A Hybrid Approach of Deep Semantic Matching and Deep Rank for Context Aware Question Answer System

Shu-Yi Xie, Chia-Hao Chang†, Zhi Zhang, Yang Mo,
Lian-Xin Jiang, Yu-Sheng Huang, Jian-Ping Shen
AI Department
Ping An Life Insurance of China, Ltd.
{xieshuyi542,zhangjiahao206†,zhangzhi600,moyang853,jianglianxin769,
huangyusheng112,shenjianping324}@pingan.com.cn
Correspondence†:strategist922@gmail.com

Abstract

Most of the existing Question Answer Systems focused on searching answers from the Knowledge-Base (KB), and ignore context aware information. Many Question Answer models perform well on public data-sets, but too complicated to be efficient in real world cases. Effectiveness, concurrency and system availability are equally important in industry which have large data and requests, we propose a Context Aware Question Answer System based on the Information Retrieval with Deep Semantic Matching and Deep Rank. It has been applied to the online question answer system for insurance Question Answer. By these means, we achieve not only high QPS but also effectiveness. Our approach improves the system's ability to understand the question with context aware coreference resolution, subject completion, and the long sentence compression. Recalls the matching questions from the ElasticSearch, Siamese CBOW (representing features by combining character embedding, word embedding and phrase embedding and using AM-Softmax as the loss function) and KBQA and then filter some unreasonable ones by entity alignment. After sorting by the deep rank model with co-occurrence words and semantic features, our system does clarification or answer output. Finally, for those questions that we are unable to provide answers, a dialogue mining module as part of our Smart Knowledge-Base Platform is developed. This results in more than 10 times improvement in terms of efficiency for manpower involved in data labeling process.

KEYWORDS

Question Answering, Coreference Resolution, Error Correction, Sentence Compression, Deep Semantic Match, Deep Rank, Knowledge-Base Management, Insurance Domain

All authors contributed equally to this manuscript.

1 INTRODUCTION

The question answer system has been widely used in intelligent customer service, personal assistants, and dialogue robots. In 2018, the pretrain techniques based on a massive corpus pre-training model have made breakthroughs in multiple NLP tasks including Semantic Match. Representative models are Elmo[9], GPT[10], BERT[8]. Higher accuracy, compared with the Siamese CBOW, can be achieved by fine-tuning BERT on downstream tasks, but the model makes inference time much longer, the running efficiency does not meet the requirements of our online products. We propose a high-efficiency contextual referential solution based on syntax analysis to solve the problems of subject missing and pronoun resolution in the question-and-answer scenario in insurance industry that achieved good results. The voice input brings convenience to users but at the same time introduces typos in the results after the text processing. We use the insurance specific noun dictionary with the error correction model of Transformer[7] to improve the input from ASR. For the purpose of increasing the accuracy of matching sentences of the user's input with terms from Knowledge-Base, we use an efficient sentence compression algorithm, which can filter some insignificant content and retain some core content of the insurance industry. We rank all the answers from the retrieval module and do answer output finally. Our contributions are following:

- Propose novel and efficient error correction, sentiment analysis, coreference resolution, sentence compression and other methods to enhance question comprehension ability especially in insurance domain.
- Using ElasticSearch, deep semantic matching and KBQA combined the IR method to quickly recall matching questions. Improve the accuracy of the QA through deep learning rank while ensuring the overall efficiency of the system.
- Proposed a number of new industry test set construction methods and the QA evaluation methods.
- Full-life processing management and optimization for the QA knowledge including question type identification, clustering and annotation dispatch for no answer questions.

2 RELATED WORK

Most of the existing professional domain question answering systems search for the most matching questions (question in KB and user query similarity matching) from the Knowledge-Base through information retrieval. Some existing question-and-answer systems such as the

Ali Xiaomi and the Baidu AnyQ are single-round questions and answers that do not consider the context information. The AliMe from Alibaba, which combines the Knowledge-Base search and Seq2Seq generation, makes achievements in the e-commerce domain[2]. We use the same method as the AliMe and the Baidu AnyQ to match the question and user query similarity and consider context chat history at the same time.

