BiDirectional Attention for Machine Comprehension

Anand Venkatesan
Electrical Engineering
Stanford University
Stanford, CA 94305
anand95@stanford.edu

Ananthakrishnan Ganesan
Institute for Computational and Mathematical Engineering
Stanford University
Stanford, CA 94305
ananthg@stanford.edu

Abstract

Machine comprehension in general, and question answering in particular, is a complex task. We implemented a neural network architecture to predict answers to questions based on a contextual paragraph. In particular, we used a BiDirectional Attention Flow mechanism to meaningfully extract interactions between context and question words, thereby facilitating the model to easily identify the correct answers. Our model achieved metrics comparable to the state-of-the-art on the Stanford Question Answering Dataset. We analyzed the attention plots as well as some erroneous predictions to understand how the model could be improved. We also depicted examples where the model predictions are more sensible than the true answer.

1 Introduction

Question Answering from Reading Comprehension continues to be an interesting topic for study, given the ever-changing context and the different kinds of questions we could pose to our machine learning models. Successful implementation of question answering can lead to less human intervention and can make most of the tasks simpler. As languages are a human abstraction, training machines to answer a question is a non-trivial task.
For instance, a particular context can appear several times in the paragraphs and hence it is extremely difficult to make the machine identify the correct part in the text. The machine must be trained to answer questions by making causal relation between the question and the given context [1]. Earlier this was done only by humans but recently, with the availability of SQuAD [2], NLP researches have aimed at solving this problem using tools from deep learning.

2 Literature Review

From our literature study, we gather that the Question Answering is a popular research problem and was explored by many research groups. Initially it started with the traditional use of Natural Language Processing Techniques like parsing, parts of speech tagging [3] etc. The advent of Neural Network gave a breakthrough where many researchers aimed at different usages [4].
One interesting work in this field was done using Dynamic Co-attention Network [5] and bilateral multi-perspective matching (BiMPM) model [6]. This model consists of encoders for the context, document, co-attention and for the dynamic pointer. The co-attention encoder interacts between the question and the context in its encoded form and this acts as an attention mechanism for this model. The process is repeated iteratively, taking the previous information to process the current prediction and this prevented the model from falling to local maxima.
Another very popular model for Question Answering is by using the Bidirectional Attention Flow model (BiDAF). It consists of a bidirectional attention flow mechanism for the reading comprehension where one corresponds to the question to context attention and another consists of context

to query attention [7]. This model is detailed in this work and we have built this model from the scratch.

3 SQuAD Dataset

The dataset for this task is the Stanford Question Answering Dataset (SQuAD) [2]. It is a prevalent reading comprehension dataset built on a set of Wikipedia articles by the cross workers. It has around 100,000 question-answer pairs on 500+ articles. The data is split in the ratio 80:10:10 where 80% is the training set, 10% is the development set, and 10% is the test set. The question, answer and the context lengths of the training set are plotted as a histograms in fig. 2.
It is seen that the data varies according to these lengths. Over 90% of the questions have short answers (less than 10 words). Most of them are even single word meaning that there can be instances where a model just predicts a single word yet have a high EM/F1 score. The dataset has a varied length of questions although most of them lie in the 5-15 word range. Proper values of these hyper-parameters are needed to decrease the training time, and improve the performance.
The SQuAD is tested against many criterion and structured engineered features. The current best model in the official public leaderboard has EM scores which surpasses the human level intelligence. The F1 score is slightly behind the human level and this remains as a challenge in the present research to beat the human performance.

