Question Classification with Deep Contextualized Transformer

Robin Luo
rluo@splunk.com
Splunk
Global Security Solution
Ningwei Liu
nliu@splunk.com
Splunk
Global Security Solution
Charles Feng
charlesfeng99@gmail.com
Splunk
Global Security Solution

Abstract

The latest work for Question and Answer problems is to use the Stanford Parse Tree. We build on prior work and develop a new method to handle the Question and Answer problem with the Deep Contextualized Transformer to manage some aberrant expressions. We also conduct extensive evaluations of the SQuAD and SwDA dataset and show significant improvement over QA problem classification of industry needs. We also investigate the impact of different models for the accuracy and efficiency of the problem answers. It shows that our new method is more effective for solving QA problems with higher accuracy

Keywords: QA Classification, NLP, Self-learning, Self-attention

1. Introduction

The Question and Answer system (QA) is widely used in the industry. Every week, one company faces hundreds and thousands of questionnaires for the products they publish. QA is a massive problem in Natural Language Processing (NLP), with the application of problem answering, sentence recognitions, etc. There are several types of problems, such as Wh-questions, statement questions, statements , etc. Each type of question has a corresponding label for a question or statement.

Earlier work in this field mainly used the Bag-of-words (BoW) to classify sentence types. Many recent works have adopted supervised and deep-learning methods on the question classification and have shown promising results (Lee and Dernoncourt, 2016). However, most of these approaches have treated the sentence as a text classification. Furthermore, the treatment has been isolated from sentence to sentence; therefore, it is unable to reflect conceptual dependencies of the words in the sentences. In reality, the different order of the same words in a sentence can have very different meanings.

The work draws some recent advances in NLP research, like BERT (Jacob et al., 2018) and Elmo (Peters et al., 2018) to produce a sentence classification model to quickly and correctly pick out the question sentence from the target text. Compared with regular algorithms for treating the QA problems, the self-learning algorithm can perform contextualized word representation to get the contextualized word meaning in the sentences. Specifically, we use the hierarchical deep neural network with the self-learning algorithm to model different types of question text, including statement questions, which are a specific type of question in the questionnaires. The research works to achieve state-of-the-art outcomes for classifying the QA problem. We demonstrate how performance could be improved with a combination of different levels of models: the hierarchical deep neural network for classification, self-learning and self-attention model like BERT for the single word embedding, and previous label of the training data with the SQuAD dataset. Finally, we explore different methods to find an effective method toclassify the QA problem.

2. Related Work

We focus on two primary methods used in recent research. One treats text as text classification, in which each utterance is classified in isolation, while another one treats the text using Contextualized Word Representation Algorithms, such as BERT with self-attention or Elmo.

Text Classification: Lee and Dernoncourt (2016) build a vector representing each utterance and use either RNN or CNN to predict the text details to classify the sentence type.

Self-learning: Jacob et al. (2018) used the BERT, and Peters et al. (2018) used Elmo to embed the text into the vector to give the contextual relationship of the sentence for each utterance. Along with these two tools, we use RNN-based or CNN-based hierarchical neural networks to learn and model multiple levels of utterance.

3. Model

The task of QA classification takes the sentence S as an input, which varies the length sequence of the utterance U= $\{u_1,\ u_2,\ u_3,\ \ldots,\ u_N\}$. For each utterance $\mathrm{u}_1 \in$ U, there has a length value of $l_i \in$ L and a corresponding target label $y_i \in$ Y, which represents the QA's result associated with the corresponding sentence.

Figure 1 shows the overall architecture of the model, which involves several main components. (1) A self-learning Algorithm to encode the sentence with the self-attention, (2) A Combination-level RNN to handle the output of the encoding and to classify the label of the sentence. We describe the details below.