3 SYSTEM OVERVIEW

Our overall system architecture is shown in Figure 1. The user's question (that is query) is used as input. If it is a voice, it will be converted into text first. The context information is passed to the pre-processing module. After error correction and coreference resolution, the processing is passed to the retrieval module. It returns the best matching with the user problem respectively from ElasticSearch based on text retrieval, the semantic retrieval based on the Siamese CBOW and the KBQA based on knowledge graph. The question list is passed into the sorting module, and the multi-way matching list is merged, and some unreasonable matching questions are removed through the entity alignment, and the final related question list will be generated through deep learning sorting. Finally, the answer will be returned to the user according to the matching question with business type.

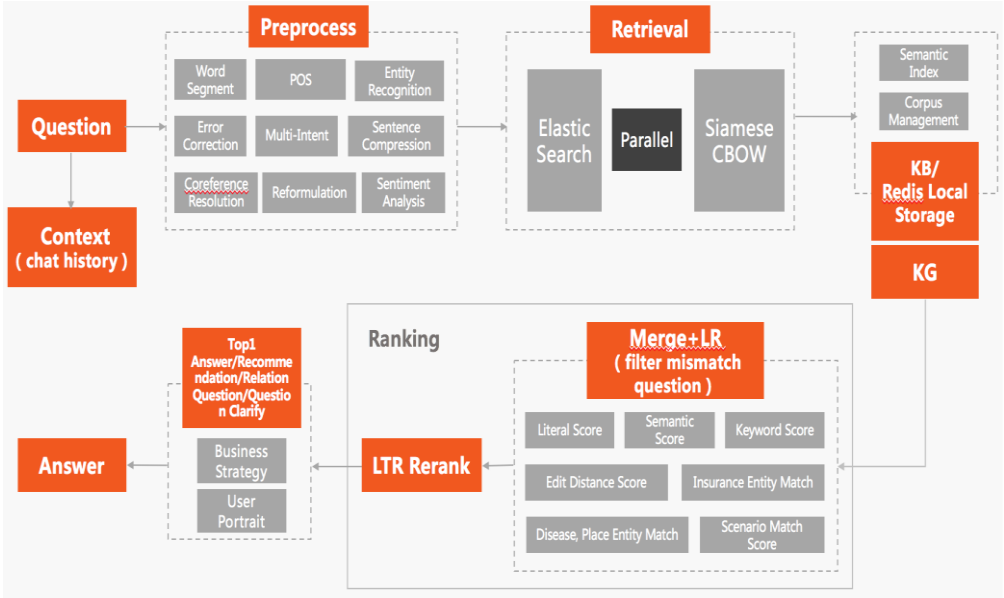

Figure 1: The overall architecture of our QA system

3.1 Pre-processing Module

We use open sourced NLP Tools with the insurance terminology dictionary for word segmentation, part-of-speech tagging and entity recognition. The multi-intention detection uses the method of splitting the sentence by punctuation and then classifying it. The question rewriting is mainly for the insurance product name, and the sentiment analysis is used to judgment the intents of the user's affirmation, negation and double negation. Following we describe more detail implements.

3.1.1 Long Sentence Compression

Step1: Divide the long sentence into several short sentences by punctuation or space, then classify the short sentences and remove the saliva statement

Step2: Based on the sentence compression scheme of probability and syntax analysis, we only retain the core sentence components. Combined with the insurance keyword dictionary to ensure the keywords are retained.

Example: Hello, I bought an insurance for my son in 2006 and I only paid 581 yuan for a year, however I didn’t pay for it after that. Now I want the customer service to refund my money.

Compress result: I bought an insurance in 2006. Now I want to refund my money.

3.1.2 Error Correction

Two solutions are used for business selection. The simple solution is based on the error correction of the insurance noun dictionary. According to the results of the previous word segment and syntactic analysis, the possible nouns are converted into PinYin and compared with the proper nouns in the dictionary for error correction. The general solution is the Transformer model with a special noun dictionary, the training datasets use about 32 million universal corpora from public news and the PinYin dictionary that from insurance domain. The input of encoder in the model is non-dictionary Chinese PinYin and Chinese word characters in the dictionary. The Decoder’s output is a pure Chinese character, where the Chinese characters in the input dictionary do not participate in the prediction, then directly generated.