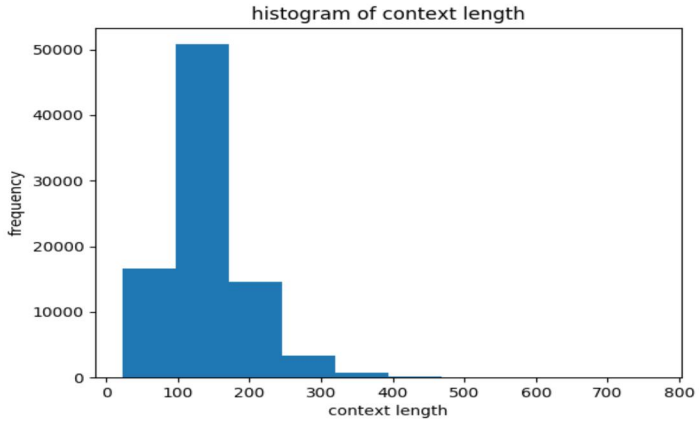
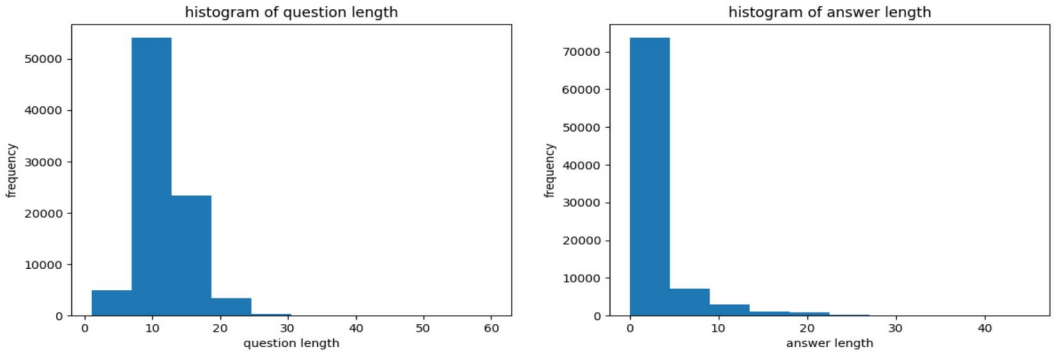


