Improvements based on NER and Self-attention mechanism

Yixian Wang {yixianwang@utexas.edu}

Abstract

In this paper, I use two methods to improve the Baseline DrQA model. The first improvement is based on the Self-attention mechanism; the second improvement is based on the NER tag system. For the approach of Self-attention, I apply the Self-attention on both the passage and the question encoding part to extract more context information. It has trivial improvement pushing the EM score to 48.84%(-0.02% point absolute improvement) and F1 score to 61.38%(0.07% point absolute improvement) on the SQuAD dev dataset. For the approach of the NER tag system, I combine the NER tag system with the Self-attention mechanism and then apply both methods to the passage encoding part and also apply Self-attention to the question encoding part. It has obvious improvement pushing the EM score to 51.21%(2.35% point absolute improvement) and F1 score to 63.62%(2.31% point absolute improvement) on the SQuAD dev dataset.

1 Introduction

This paper focuses on the problem of extracting more context information from passages and questions to improve the baseline model. The baseline model uses the **general** attention mechanism between passage and questions to get useful context information around each token. It also applies bidirectional RNN in order to extract more **local** context information from the passage and questions.

However, the baseline model assumes that combining bidirectional RNN and the general attention mechanism can extract enough information. In this final project of the baseline QA system, it requires model to answer questions based on the given passage. This system doesn't need to generate new words to answer questions. Instead, the only thing this system needs to do is to define a **range** within the passage for questions. For the purpose of improving the accuracy of getting **subtext** of passage, models can achieve better performance if they can extract more useful context information on passage and on questions.

In this paper, I show how **Self-attention mechanism** is needed and what performance **NER tag system combined with Self-attention mechanism** have on the SQuAD Dev dataset, Newsqa dataset, and adversarial SQuAD dataset. My improvements for the Baseline DrQA model composed of: (1) Self-attention, an extra layer before RNN input layers, given a word embedded passage(or question) matrix, efficiently return a new word embedded matrix which containing local context information for words of the passage. (2) NER tag system, Named-entity recognition, seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc.

My experiments show that even if the Self-attention mechanism only gets trivial improvements, it is also needed for further exploration with other information extraction methods(e.g. NER). NER tag system can significantly improve baseline model accuracy on both exact match and F1 scores. Finally, my improved system is evaluated with several benchmarks. In addition, I also show my model's drawback that why performance becomes worse on the adversarial SQuAD dataset and Newsqa dataset through the use of the Self-attention and NER tag system compared to the original baseline model.

2 Related Work

My first motivation to apply Self-attention mechanism to baseline model is from (Vaswani et al., 2017) and (Devlin et al., 2019).

And my motivation that to use NER tag system is from (Chen et al., 2017) and (Ratinov and Roth, 2009).

Model Architectures	EM	F1
Baseline model	48.86	61.31
Baseline model + self-attention(passage)	44.84	57.45
Baseline model + self-attention(question)	44.28	57.11
Baseline model + **self-attention(passage and question)**	**48.84**	**61.38**
Baseline model + NER(passage)	50.89	63.54
Baseline model + **NER(passage)+ self-attention(passage and question)**	**51.21**	**63.62**

Table 1: Results on SQuAD Dev Dataset. % of accuracy on EM and F1 score. EM: Exact match. F1 score: harmonic mean of the precision and recall. (passage): only applied on passage. (question): only applied on question. (passage and question): applied on both passage and question.

3 Improved Model Architecture

In the following, I describe my improved system based on baseline DrQA system, which consists of two components: (1) Self-attention mechanism for extracting local context information for passage and questions and (2) NER tag system for extracting structure context information of a given passage.

3.1 Self-Attention mechanism

The self-attention mechanism can relate different positions of a single sequence in order to compute a new representation of the same sequence. It enables me to learn the correlation between the words. Self-attention is slightly different from general attention. General attention computes correlation between **passage and question** whereas Self-attention computes correlation between **words** within a passage or a question. My best performing system uses the Self-attention mechanism while preserving speed and memory efficiency.

I use the Self-attention mechanism applied on both passage and question as to the first part of my full improved model.

3.2 NER tag System

I use the Spacy package to obtain my NER **token level** tags for each input passage. (The reason I choose Spacy is that Spacy is almost eight times faster than Stanza. It has optimized on C language.) My NER tag system works as follows:

Recompute tokens for passages, questions and answers The motivation of recomputing these tokens with Spacy is to make sure that NER tags are consistent with passage tokens. Recomputing tokens for question and answer has a similar reason. However, applying Spacy to recompute tokens introduces a new issue that can lead to few answers used for training that can't match with passage on the token level. To solve this issue I hardcode these unmatched words in the passage to separate corresponding tokens properly. The reason I still use recomputed tokens in my final version of the improved system instead of original tokens from data is that recomputed tokens can achieve better consistency compared to original data tokens.

Obtain NER tags and adding to vocabulary I use Spacy to obtain NER tags and then add these tags to vocabulary which means I can get the index for each NER tag after adding them to vocabulary.

