Incorporating Part-of-Speech Tags and Named Entities into Match-LSTM

Dilsher Ahmed
Stanford University
Stanford, CA 94305
dilsher@stanford.edu

Raunak Kasera
Stanford University
Stanford, CA 94305
raunak@stanford.edu

Abstract

Reading Comprehension is an extremely important machine learning task with applications for Information Extraction and Question Answering. In this paper we seek to incorporate Part-of-Speech (POS) Tags and Named Entitites (NE) into the existing Match-LSTM model for question answering. We use the Stanford Question Answering Dataset (SQuAD) and attempt to identify the answers to various questions given a context paragraph.

1 Introduction

Reading comprehension has gained further prominence in recent years and the challenge of allowing machines to comprehend text is one of the main goals of natural language processing (NLP). To this end, several methods of evaluation have been utilized with question answering being one which has proved popular.

Question answering typically involves some piece of text being provided to the machine along with a question relevant to the text. Models seek to answer the question in a realistic human manner i.e. not providing overly long or too short answers. There are various datasets for question answering (Richardson et. al, 2013), (Hill et. al, 2016), (Rajpurkar et. al, 2016), some of which use multiple-choice questions and others which do not. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et. al, 2016), provides a context paragraph from Wikipedia with relevant questions having been added to the dataset through crowd-sourcing. As SQuAD is a relatively recent dataset, it has not yet been explored as much as other datasets and we seek to incorporate features that have proved useful on existing datasets into models created for SQuAD in the hopes of further improving these existing models. A sample input tuple has been described in 1.

Note that the answers in SQuAD will always necessarily be contiguous subsequences of words from the context paragraph. As can be noticed, the answer to the question can henceforth be uniquely determined by expressing its boundary indices in the context paragraph.

1.1 Analysis

We conducted various analyses over our dataset and generated histograms over the context, question and answer lengths as shown in Graphs 1.1, 1.1 and 1.1. One interesting fact to notice is that answer length and the number of answers of that length are inversely correlated.

We hypothesized that since many answers consist of only one or two words that these answers are most likely to be simple facts and hence either nouns or numbers which provided our initial motivation for integrating POS tagging and NEs. We ran our entire dataset through a POS tagger and tagged contexts and answers. As we can see in Graph 1.1, nouns make up a total of 31% of our contexts but make up 50% of our answers. Similarly numbers make up only 3% of our contexts but almost 10% of our answers.

Table 1: A sample input context paragraph, question and its answer

DESCRIPTION	VALUE
Context Paragraph	Initially, officials were unable to contact the Wolong National Nature Reserve, home to around 280 giant pandas . However, the Foreign Ministry later said that a group of 31 British tourists visiting the Wolong Panda Reserve in the quake-hit area returned safe and uninjured to Chengdu . Nonetheless , the well-being of an even greater number of pandas in the neighbouring panda reserves remained unknown. Five security guards at the reserve were killed by the earthquake. Six pandas escaped after their enclosures were damaged . By May 20, two pandas at the reserve were found to be injured , while the search continued for another two adult pandas that went missing after the quake . By May 28, 2008, one panda was still missing. The missing panda was later found dead under the rubble of an enclosure. Nine-year-old Mao Mao, a mother of five at the breeding center, was discovered on Monday, her body crushed by a wall in her enclosure. Panda keepers and other workers other workers placed her remains in a small wooden crate and buried her outside the breeding centre .
Question	What place could officials not contact?
Answer	the Wolong National Nature Reserve
Answer Boundaries	$[7, 11]$