Figure 1: Context Length



(a) Question Length

(b) Answer Length

Figure 2: Histograms of Question and Answer Lengths

4 Model Architecture

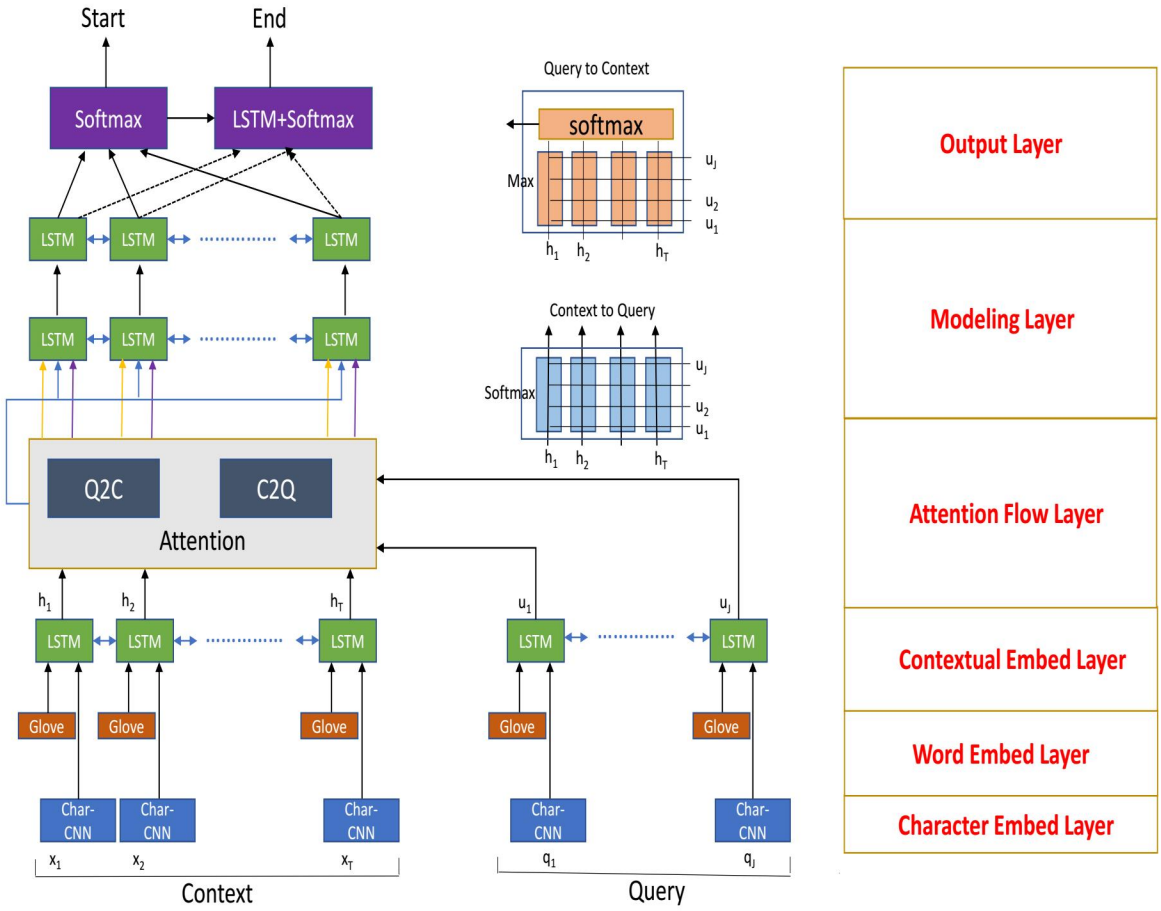


Figure 3: BiDirectional Attention Flow Model Architecture

We followed [7] and created a hierarchical multi-stage architecture depicted in fig. 3. We use D to denote the size of word embeddings, H for hidden layer size, J for the question size and T for the context size. All of our BiDirectional Long Short-Term Memory Network (BiLSTM) are followed by dropout with the same dropout probability. Our model contains the following six layers:

1. **Character Embedding Layer:** This layer performs character-level embedding for each word in a sentence to a high-dimensional vector space using a one-dimensional Convolutional Neural Network with Leaky Rectified Linear Unit activation with parameter α. The output from this layer is of dimension $D \times 9 * T$ for context ($D \times 9 * J$ for question), where we have assumed each word to contain 9 characters on average. We further learn a linear map to transform the output dimensions to $D \times T$ ($D \times J$ for questions). We found this layer to be most useful for words that did not have a pre-trained word vectors.
2. **Word Embedding Layer:** We used a pre-trained GloVe embedding for this layer, which provides $D \times T$ ($D \times J$ for question) word vectors.
3. **Contextual Embedding Layer:** We concatenate the embeddings from the previous two layers and then identify temporal interactions among the words using a BiLSTM. We use a hidden size of dimension H and concatenate the forward and the backward states to obtain $C \in \mathbb{R}^{2H \times T}$ for context ($Q \in \mathbb{R}^{2H \times J}$ for question).
4. **Attention Flow Layer:** This layer captures the importance of question word-context word pairs using C and Q as inputs. We first create a shared similarity matrix $S \in \mathbb{R}^{T \times J}$ such that $S_{ij} = w^T[C_{:i}; Q_{:j}; C_{:i} \circ Q_{:j}]$.

We then create a context to query attention $\tilde{Q}_{:i} = \sum_j \text{softmax}(S_{i:})_j U_{:j}$ to identify which query words are relevant for each context word. We also create a query to context attention $\tilde{C} = \sum_j \text{softmax}(\max_{col}(S))_j C_{:j}$, which contains a weighted sum of the most important context words with respect to query.
Finally we obtain the embedding $G \in \mathbb{R}^{8H \times T}$ by $G_{:i} = [C_{:i}; \tilde{Q}_{:}; C_{:i} \circ \tilde{Q}_{:i}; C_{:i} \circ \tilde{C}]$.

5. **Modeling Layer:** This layer captures interactions among the query conditioned context words by using two layers of BiLSTM on G to obtain M of dimensions $2H \times T$.

6. **Output:** We obtain the probability distribution on the start word using $p_{st} = \text{softmax}(w_{st}^T[G; M])$, and end word probability distribution as $p_{en} = \text{softmax}(w_{en}^T[G; M_2])$, where M_2 is obtained by passing M through a BiLSTM layer.

7. **Loss:** We minimize a cross entropy loss on logits $w_{st}^T[G; M]$ and $w_{en}^T[G; M_2]$.

5 Hyper Parameter Tuning

We tuned the Leaky ReLU parameter α (which denotes the fractional effect of negative values), the dropout probability, and the size of the BiLSTM hidden layer. Since training the complete model was time consuming, we used the loss on the dev set after 100 iterations of training to determine the best parameters. The variations in dev-loss after 100 iterations for the three parameters are as shown in fig. 5. We could not search for higher values of hidden size than that shown in 5 due to constraints on memory.