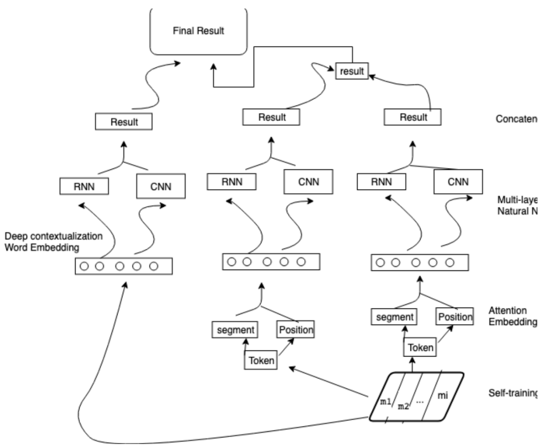

Figure 1. The graph of the model Architecture

3.1 Context-aware Self-learning

Our self-learning algorithm encodes a variable-length sentence into a fixed size. There are two types of the algorithm; one based on Self–Attention and another based on deep contextualization word representation.

3.1.1 Deep contextualization word representation

The model uses the BiLM to consider the different position of utterances within the sequence. Inspired by Peters et al. (2018), we use PCA and t-SNE to reduce the dimensions from a higher level to a reduce the dimensions from a higher level to a lower level. Then we use the Combination-Level RNN (Section 3.2) which provides us with the previous hidden state of the encoded utterance. It provides us the contextual relationship in the sentences and combines all hidden states of words in sentences. After that, the deep our modifications contextualization word representation encoder encodes the combination into the 2-D vectors of each sentence. We follow the instruction of Peters at el. (2018) to explain below.

An utterance t_i, which is the sequence of the sentence, is mapping into the embedded layer. The deep contextualization representation uses BiLM to combine the forward and backend LM. The formulation of the process is as follows:

$$\sum_{k=1}^{N} \left(\log p(t_k \mid t_1, \ldots, t_{k-1}; \Theta_x, \overrightarrow{\Theta}_{LSTM}, \Theta_s) \right.$$
$$\left. + \log p(t_k \mid t_{k+1}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \right) .$$

Moreover, we weigh the perform of the model with computing as indicated here:

$$E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM} . \quad (1)$$

In (1), the s_j^{task} is softmax-normalized weights, and the scalar parameter γ^{task} allows the task model to scale the entire vector. In the simple case, the representation would choose the top layer and E(R_k) = $\mathbf{h}_{k,j}^{LM}$.

3.1.2 Self-Attention

For each word in the utterance, we would use some Self-Attention model to encode them. The most popular Self-Attention model base is on BERT (Devin et al. 2018). The model will encode a variable-length sequence using an attention mechanism that considers the different position, token, and segment within the sequence. Inspired by Devin et al. (2018) and Tran et al. (2017), we apply the Combination-Level RNN (Section 3.2) into a self-attractive encoder (Lin et al. 2017). We use the 24 layers and 1024 Hidden Uncased BERT also with the RobertaBERT as the base of the embedding to encode the context to the 3D tensor. We follow the instruction of Vipuls Raheja and Joel Tetreault (2019) and Joel Tetreault and Liu et al. (2019) to explain the modification mentioned below.

The utterance t_i is also mapped into the embedding layer and results in s-dimensional embedding for each word in the sequence based on the Transformer (Vaswani et al. 2017). Then the embedding is put into the bidirectional-GRU layer.

Vipul Raheja and Joel Tetreault (2019) describe the contextual self-attention score as:

$$S_i = W_{s2} tanh(W_{s1} H_i^T + W_{s3} \overrightarrow{g_{i-1}} +$$

(2)

Here W_{S1} is a weight matrix, W_{S2} and W_{S3} is a matrix of parameters. b is a bias of the vector represented in Equation 2. This can be treated as a 2-layer MLP with bias, and d_a with a hidden unit.

3.2 Combination-level RNN

The utterance representation hi from the past two models are passed into the combination-level RNN. Based on Figure 1, we would pass all of the hidden layers concatenated into a final representation Ri of each utterance. This process is based on the requirements of the problem. We would fine-tune the algorithm.

