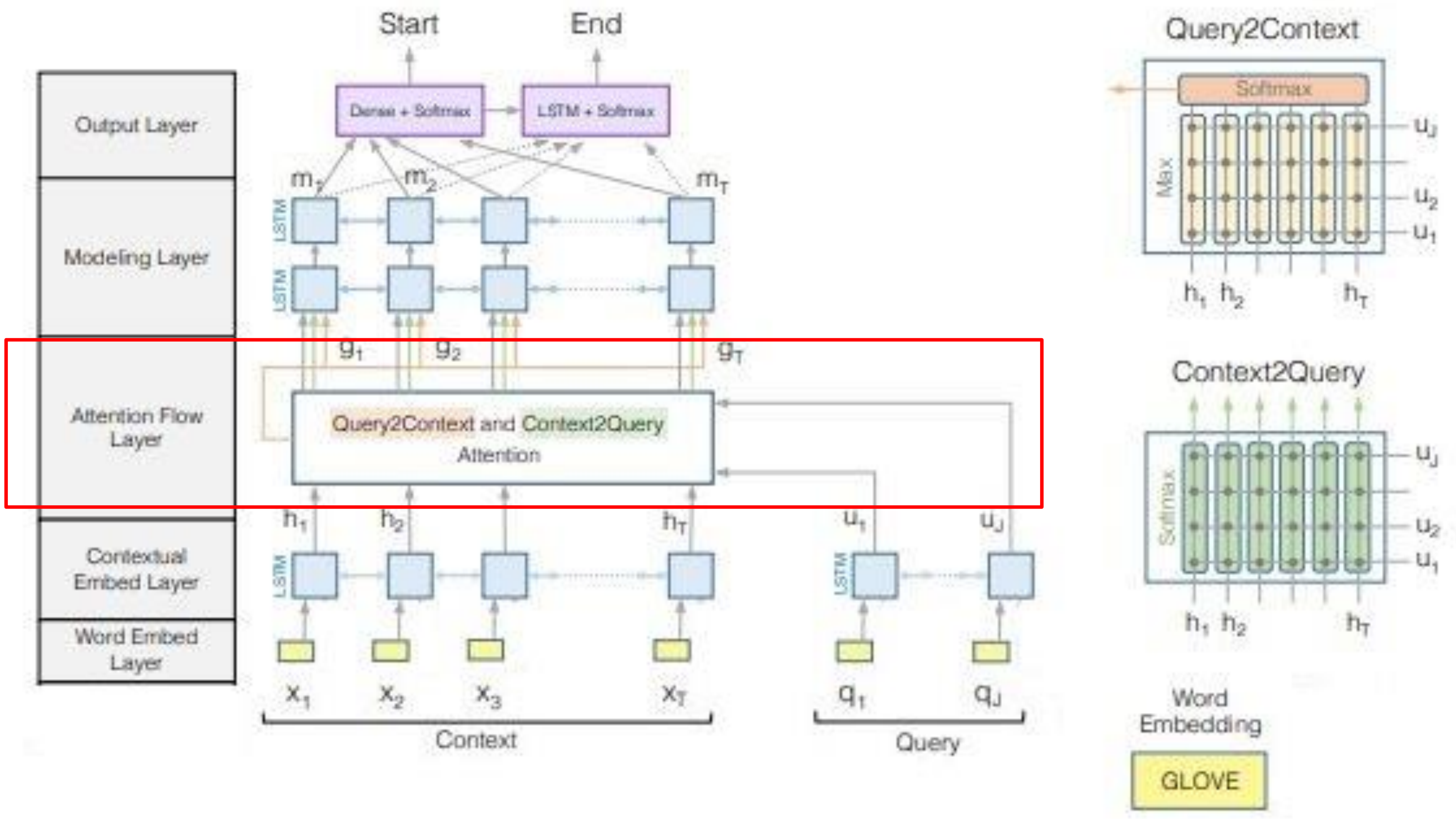


Figure: Feeding word embeddings to RNN encoder

Attention Flow Layer



Attention Layer: What is Attention ?