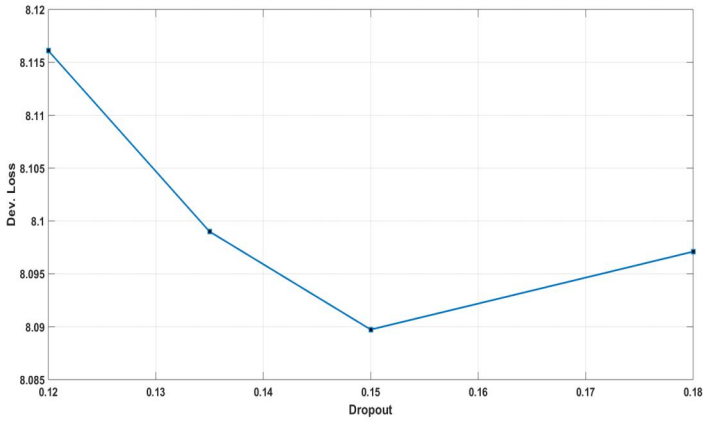
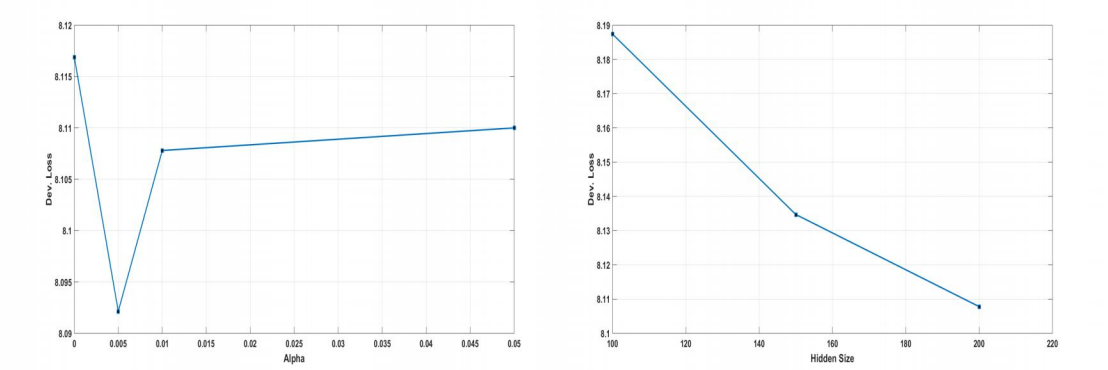


Figure 4: Hyper-parameter tuning: Dropout probability vs Dev. Loss

Table 1: Model Performance

Model	F1 Score	EM Score
Baseline	43.615	34.551
Logistic Regression[2]	51.0	40.4
Our Model	67.543	5.775
BiDAF	81.1	73.3



(a) α vs Dev. Loss (b) Hidden Size vs Dev. Loss

Figure 5: Hyper-parameter tuning

6 Results

After tuning hyper-parameters, we simulated our model for about $12k$ iterations using the training partition of the SQuAD. Table 1 contains the standard F1 and EM metrics of our model, along with that of the provided baseline and that from (BiDAF) [7]. The metrics from our implementation are comparable to that of BiDAF, and we have achieved significant improvement over the given baseline.

7 Attention visualization

In order to understand the role of the output layer as well as the attention layer, we have plotted the context to query attention along with the start and the end logits in fig. 6 for the context and question given below. From the attention map, it is evident that the attention mechanism correctly maps the word "long" in the question maps to the word "long" (82^{nd} word in 0 index) in the context. Therefore the start and the end logits suitably cover this word "long", and results in a highly accurate prediction.

Example Context:
in the centre of basel , the first major city in the course of the stream , is located the " rhine knee " ; this is a major bend , where the overall direction of the rhine changes from west to north . here the high rhine ends . legally , the central bridge is the boundary between high and upper rhine . the river now flows north as upper rhine through the upper rhine plain , which is about 300 km long and up to 40 km wide . the most important tributaries in this area are the ill below of strasbourg , the neckar in mannheim and the main across from mainz . in mainz , the rhine leaves the upper rhine valley and flows through the mainz basin .
Question: how long is the upper rhine plain ?
True Answer: 300 km long
Predicted Answer 300 km long
F1 Score Answer: 1.000

