Golden Retriever: Question Retrieval System

Hamman W. Samuel, Mi-Young Kim, Sankalp Prabhakar, Mohomed Shazan Mohomed Jabbar
Department of Computing Science, University of Alberta, Edmonton, Canada
Email: {hwsamuel,miyoung2,sankalp,mohomedj}@ualberta.ca

***Abstract*—Duplicate questions get posted on Q&A online forums because users may not be aware of similar questions. Our proposed system, Golden Retriever, can recommend existing questions that are semantically related to incoming questions. Compared with other existing techniques such as Latent Semantic Indexing, Language Model and Semantic Similarity, our approach shows good results for the ICHI Healthcare Data Analytics Challenge dataset using normalized TF-IDF, relevance heuristics, and semantic relatedness.**

I. Introduction

In recent years Question and Answer (Q&A) forums about healthcare information have become quite popular. Large numbers of questions and answers archived in these forums provide a valuable resource to patients and caregivers. One major challenge with such resources is that users may ask the same question repetitively. This may be because the user is unable to spend time to find similar questions on their own, or due to lack of domain knowledge, leading to a significant increase in repeated questions. We develop a system called "Golden Retriever" to address this challenge of question duplication by retrieving and recommending semantically related questions for a new incoming question, thereby facilitating users in discovering previously posted questions that are similar to their intended queries. Golden Retriever is evaluated using an expert-curated dataset of online discussions from patient forums related to Type II diabetes, provided as part of the ICHI Healthcare Data Analytics Challenge. Retrieval of similar questions in community-based question answering has been studied extensively in [1], [2]. However, the characteristics of our dataset questions are different from [1]: a question consists of many sentences. In [2], the assumption is that a question and its paired answer share the same topic distribution, but our dataset does not contain corresponding answers. Moreover, we note that a question-answer pair does not necessarily share the same prior distribution over topics.

II. Methodology

Our approach approximates users' information seeking behaviour [3], which includes domain experts and non-experts. Given an incoming question, the expert searches for similar questions in the corpus by determining relevant keywords from the incoming question. Corpus questions that contain these keywords are candidates for matches. The expert also has a sense of semantically related words, such as synonyms and abbreviations, that could be matched in the corpus. Finally, the expert has knowledge of the context of the incoming question and what exactly is being asked. A high level overview of the system is given in Figure 1.

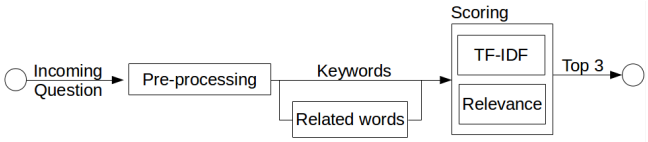

Figure 1. System Architecture

A. Pre-processing

As part of pre-processing, existing questions are indexed to create a corpus index. Regular expression tokenization is applied to each question and stop words are filtered using a modified Glasgow Stop Words List. New incoming questions are pre-processed by tokenization, stop words filtering, and expanding any keywords that are medical abbreviations to their full forms using a subset of the list from [4]. The incoming question's tokenized keywords are then matched with the corpus index and scored.

B. Scoring

For keyword matching, we compute the normalized TF-IDF for each keyword in the incoming question using the corpus index. The Term Frequency-Inverse Document Frequency (TF-IDF) measure is well-known in literature as a weighting factor for word relevance in documents. Given an incoming question, a matching score is determined for each corpus question, $score(C_j) = \sum \frac{\texttt{tfidf}(k_i, C_j)}{\max(\texttt{tf}(w, C_j))}$, where $\texttt{tf}$ is the term frequency of a keyword w in the corpus question C_j, while k_i represents the keywords in the incoming question. The best matches in the corpus are then determined by sorting the scores. The score is also adjusted using relevance heuristics and semantic relatedness.

C. Relevance heuristics

We note that the title of the question is a summarization of the question. Consequently, the keywords extracted from the incoming question title are given a higher score by using a constant relevance weight. Also, keywords matched with the corpus question title are given an additional higher score. For

determining relevance within the question's body keywords, we observe that various words are not relevant even after stop words are filtered. We apply another filter by removing all non-medical terms. The medical terms are determined by using Merriam-Webster's Medical Dictionary API [5]. We use the Lancaster Stemming algorithm to compare keywords to medical words and take into consideration different word forms, such as singularization/pluralization.