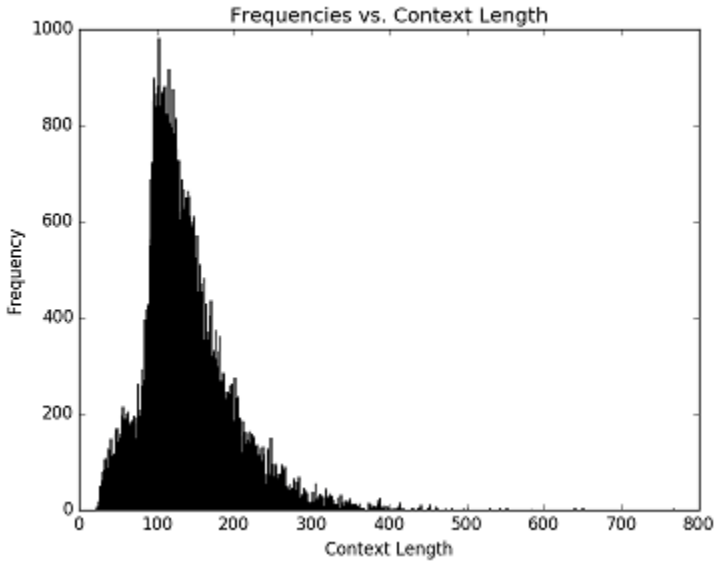

Figure 1: Histogram of Context Lengths vs. Frequencies

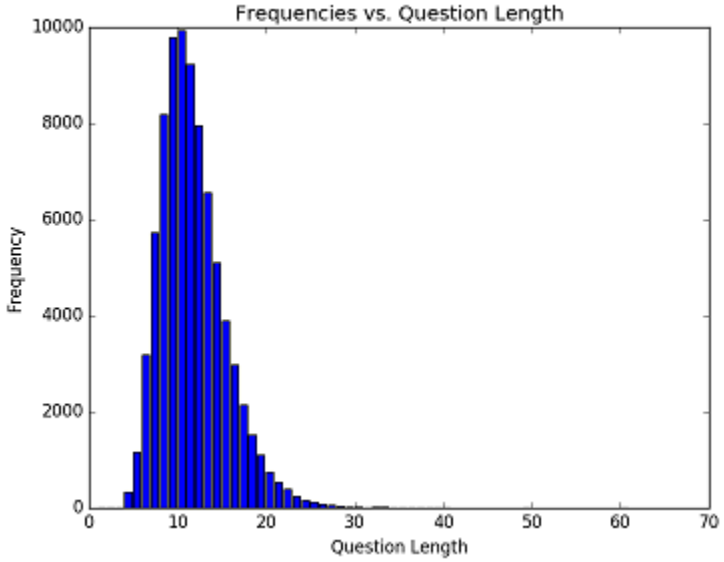

Figure 2: Histogram of Question Lengths vs. Frequencies

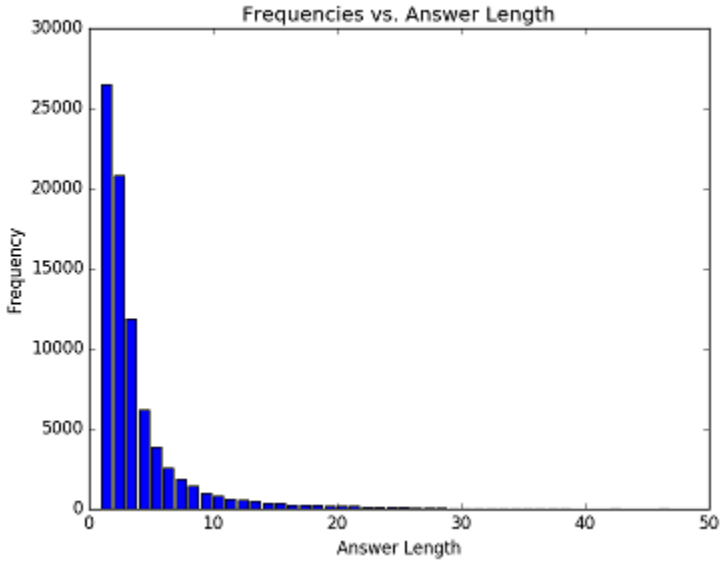

Figure 3: Histogram of Answer Lengths vs. Frequencies

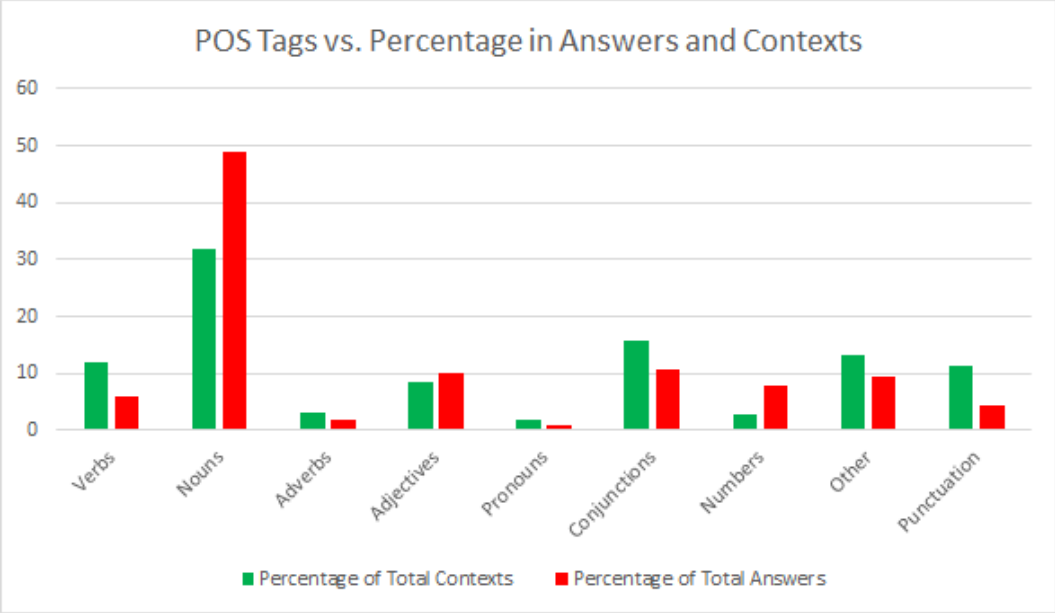

Figure 4: Histogram Comparing the Percentage of Context Words and Percentage of Answer Words for Different POS Tag Groups

2 Model

In this section, we will review Match-LSTM (Wang & Jiang, 2016) and the modifications we made to it as we present our full architecture for this problem.

