International Journal of Computer Applications (0975 - 8887)
Volume 183 - No.23, September 2021

Question Expansion in a Question-Answering System in a Closed-Domain System

Haniel G. Cavalcante, Jéferson N. Soares, José E. B. Maia
Centro de Ciências e Tecnologia - CCT
Universidade Estadual do Ceará - UECE
60714-903 Fortaleza-CE Brasil

ABSTRACT

A great challenge in Information Retrieval Systems (IRS) is to extract the information intention of the user from a command line interface query, so it can recover relevant documents. This problem gets worse in Question-Answering Systems (QAS) in a Closed Domain, for in this scenario, there's a higher divergence between the open language available for the user to elaborate questions and the limited vocabulary in the document collection available in the system (which is usually small). This work proposes and evaluates a system of Query Expansion (QE) for a closed domain QAS based on the semantic similarity between terms of the Word Net and a previously built semantic model using the system's knowledge base. The tests are made by answering questions about the two closed collections of documents showed this method is effective in improving performance of the Closed Domain QAS.

Keywords

Query Expansion, Question-Answering System, Closed Collection of Documents, Information Retrieval.

1. INTRODUCTION

Question-Answering Systems (QAS) are programs in Naural Language Processing area which aim to automatically extract answers to questions formulated by the user, which means no participation of a human on the answering. One of the classifications used for these systems has to do with the knowledge universe it intends to answer the questions [14]. Closed Domanin QASs aim to answer general questions based on open databases such as the Web, Restricted Domain QAS (RD-QAS) address answers to sector questions, such as biomedical, and Closed Domain QAS are confined to answering questions about a closed collection of documents, which is usually small. Examples of this last type would be a QAS to answer questions about a consumer protection code or customs laws of a state.

It is known [13, 11] that a great challenge in Information Retrieval Systems (IRS) is to extract the information intention of a user from a command line interface query, so it can correctly retrieve the relevant documents. This problem gets worse in Question-Answering Systems (QAS) in a Closed Domain, for in this scenario, there's a higher divergence between the open language available for the user to elaborate questions and the limited vocabulary in the document collection available in the system (which is usually small) [1].

Question classification and expansion before the information retrieval process can be used to mitigate these problems [15, 1]. Question Classification reduces the scope on the search for relevant documents and usually enhances performance of the QAS. However, classification does not help in the problem that usually the question is a text of few words, usually ambiguous, which are hard to extract the information of intention or need of the user. To mitigate this problem, query reformulation with or without expansion is used.

Query Expansion (QE) is the process of reformulating or adding terms to a given question to improve performance of information retrieval, particularly in the sense of better expressing intentions and needs of the user [17]. The core of this problem is when selecting the expansion terms based on the terms used in the original question. When available, context information can be used [8, 2].

A previous classification of question by its subject or type can be of much help in the query expansion [16]. For example, a question which has a type "where" and not "who" or "when" can help filtering the set of terms suggested by consulting Wordnet. As the classification, a context information can also be used. For example, knowing that this is the third question of a series about the local soccer team can eliminate many candidate words to be used in the expansion.

A frequently used technique for query expansion is usually to use lexical database such as WordNet [6]. These lexical databases can offer an array of other words related to a certain word, such as synonyms, hypernyms and hyponyms. However, if applied alone in the context of QAS in a Closed Domain it can be ineffective as many of the added words do not exist in the knowledge base. For that reason, in this context, it is necessary that a semantic filtering procedure of the suggested terms provided by WordNet is done before adding them to the expanded query.

This work proposes and evaluates a system of Query Expansion (QE) for a QAS in a Closed Domain (QAS-CD) based on semantic similarity between terms in the WordNet and the ones of a semantic model previously built from the system's knowledge base. The main idea is that as the collection of documents is small, a semantic tree model is built and a dictionary of terms of the knowledge base. This structure is used to filter the suggested expansion.

After this brief introduction, the rest of this work is organized in 4 sections. Section 2 reviews a group of works more closely related with this research and section 3 describes the used methods. Section 4 presents the experiments as well as the results and section 5 concludes the work.