This is more suitable for the problem classification to put the layers with the proper percentages in the final representation. Then we put the result into the CRF layer to figure out the relationship between the label and the context of the utterances. This method is not independently decoding the label of the utterances; it should consider all of the relationships of the sentences. Then, it should determine the most related decoder to decode them to the related labels. The combination-level RNN would also have the function to supervise the labels and fix them.

3.3 Super-attractive

The model that we use combines the final representative of the combination for hidden layers via self-learning and self-attention. It can help us figure out what the labels those utterances are and produce the results. The score we compute for the algorithm is to calculate the accuracy of the correct labels in the classifications as Hossin M. and Sulaiman M.N. (2015) suggests. Also, we apply an advanced check for the question and answer problem. For sentences without clear results, we put them into the parser tree for another classification. The parser tree we use is based on Huang (2018). We use its Tensor Product Representation to rebuild our parser tree for our model. The original Stanford Parser Tree (2008) is good to classify the relationship of the sentences. However, in our model, we use the Bi-LSTM with the attention algorithm to rebuild the parser tree and get the tree graph with POS tags. This is useful to calcify the structure of the sentence. After that, we use the graph we obtain to analyze the structure of utterances and produce the classification of the unsure sentence in the document. Finally, we determine the combination result for the users to check the question and answer problems.

4. Data

We evaluate the accuracy of the classification model with one standard dataset - the Switchboard Dialogue Act Corpus (SwDA) (Jurafsky et al., 1997) consisting of 43 classes, and make the word extension with the Stanford Question Answering Dataset to use self-attention for the task. The Natural Language Toolkit Dataset (NLTK) (Steven Bird and Edward Loper, 2002) is another significant resource for the test case. We then use the training, validation, and test splits as defined in Lee and Dernoncourt (2016).

Table 1 shows the statistics for both datasets. There are many kinds of labels of the class to classify the kind of sentences they are. There are some special DA classes in both datasets, such as Tag-Question in SwDA and Statement-Question in NLTK. Both datasets make over 25% of the question type labels in each set.

Dataset	Train	Validation	Test	|T|	|N|
SwDA+SQuAD	87k	10k	3k	43	100k
NLTK	8.7k	1k	0.3k	15	10k

Table 1. Number of Sentences in the Dataset. |T| represents the number of classes and |N| represents the sentence size

5. Result

We have compared the classification accuracy of our model with several other models (Table 2). For methods using attention and deep contextualization word representation in some approaches to model the sentence of questionnaire documents, even some of them use the self- attention for the task. However, they did not perform as well as our model. All models and their variables were trained eight times, making an average of the performance as a result. And we find these previous algorithms did not perform as well as our model. Our model is better than Vipul and Joel (2019) by 0.4% in SwDA dataset with measure its accuracy score and 3.9% for the Li and Wu (2016) methods in SWQA dataset. It also beats the TF-IDF GloVe baseline by 17.2% in SwDA.

Model	SwDA+SQuAD	NLTK
TF-IDF GloVe (2014)	66.1	70.3
Li and Wu (2016)	79.2	-
Peters et al. (2018)	76.3	-
Vipul Raheja and Joel Tetreault (2019)	82.7	85.8
Lee and Dernoncourt (2016)	75.9	77.4
Our Method	83.1	85.5
RoBERTa	82.2	84.7

Table 2. QA Classification Accuracy of the different approaches

The improvements based on our model has a significant meaning for other modelsl. However, the performance in NLTK is still not good as that of the Vipul and Joel (2019). The reason for the lower accuracy is dependent on the contextual details and label noise of the dataset. The context in the NLTK dataset indicated the existence of some data not easily readable for the machine, such as some error codes. Also, the label in the NLTK dataset is only 35% of the label for the SwDA ones. As a result, due to the label noise and the contextual details, the performance of NLTK did not show significant gains over that of SwDA.