- It describes how closely the words are co-related to each other
- Example,

  She is eating a green apple

  When we see eating we expect a food word very soon. The color describes the food, but probably not so much with eating

  So (eating, apple) and (green, apple) have high attention while (eating, green) has low attention

Attention Layer : Dot Product Attention

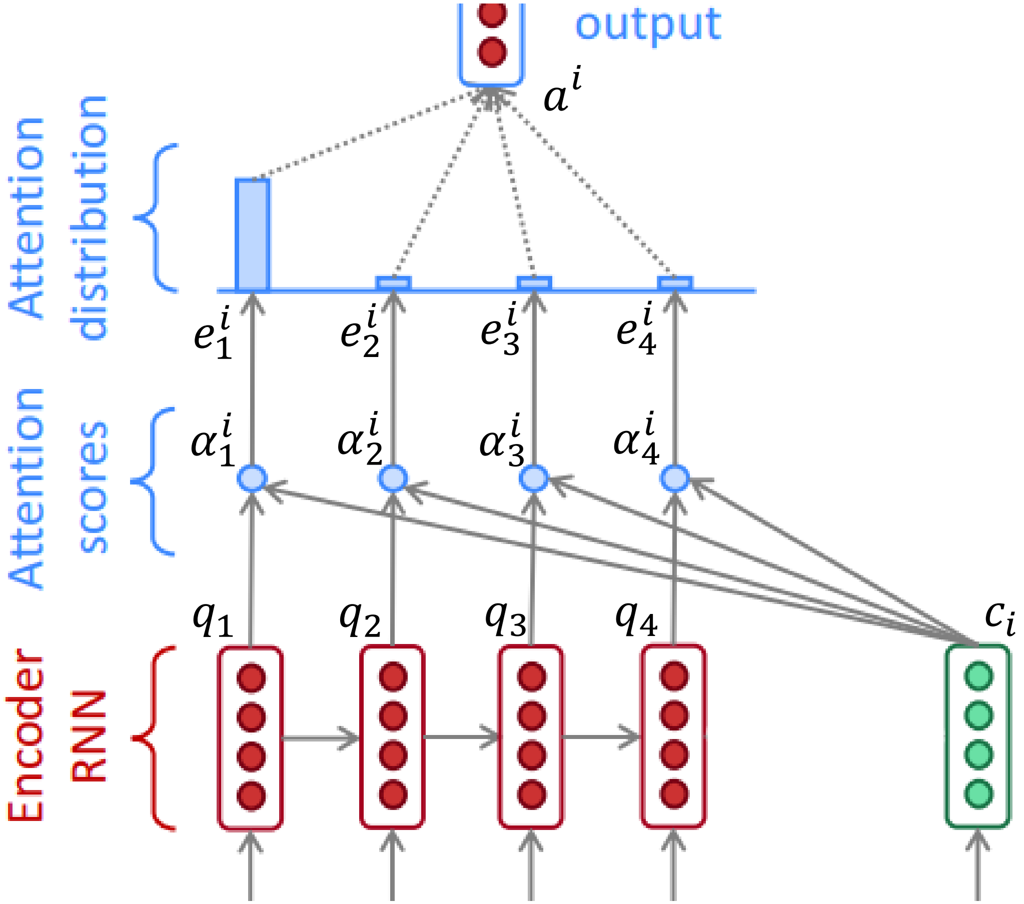


Figure: Attention Visualization

- In QA systems, we need to find given the question which words in the context should I "attend" to
- The equations for dot product attention is

$$e^i = \left[c_i^T q_1, \ldots, c_i^T q_m\right] \; \epsilon \; R^M$$

$$\alpha^i = softmax\left(e^i\right) \epsilon \; R^M$$

$$a^i = \sum_{j=1}^{M} \alpha_j^i q_j \;\; \epsilon \;\; R^{2d}$$

where c_i is the i^{th} context vector,

q_j is the j^{th} question vector

Attention Layer: Similarity Matrix and Context to Query Attention

- Compute a similarity matrix $S \epsilon R^{T \times J}$ which contains the similarity score S_{ij} for each pair (c_i, q_j) of context and hidden states

$$S_{ij} = w_{(s)}^T \left[c_i; q_j; c_i o\, q_j\right]$$

- Context-to-query (C2Q) attention signifies which query words are most relevant to each context word
- Context to Query Attention is similar to the dot product attention

$$\alpha^i = softmax(Si{:}) \epsilon R^J \ \forall\, i\, \epsilon\{1, \ldots, T\}$$

$$a_i = \sum_{j=1}^{J} \alpha_j^i q_j \ \epsilon R^{2d} \ \forall\, i\, \epsilon\{1, \ldots, T\}$$

