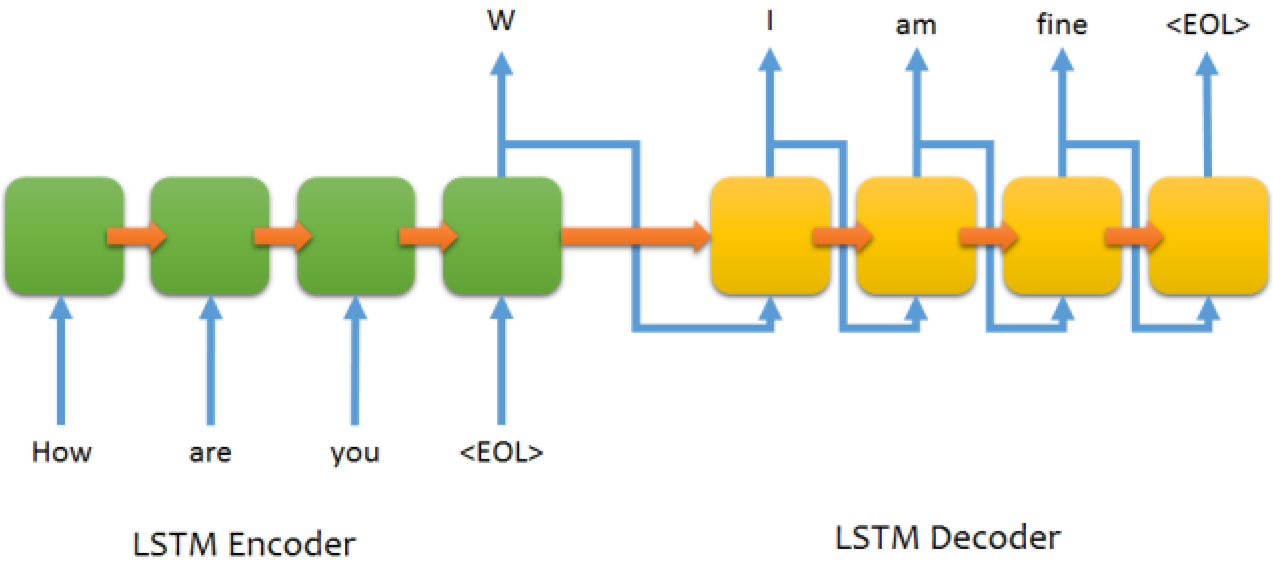
R-Seq2Seq: Relational Sequence to Sequence Learning for Question Answering