2. RELATED WORK

This section presents the general lines of some works that represent the approaches used for query expansion in QAS while situating the proposed approach. A survey of such methods can be found in [17].
The research on the work [10] aims to develop a method for analyzing why-questions. The proposed method is based on domain ontology and considers the expected types of answers. It uses a bag-of-words model of semantic entities to represent a question instead of using the terms. The method expands a question by adding semantic entities obtained in a query executed in the domain ontology. A performance comparison done with methods based on keywords and key phrases shows better performance than these in the chosen data sets.
The work of [3] proposes to integrade the rule-based aprproach with sentence classification based o an HMM (Hidden Markov Model) to classify questions and extract answers in a Closed Domain QAS. The idea behind the method is that both techniques make use of the dependency relation between question words. The experiments show a significant and superior performance compared to the TF-IDF model approach used as baseline reference. The authors also published their new annotated data set used in the experiments in a public repository.
A question reformulation model and simultaneous validation of the answer is proposed in [9] with the objective of taking advantage of the natural interaction between these two modules of the QAS. In the question reformulation procedure the method uses the syntactical and semantical relations between the question terms as well as the similarity between the whole or parts of the current question based on the question-answer knowledge. A sub classification of the 5Q typology is used to classify the question in the initial stages of the process. The proposed method is evaluated with a survey on the users with satisfactory results, and it doesn't make comparison with other works.
In [7] the authors developed a QAS in a Closed Domain to answer questions about the Koran, holy book of the muslim religion. The question classification is based on n-grams and uses a neural network and the set of classes is highly restricted and dependent on the domain, made up of two classes: fasting and pilgrimage. The question expansion is based on the Word Net in English, which benefits from a collection of Islamic terms so that only the most meaningful words remain.
The authors of [5] developed a project of a QAS to answer queries from customers in a Closed Domain of services of a telecommunications company. The project is developed manually and does not contemplate query classification or expansion. Its methodology consists in aggregating semantic information in the knowledge base through keywords and headwords to improve the information retrieval module.
Although some of the ideas presented in this section are used, as is the case with the ideas of keyword, keyphrase and ontology, the proposed approach of this work should not be confused with them. Specifically, WordNet is used to generate an initial set of candidate words, which is filtered using a semantic model in the target Closed Domain, resulting in the expanded query. The semantic tree model used makes possible to choose the terms with greater specificity.

3. METHODS

Pre-trained deep learning models in conjunction with QA systems frameworks are state of the art in the development of question answering (QA) systems. However, fine-tuning these models can require dataset sizes that are not available in closed domain QA projects. This is the case of this project where it is believed that the project starting from scratch is advantageous.

3.1 Application Context

The Question expansion task in this work is in the context of developing a QAS in a closed domain (QAS-CD) which has the functional diagram shown in figure 1. In a view of greater blocks, an QAS is made up of three modules: question processing, document processing (information retrieval) and answer processing. After receiving and preprocessing a question with the pipeline adjusted to remove stop words, do stemming and uniformization, the question processing module makes two relevant operations: question classification and question expansion, which is the object of this work, to be detailed in the next section.

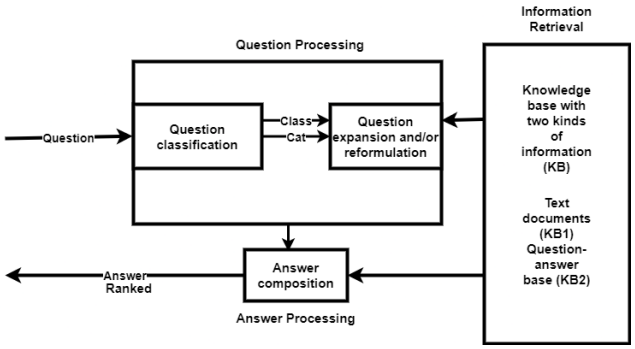

Fig. 1. Functional block diagram of the question-answer system in a Closed Domain (QAS-CD) being developed. This paper focuses on the "Question expansion and/or reformulation" block.