Attention Layer: Query to Context Attention

- Query to Context attention signifies which context words have the closest similarity to one of the query words
- For each context location i take the max of the corresponding row of the similarity matrix to create a vector m

$$m_i = \max_j S_{ij} \ \epsilon \ R^T$$

$$\beta = softmax(m) \ \epsilon \ R^T$$

$$c' = \sum_{i=1}^{T} \beta_i c_i \ \epsilon \ R^{2d}$$

- Output of attention layer $b_i = [c_i; a_i; \mathrm{c_i \ o \ a_i}; \mathrm{c_i \ o \ c_i'}] \ \epsilon R^{8d}$

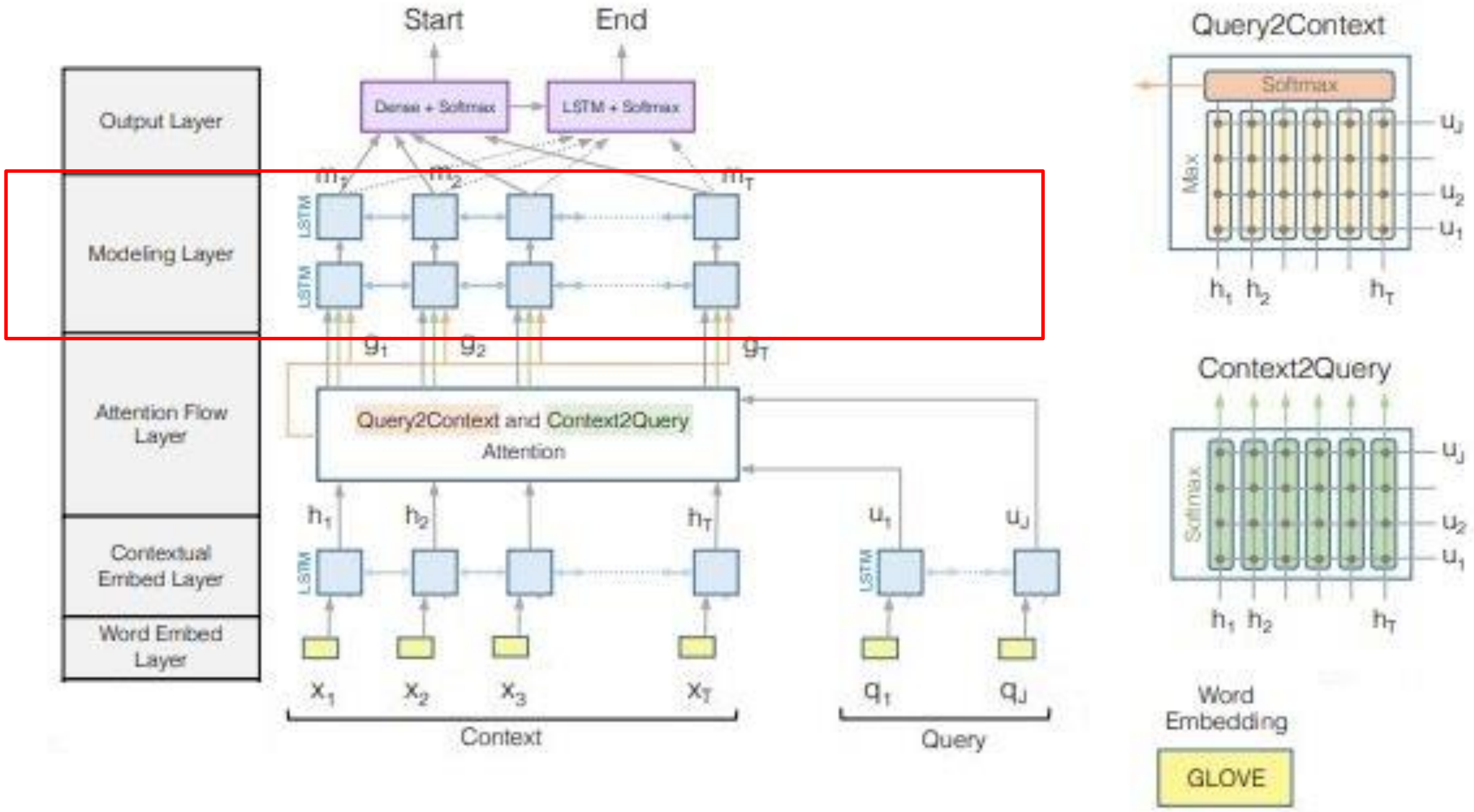
c_i - Representation of the ith context word