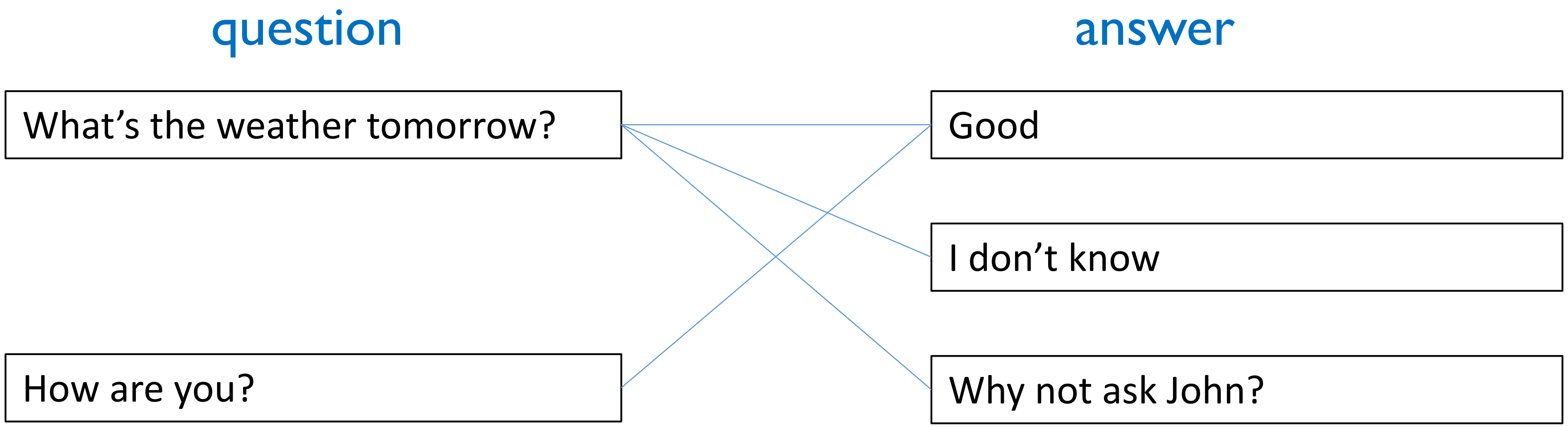
Xiaodong Gu

2017.08.30

Sequence to Sequence Learning



Consider Relations between Q and A



question

answer

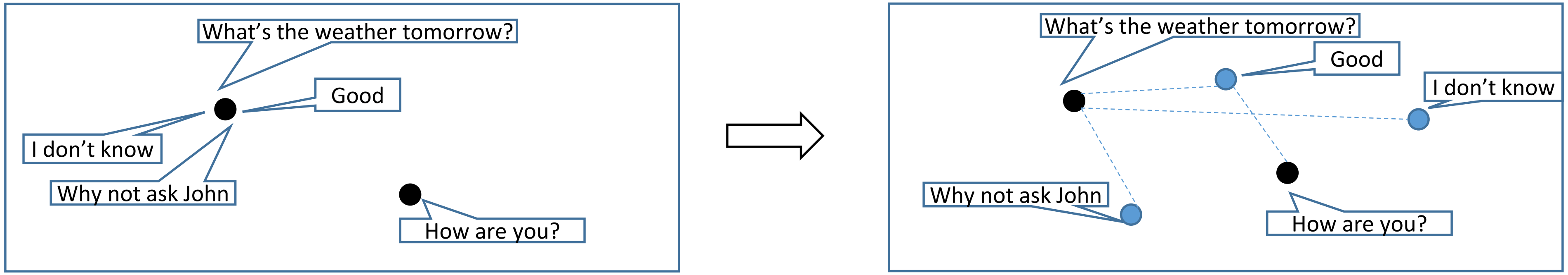
What's the weather tomorrow?

How are you?

