P9: Automated query processing from passages
Group No. 5

Gautam Sagar, 21CS60R15
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
West Bengal, IN
`gautamsagarnsit@kgpian.iitkgp.ac.in`

Yashas Rohan R, 19EE10066
Department of Electrical Engineering
Indian Institute of Technology Kharagpur
West Bengal, IN
`rohanryashas@gmail.com`

Razorshi Prozzwal Talukder, 21CS60A03
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
West Bengal, IN
`razorshi.prozzwal@kgpian.iitkgp.ac.in`

Vaibhav Vishal, 18EE10055
Department of Electrical Engineering
Indian Institute of Technology Kharagpur
West Bengal, IN
`rapbtycoon256@gmail.com`

Abstract

Knowledge base question answering (KBQA) is an important task in Natural Language Processing. Question Answering (QA) systems enable users to retrieve exact answers for questions posed in natural language.Question Answering (QA) over Knowledge Base (KB) aims to automatically answer natural language questions via well-structured relation information between entities stored in knowledge bases.In this project we aim to implement traditional and modern State of the Art question answering system over a long passage. The Passage chosen is taken from a history text book. The length of the passage chosen is more then 7000 words. This project demonstrates the ability of traditional QA systems and modern Deep learning based QA systems.

1 INTRODUCTION

Question Answering (QA) systems have emerged as powerful platforms for automatically answering questions asked by humans in natural language using either a pre-structured database or a collection of natural language documents. In other words, QA systems make it possible asking questions and retrieve the answers using natural language queries and may be considered as an advanced form of Information Retrieval (IR). Knowledge Base QA Find answers from structured data source (a knowledge base) instead of unstructured text. Standard database queries are used in replacement of word-based searches. This paradigm, make use of structured data, such as ontology. An ontology describes a conceptual representation of concepts and their relationships within a specific domain. Ontology can be considered as a knowledge base which has a more sophisticated form than a relational database. The fundamental problem for retrieval-based QA systems is to retrieve the most similar question from the QA knowledge base given a query, so as to provide the respective answer. Such a text (i.e., query-question) matching problem can be represented as Paraphrase Identification (PI) or some form of Natural Language Inference (NLI).When dealing with such text matching problems in the realworld industrial-scale QA applications, we are facing two prominent challenges, i.e., i) the lack of abundant data to learn a model with high accuracy and ii) the requirement of high inference speed for online model serving. Recent advances on text matching rely heavily on the flourishment of deep learning models. On the one hand, those deep models are proven to be effective when rich in-domain labeled data is available. However, in real-world applications, it is challenging to obtain a sufficient amount of labeled data for every domain of interest, as data annotation is commonly time-consuming and costly. On the other hand, high Query-Per- Second (QPS) requirements for seamless online serving demand the deployed models to be light-weight. Thus the trained models have to be either designed to be simple in structure but effective in performance, or compressed if well-performed large models are originally trained. With recent emergence if language models like BERT which have Billions of parameters the main question that arises is how do we utilize pre-trained models together with other data sources from different domains to facilitate knowledge transfer that is effective in performance and efficient in serving?

In this project we have implemented BERT based question answering system which takes in questions and a passage supplied by the user and answer the question. One traditional method of QA is also demonstrated and finally both the methods are compared.

2 CASE STUDY

2.1 Modern SoTA Methods

2.1.1 BERT

2.1.2 XL-Net

2.1.3 ALBERT

2.2 Traditional QA Methods

2.2.1

2.2.2

3 BERT based QA

To implement BERT based QA we used HuggingFace library. Transformer is their NLP Library that contains state-of-the-art transformer models. To feed question and answer into the BERT model the two pieces of text i.e. question and the answer are joined by [SEP] token. This combined text is then encoded into appropriate form. After input tokenization and encoding we get input encoding in which each token is attached with a token id. From the HuggingFace model for BERT question answering we get start and end token index. Score for these answers are calculated separately. Based on the start and end index we construct the answer from the passage and return it. The in-built model has a limitation of 512 tokens at the input so we have break the passage into 500 tokens and slide the window of 500 tokens by 250 tokens at each step. This increases the inference time for the final answer but it also increases the accuracy of the answer. The answer with highest score is chosen as the final answer. To use our implementation, a user has enter question and wait for the answer. Final answer may appear after 3-4 min.. To end the program user will have enter "END" in the question prompt. User also has the option to save his question in a text file named as "questions.txt" and uncomment the appropriate portion of the code which is also indicated in the notebook. The passage is always supplied to the program in the form of a text file named as "Passage.txt". Some examples are present in the notebook itself.

4 Traditional QA System

5 Results

Table 2 Shows the results for PI and NLI under the single teacher setting.

6 Conclusion

1. Domain-Aware Knowledge Distillation (DAKD), enhances the teacher-student paradigm to facilitate cross-domain transfer learning, where teacher and student tasks belong to heterogeneous

Table 1: Results for PI and NLI under the single teacher setting

Dataset	PI		NLI	
	ACC	AUC	ACC	AUC
Base Model	0.846	0.869	0.730	0.766
FS	0.849	0.871	0.745	0.804
SP	0.854	0.875	0.727	0.798
KD	0.860	0.871	0.751	0.805
DAKD	0.867	0.881	0.755	0.811

domains, with the goal to improve the student model performance of the target domain.
2. DAKD framework considers both the "dark knowledge" from teacher models and adaptive hints to alleviate domain differences.
3. Extensive experiments on two benchmark datasets show the proposed method has better performance than baselines. We have also deployed our method.
4. DAKD is deployed in an online production system and observed significant improvements.

References

[1] Che Chen, Chengyu Wang, Minghui Qiu, Dehong Gao, Linbo Jin, & Wang Li (2001) Cross-domain Knowledge Distillation for Retrieval-based Question Answering Systems. In WWW, pp. 2613–2623.

[2] Geoffrey E. Hinton, Oriol Vinyals, & Jeffrey Dean. (2015). Distilling the Knowledge in a Neural Network.arxiv:1503.02531

[3] Soare Calijorn, A. M., Parreiras, & F. S. (2018). A literature review on question answering techniques, paradigms and systems. Journal of King Saud University - Computer and Information Sciences Advance online publication. doi:10.1016/j. jksuci.2018.08.005