a_i - Context to query attention for the ith context word

$\mathrm{c_i'}$ - Query to context attention for the ith context word

Bidaf Attention Layer Tensorflow implementation

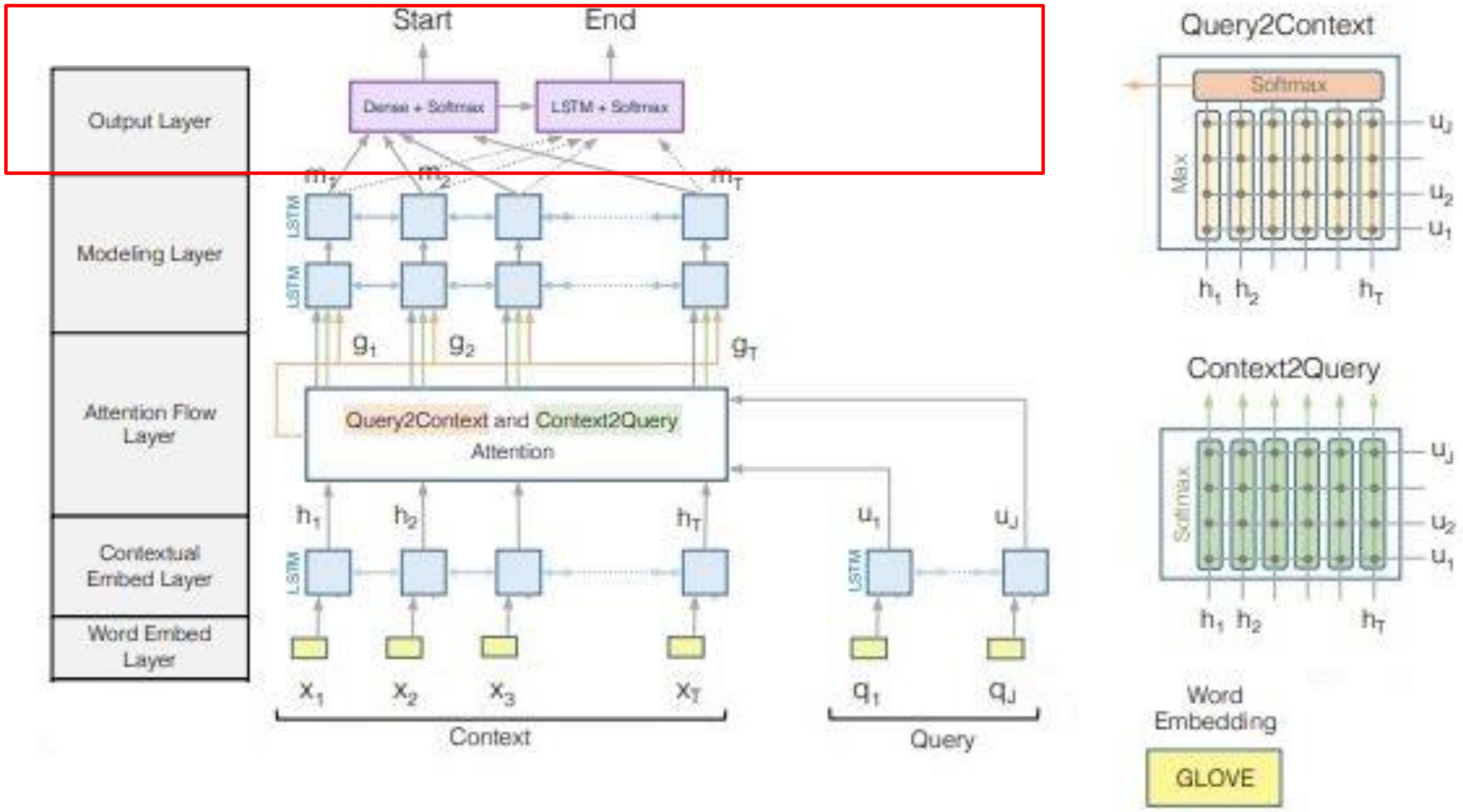
Modeling Layer



Modeling Layer

- Output of the attention layer is input to the modelling layer
- It captures the interactions between the context words conditioned on the query
- Consist of 2 layers of bi-directional LSTM with an output size of d in each direction
- Output of this will be $M \epsilon R^{2d \times T}$

Output Layer



Output Layer

- The task requires to find a sub-phrase of the paragraph to answer the query
- The phrase can be found by predicting the start and end index
- Output of the attention layer G is concatenated with the modelling layer output M
- The start word index is found by finding probability distribution of the start index

$$p^1 = softmax({w^t}_{p^1}[G; M])$$

- For the end index of the answer phrase we pass modelling layer output M through another LSTM layer with output M^2

