Expert Systems With Applications 91 (2018) 418–433

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval

Mladen Karan*, Jan Šnajder*

University of Zagreb, Faculty of Electrical Engineering and Computing, Text Analysis and Knowledge Engineering Lab, Unska 3, 10 000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 14 June 2017
Revised 11 September 2017
Accepted 12 September 2017
Available online 12 September 2017

MSC:
00-01
99-00

Keywords:
Question answering
FAQ retrieval
Learning to rank
ListNET
LambdaMART
Convolutional neural network

ABSTRACT

A frequently asked questions (FAQ) retrieval system improves the access to information by allowing users to pose natural language queries over an FAQ collection. From an information retrieval perspective, FAQ retrieval is a challenging task, mainly because of the *lexical gap* that exists between a query and an FAQ pair, both of which are typically very short. In this work, we explore the use of supervised learning to rank to improve the performance of domain-specific FAQ retrieval. While supervised learning-to-rank models have been shown to yield effective retrieval performance, they require costly human-labeled training data in the form of document relevance judgments or question paraphrases. We investigate how this labeling effort can be reduced using a labeling strategy geared toward the manual creation of query paraphrases rather than the more time-consuming relevance judgments. In particular, we investigate two such strategies, and test them by applying supervised ranking models to two domain-specific FAQ retrieval data sets, showcasing typical FAQ retrieval scenarios. Our experiments show that supervised ranking models can yield significant improvements in the precision-at-rank-5 measure compared to unsupervised baselines. Furthermore, we show that a supervised model trained using data labeled via a low-effort paraphrase-focused strategy has the same performance as that of the same model trained using fully labeled data, indicating that the strategy is effective at reducing the labeling effort while retaining the performance gains of the supervised approach. To encourage further research on FAQ retrieval we make our FAQ retrieval data set publicly available.

1. Introduction

Frequently asked questions (FAQ) collections are document collections composed of manually constructed question-answer documents (henceforth: FAQ pairs). FAQ collections provide an effective way to represent information and are particularly popular with large-scale service-providing companies and agencies for presenting readily available information to their customers. Typically, such collections focus on a very narrow domain of interest (products and services of a telecom company, services of a governmental agency, etc.).

While users can navigate FAQ collections themselves, using an *FAQ retrieval system* is often a much faster and more effective alternative. An FAQ retrieval system provides a natural language interface for querying an FAQ collection: based on a user's query, the system produces a list of FAQ pairs ranked by relevance. An FAQ retrieval system can be used at several points in the customer service workflow. First, it provides customers with efficient access to information regarding a company's products and services. Second, it provides customer service agents with rapid access to internal FAQ collections, increasing the quality and efficiency of customer service. Third, it allows partial automation of some customer service tasks, e.g., sending automated e-mail answers to the most typical user queries (Malik, Subramaniam, & Kaushik, 2007; Sneiders, 2010).

From an information retrieval (IR) perspective, FAQ retrieval is a profoundly challenging task. The main reason is that texts are short, making it harder to bridge the *lexical gap* – the word mismatch between a user's query and the text in an FAQ pair (Berger, Caruana, Cohn, Freitag, & Mittal, 2000; Lee, Kim, Song, & Rim, 2008).[1] Consider the following example:

Query: *"How can I seal a hole in a gas tank of a car?"*

* Corresponding authors.
E-mail addresses: mladen.karan@fer.hr (M. Karan), jan.snajder@fer.hr (J. Šnajder).
URL: http://www.takelab.fer.hr/mladen (M. Karan), http://www.takelab.fer.hr/jan (J. Šnajder)

[1] Not to be confused with *lexical gap* from linguistics, which refers to the "the lack of a convenient word to express what (the speaker) wants to speak about" (Lehrer, 1974).

http://dx.doi.org/10.1016/j.eswa.2017.09.031
0957-4174/

FAQ Q: *"How to patch a leak in the fuel cell of my automobile?"*
FAQ A: *"There are several ways to approach this..."*

In this example, even though the FAQ pair is highly relevant to this query, the word overlap is rather low. Generally, the lexical gap widens for shorter documents. The gap is also widened by a specific language that the users typically use for writing queries, as investigated by Barr, Jones, and Regelson (2008).

Much research in the IR community has focused on mitigating the lexical gap problem. According to Jeon, Croft, and Lee (2005), the approaches can be categorized as follows: those that leverage knowledge bases (e.g., Burke et al., 1997), those based on manually crafted rules (e.g., Sneiders, 2002), and statistical approaches (e.g., Berger et al., 2000). Statistical approaches can further be divided into unsupervised and supervised approaches. Statistical approaches have shown to be the most promising, especially those based on supervised machine learning, as such models can effectively learn to assess relevance based on the matches between specific words from the query and the document. Considerable effort has been devoted to developing supervised *learning-to-rank* models (Agarwal et al., 2012), particularly for the task of community question answering (CQA) – a task similar to FAQ retrieval but typically with much larger data sets available. The state-of-the-art CQA models typically utilize a rich feature representation of text, such as the neural network-based models (dos Santos, Barbosa, Bogdanova, & Zadrozny, 2015; Severyn & Moschitti, 2015). However, to the best of our knowledge, there is no study on applying such models to FAQ retrieval, where the data are domain-specific and the relevance labels are scarce or even nonexistent – a situation typical for FAQ collections of large service providers.