The second module is the document processing module, where the Information Retrieval (RI) is used t build an answer for the user. In this project the Knowledge Base (KB) is made up of two sets of documents in XML format: text documents (KB1), which define the closed domain that is object of the QAS and a questions and answers base (KB2) properly validated. A sentence (parts of the text) retrieval strategy or answers that meet the information needs of the user must be developed based on these two information sources.
In this text, the initial knowledge base will be called KB and KBs the semantically improved knowledge base made up of KB1 (domain documents) and KBs2 (question-answer document). Regarding the project development, KB2 can be a set of question-answer pairs elaborated by a specialist in the domain for the project purpose, it can be data from an existing FAQ (Frequently Asked Questions) or data collected from the Web for this purpose.
Finally, the third module is the answer processing module. It receives a set of texts from the RI module, possibly ranked by some criteria and must fuse this information with other information received from the question analysis module to finally select and format one or more answers to the user. Note that a correct classification and expansion of the question is an essential and critical stage which determines the success of the QAS-CD. It is critical because it's it's difficult for the following modules to recover a classification or expansion error in a question. The next section presents the expansion method proposed in this work.

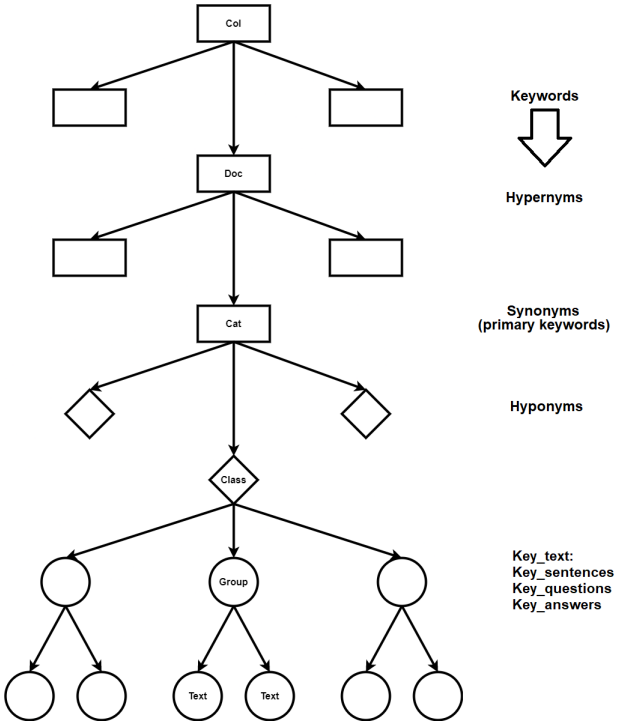

Fig. 2. Semantic model of the knowledge base used in the implementation of the QAS-CD. Trough the decomposition, each text cannot be part of two groups at the same time.

3.2 Query Expansion

In an IR-based QA system a question must be reformulated and turned into a query to the knowledge base. The proposal in this paper does not consider interaction between questions in a sequence. The reformulation dealt with here consists of an expansion of the question with a view to generating semantic intersection with the knowledge base in order to retrieve relevant excerpts (passages) or answers.

Figure 2 presents the general idea of the semantic model used for the knowledge base of the QAS-CD for both document collection and the question-answer base, which is also considered, as a whole, a document.

The model is a semantic tree where the root is a collection (col) made up of documents (doc) with its internal units, such as chapters and sections, grouped coherently in their semantic categories (cat). A category can be made up of a set of classes with some affinity in the perspective of the domain knowledge and of user language. Categories, in fact, are wide classes. A class (class), in turn, are formed by groups of texts (groups), each dealing with a different subject, but sharing some propriety. Texts are the smallest units of information retrieval used in the model. The modeling procedure must take care of the segmentation of texts so they won't result in ambiguity, belonging to more than one group at the same time.

The modeling procedure consists in segmenting the documents in basic units of text, which were called **passage**, group them in mutually excluding groups, organize them in the semantic hierarchy and associate **key words** or **key phrases** to each level of the semantic tree. Then, sets of classes and categories to cover KB1 and KB2 must be created by a specialist to finish the model.

