Knowledge-Based Systems 93 (2016) 75–83

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Learning semantic representation with neural networks for community question answering retrieval

Guangyou Zhou [a,*], Yin Zhou [a], Tingting He [a], Wensheng Wu [b]

[a] School of Computer, Central China Normal University, Wuhan 430079, China
[b] Computer Science Department, University of Southern California, Los Angeles, CA 90089, USA

ARTICLE INFO

Article history:
Received 29 May 2015
Revised 2 November 2015
Accepted 3 November 2015
Available online 10 November 2015

Keywords:
Community question answering
Question retrieval
Text mining
Yahoo! Answers

ABSTRACT

In community question answering (cQA), users pose queries (or questions) on portals like Yahoo! Answers which can then be answered by other users who are often knowledgeable on the subject. cQA is increasingly popular on the Web, due to its convenience and effectiveness in connecting users with queries and those with answers. In this article, we study the problem of finding previous queries (e.g., posed by other users) which may be similar to new queries, and adapting their answers as the answers to the new queries. A key challenge here is to the bridge the lexical gap between new queries and old answers. For example, "company" in the queries may correspond to "firm" in the answers. To address this challenge, past research has proposed techniques similar to machine translation that "translate" old answers to ones using the words in the new queries. However, a key limitation of these works is that they assume queries and answers are parallel texts, which is hardly true in reality. As a result, the translated or rephrased answers may not look intuitive.

In this article, we propose a novel approach to learn the semantic representation of queries and answers by using a neural network architecture. The learned semantic level features are finally incorporated into a learning to rank framework. We have evaluated our approach using a large-scale data set. Results show that the approach can significantly outperform existing approaches.

1. Introduction

With the development of Web 2.0, community question answering (cQA) services like Yahoo! Answers,[1] Baidu Zhidao[2] and WkiAnswers[3] have attracted both academia and industry great attention [1–3]. In cQA, anyone can ask and answer questions on any topic, and people seeking information are connected to those who know the answers. As answers are usually explicitly provided by human, they can be helpful in answering real world questions.

One fundamental task for reusing content in cQA is finding the existing answers for queries, as query-answer pairs are the keys to accessing the knowledge in cQA. Many studies have been done along this line [1,2,4–10]. One big challenge for community question answering retrieval is the lexical gap between queries and answers in the archives. Lexical gap means that the queries may contain words that are different from, but related to, the words in the answers. For example, if an user's query contains the word "company" but a target answer contains the word "firm", then there is a lexical gap and the target answer may easily regarded as an irrelevant one. This lexical gap has become a major barricade preventing traditional IR models (e.g., BM25 [11]) from retrieving the target answers in cQA.

To solve the lexical gap problem, previous work in the literature proposed a method to leverage query-answer pairs and learn translation models to improve traditional IR models [1,2]. The basic assumption is that query-answer pairs are "parallel texts" and relationships of words (or phrases) can be established through word-to-word (or phrase-to-phrase) translation probabilities [1,2,8]. Experimental results show that translation models obtain state-of-the-art performance for community question answering retrieval in cQA. However, query-answer pairs are far from "parallel" in practice, there are large number of unaligned words in query-answer pairs than in bilingual pairs, which confuses the word alignment tools [9].

In this paper, we study how to learn query and answer representations for community question answering retrieval. Specially, we head for modeling queries and answers in a more natural way. The model should not only highlight the instinct heterogeneity of queries and answers, but also be flexible enough to take other answers rather than the best answers into account. To this end, we propose a novel supervised approach to automatically learn semantic representations for queries and answers. The underlying assumption is that although

* Corresponding author. Tel.: +86 13720160306.
E-mail address: gyzhou@nlpr.ia.ac.cn (G. Zhou).

[1] http://answers.yahoo.com/
[2] http://zhidao.baidu.com/
[3] http://wiki.answers.com/

http://dx.doi.org/10.1016/j.knosys.2015.11.002
0950-7051/

questions and answers are heterogeneous in many aspects, they share some equivalences in the semantic level. Thus, we can learn the unified representations for query-answer pairs using an approach based on neural networks. In details, the procedure of using a deep neural network (DNN) to rank a set of answers for a given query is as follows. First, a non-linear projection is performed to map the query-answer pairs to a common semantic space. Then, the relevance of each answer given the query is calculated as the cosine similarity between their vectors in that semantic space. The neural network models are discriminatively trained using the query-answer pairs such that the cosine similarity of the best answer given the query is maximized. Finally, this semantic level feature is incorporated into a learning to rank (LTR) framework, which also includes a rich set of statistical-based features described in [12]. The relative importance of each feature is learned via a SVMRank algorithm [13] that utilizes a large-scale query-answer pairs in cQA archive. Different from the previous semantic models that are learned in an unsupervised fashion [14], our models are optimized directly for queries and the corresponding best answers, and thus give superior performance.