D. Semantic relatedness

Some words are semantically related but not similar, such as *libido* and *impotence*, or *sugar* and *glucose*. The domain expert has an understanding of these words as related, but most dictionaries, ontologies and thesauri we investigated do not group such words together. We ultimately used Moby Thesaurus [6] for synonym lookups because both semantically similar and semantically related words are grouped together. The Moby Thesaurus contains over 30,000 root words, 2.5 million synonyms and related words. Synonyms are used only in situations where no matches are found for an incoming question's keyword within the corpus. Moby Thesaurus is used to look up synonyms, and the TF-IDF for the synonyms is used as a score for the original keyword. The number of synonyms that match corpus questions is used as an offsetting weight. This is necessary to balance the possible inflation of scores for a keyword having many synonyms. As an additional relevance heuristic, synonyms are also filtered by removing all non-medical keywords that are not in the corpus titles.

III. Results

For evaluation, we use self-curated annotations for each incoming question by manually selecting the best possible matches (up to 3) from the corpus questions set. Using the annotations as baseline, recall is calculated for each incoming question based on the number of matches with the baseline, which is then averaged out for all questions in the incoming questions set. We obtain the best performance with the Golden Retriever system, referred to as the Current Model (CM). Using the CM, we get a recall of 71.8% for the ICHI Healthcare Data Analytics Challenge dataset. We also compare our approach with other popular pattern matching techniques: Latent Semantic Indexing (LSI) [7], Language Model-based Information Retrieval (LM) [8], and Semantic Similarity (SS) [9]. In some cases these models work well in detecting accurate patterns, but our approach shows the best results, as demonstrated in the recall rate comparison graph in Figure 2.

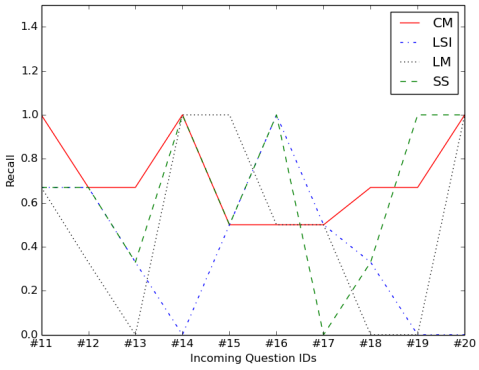

Figure 2. Recall Comparison

IV. Conclusion

A lot of duplicate questions are posted on Q&A forums because users may not be able to spend time to find similar questions, or they lack domain knowledge. Our proposed system, Golden Retriever, finds existing questions that are similar to new questions. Compared with other popular approaches, our system shows good results using normalized TF-IDF, relevance heuristics, and semantic relatedness.

References

[1] J. Luo, G. Q. Zhang, S. Wentz, L. Cui, and R. Xu, "SimQ: Real-Time Retrieval of Similar Consumer Health Questions," *Journal of Medical Internet Research*, vol. 17(2), 2015.

[2] Z. Ji, F. Xu, B. Wang, and B. He, "Question-answer topic model for question retrieval in community question answering," in *Proceedings of the 21st ACM international conference on Information and Knowledge Management*, 2012, pp. 2471–2474.

[3] D. Ellis and M. Haugan, "Modelling the Information Seeking Patterns of Engineers and Research Scientists in an Industrial Environment," *Journal of Documentation*, vol. 53, no. 4, pp. 384–403, 1997.

[4] Wikipedia, "List of Medical Abbreviations," https://en.wikipedia.org/wiki/List_of_medical_abbreviations, Retrieved July 4, 2015.

[5] M.-W. Inc., "Merriam-Webster's Medical Dictionary with Audio," http://www.dictionaryapi.com/products/api-medical-dictionary.htm, Retrieved June 20, 2015.

[6] G. Ward, "Moby Thesaurus," 1996, http://icon.shef.ac.uk/Moby/mthes.html, Retrieved July 7, 2015.

[7] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by Latent Semantic Analysis," *JAsIs*, vol. 41, no. 6, pp. 391–407, 1990.

[8] J. M. Ponte and W. B. Croft, "A Language Modeling Approach to Information Retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1998, pp. 275–281.

[9] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," *WordNet: An Electronic Lexical Database*, vol. 49, no. 2, pp. 265–283, 1998.