In principle, supervised ranking models, including the FAQ retrieval models, leverage two kinds of redundancy in data to learn the matching between user's query and a document (FAQ pair):

1. *Document redundancy (DR)* – Having observed several different relevant documents for a given query, the model can infer which terms (and the corresponding concepts) are important to match, and which can be ignored in matching. For instance, consider the following query: *"How do I change the light switch in my bedroom?"*. The relevant FAQ pairs contain (in their question part) *"Replace light switch in kitchen"* and *"Change wall switch in bathroom"*. The non-relevant FAQ pairs contain *"Install TV in bedroom"* and *"Turn off electricity in bedroom"*. The model can infer that, for the information need expressed by the query, a relevant document should match the concept of *change/replace* and the concept of *switch*, while *room type* is less important;
2. *Query redundancy (QR)* – Having observed several paraphrases of the same query, the model can infer which terms (and the corresponding concepts) are truly salient for the expressed information need (and thus important for matching) and which are merely noise. For instance, consider the following query: *"How do I remove a sticker from a window?"*. The query may be paraphrased as *"Removing a label from an empty jar of peanuts"*, *"Getting duct tape off of a glass surface"*, *"Removing sticker from windshield"*, or *"Taking off a tag that my friend stuck to our mirror"*. The model can infer that, for the information need described by the paraphrases, the most salient concepts are (1) *something adhesive*, (2) *a glass surface*, and (3) *removal*, while concepts such as *peanuts* and *friend* are less salient and should be considered less important during matching.

To illustrate the influence of both redundancy types on ranking models, consider a set of paraphrased queries, Q_{Ai}, all of which express the same underlying information need I_A and which are associated with a set of relevant documents $R_A = \{D_{A1}, \ldots, D_{AN}\}$. Moreover, consider another set of paraphrase queries, Q_{Bi}, corresponding to another information need I_B and associated with a set of relevant documents $R_B = \{D_{B1}, \ldots, D_{Bn}\}$. The situation may be depicted as a graph, as shown in Fig. 1, with the presence of an edge between two nodes indicating relevance of a document for a query. A supervised ranking model essentially exploits the information conveyed by the edges of this graph. Most information is conveyed by the edges that are *present* in the graph – to a supervised model they provide positive instances, i.e., query-document pairs where the document is relevant. However, to a lesser extent, the *absent* edges also convey some information, as they provide to a supervised model negative instances, i.e., query-document pairs where the document is *not* relevant. In any given domain, most query-document pairs will represent negative instances and are available in abundance. The supervised learning-to-rank model must, essentially, learn how to successfully differentiate a few positive instances from a multitude of negative ones. To properly learn this task, having an appropriate number of positive instances is crucial, which translates into increasing the number of edges in the graph. This can be done by either labeling more query paraphrases (query redundancy) or finding more relevant documents for reach information need (document redundancy).

While both types of redundancy are present, to a certain extent, in all IR problems, query redundancy is often more difficult to exploit directly. Namely, in most IR scenarios the query paraphrases are not explicitly available, i.e., while some queries may well be paraphrases of each other, this information is seldom explicitly encoded in the data set. Furthermore, in cases where this information is encoded, e.g., in the CQA collection of Hoogeveen, Verspoor, and Baldwin (2015), the number of paraphrases is rather small, because the policies of most large-scale community QA sites actively discourage duplicates.

Domain-specific FAQ retrieval is different from general IR in two important ways. First, the limited domain and the relatively small size of an FAQ collection make the set of unique information needs relatively small. This makes it *feasible* – in terms of the labeling effort – to provide query paraphrases for each of the expected information needs. Second, the information needs of users are relatively static, making query paraphrases very *practical*, as they will not have to be updated very often. In these respects, FAQ retrieval is more similar to CQA, though it does differ from general CQA in that FAQ collections tend to be focused on a much more limited domain and that FAQ questions tend to be much shorter than CQA questions.

The above observations motivate us to define three types of data labeling strategies with respect to the overall required effort:

- *Comprehensive* – Produces sizable amounts of both relevance judgments and query paraphrases, yielding both *document redundancy* and *query redundancy*, as shown in Fig. 1d. With respect to labeling effort, this type of strategy is the most demanding;
- *Relevance-focused* – Produces few or no query paraphrases but large amounts of relevance judgments (yielding more *document redundancy*, as shown in Fig. 1b. This strategy type requires slightly less effort than the *comprehensive* type but is still very time consuming as it requires relevance judgments for many FAQ pairs;
- *Paraphrase-focused* – Produces a small number of relevance judgments but a comparatively large number of query paraphrases (yielding more *query redundancy*), as shown in Fig. 1c.

All three strategies result in a labeled data set, which serves as training data for supervised ranking models. Furthermore, all three strategies rely on both types of redundancy, however they differ in the proportion of document vs. query redundancy.

Ranking models trained using data labeled via the comprehensive strategy should perform the best; yet, for most practical pur-