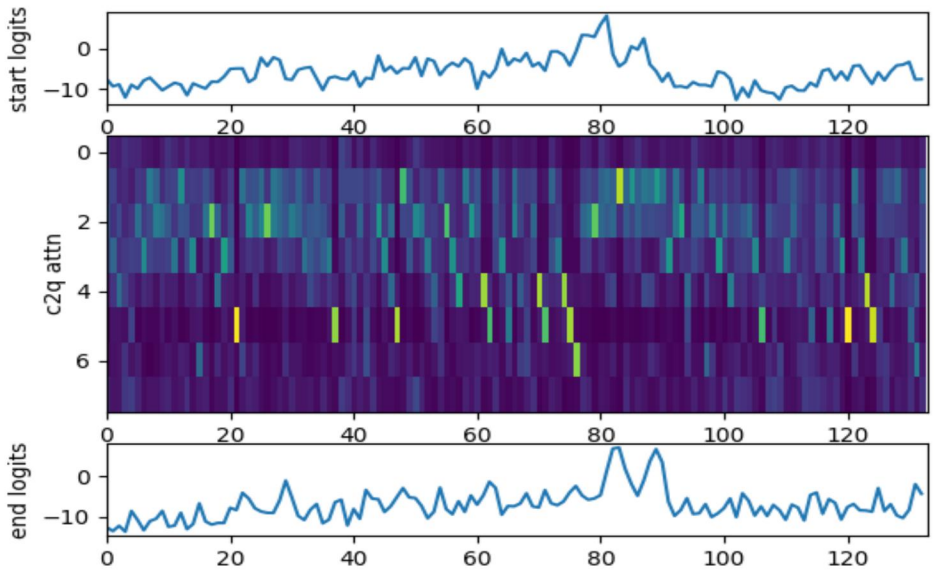


Figure 6: Context to Query Attention with Start/End Logits

EM Score: True

8 Error Analysis

We analyzed the errors made by our model and attempted to broadly classify them into a small number of categories.

8.1 Wrong Answer in Dataset

In the following example, the true answer is clearly wrong. However, our model predicted the correct answer and was able to associate the phrases "ran out of gasoline" in the question to "no fuel" in the context. This example highlights the role automated machine comprehension systems can play in the future.
Example Context:
in 1973 , nixon named william e. simon as the first administrator of the federal energy office , a short-term organization created to coordinate the response to the embargo . simon allocated states the same amount of domestic oil for 1974 that each had consumed in 1972 , which worked for states whose populations were not increasing . in other states , lines at gasoline stations were common . the american automobile association reported that in the last week of february 1974 , 20 % of american gasoline stations had no fuel .
Question: according to the aaa , what is the percentage of the gas stations that ran out of gasoline ?
True Answer: last week of february 1974 ,
Predicted Answer: 20 %
F1 Score Answer: 0.000
EM Score: False

8.2 Imprecise answer boundaries in the dataset

In a few examples, our model predicted precisely but the true answer had unnecessary words which penalized our model very badly. Though our model predicted the answer perfectly in the example

below, the F1 score is very low and the EM score is False. Such errors occurred because the model was trained on a dataset containing predominantly short answers, and hence looks for the shortest possible answer.
Example Context:
The iroquois sent runners to the manor of william johnson in upstate new york . the british superintendent for indian affairs in the new york region and beyond , johnson was known to the iroquois as warraghiggey , meaning " he who does great things . " he spoke their languages and had become a respected honorary member of the iroquois confederacy in the area . in 1746 , johnson was made a colonel of the iroquois . later he was commissioned as a colonel of the western new york militia . they met at albany , new york with governor clinton and officials from some of the other american colonies . mohawk chief hendrick , speaker of their tribal council , insisted that the british abide by their obligations and block french expansion . when clinton did not respond to his satisfaction , chief hendrick said that the " covenant chain " , a long-standing friendly relationship between the iroquois confederacy and the british crown , was broken .
Question: what was william johnson 's iroquois name ?
True Answer: warraghiggey , meaning " he who does great things . "
Predicted Answer: warraghiggey
F1 Score Answer: 0.250
EM Score: False

