|| Volume 3 || Issue 5 || May 2018 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

System to Rank Answers in Community Question

Sana Kadar Sayyad[1], Dr. Pankaj M. Agarkar[2]
Student, ME Computer, DYPCOET, Department Of Computer Engineering, Pune[1]
Assistant Professor, DYPCOET, Department Of Computer Engineering, Pune[2]

Abstract— **Community question answering system-(cQA), one of the fastest-growing user-generated content (UGC) portals, has raised as an enormous market, so to speak, for the fulfilment of complex information needs. cQA enables users to ask/answer questions and search through the archived historical question-answer (QA)-pairs. Propose system present a novel scheme for answer selection in cQA settings. It comprises of an offline-learning and an online search component. In the offline-learning component, instead of time-consuming and labor-intensive annotation, we automatically construct-the positive, neutral, and negative training samples in the forms of preference pairs guided by our data-driven observations. We then propose a robust pairwise learning to rank model to incorporate these three types of training samples. In the online search component, for a given question, we first collect a pool of answer candidates via finding its similar questions. We then employ the offline learned model to rank the answer candidates via pair wise comparison. We have conducted extensive experiments to justify the effectiveness of our model on one general cQA dataset and one vertical cQA dataset.**

Keywords: *Answer Selection, Community Question Answering(cQA), Pairwise Comparison*

I INTRODUCTION

Question Answering (QA) is the task of automatically generating answers to natural language questions from humans. It provides a natural interface (often via text, image or speech) for human computer interaction (HCI), with the goal of satisfyingly answering as many questions as possible [10]. Question answering is one of the few natural language tasks most humans perform daily, among other common tasks such as natural language understanding and generation [1]. There are a wide variety of types of questions asked every day:

1. Fact-seeking (factoid), questions about general world knowledge. Usually they come with standard answers and can be judged as either correct or incorrect [2].

For instance:

- How old is the earth? (answered with a single short phrase)
- What are the planets in the solar system? (answered with a list of phrases)
- What is a planet? (answered with a definition)

Why is water essential to life? (answered with an explanation)

2. Opinion-seeking, questions about subjective belief. Usually they do not have definite answers but they can be judged as either relevant/acceptable or irrelevant/ unacceptable [2],[3].

For instance:

What's the most epic photo ever taken? (most popular question on Quora.com, a community-based QA website)

What if the chicken didn't cross the road? (hypothetical)

Which dress should I pick? (gentlemen, be careful with the answer)

How to grow a garden? (with a process answer)

In our proposed scheme, we have three main contributions:

1. Inspired by our user studies and observations, we present a novel approach to constructing the positive, neutral, and negative training samples in terms of preference pairs. This greatly saves the time-consuming and labor-intensive labeling process.

2. We propose a pairwise learning to rank model for answer selection in cQA systems. It seamlessly integrates hinge loss, regularization, and an additive term within a unified framework. Different from the traditional pairwise learning to rank models, ours incorporates the neutral training samples and learns the discriminative features. In addition, we have derived its closed-form solution by equivalently reformulating the objective function into a smoothed and differentiable one.

3. We have released the codes and datasets to facilitate other researchers to repeat our work and verify their ideas.

Question answering is such an amazing application for AI and NLP that people have imagined creating knowledgeable and/or conversational robots in so many fictional novels and movies [4],[11]. But it is still unsolved. This is the most fundamental motivation.

II MOTIVATION

Most existing Question Answering systems classify new questions according to static ontologies. These ontologies incorporate human knowledge about the expected answer (e.g. date, location, person), answer type granularity (e.g. date, year, century), and very often semantic information about the question type (e.g. birth date, discovery date, death date). While effective to some degree, these ontologies are still very small, and inconsistent. Considerable manual effort is invested into building and maintaining accurate ontologies even though answer types are arguably not always disjoint and hierarchical in nature (e.g. "Where is the corpus callosum?" expects an answer that is both location and body part). The most significant

|| Volume 3 || Issue 5 || May 2018 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

drawback is that ontologies are not standard among systems, making individual component evaluation very difficult and re-training for new domains time-consuming [1],[5].