3.1.3 Coreference Resolution

We use context chat history as Coreference Resolution reference. Our implementation ideas are word segmentation, part-of-speech tagging, dependency syntax analysis, subject-predicate extraction, entity substitution. For example:

(Question) What is the price of life insurance? (Answer) 300 yuan per year.

(Question) How about car insurance? (Coreference resolution result) What is the price of car insurance?

3.2 Retrieval Module

The retrieval module includes keyword search, deep semantic matching and KBQA recall, using the advantages of each of these three methods to increase the number and diversity of recall answers. The keyword search is retrieved using the open source ElasticSearch (ES) engine. As for the deep semantic retrieval, we use the deep semantic model to perform semantic vector representation on the user query and the knowledge in the knowledge-base (standard question and extension question), and use the Annoy algorithm to quickly find and match the semantic vector. The deep semantic model is modeled using the siamese network [5]. For each query, the similarity of the annotations is used as the positive sample, the negative samples are generated by the random sampling method, and random sampling is performed for each iteration, which greatly increases the randomness of the training data and improves the generalization ability of the model. Inspired by the idea from the loss definition of face recognition, we use AM-Softmax as the loss function and achieved the best results.

$$X_i = normalize(\sum_{w\ in\ s_i} embed_w)$$

$$\cos\theta_{y_i} = X_{\mathrm{q}}.X_i^T$$

$$L_{AMS} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s.(\cos\theta_{y_i}-m)}}{e^{s.(\cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c}e^{s.\cos\theta_j}}$$

The sentence vector X_i is obtained by normalizing the summation of word embedding in sentence S_i. The $\cos\theta_{y_i}$ is the similarity of user query vector and question i's vector. Both s and m are hyperparameters where s is the scale factor and m determines the classifier's boundary size.

In terms of feature extraction, for the purpose of extracting local word order relationships and context information better, we use LSTM, CNN, BERT and other networks to extract features. BERT performs best, but it takes a long time for online inference. Due to the limited quality of large-scale industrial corpus annotation, some data noise exists. The more complex models the more noise is fitted so the generalization ability is not as good as the simple model. Therefore, we chose the CBOW[4] model for feature extraction. Considering that Chinese word segmentation has limited effect in specific fields. To reduce the influence of word segmentation errors, we also use multiple dimensions of pre-training vectors to build our model, including: character embedding, word embedding, high-frequency phrase vector, where character embedding can solve literal matching, word embedding can represent the semantics of words, and phrase vectors can capture local-level word order relationships and achieve the best results.

We have done some benchmarks by using insurance domain dataset in different models also, the result show as following:

Method	Siamese LSTM	Siamese CNN	Siamese CBOW	BERT
Recall	80.6%	83.5%	85.2%	88.9%

Table 1: Benchmark results of Deep Semantic models

In KBQA, it receives the pre-processed question information, characterized by the context information, the entity type, and the entity relationship, and predicts the subject entity to be queried through the question recognition model[1], and the neighboring nodes centered on the entity from the KG.

3.3 Ranking Module

The ranking module includes a deep ranking model and a rule sorting. The deep ranking model is mainly used to merge and score the answers of multiple recalls. The rule sorting is mainly used to verify the rules of the sorted answers again to ensure not only the stability but also reasonability of the sorted answers. In the choice of deep ranking model, we use the commonly used pair-wise ranking model. Owing to the model is less difficult for data collection, we define the format of the input sample as the pair of ***<user query, candidate queries>*** when modeling. By constructing the scorer, the scores of the correctly matched samples are as high as possible (normalized to [0,1]) and the scores of the mismatched samples are as low as possible. The deep ranking model uses the interaction model, which not only considers the semantic vectors

of these two parts but also considers the calculation of the interaction information of these two parts so that it could get more accurate matching. In addition to semantic features, our model uses co-occurrence words in ***<user query, candidate queries>*** to model literal features. To better match the user's intention, we construct an intent classifier in the insurance industry, and perform intent feature extraction on the user query and the candidate queries respectively as input to the sorting model. In addition, we have made some attempts on the sentence features and get good results. As the Figure2 shows. We have:

$$\text{score}(\text{q}, \text{d}) = \text{FN}([X_d, X_q, X_{addtion}])$$

$$\text{L}(\text{q}, d^+, d^-; \theta) = \max(0, 1 - \text{score}(\text{q}, d^+) + \text{score}(\text{q}, d^-))$$

