Question Answering System with Bidirectional Attention Flow

Junjie Ke, Yuanfang Wang, Fei Xia
CS224N Final Project
Stanford University

Stanford | ENGINEERING
Computer Science

Background

- Question Answering (QA) with provide context for Machine Comprehension (MC)
- End-to-end deep neural network with bi-directional attention flow

Problem Statement

Dataset: SQuAD[1],100K question-answer pairs, along with a context paragraph

Problem:
Given word sequence of context with length *m*, ***p*** = {p_1, p_2, …, p_m} and question with length *n*, ***q*** = {q_1, q_2, …, q_n}, the model needs to learn a function *f*: (***p***, ***q***) → {a_s, a_e}, where the answer is a pair of scalar indices pointing the start position (a_s) and end position (a_e) of the answer to the question ***q*** in context ***p***.

Method

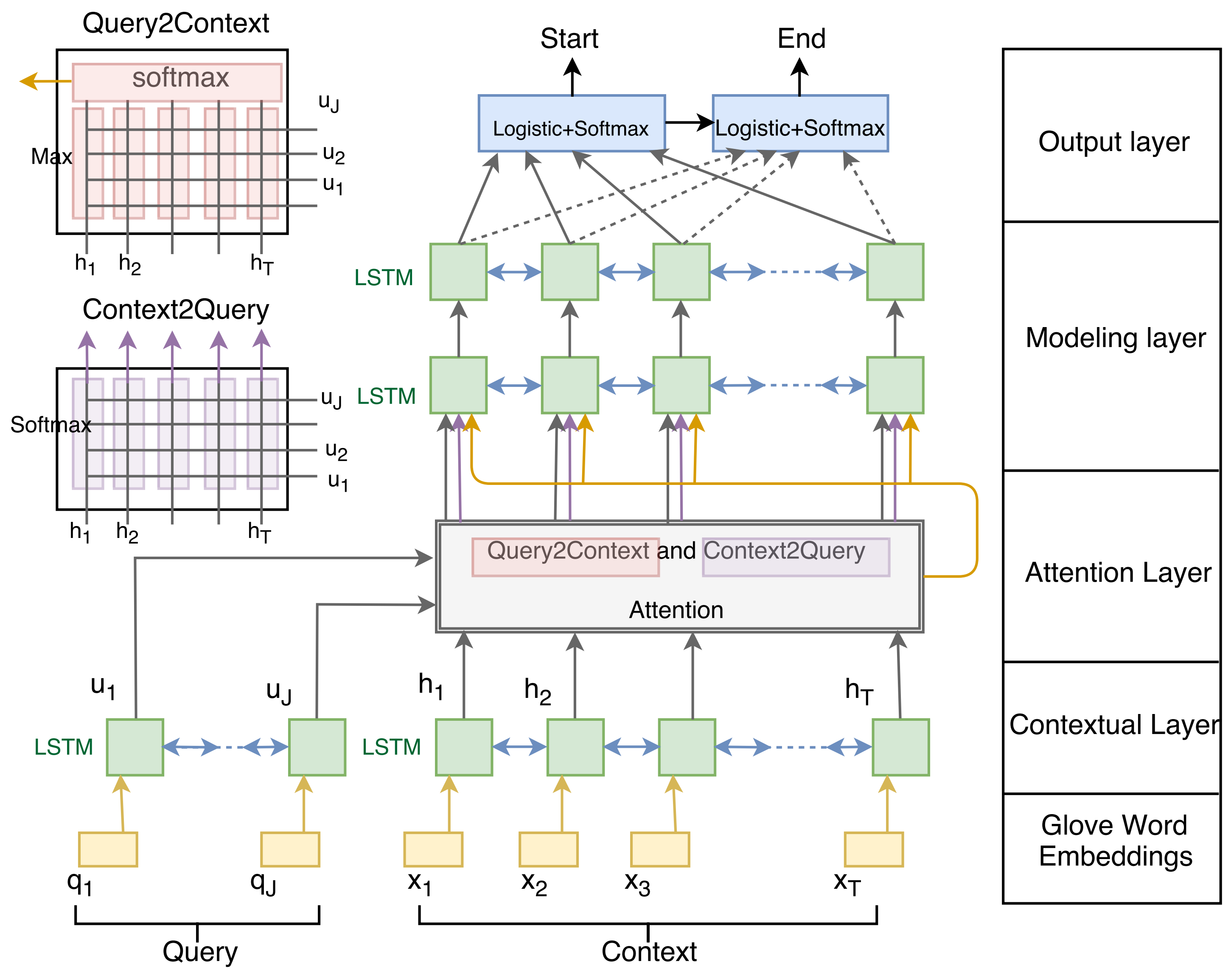

Fig 1. Architecture of our model

Baseline:
Contextual layer and Q2C attention layers.

Incrementally development with components of BiDAF[2]:
- **Embedding layer**: word embeddings only, which are fed into contextual layers without highway network
- **Attention layer**: of both Context2Query and Query2Context attention.
- **Modeling layer**: two layers of Bi-directional LSTM, same as BiDAF.
- **Output layers**: used W_1M_1 and W_2M_2, instead of using $W_1[G, M_1]$, $W_2[G, M_2]$ as logit before softmax. No Bi-directional LSTM between start and end prediction.

Additional skills for performance boosting:
- Analyzed the overfitting issue of BiDAF model and the effect of dropout and weight decay.
- Tuned the model with different OOV(out of vocabulary) strategy and got different results.