$$p^2 = softmax({w^t}_{p^2}[G; M^2])$$

Variation in BiDaf

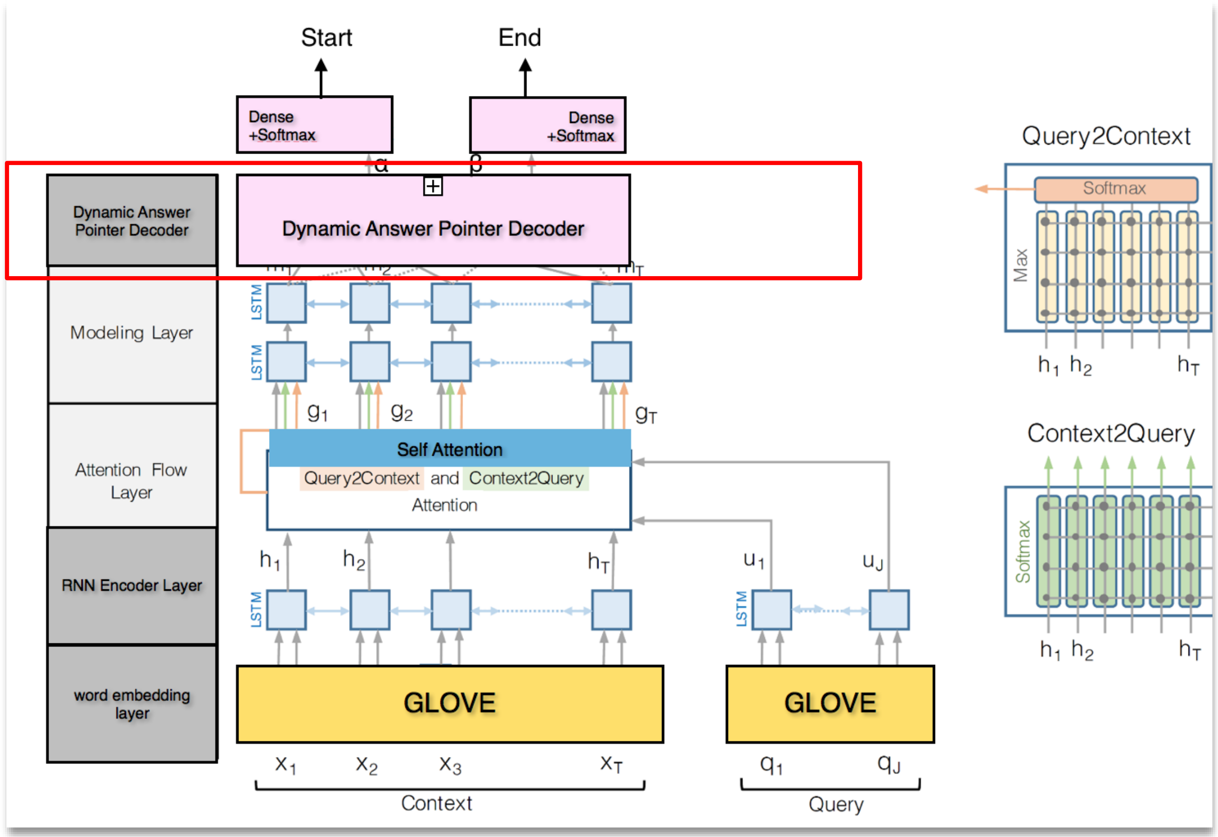


Figure: Final model for QA system

- The model is a variant of the bi-directional attention flow(bidaf) paper
- Have added a self attention layer inspired by R-NET paper from Microsoft
- The decoder of the bidaf is replaced by a dynamic pointer decoder from Co-attention in Dynamic Networks
- We will go over each of the layers