8.3 Slight mismatch in the attention

The linear attention model was found to work really well in most cases, but in a handful cases, this linear attention was found to be insufficient as is evident in the example below. Adding a more complex attention might resolve erroneous predictions such as that shown below, but might significantly increase the computational cost for very little gains.
Example Context:
the plague theory was first significantly challenged by the work of british bacteriologist j. f. d. shrewsbury in 1970 , who noted that the reported rates of mortality in rural areas during the 14th-century pandemic were inconsistent with the modern bubonic plague , leading him to conclude that contemporary accounts were exaggerations . in 1984 zoologist graham twigg produced the first major work to challenge the bubonic plague theory directly , and his doubts about the identity of the black death have been taken up by a number of authors , including samuel k. cohn , jr. (2002) , david herlihy (1997) , and susan scott and christopher duncan (2001) .
Question: what did shrewsbury note about the plague ?
True Answer: rates of mortality in rural areas during the 14th-century pandemic were inconsistent with the modern bubonic plague
Predicted Answer: contemporary accounts were exaggerations
F1 Score Answer: 0.000
EM Score: False

8.4 Long Answers in the Predictions

Another reason our model did not perform well is because the predicted sentences were long but the true answers were comparatively short. Our model was correctly able to identify the sentence containing the answers but was falling behind in identifying correctly which part of the sentence the true answer to the question remained. This resulted in a comparatively reasonable F1 score but low EM score. Adding an additional LSTM layer might solve this problem.
Example Context:
as opposed to broadcasts of primetime series , cbs broadcast special episodes of its late night talk shows as its lead-out programs for super bowl 50 , beginning with a special episode of the late show with stephen colbert following the game . following a break for late local programming , cbs also aired a special episode of the late late show with james corden .
Question: what other cbs talk show played , after the main one that began immediately after super bowl 50 ?
True Answer: the late late show with james corden
Predicted Answer: cbs also aired a special episode of the late late show with james corden
F1 Score Answer: 0.667
EM Score: False

8.5 Inaccuracy in the Predicted boundaries

Our model was either adding or omitting one or two words from the context when compared to the true answer in some cases. This can be owed due to the perspective of answering a question. These errors are highly challenging and they are almost impossible to eliminate them.
Example Context:
huguenot numbers peaked near an estimated two million by 1562 , concentrated mainly in the southern and central parts of france , about one-eighth the number of french catholics . as huguenots gained influence and more openly displayed their faith , catholic hostility grew , in spite of increasingly liberal political concessions and edicts of toleration from the french crown . a series of religious conflicts followed , known as the wars of religion , fought intermittently from 1562 to 1598 . the wars finally ended with the granting of the edict of nantes , which granted the huguenots substantial religious , political and military autonomy .
Question: when were the wars of religion fought ?
True Answer: from 1562 to 1598
Predicted Answer: 1562 to 1598
F1 Score Answer: 0.857
EM Score: False

9 Conclusions and Future Work

We found that our implementation of the BiDirectional Attention Flow mechanism for machine comprehension task performed comparably to the state-of-the-art. There were a few systematic errors that could have been rectified with a slight modification of the neural network architecture. However, it was encouraging to observe examples where the true answer was certainly wrong but the machine made accurate predictions, suggesting that machine comprehension systems can soon assist/even operate independently on comprehending new data.
Based on our analysis of model prediction errors, a more complex attention mechanism might rectify errors in situations where the model prediction is off by one word, with the meaning more or less preserved. Additionally, using the softmax probability distribution to compute the cross entropy loss as opposed to the logits would could have potentially made the model more robust to small perturbations in logits.

Acknowledgments

We would like to thank the instructor and TAs for their invaluable suggestions and feedback about improving the models. We also want to thank them for providing us the access to Azure which motivated us to do more in the project quickly.

References

[1] S. Sugawara and A. Aizawa, An analysis of prerequisite skills for reading comprehension, EMNLP 2016, p. 1, 2016.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, ArXiv e-prints, Jun. 2016.

[3] D. A. Ferrucci, "Introduction to "This is Watson"," IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 1:1-1:15, 2012.

[4] S. Wang and J. Jiang, Machine comprehension using match-lstm and answer pointer, arXiv preprint arXiv:1608.07905, 2016.

[5] C. Xiong, V. Zhong, and R. Socher, "Dynamic Coattention Networks For Question Answering," arXiv preprint arXiv:1611.01604, 2016.

[6] Z. Wang, W. Hamza, and R. Florian, "Bilateral Multi-Perspective Matching for Natural Language Sentences," arXiv preprint arXiv:1702.03814, 2017.

[7] Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." arXiv preprint arXiv:1611.01603 (2016).