Training Strategy

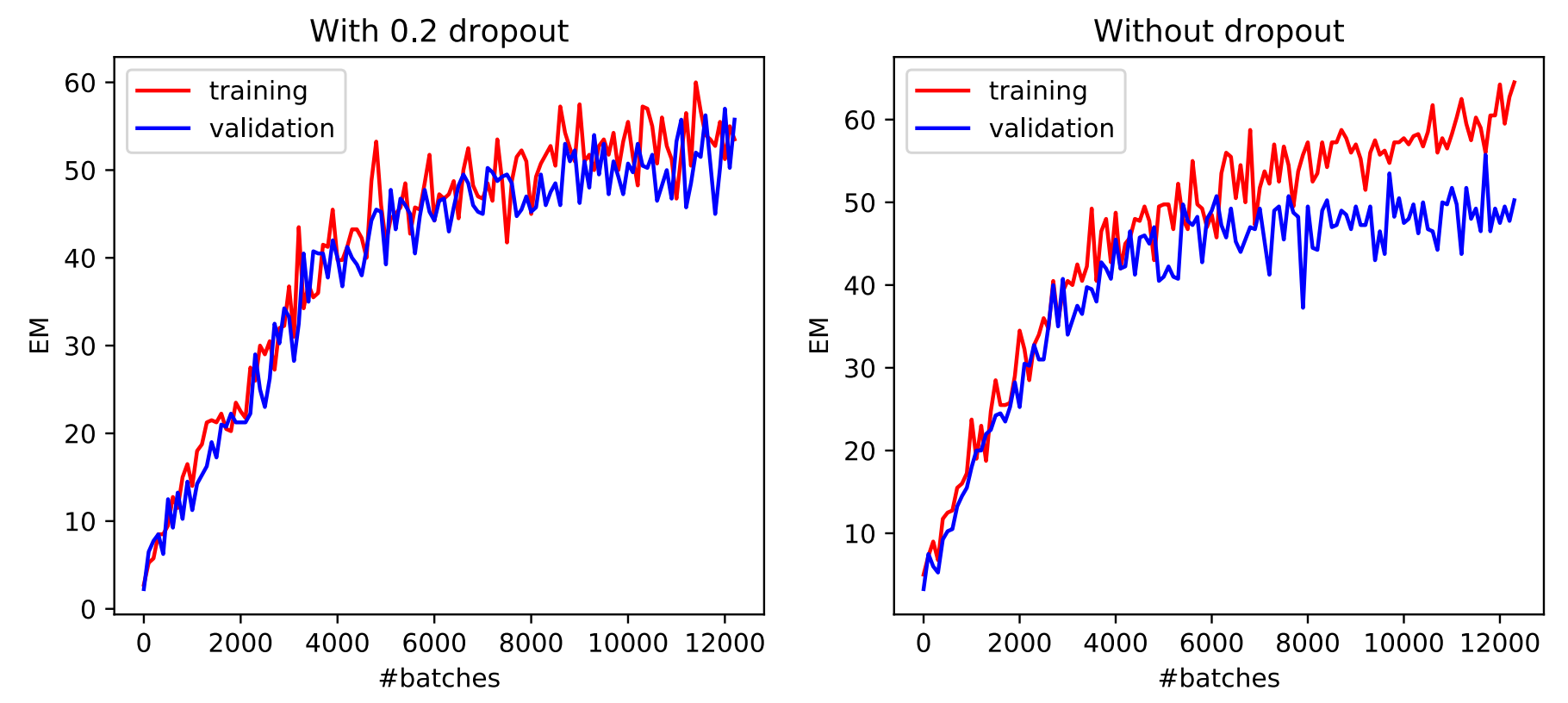

Fig 3. EM performance with and without dropout

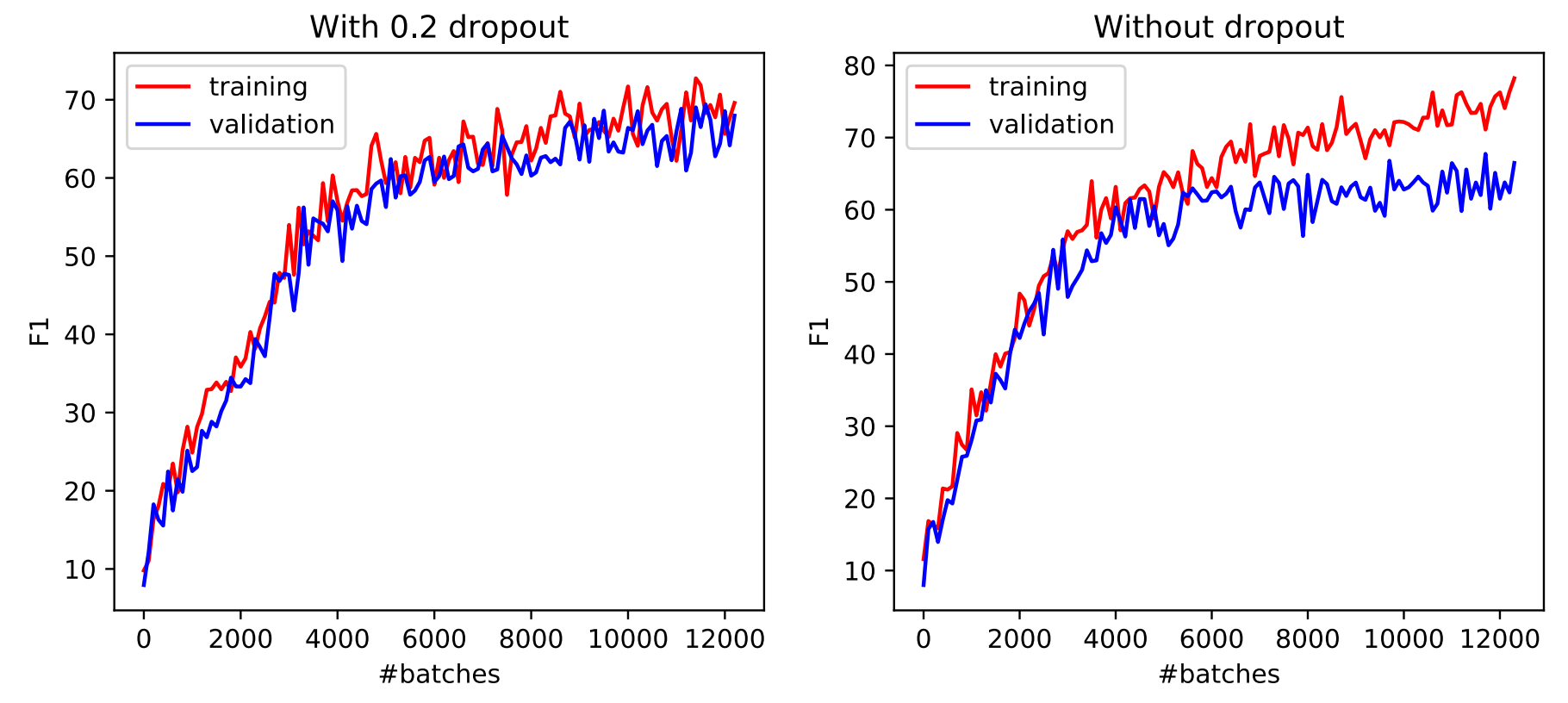

Fig 4. F1 performance with and without dropout

This section is about hyper-parameter tuning and tricks to speed up training.
- **Padding Strategy**: We sort all the contexts by length and randomly sample a batch size window. We then pad each batch of the context to its longest length. Each epoch takes 30min and we usually train 5-7 epochs for a single model.
- **Optimizer**: We used Adam optimizer with initial learning rate 0.048.
- **Weight Decay Rate**: We use a weight decay rate of 0.9999.
- **Dropout rate**: As in Fig 3 and Fig 4, even with weight decay, the model overfits quickly. So we applied a dropout to input gate of all LSTM cells with dropout rate 0.8.

Table 2. Performance comparison with different OOV handling

	F1	EM
OOV set to glove/random	**72.2**	**61.22**
OOV set to random	71.9	60.8
OOV set to zero	71.2	60.2
No OOV handling	63.1	48.9

- **OOV handling**: OOV handling is important as test time. We use different OOV handling methods: 1)set embeddings of OOV to zeros. 2) set embeddings of OOV to a random vector, 3) set embeddings of OOV to glove embedding, if not in glove, set to random. The results can be found in Table 2.

Quantitative Results

Table 1. Performance comparison with other methods

	F1 Score	EM	F1 Score	EM
	Dev set		Test set	
BiDAF(our implementation, single)	72.1	61.0	-	-
BiDAF(our implementation, ensemble)	75.8	65.3	76.5	66.3
BiDAF(reference implementation, single model)	77.3	67.7	77.3	68.0
BiDAF(reference implementation, ensemble)	80.7	72.6	81.1	73.3
MPCM[3]	-	-	75.1	65.5
Dynamic Coattention	-	-	75.9	66.2
r-net	-	-	77.9	69.5

The part marked in blue is our implementation, we achieved F1 score **75.8** and EM **65.3** on dev set, and F1 score **76.5** and EM **66.3** on test set.

From the results we can se that the performance is comparable to state of the art machine comprehension methods. The performance gap with reference implementation of BiDAF can be explained by lack of character level embeddings.