It wasn't developed in this project a procedure for automatic class and category generation. In closed domains, this definition is viable ad hoc through a specialist. A good practice, that has shown to be very effective in the project is to initially associate relevant key words to the class level and use hypernyms and hyponyms of these words to levels of category and group respectively. In the current level of the project, **key words** and **key phrases** are also added based on knowledge of specialists.

The model is completed building a dictionary of words from the knowledge base *dic()* which will be used in the expansion procedure. The dictionary contains all possible words to be used in the question expansion. It is obtained in an interactive procedure of trimming an initial list of all present words in the KB.

The expansion procedure is described in the Algorithm 1. Initially, the algorithm pre processes and extracts the question terms. The following steps are based on the formation and processing of three sets: tp represents the set of terms of p, te the set of candidate terms to join the expansion, originated from the Wordnet, and tpe represents the set of expanded question terms. The procedure consists of using a *wordnet()* to acquire candidate terms to join the expansion, then restrict this set by using the dictionary *dic()*.

Algorithm 1 Pseudo-algorithm of the question expansion procedure using semantic information.

Input: A question to be expanded p, a dictionary of relevant domain terms dic and an API for querying Wordnet .
Output: The sets of terms of the expanded query, tpe.
$p \leftarrow preprocess(p)$
$tp \leftarrow terms(p)$
for $t_i \in tp$ **do**
  $te \leftarrow wordnet(t_i)$
  for $t_j \in te$ **do**
    if $dic(t_j)$ **then**
      $tpe \leftarrow append(tpe, t_j)$
    end if
  end for
end for

4. EXPERIMENTS AND RESULTS

The method was evaluated in two experiments with different knowledge bases. The first is a question and answer base about Covid-19 publicly available. The second is a knowledge base from a project the authors were working on.

Two metrics were used to evaluate performance which are P@X and MRR (*Mean Reciprocal Rank*) [12]. P@X measures precision for a question among the X best ranked documents retrieved. For example, for a test with N questions P@10 is givern by:

$$P@10 = \frac{\sum_{i=1}^{N} P@10_i}{N},$$

where $P@10_i$ is equal to 1 if the answer for this question is found in the TOP-10 documents and 0 otherwise. MRR, in turn, is defined as:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank(i)},$$

where $rank(i)$ is the position of the first relevant answer for the question i.

The base of passages generated from the domain documents (KB1) isn't naturally, labeled. And even for the question-answer base, although there is an associated answer for each question, it is possible to have an answer that answers another question that isn't associated with it. This happens because a same question can be formulated in different ways. Therefore, the ground truth to evaluate performance was done by the specialist inspecting recovered passages and answers.

4.1 Covid-Q

Covid-Q is a dataset of questions and answers about the virus and syndrome of Covid-19 [18] gathered from many open sources such as Quora, Google and Yahoo. They're questions asked by humans and answered by humans in English. The original file contains inconsistencies, such as questions without an answer and also question-answer pairs that are not categorized. After removing these two different cases, the data set used in this experiment ended up as shown in table 1: 16 categories with each category divided in classes, with a total of 93 classes with 693 question-answer pairs. The experiment used only the category level. For classification in tis level, the author in [18] explains that BERT classifier [4] reaches 58.1 % accuracy when trained with 20 examples for each category. This makes it a challenging data set to work on.
Note, however, that the categorization of question-answer pairs does not mean that an answer really provides what the question asks nor will the registered answer for a question will be satisfactory to provide what another question asks. The same question can be expressed in many ways. The questions were answered by humans and a correct content criteria for answers wasn't used in the data set. The tests were done using leave-one-out. The table 2 shows the result of experiments.