The remainder is structured as follows. Section 2 reviews the related work. Sections 3 introduce the proposed work of system, followed by the conclusion and future work in Section 4.

III RELATED WORK

In existing system, finding similar questions from historical archives has been applied to question answering, with well theoretical underpinnings and great practical success. Nevertheless, each question in the returned candidate pool often associates with multiple answers, and hence users have to painstakingly browse a lot before finding the correct one.

Current QA Systems are capable of evaluating answers from complex system of data. Many of the present QA systems are for a particular domains that is, specific topic such as scientific topics, or for limited types of questions only, such as descriptive questions. Any problem with the present QA system is that they suffer from low recall. The answer to question is also limited to pre-defined categories [1].

Use a wide-coverage statistical parser which aims to produce full parses. The constituent analysis of a question that it produces is transformed into a semantic representation which captures dependencies between terms in the question. [2],[6] the current trend in Question Answering focus on open domain, which has been largely driven by the TREC-QA Track. Nonetheless, QA system of open domain is lacking to treat the special domains for all question types, because no restriction is imposed either on the question type or on the user's special vocabulary and it is very hard to construct a common knowledge (ontology) base for open domain.

A. Instance-Based Question Answering:

This paper presents a data driven, instance-based approach for Question Answering. We adopt the view that strategies required in answering new questions can be directly learned from similar training examples (question-answer pairs). Consider a multi-dimensional space, determined by features extracted from training data [7]. Each training question is represented as a data point in this space. Features can range from lexical n-grams to parse trees elements, depending on available processing.

B. Feature-driven Question Answering:

Treating question answering as a machine learning problem, the most fundamental challenge is capturing the most useful signals of the answer to the question. Or put it another way, question answering is a pattern recognition problem for answers. However, answer patterns do not come out of the box automatically: they need to be produced. This production process usually requires a lot of linguistic insight, and years of experience. One central challenge for this dissertation is to design methods for generating these answer patterns, then recognizing them, both in an automatic way. largescale discriminative training: the sequence tagging CRF model is a powerful tool [7],[8]. But how does it scale up, especially when we have tens of thousands questions for training? It also compares systematically several o_-the-shelf bilingual and monolingual aligners in the task of question answering.

IV PROPOSED WORK

We present a novel scheme to rank answer candidates via pair wise comparisons. In particular, it consists of one offline learning component and one online search component. In the offline learning component, we first automatically establish the positive, negative, and neutral training samples in terms of preference pairs guided by our data-driven observations [1]. We then present a novel model to jointly incorporate these three types of training samples. The closed-form solution of this model is derived. In the online search component, we first collect a pool of answer candidates for the given question via finding its similar questions. We then sort the answer candidates by leveraging the offline trained model to judge the preference orders.

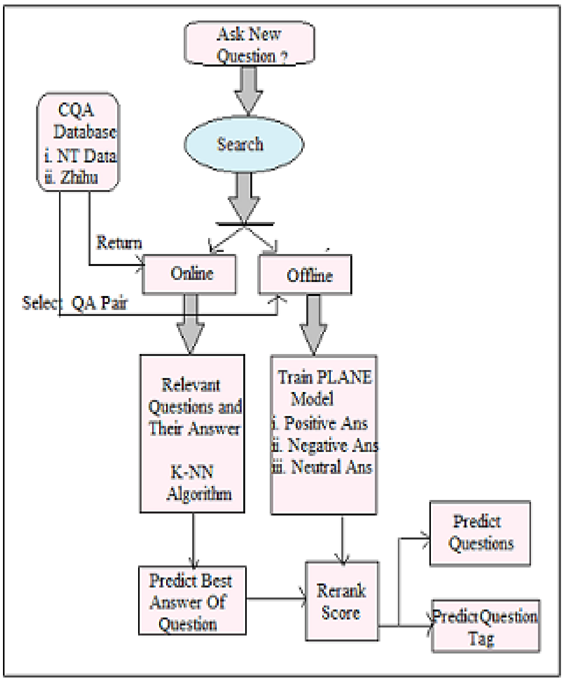

Figure 1 System Architecture

Advantages of Proposed System:

1. Our model can achieve better performance than several state-of-the-art answer selection baselines [1].
2. Our model is non-sensitive to its parameters.
3. Our model is robust to the noises caused by enlarging the number of returned similar questions.

4. The pair wise learning to rank models including our proposed PLANE are very sensitive to the error training samples [9].

V CONCLUSION AND FUTURE WORK

In this work, we present a novel scheme for answer selection in cQA settings. We can conclude that our model can achieve better performance than several state-of-the-art answer selection baselines, our model is non-sensitive to its parameters, our model is robust to the noises caused by enlarging the number of returned similar questions, and the pairwise learning to rank models including our proposed PLANE are very sensitive to the error training samples.

Beyond the traditional pairwise learning to rank models, our model is able to incorporate the neutral training samples and select the discriminative features. It, however, also has the inherent disadvantages of the pairwise learning to rank family, such as noise-sensitive, large-scale preference pairs, and loss of information about the finer granularity in the relevance judgment. In the future, we plan to address such disadvantages in the field of cQA.

ACKNOWLEDGEMENT

I express my sincere thanks to my project guide Dr. Pankaj M. Agarkar who always being with presence & constant, constructive criticism to made this paper. I express our deepest gratitude to all of them for their kind-hearted support which helped us a lot during paper work. At the last I thankful to my friends, colleagues for the inspirational help provided to me through a paper work.

REFERENCES

[1]. Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang, “Data-Driven Answer Selection in Community QA Systems,” *IEEE Transactions on Knowledge and Data Engineering, Vol*. 29, *No*. 6, June 2017.

[2]. A. Agarwal, et al., “Learning to rank for robust question answering,” *in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, *pp*. 833–842.

[3]. D. Savenkov, “Ranking answers and Web passages for non-factoid question answering: Emory university at TREC LiveQA,” *in Proc. 24th Text REtrieval Conf.*, 2015.

[4]. X. Li, Y. Ye, and M. K. Ng, “MultiVCRank with applications to image retrieval,” *IEEE Trans. Image Process., vol*. 25, *no*. 3, *pp*. 1396–1409, *Mar*. 2016.

[5]. M. K. Ng, X. Li, and Y. Ye, “MultiRank: Co-ranking for objects and relations in multi-relational data,” *in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, *pp*. 1217–1225.

[6]. L. Nie, Y. Zhao, X. Wang, J. Shen, and T. Chua, “Learning to recommend descriptive tags for questions in social forums,” *ACM Trans. Inf. Syst., vol. 32, no. 1, 2014, Art. no*. 5.

[7]. Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” *in Proc. 31st Int. Conf. Mach. Learn., 2014, pp.* 1188–1196.

[8]. Q. H. Tran, V. D. Tran, T. T. Vu, M. L. Nguyen, and S. B. Pham, “JAIST: Combining multiple features for answer selection in community question answering,” *in Proc. 9th Int. Workshop Semantic Eval., 2015, pp*. 215–219.

[9]. W. Wei, et al., “Exploring heterogeneous features for queryfocused summarization of categorized community answers,” *Inf. Sci., vol. 330, pp.* 403–423, 2016.

[10]. L. Yang, et al., “CQArank: Jointly model topics and expertise in community question answering,” *in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., 2013, pp.* 99–108.

[11]. B. Li and I. King, “Routing questions to appropriate answerers in community question answering services,” *in Proc. 19th ACM Int. Conf. Inf. Knowl. Manage., 2010, pp*. 1585–1588.