$\text{score}(\text{q}, \text{d})$ is the matching score of query and document ,while $X_d, X_q, X_{addtion}$ are the inputs of the neural network; $\text{L}(\text{q}, d^+, d^-; \theta)$ is the hinge loss of a train sample pair. Taking into account the professional requirements of the question-and-answer in the insurance field and the fact that sorting model cannot achieve 100% accuracy, we have added a priori knowledge of the insurance industry in the ordering of rules to ensure the professionalism of question-and-answer. Rule sorting mainly considers the alignment of professional entity information between the user query and the candidate question, and the best matching question should be consistent with the entity described by the user query and avoiding give an irrelevant answer.

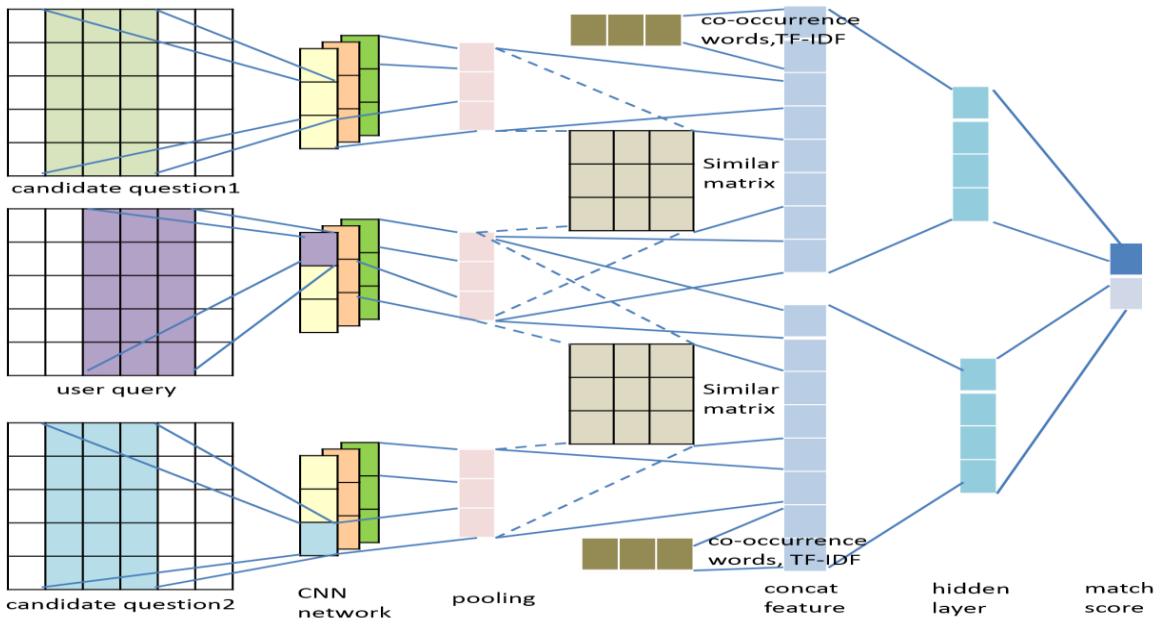

Figure 2: Our Deep Rank Architecture

3.4 Output Module

It gets the matching question list from rank module. If the confidence level is lower than the preset threshold, it will response a question to have user clarification and let the user to confirm the question he wants to ask and make a related question. If the confidence is high, the answer corresponding to the top one matching question or the recommendation question is returned according to the business rule.

3.5 Intelligent Knowledge-Base

The intelligent Knowledge-Base is a behind-the-scenes role in the Q&A system. In addition to providing the FAQ engine with raw materials, it also manages and optimizes the life-cycle of the question-and-answer knowledge. The specific process can be seen in Figure 3.