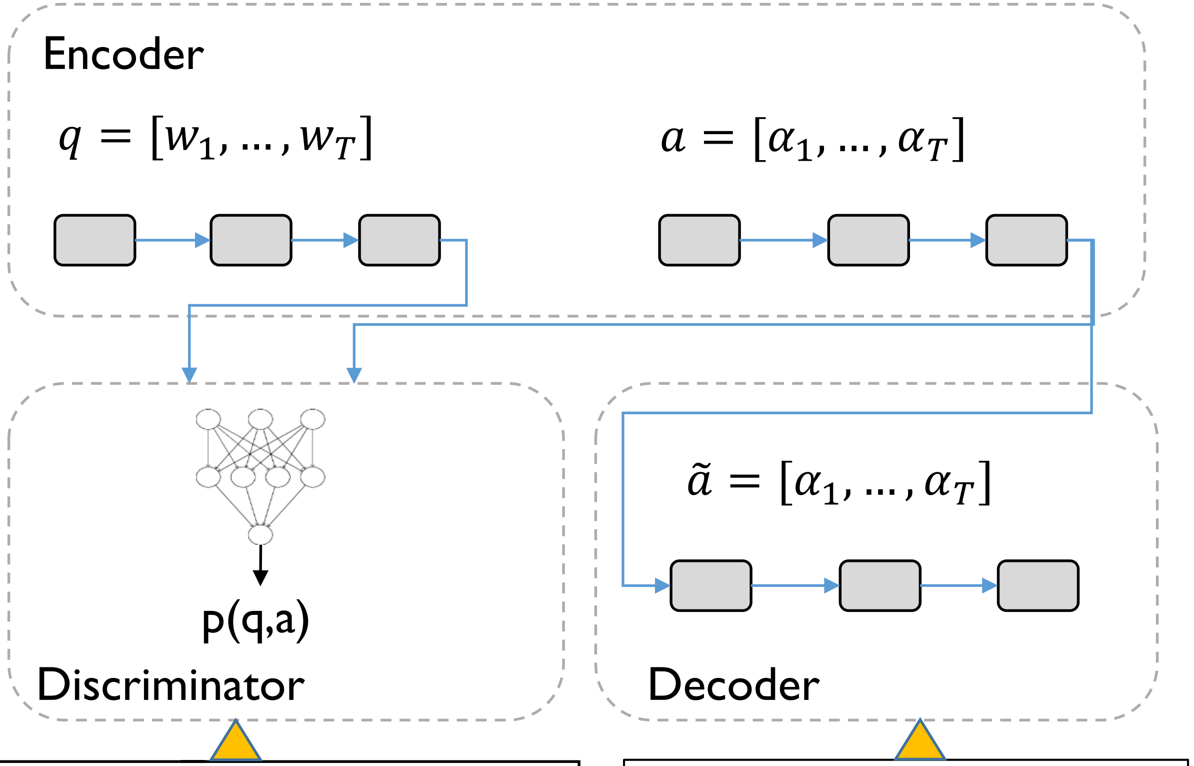
Good

I don't know

Why not ask John?



Proposed Model

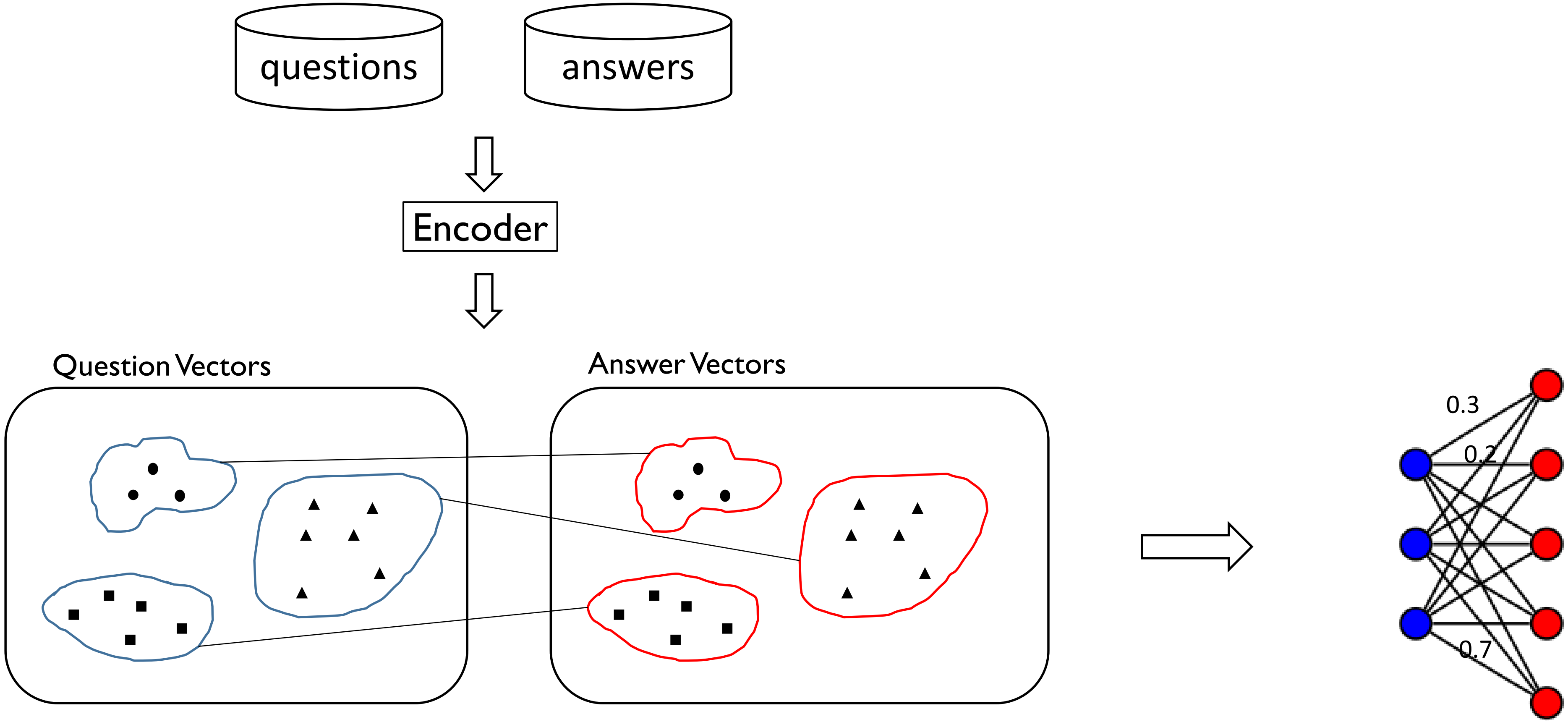

Encoder
$q = [w_1, \dots, w_T]$
$a = [\alpha_1, \dots, \alpha_T]$
p(q,a)
Discriminator
$\tilde{a} = [\alpha_1, \dots, \alpha_T]$
Decoder

$$\begin{aligned}\mathcal{L}_{dis} = y \log p(a|q) \\ +(1-y)\log(1-p(a|q))\end{aligned}$$

$$\mathcal{L}_{dec} = -\sum_{i=1}^{T} \log p(\alpha_i|q)$$

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{dec}+\lambda\mathcal{L}_{dis}$$