Dynamic Pointer Decoder

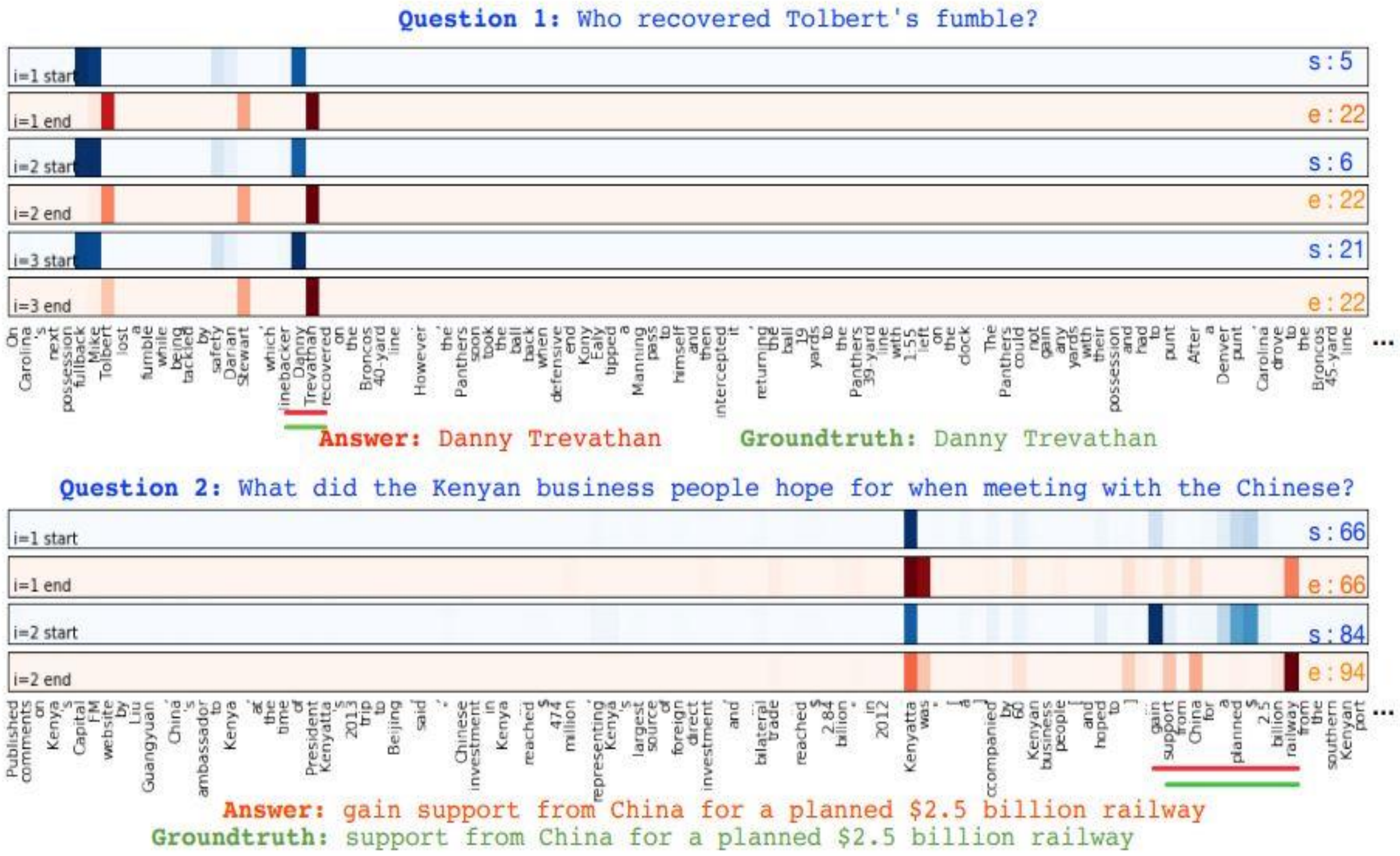


Figure: Examples of start and end distributions produced by dynamic decoder

Given a question-answer document there can be several answer spans corresponding to local maxima

The dynamic answer pointer decoder uses an iterative technique to predict the start point and end point

This allows the model to recover from initial local maxima corresponding to incorrect answer span

As shown in the figure for question 1 the start index iteratively updates from index 5 to final index 22

Agenda

- Motivation for Question Answering System
- Reading Comprehension using Bi-Directional Attention Flow (BiDaf)
- Detail Explanation of Each Layer
- Training Loss and Hyperparameters
- Next Steps

Loss Function

- The training loss is defined as sum of negative log probability of the true start and end indices

$$L(\theta) = -\frac{1}{N}\sum_{i}^{N} \log\left(p_{y_i^{start}}^{start}\right) + \log(p_{y_i^{end}}^{end})$$

where

θ is the set of all trainable weights

N is the number of samples in the training set

y_i^{start} and y_i^{end} are the true start and end indices of the i^{th} sample

p_k indicates the k^{th} value of vector p

Hyperparameter Tuning

- The table below shows the different hyperparameters and their range

Hyperparameters	**Values**
Dropout	(0.15, 0.2, 0.3, 0.4)
Hidden Layer Size	(75, 100, 200)
Learning Rate	(0.0001,0.0005, 0.001, 0.005, 0.01)

- Adam optimizer used for training

Question Answering System Demo