The performance of our model is more sensitive than the model used commonly for the problems, including the error code. However, it has a higher accuracy considering the complete problem classification. In future research, we should improve our algorithm,
which has a higher ability to handle the problem of the label
noise and context detail that are not clear.

6. Conclusion

We developed a new model which carefully performed the QA classification and made comparisons with common-use algorithms by testing the SwDA dataset. We used different utterance representation methods and determined that the context details depend highly on the classification performance. For example, the reason of NLTK is not as good as Vipul and Joel (2019) results was because there were too many label noises and the context details were not so easy to read. Working with attention and combination level to the classification, which has not been previously applied in this kind of task enables the model to learn more from the context and get more real meaning of the words in utterances than previously. It helps to improve the performance of the classification for these kinds of tasks.

In our future work, we will try more attention mechanisms, such as block self-attention (Shen et al., 2018b), or hierarchical attention (Yang et al., 2016) and hypergraph attention (Song et al. 2019). They can incorporate the information from different representations for the various positions and can capture both local and long-range context dependency. Also, this approach should help with the problem of the hard-readable context, such as the problem of the NLTK dataset that causes accuracy to become lower than usual. We will seek more dataset combinations to do the question classification work. We will use RACE (Lai et al., 2017) and GLUE (Wang et al., 2019) datasets to do more test work and make more stable algorithms to solve the question classification issues. work and make more stable algorithms to solve the question classification issues.

Reference

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 515–520. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489. Association for Computational Linguistics.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discoursefunction annotation coders manual, draft 13. Technical report, University of Colorado at Boulder Technical Report 97-02.

Rajpurkar P, Zhang J, Lopyrev K, Liang P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv [cs.CL].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. July 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv [cs.CL]

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 428–437. Association for Computational Linguistics.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In International Conference on Learning Representations 2017 (Conference Track).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NIPS 2017, 4-9 December 2017, Long Beach, CA, USA, pages 6000–6010.

Song Bai, Feihu Zhang, and Philip H.S. Torr. Jan 2019. Hypergraph Convolution and Hypergraph Attention. arXiv [cs.CL].

Wei Li and Yunfang Wu. 2016. Multi-level gated recurrent neural network for dialog act classification. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1970–1979. The COLING 2016 Organizing Committee.

Hossin M. and Sulaiman M.N. March 2015, A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS . International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543. Association for Computational Linguistics.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. 2018b. Bi-directional block selfattention for fast and memory-efficient sequence modeling. In International Conference on Learning Representations.

Steven Bird and Edward Loper. July 2002. NLTK: The Natural Language Toolkit. arXiv [cs.CL]

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In International Conference on Learning Representations (ICLR).

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Marie-Catherine de Marneffe, Christopher D. Manning. Sept 2008. “Stanford typed dependencies manual”. Stanford University NLP group.

Qiuyuan Huang, Li Deng, Dapeng Wu, Chang Liu, and Xiaodong He. Feb 2018. Attentive Tensor Product Learning. arXiv [cs.CL].

Vipul Raheja and Joel Tetreault. May 2019. Dialogue Act Classification with Context-Aware Self-Attention. arXiv [cs.CL]

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Appendix for Question Classification with Deep Contextualized Transformer

A. Finetuning Hyperparameters

Hyperparam	SQuAD
Learning Rate	1e-5
Weight Decay	0.1
Epochs	7
Batch Size	8k

Table 3：Hyperparameters of Finetuning RoBERTa on SQuAD

B. Pretraining Hyperparameters

Hyperparam	RoBERTa	BERT
No. of Layers	24	24
Hidden Size	1024	1024
FNN Inner Hidden	4096	-
Attention Heads	16	16
Attention Head size	64	64
Dropout	0.1	0.1
Batch Size	8k	8k

Table 3：Hyperparameters of Pre-training RoBERTa and BERT