Relation Summarization



questions

answers

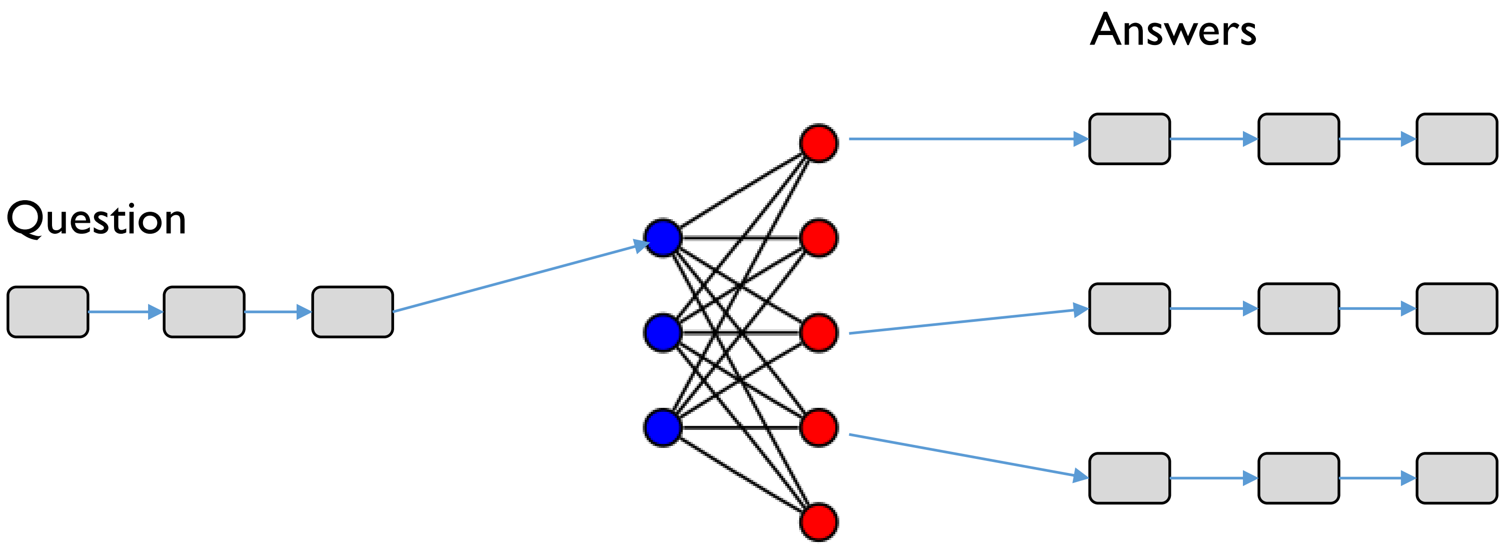
Encoder

Question Vectors

Answer Vectors

Q-A relations

Answer Sampling



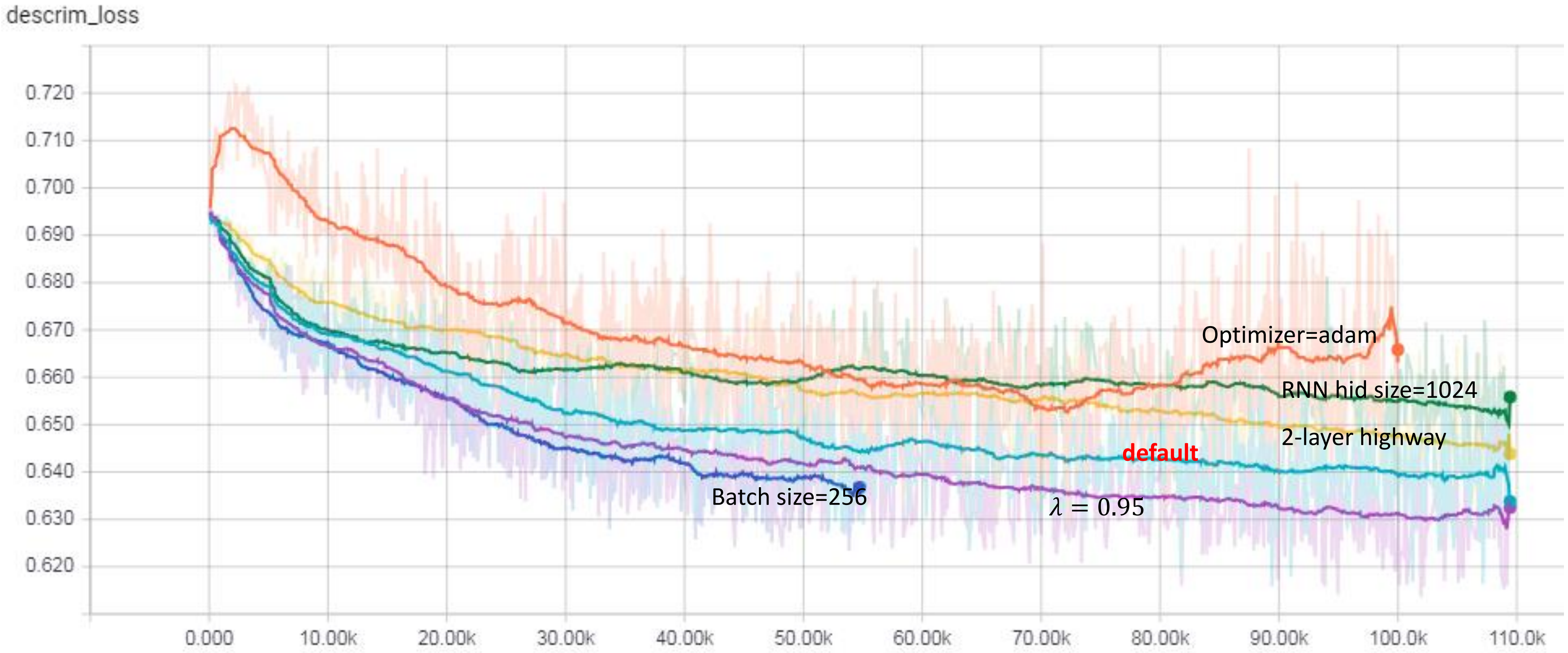
Datasets

- Opensub
- Ubuntu Dialog Corpus
- Cornell Movie Dialogs Corpus
- Twitter Chat log