- http://ec2-18-191-73-4.us-east-2.compute.amazonaws.com:8000/

Evaluation Metrics

- There are three golden answers per question
- We will consider the following two metrics
  - Exact Match(EM) Score: 1/0 accuracy on whether the system matches one of the three answers
  - F1 score: Consider each of the golden answers as bag of words. F1 score is the harmonic mean of precision and recall score.
  - Net F1 score is average of per question F1 scores

Results

- The final model was built in incremental steps from the baseline
- Summary of the results is shown below

Model	Hidden Size	Dev EM	Dev F1	Test EM	Test F1
Baseline	200	36	45	-	-
BiDaf	200	64.42	74.44	65.59	75.12
BiDaf	100	62.5	73.2	-	-
Bidaf with Variation	100	63.95	74.05	64.9	74.8

Next Steps

- Bi-directional Attention Flow paper was published in ICLR 2017
- The paper serves as a good starting point for building a complex neural network and tries to emphasize how attention mechanism works
- Solving QA task is a research problem and many different papers have been published since then
- The newer models have more emphasis on attention mechanism like BERT

References

- Bidirectional Attention Flow for Reading comprehension https://arxiv.org/pdf/1611.01603.pdf
- DynamiC Coattention Network for Question Answering https://arxiv.org/pdf/1611.01604.pdf
- CS224n – Natural Language Processing using Deep Learning Stanford

Contact

- Email : akshaynavalakha@gmail.com
- LinkedIn: https://www.linkedin.com/in/akshaynavalakha/

Questions