We evaluate the contribution of our semantic level features for community question answering retrieval under two settings: (1) large-scale automatic evaluation over query-answer pairs in cQA archives; (2) manual evaluation of the top retrieved answers for a set of test queries. We compare our approach to a state-of-the-art LTR model that utilizes only statistical-based features. The performance improved significantly in query-answer ranking by both evaluations when the semantic level features are incorporated, demonstrating the potential of our learn scheme for community question answering retrieval.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 describes neural network based approach for question and answer representation. Section 4 presents the learning to rank scoring model. Section 5 presents the experimental results. Finally, we conclude the paper in section 6.

2. Related work

2.1. Question retrieval in cQA

In recent years, with the flouring of community question answering (cQA) archives, significant research efforts have been conducted in attempt to improve question retrieval in cQA [1–9,15–20]. Particularly, the effectiveness of methods based on language model is proven. Most cQA researchers focus on leveraging metadata in cQA to improve the performance of the traditional language models for question retrieval [9]. Basically, there are five groups of work. The first group considers leveraging categories of questions. For example, Cao et al. [21] and Cao et al. [7] proposed a language model with leaf category smoothing in which they estimated a new smoothing item for language models from questions under the same category. Cai et al. [22] proposed a topic model incorporated with the category information into the process of discovering the latent topics in the content of questions. Then they combine the semantic similarity based latent topics with the translation-based language model [2] into a unified framework for question retrieval. Zhou et al. [17] proposed a faster and better retrieval model by leveraging category to filter certain amount of irrelevant questions under a wide range of leaf categories. Zhou et al. [3] proposed a novel approach called group non-negative matrix factorization with natural categories for question retrieval. This is achieved by learning the category-specific topics for each category as well as shared topics across all categories via a group non-negative matrix factorization framework. Recently, Zhou et al. [20] proposed to learn continuous word embeddings with metadata of category information within cQA pages for question retrieval. The basic idea is that category information encodes the attributes or properties of words, from which we can group similar words according to their categories. Thus the category information benefits the word embedding learning for question representation.

The second group leverages question–answer pairs to learn various translation models to bridge the lexical gap problem. For example, Jeon et al. [1] proposed a word-based translation model which exploits the semantic similarity among answers of existing questions to learn translation probabilities, which allows them to match semantically similar questions despite lexical gap. Xue et al. [2] proposed a word-based translation language model for question retrieval with a query likelihood model for the answer. Experiments consistently reported that the word-based translation model could yield better performance than the traditional methods (e.g., VSM, BM25 and LM). However, these word-based translation models are considered to be context independent in that they do not take into account any contextual information in modeling word translation probabilities. In order to further improve the word-based translation model with some contextual information, Riezler et al. [23] and Zhou et al. [8] proposed a phrase-based translation model for question and answer retrieval. The phrase-based translation model can capture some contextual information in modeling the translation of phrases as a whole, thus the more accurate translations can better improve the retrieval performance. Furthermore, Singh [15] addressed the lexical gap issues by extending the lexical word-based translation model to incorporate semantic information (entities). However, since it is possible for unimportant words (e.g., non-topical words, common words) to be included in the translation models, a lack of noise control on the models can cause degradation of retrieval performance. Lee et al. [5] investigated a number of empirical methods to eliminate unimportant words in order to construct compact translation models for retrieval purpose. Bernhard and Gurevych [6] proposed to use as a parallel training data set the definitions and glosses provided for the same term by different lexical semantic resources. Besides, Zhou et al. [24] proposed to use of translated words to enrich the question representation, going beyond the words in the original language to represent a question. Zhou et al. [25] proposed to employ statistical machine translation to improve question retrieval and enrich the question representation with the translated words from other languages via matrix factorization. Zhang et al. [19] explored a pivot language translation based approach to derive the paraphrases of key concepts.

The third group applies topic modeling techniques for information retrieval. In recent years, probabilistic topic models have also been introduced to cQA. For example, Cai et al. [22] incorporated question category into the traditional topic model and combined the topic model with a translation-based language model. Ji et al. [26] proposed a question–answer topic model to learn the latent topics aligned across the question–answer pairs to alleviate the lexical gap problem, with the assumption that a question and its paired answer share the same topic distribution. Zhang et al. [9] proposed a supervised question–answer topic modeling approach, which assumes that questions and answers share some common latent topics and are generated in a question language and answer language. Besides, other researchers also applied the topic models for the related tasks in cQA. Guo et al. [27] proposed a generative model to simulate user behaviors in cQA, for both question asking and answering, and then simultaneously obtain topic analysis of questions/answers and users. Then they recommended answer providers for new questions according to discovered topic as well as term-level information of questions and users. Zhou et al. [18] proposed a topic-sensitive probabilistic model by taking into account both the link structure and the topical similarity among users for expert finding.

The fourth group employs the syntactic information for question retrieval. For example, Duan et al. [4] first detected question topic and question focus by using a tree cut method and syntactic parser. They then proposed a new language model to capture the relation between question topic and question focus for question retrieval. After that, Wang et al. [28] proposed a syntactic tree matching model to finding