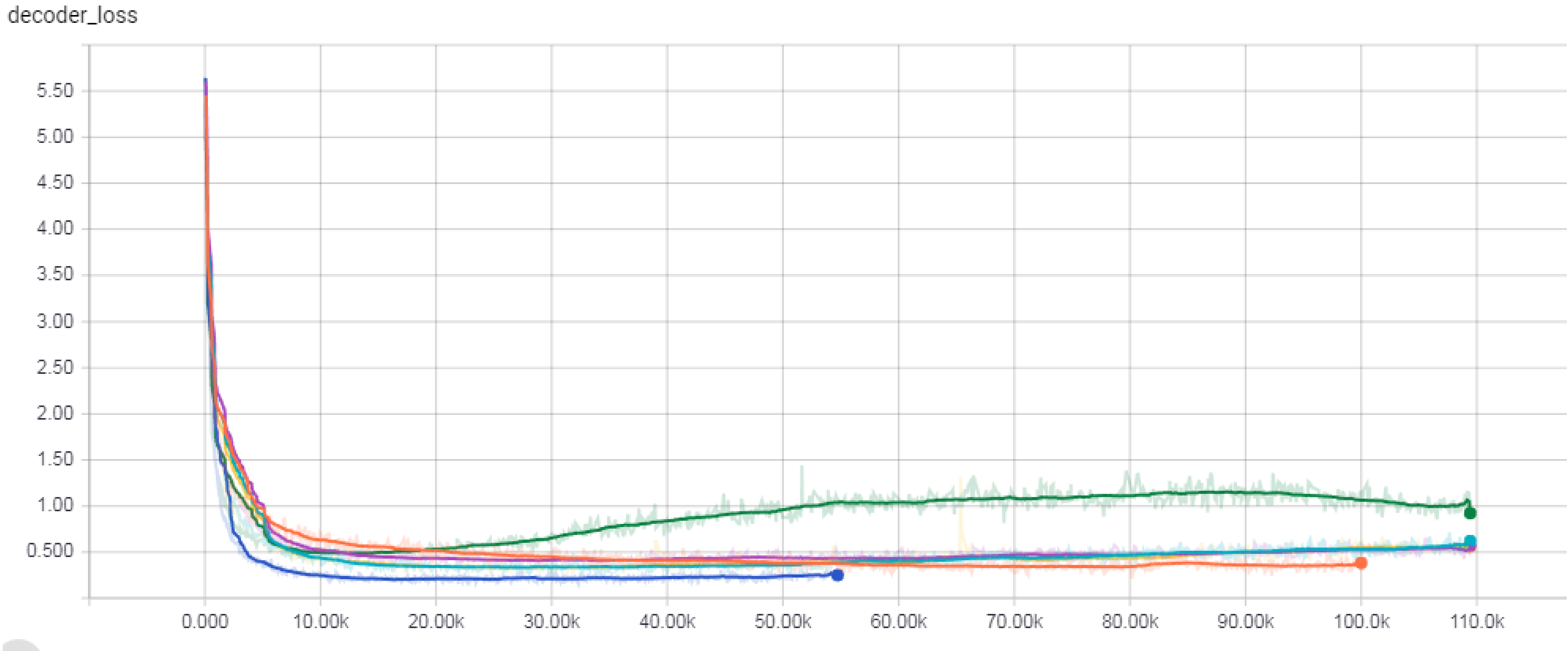
Training

- Optimizers:
  - Embedder: Adam
  - Encoder: Adam
  - Discriminator: Adadelta
  - Decoder: Adam
- Hyper Parameters
  - Batch Size: 128
  - λ: 0.9
  - RNN Hidden Size: 512
  - Word Embedding Size: 100
  - Learning rate=0.001

Training


descrim_loss
0.720
0.710
0.700
0.690
0.680
0.670
0.660
0.650
0.640
0.630
0.620
0.000
10.00k
20.00k
30.00k
40.00k
50.00k
60.00k
70.00k
80.00k
90.00k
100.0k
110.0k
Optimizer=adam
RNN hid size=1024
2-layer highway
default
Batch size=256
$\lambda = 0.95$

Training



decoder_loss
5.50
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.500
0.000
10.00k
20.00k
30.00k
40.00k
50.00k
60.00k
70.00k
80.00k
90.00k
100.0k
110.0k