2.1 Match-LSTM

Match-LSTM is a model for predicting textual entailment. Given two pieces of text, a hypothesis and a premise, it attempts to determine whether the premise entails the hypothesis. At a high level this can be applied for question answering by viewing the question as our premise and we are attempting to identify a region of our context as the hypothesis. Match-LSTM uses an attention mechanism to obtain a weighted representation of the premise at each position in the hypothesis. This forms a weighted vector which is combined with a vector referring to the current position of the hypothesis which is fed into an LSTM. The final results are the aggregation of the resultant vectors returned from all LSTMs which are used to predict the answer.

2.2 Our Method After Modifications

Our problem can formally be stated as follows. We are given a context paragraph C with maximum length m_C and a question Q with maximum length m_Q. We seek to identify the boundaries of the answer to that question within the context paragraph i.e. we are looking for two indices a_s and a_e such that the words from $C[a_s]$ to $C[a_e]$ form our answer.

We start by converting the context paragraph into a matrix representation $\mathbf{C} \in \mathbb{R}^{d \times m_C}$ which is formed by converting the words in the context paragraph into their GloVe embeddings (with dimensionality d) and padding the representation till it reaches length m_Q. Similarly we compute a matrix representation $\mathbf{Q} \in \mathbb{R}^{d \times m_Q}$ for our question.

The next step is to run POS tagging on the question and the context paragraph using Stanford NLTK. Stanford NLTK will generate 1 of 36 potential tags for every word and we have added 2 new tags for Punctuation and for Padding. These tags are then grouped so that they fall into the groups as described in 1.1. This gives us a binarized representation for the POS tags for both the question and the context. These representations are then concatenated onto $\mathbf{C}$ and $\mathbf{Q}$ to form our final representa-

tions. NEs are similarly incorporated with a binarized representation distinguishing between named entities and non-named entities.

We follow the same idea as Match-LSTM to encode the questions using an LSTM cell but we chose to use bidirectional LSTM cells. The Answer Pointer layer of Match-LSTM is copied as is with no modifications. After the completion of this layer we have a probability distribution for each context index for both starting and ending boundaries. The starting boundary is picked to be the index which maximizes the probability in its distribution whereas the ending index is chosen in the same way after filtering out indices smaller than the starting index.

One final change in our model is that it uses an Adam optimizer with decaying learning rate as opposed to the Adamax optimizer utilized by Match-LSTM.

3 Evaluation

We evaluate the model by computing the F-1 Score of our predicted answers vs. the actual answers. The precision for a given answer is computed as the ratio between the words present in both our prediction and the actual answer, and the total number of words in the actual answer. Recall is computed as the ratio between the total number of predicted words that are part of the actual answer and the total number of words that are part of the actual answer.

The exact match score is the percentage of answers that we get completely correct.

Our results are as found in 2

Table 2: A table showing the results of our models

Model Description	F1-scores	Exact-Match scores	# of Epochs Trained
Basic Model (No POS)	$0.50/0.48$	$0.31/0.31$	20
Basic Model (POS)	$0.75/0.57$	$0.58/0.40$	6
POS Model (High learning rate + decay)	$0.83/0.53$	$0.69/0.39$	4
POS Model (Low learning rate + decay)	$0.56/0.44$	$0.36/0.28$	9

Note that we were unable to try many values for the various hyper-parameters and hence the F1-scores of the model are relatively low. However, we noticed that just a simple change of adding POS to the basic model had improved the F1-score drastically in both training and test datasets and so we regard our hypothesis as confirmed.

3.1 Proposed Improvements

We noticed that our model tends to occasionally predict extremely large answer spans which affects our F1-Score quite negatively. It would be interesting to integrate the histogram of answer lengths (refanswer) into the probability distribution in order to ensure that shorter answers are predicted more often. However, this integration would require the answer length probabilities to be normalized in order to ensure that the model does not end up predicting all answers as having length 1.

4 Conclusions

As we hypothesized, POS tagging and NEs prove to be useful in improving the accuracy of simple Match-LSTM based models for Question Answering. Additional tuning of the hyper-parameters however is more important as several of our models either overfit or stopped learning altogether

after a few epochs. Exponential decay of learning rate was only slightly useful in alleviating this problem.

5 Acknowledgements

We would like to thank Professor Manning, Professor Socher and the Course Assistants for CS224N for doing a great job in what was undoubtedly a logistical nightmare of a course to deal with.

[1] Richardson, M., Burges, C. J., & Renshaw, E. (2013, May). MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In EMNLP (Vol. 3, p. 4).

[2] Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. arXiv preprint arXiv:1511.02301.

[3] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

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