Concatenate NER tags after passage tokens I use NER tags and concatenate them after passage tokens to create a new passage token level input sequence. Because I have added NER tags to vocabulary, I can convert the whole new passage token level input sequence into a numerical representation. Finally, I use this new numerical passage input sequence as my input for the RNN model to train my improved model.

My final version of the improved system which has the best performance combined NER tag system and Self-attention mechanism. (Table 1 shows performance for different model architectures.)

4 Experiments

This part first presents evaluations of the Self-attention mechanism and NER tag system and then describes tests of their combination on SQuAD Dev Dataset and Adversarial SQuAD Dev Dataset.

4.1 Self-attention mechanism Evaluation on SQuAD Dev Dataset

I first examine the performance of my system only with the Self-attention mechanism on the SQuAD Dev dataset. Table 1 compares the performance of the three approaches described in Section 2.1

Dataset	Model	EM	F1
SQuAD Adversarial	Baseline model	**37.33**	**47.96**
SQuAD Adversarial	Baseline model + NER + Self-attention	36.25	46.64
Newsqa Dev	Baseline model	**20.56**	**32.45**
Newsqa Dev	Baseline model + NER + Self-attention	16.05	27.39

Table 2: Results of models comparisons on SQuAD Adversarial and Newsqa Dev datasets. % of accuracy on EM and F1 score.

for the task of extracting more local context information from passage and question. I compute the performance of combining the baseline model with Self-attention that only applied to passage or question. Results on both of which performed worse. I also compare applying Self-attention to both passage and question at the same time. The result presented in Table 1 indicates that it outperforms the original baseline model, but it only has trivial improvement. The reason is that the original bidirectional RNN model has the ability to extract enough context information from the input and thus the Self-attention has a small contribution. However, Self-attention is still needed for my final version of the system, which I will discuss in detail in Section 4.3.

4.2 NER tag system Evaluation on SQuAD Dev Dataset

Next, I evaluate my improved system that only with NER tag system on the SQuAD Dev dataset. As described in section 2.2, I reform all the inputs of the system. I apply the Spacy toolkit for new tokenization and also generating NER tags. I also rescale the vocabulary to include NER tags.

Result and analysis Table 1 presents my evaluation result on SQuAD Dev Dataset. The improved system only with NER tag system can achieve 50.89% exact match and 63.54% F1 scores, which surpass all experiments based only on the Self-attention mechanism. NER plays a key role by identifying and classifying entities in a text and enabling the system to understand more from a text. I concatenate the NER sequence after the corresponding text, and then the new sequence will including important information such as person or location which is helpful to define the range of the answer.

4.3 NER + Self-attention Evaluation on SQuAD Dev Dataset and Other Datasets

Finally, I assess the performance of my fully improved system using SQuAD Dev Dataset, Adversarial SQuAD Dev Dataset, and Newsqa Dev Dataset without fine-tuning. My final version of the full improved system uses the NER tag system applied on passage and the Self-attention mechanism applied on both passage and question at the same time. I find that while the Self-attention mechanism itself doesn't help improve performance, it does help when applied NER and Self-attention at the same time on SQuAD Dev Dataset. However, when I test my final version system on Adversarial SQuAD Dev Dataset and Newsqa Dev Dataset, the system performs worse than the original baseline system. Especially that the result of the final version system that testing on Newsqa Dev Dataset is the worst.

Result and analysis Table 1 presents the performance of the final version of the improved system.

I am interested in the full system that can extract more information given any passage or question. Table 1 presents performance of final system. The final version of the system trained only on SQuAD is outperformed compared with the original baseline system, and it has obvious improvement pushing the EM score to 51.21%(2.35% point absolute improvement) and F1 score to 63.62%(2.31%point absolute improvement) on the SQuAD Dev dataset. The self-attention mechanism can obtain extra performance when it works with the NER tag system. It indicates that NER and Self-attention mechanism play a complementary role in the task, and Self-attention can help the NER tag system to exploit more local context information to get better performance compared to the system only applied NER tag system.

Table 2 presents performance on Adversarial SQuAD Dev Dataset and Newsqa Dev Dataset. Despite improved system with NER and Self-attention mechanism has the approach to extract more in-

formation from input, it can't provide reasonable performance across other datasets. This indicates that my final system is overfitting on the SQuAD training dataset, and it doesn't have the ability of generalization compared to the original baseline system.

5 Conclusion

I studied the task of improving the performance of the DrQA baseline system. My results indicate that the NER tag system is a key factor that can achieve better performance on the SQuAD dataset and the Self-attention mechanism plays a complementary role for the NER tag system for boosting its performance further.

The drawback of my research is the generalization issue. My final version system can't perform better on Newsqa and adversarial SQuAD dataset.

However, it does have better performance for the purpose of extracting more useful local context information to define range within a given passage and question.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. arXiv: 1704.00051v2. Version 2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805v2. Version 2.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.Gomes, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv: 1706.03762v5. Version 5.