Table 1. Information about the Covid-Q data set used in the experiment of query expansion.

num	categoria	# class	# pares perg-resp
1	Individual Response	5	27
2	Economic Effects	0	0
3	Nomenclature	4	22
4	Comparison	4	30
5	Speculation	1	43
6	Reporting	4	36
7	Societal Effects	3	101
8	Symptoms	3	13
9	Origin	7	11
10	Prevention	15	44
11	Testing	7	28
12	Having COVID	4	21
13	Treatment	6	47
14	Transmission	22	215
15	Societal Response	6	43
16	Other	2	12
Total	16	93	693

Discussion: Due to the laborious procedure of inspection of returned passages in the information retrieved passages in the information retrieval, in order to identify which ones correspond to to each question, it was used a small set of 21 queries in this test. The most relevant comparison is between the first and last lines of the table. It can be noticed that all the values improve significantly with $P@5$ almost doubling, increasing from 0,33 to 0,57. The value $P@10$ of 0,86 means that the answer processing module in the diagram in Figure 1 can work with the 10 best results ranked. MRR improved from 0.19 to 0.24, which means that a greater number of correct answers appear closer from the top of the list. The tests with KBs1 and KBs2 aim to measure the partial effects of each part of the KBs. It is noteworthy that both contribute and that for obtaining better results, both must be used. From the last, results showed that $P@1$ starts to be different from zero, indicating correct documents are in the top of the list (2, in this case).

Table 2. Performance of the query expansion using the knowledge base for the Covid-Q data set.

Knowledge base	$P@1$	$P@5$	$P@10$	MRR
baseline (KB)	0	0,33	0,67	0,19
KBs1	0	0,38	0,71	0,19
KBs2	0	0,43	0,76	0,20
KBs	0,095	0,57	0,86	0,24

4.2 Regimento MACC (RgMacc)

The second experiment is part of a project of a QAS in development by the authors as described in section 3. The document base is the statute of MACC (Mestrado Acadêmico em Ciência da Computação, or Academic Masters in Computer Science in English) and the QAS must answer questions from students and candidates during the selective process. The statute was formatted as a XML file with each chapter, article, paragraph or caput begin considered as a candidate document for answer retrieval to a question. For the tests and creation of the KB2 it was asked from students and candidates the elaboration of 200 free questions that were processed, labeled and answered so they could be used in the modules of classification, expansion, information retrieval and answer processing in the project. Table 3 shows the classes and numbers of documents, including questions, answers or paragraphs of the statue present in the knowledge base. Out of these, 20 % were set aside for testing, chosen randomly. The tests were done using leave-one-out. Table 4 shows the experiment results.

Table 3. Information about the data set RgMacc used in the query expansion experiment.

id	cat	class	# pares perg-resp	# textos
1	adm	Organização	14	8
2	adm	Colegiado	13	20
3	adm	Coordenação	16	21
4	adm	Comitê	17	34
5	aca	Docentes	3	1
6	aca	ProcessoSeletivo	13	24
7	aca	Curso	9	16
8	aca	PlanoCurricular	20	32
9	aca	Avaliação	7	9
10	aca	Conclusão	10	23
11	aca	Título	4	8
Total	2	11	126	196

Discussion: Also in this case, due to the slow process of inspection of returned passages in the IR, to build the ground truth 21 questions were used for testing. A comparison of the numbers in lines 3 and 4 with line 2 shows that each part of the KBs contributes for an improvement in the result. It can also be seen that the semantic improvement of the KB with the proposed procedure in Algorithm 1

elevated $P@10$ from 0.57 to 0.91. This value $P@10$ of 0.91 means that the answer processing module in figure 1 diagram can work with the 10 best ranked results. MRR improved from 0.17 to 0.25 also points to good answers found earlier in the ranked list processing. Finally, only when using the semantically improved KB it was possible to obtain some of the correct answers in the top of the list (3, in this case).

Table 4. Performance of the query expansion using the dataset RgMacc.

Knowledge base	$P@1$	$P@5$	$P@10$	MRR
baseline (KB)	0	0,29	0,57	0,17
KBs1	0	0,38	0,76	0,19
KBs2	0	0,33	0,71	0,18
KBs	0,143	0,67	0,91	0,25

5. CONCLUSION

Query expansion aims to fill the gap between user information intention when writing a question and the question itself written in a few words. A good expansion is decisive for recovering useful documents in any context of Information Retrieval (IR).