Error Analysis

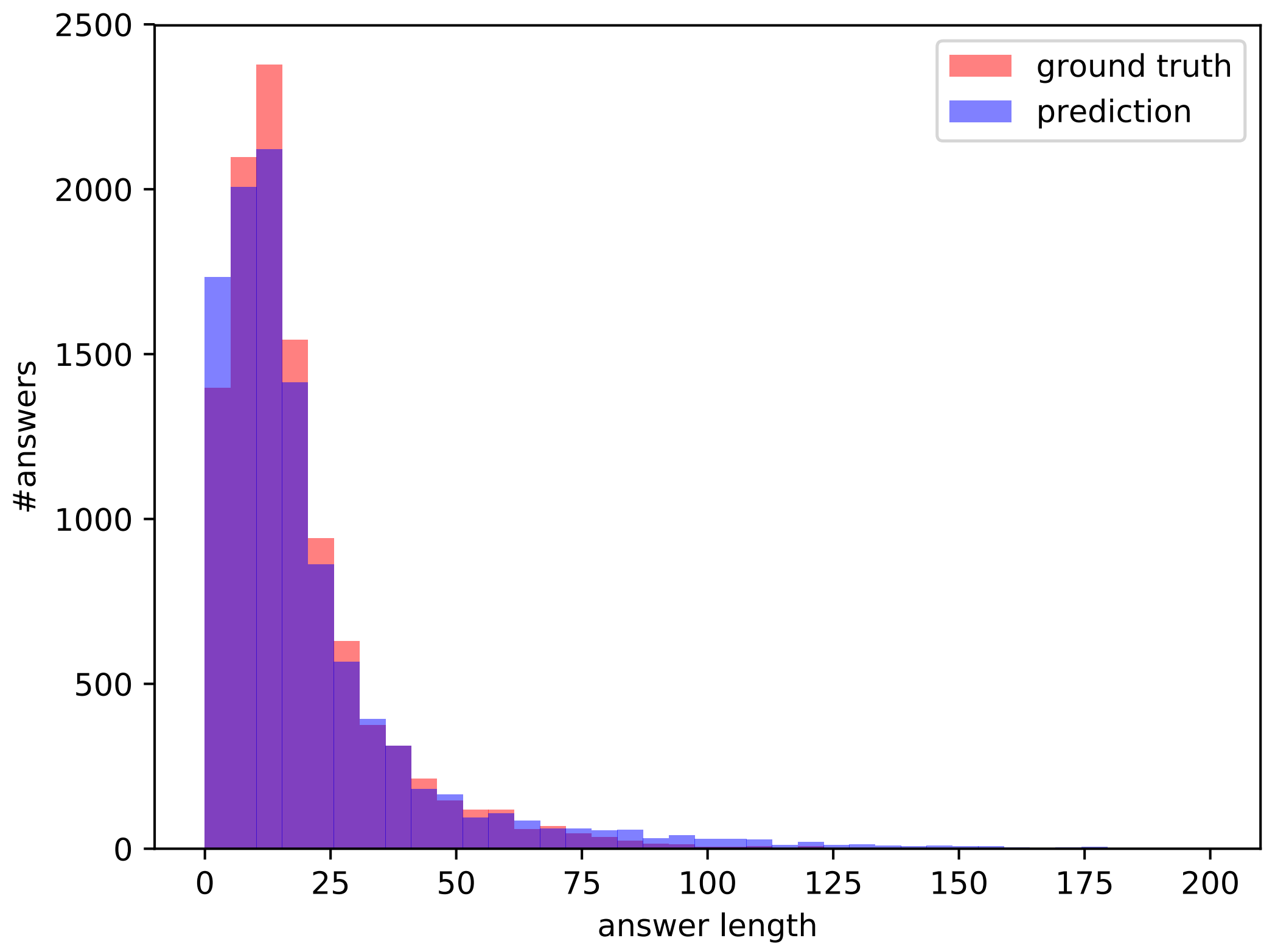

Fig 2. Distribution of ground truth answer length and predicted answer length

1. Imprecise answer boundaries

In most of the incorrect cases, the model gives a inaccurate position around the boundries. 2-3 words around the answer span may be mistakenly omitted or included.

Context: Rugby is also a growing sport in southern California, particularly at the high school level, with increasing numbers of schools adding rugby as an official school sport.
Question: At which level of education is this sport becoming more popular?
Prediction: 'high school level'
Answer: ['high school', 'high school', 'high school']

2. Long tail of predicted answers

By comparing the distribution of answer length (Fig.2), we find that our model tends to give a longer answer than ground-truth. Often times it's able to correctly predict the starting position but has a long tail of many irrelavent words. This shows that the model may not learned that short and concise answers are preferred in this case.

Context: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.
Question: Where was the practice place the Panthers used for the Super Bowl?
Prediction: 'San Jose State practice facility and stayed at the San Jose Marriott'
Answer: ['San Jose', 'the San Jose State practice facility', 'San Jose State']

References

[1] Rajpurkar P, Zhang J, Lopyrev K, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv:1606.05250, 2016.
[2] Seo, Minjoon, et al. "Bidirectional Attention Flow for Machine Comprehension." *arXiv preprint arXiv:1611.01603* (2016).
[3] Wang, Zhiguo, et al. "Multi-Perspective Context Matching for Machine Comprehension." *arXiv preprint arXiv:1612.04211* (2016).