Training – Clustering

- Minibatch K-Means
  - N_clusters: 10000
  - batch_size: 10001

Instance Table

ID	Vector	clus
0	[0.2,0.5,...0.8]	0
1		12
...		
n		

Questions

ID	Vector	clus
0	[0.1,0.4,...0.8]	11
1		7
...		
n		

Answers

Prob
0.23
0.45
...
0.77

probs

Cluster Table

ID	Center Vector	instances
0	[0.1,0.5,...0.8]	33,58,99
1		
...		
n		

Question Clusters

ID	Center Vector	instances
0	[0.8,0.5,...0.7]	12,45,27
1		55,1004
...		
n		

Answer Clusters

Relation Table

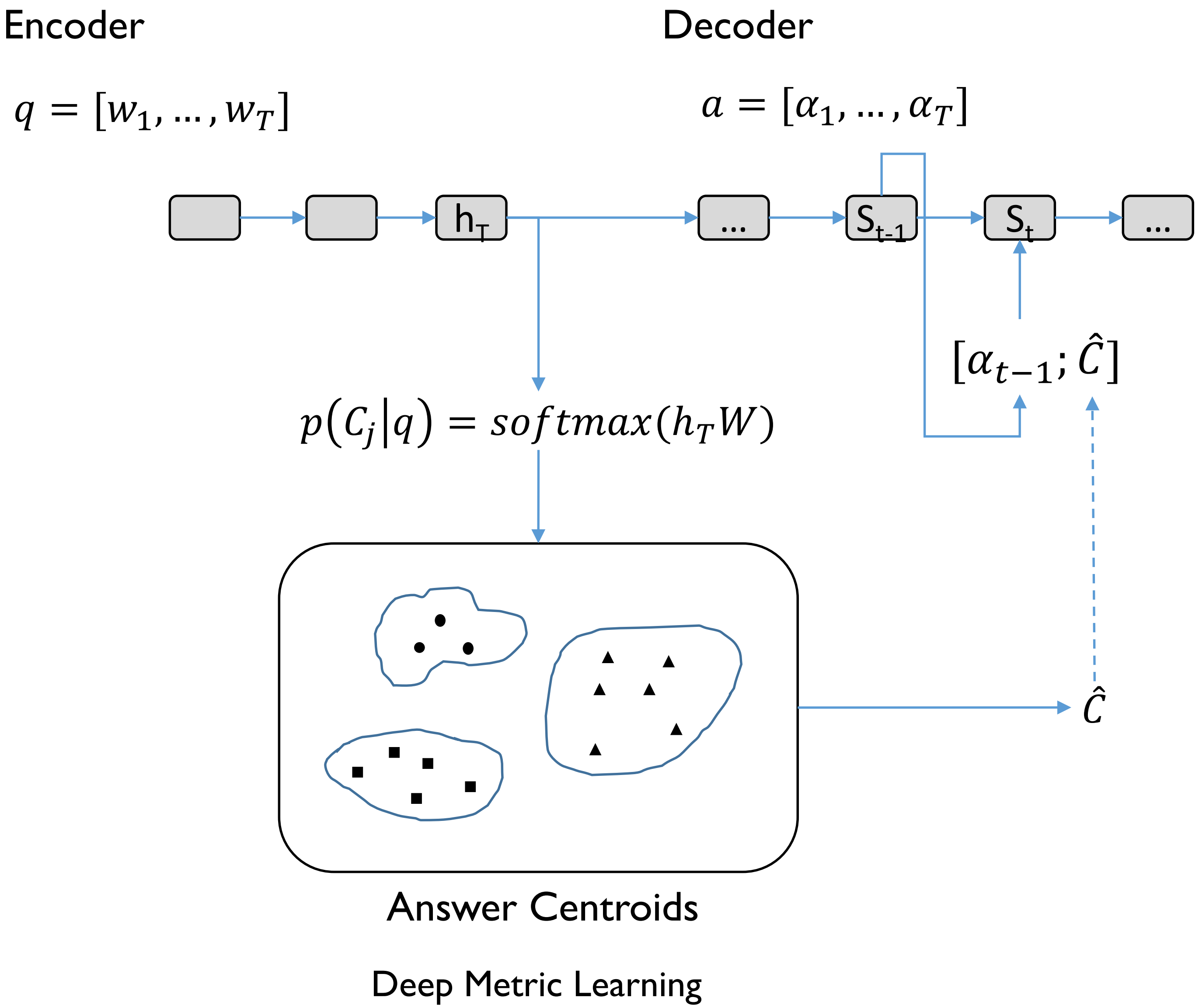
Q_clus	A_clus	Prob
0	15	0.9
0	21	0.75
0	53	0.6
1	..	
1	..	
1	...	
...		

Q-A relations

Challenges & Future Plan

- Think of better loss term for Discriminator Training
- Try to separate decoder (store the encoded vector for each answer)
- Introduce Random Noise to the Decoder (Variational Auto-Encoder)
- Applying to Source Code (API sequence) Generation

Conditional Sequence to Sequence



Encoder

Decoder

$q = [w_1, \dots, w_T]$

$a = [\alpha_1, \dots, \alpha_T]$

h_T

...

S_{t-1}

S_t

...

$[\alpha_{t-1}; \hat{C}]$

$p(C_j|q) = softmax(h_T W)$

$\hat{C}$

Answer Centroids

Deep Metric Learning

1. Clustering all answers with their word embeddings (mean of word vectors in each answer sentence)
2. Training Encoder-Decoder model with <q,a,C> triples, where q=question, a=answer and C=answer cluster.
   a) Encoder: an RNN to encode q into a vector h_T
   b) Cluster label predictor: given h_T, predict the potential cluster label $\hat{C}$
   c) Decoder: an RNN to decode answer a according to h_T and answer cluster centroid/label
   d) Train to minimize two objectives
      - Accuracy of cluster prediction
      - Accuracy of decoded answer

$$loss = -\log p(C_j|q) - \sum_{t=1}^{N} \log p(\alpha_t|q, C_j)$$

3. Prediction: given q, first predict C, then decode a using encoded q as well as C.

Q & A