This work examined the task of expanding queries in the development of a QAS in a Closed Domain CD when the database is made up of two kinds of information: some documents and a question-answer database. QAS-CD becomes this challenging task due to its knowledge base being too limited, not having enough diversity of sufficient cases for the expansion based on wide vocabulary to be efficient.

The proposed approach consisted in modeling a knowledge base as a semantic tree and associate keywords and key questions in the many categories of subjects both in texts and question-answer pairs. The results of the used datasets in the test showed that these tools generated more than 20 % in accuracy over the performance of the baseline method, either in MRR, P@1, P@5 or P@10.

Question syntactic analysis techniques based on grammar have been already used in other works [9] aiming to improve the question generation. In future works the goal will be to experiment and bring these techniques to this expansion method used and evaluate its improvement in a QAS-CD. Another work being developed is a procedure based on topic analysis (semi automatic) to generate sets of classes and categories for small size knowledge bases.

6. REFERENCES

[1] Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–199, 2000.

[2] Fabiano Tavares da Silva and José EB Maia. Query expansion in text information retrieval with local context and distributional model. *J. Digit. Inf. Manag.*, 17(6):313, 2019.

[3] Caner Derici, Kerem Çelik, Ekrem Kutbay, Yiğit Aydın, Tunga Güngör, Arzucan Özgür, and Günizi Kartal. Question analysis for a closed domain question answering system. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 468–482. Springer, 2015.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[5] Hai Doan-Nguyen and Leila Kosseim. Improving the precision of a closed-domain question-answering system with semantic information. In *Coupling approaches, coupling media and coupling languages for information retrieval*, pages 850–859. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2004.

[6] Christiane Fellbaum and K Brown. Encyclopedia of language and linguistics. 2005.

[7] Suhaib Kh Hamed and Mohd Juzaiddin Ab Aziz. A question answering system on holy quran translation based on question expansion technique and neural network classification. *J. Comput. Sci.*, 12(3):169–177, 2016.

[8] Yanli Hu, Chunhui He, Zhen Tan, Chong Zhang, and Bin Ge. Fusion of domain knowledge and text features for query expansion in citation recommendation. In *International Conference on Knowledge Science, Engineering and Management*, pages 105–113. Springer, 2020.

[9] Mohammad Reza Kangavari, Samira Ghandchi, and Manak Golpour. Information retrieval: Improving question answering systems by query reformulation and answer validation. *International Journal of Industrial and Manufacturing Engineering*, 2(12):1275–1282, 2008.

[10] AAIN Eka Karyawati, Edi Winarko, Azhari Azhari, and Agus Harjoko. Ontology-based why-question analysis using lexico-syntactic patterns. *International Journal of Electrical and Computer Engineering*, 5(2):318, 2015.

[11] Xiaowei Liu, Weiwei Guo, Huiji Gao, and Bo Long. Deep search query intent understanding. *arXiv e-prints*, pages arXiv–2008, 2020.

[12] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[13] Alaa Mohasseb, Mohamed Bader-El-Den, and Mihaela Cocea. Detecting question intention using a k-nearest neighbor based approach. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 101–111. Springer, 2018.

[14] Diego Mollá and José Luis Vicedo. Question answering in restricted domains: An overview. *Computational Linguistics*, 33(1):41–61, 2007.

[15] Xiaojun Quan, Liu Wenyin, and Bite Qiu. Term weighting schemes for question categorization. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):1009–1021, 2010.

[16] Jéferson N. Soares, Haniel G. Cavalcante, and José E. B. Maia. A question classification in closed domain question-answer systems. *International Journal of Applied Information Systems*, 12(38):1–5, July 2021.

[17] Haoming Wang, Ye Guo, Xibing Shi, and Fan Yang. Conceptual representing of documents and query expansion based on ontology. In *International Conference on Web Information Systems and Mining*, pages 489–496. Springer, 2012.

[18] Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. What are people asking about covid-19? a question classification dataset. *arXiv preprint arXiv:2005.12522*, 2020.