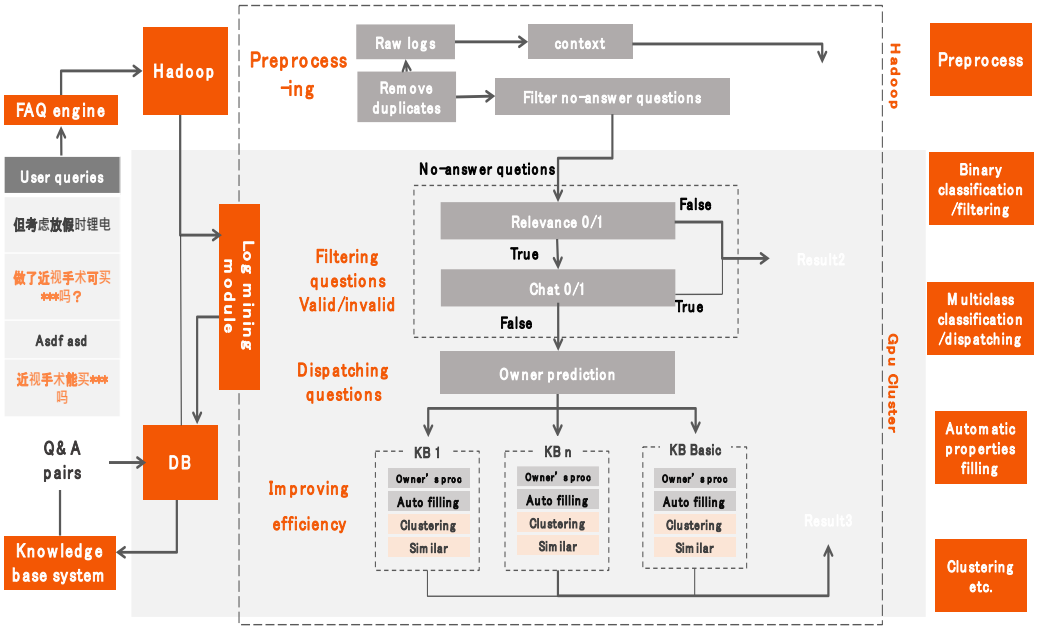

Figure 3: Our Intelligent KB Module Architecture

4 Evaluation Metrics

The Q&A assessment indicators mainly include the number of valid questions, Top1 response accuracy, Top3 accuracy, effective question response accuracy and knowledge coverage. The test set was divided into 5 categories, which are online log sampling used for the evaluation of model, the bad case collection, high frequency question mining used for algorithm regression testing, semantic test sets written according to demands fully cover the requirements, and the

corpus that delete the non-keywords, increase noise, synonym transfer and other methods to generate the literal test set to evaluate the robustness of the model. Our system achieved good results in these insurance business test sets and provided online service for one hundred million customers.

5 Conclusions

This paper proposes a context aware, error correction, coreference resolution, long sentence compression, ElasticSearch and deep semantic matching with the Siamese CBOW and deep learning sorting for the question-and-answer system. Our approaches not only have good performance in engineering but also in model accuracy. Its architecture supports high concurrency requirements in real world use cases and has high availability that fits the standard production environment. We have already applied this system in on-line intelligent customer service bot, AI assistant, AI selling bot and other human-computer interaction AI products. In the future, we hope our question-and-answer system could support multimedia interaction, such as pictures, audios and videos in addition to text and voice so that we could solve more problems for users with more intelligence.

REFERENCES

[1].Yunqi Qiu, Manling Li, Yuanzhuo Wang, Yantao Jia, Xiaolong Jin,2018, Hierarchical Types Constrained Topic Entity Detection for Knowledge Base Question Answering,ACM 2018, April 23–27, 2018 , Lyon, France.

[2] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, Wei Chu, 2017, AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine, ACL 2017 pages 498-503

[3] Kim Y. Convolutional Neural Networks for Sentence Classification[C] Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746-1751.

[4] Tom Kenter,Alexey Borisov,Maarten de Rijke, 2016, Siamese CBOW- Optimizing Word Embeddings for Sentence Representations, ACL 2016

[5] Paul Neculoiu,Maarten Versteegh,Mihai Rotaru, 2016, Learning Text Similarity with Siamese Recurrent Networks, ACL 2016 Proceedings of the 1st Workshop on Representation Learning for NLP, pages 148-157

[6] Aliaksei Severyn,Alessandro Moschitti, 2015, Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks, Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval pages 373-382

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, 2017, Attention Is All You Need, NIPS 2017

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Computation and Language 2018

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

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.