International Journal of Applied Information Systems (IJAIS) ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12 - No. 38, July 2021 - www.ijais.org

A Question Classification in Closed Domain Question-Answer Systems

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

ABSTRACT

A Question-Answer System (QAS) is a Natural Language Processing (NLP) application whose purpose is to answer questions from users by consulting one or more available knowledge bases. Classifying the question posed by a user in a class of a predetermined set has risks and advantages. On the one hand, the correct classification reduces the scope of the search for the answer, generally resulting in more correct answers and greater processing efficiency of the QAS; on the other hand, an error in the classification reduces the chances of the system recovering from this error in the later stages of processing, thus resulting, almost always, in unsatisfactory responses. This work develops and evaluates a question classification scheme for Closed Domain QAS. The experiments showed that the approaches described for defining class and question classification can be used successfully in QAS based on closed collections of documents.

Keywords

Question Classification, Question-Answer System, Closed Document Collection, Information Retrieval

1. INTRODUCTION

One of the common ways that human beings use to seek knowledge and meet their need for information is through the formulation of questions. Therefore, Question Answer Systems (QAS) are increasingly common in the context of instant communication to facilitate the interaction between the producer and the final consumer of information. There are basically two types of these systems: those in which a question submitted by a human user is answered by another human user and those in which the question is answered by the machine. In this work, a Question-Answer System (QAS) refers to a Natural Language Processing (NLP) application whose purpose is to answer questions from users formulated in natural textual language by consulting one or more available knowledge bases. A human does not interfere with the response.

There is a great effort to create efficient QAS systems which present quick, short, precise and specific responses, and are aware of the context, and which are able to validate the presented answer [19]. Studies on this topic seek to improve the return of information, that is, given a question, the system must return a coherent answer. Typically these responses are retrieved from unstructured data (i.e. data that does not have a logical organization) such as collections of texts extracted from the web or a set of local documents.

The working domain of a QAS can vary from a completely open domain system, which proposes to answer questions from the user on any subject, to closed domain systems in which the purpose is to answer questions based only on a closed collection of documents, usually small [14]. An example of the first type are systems that use the Web as a base to extract knowledge and generate responses, and an example of the latter type are QAS aimed at serving the specific users of a company's organizational processes. The Web is an open base of general subjects, constantly evolving, while the processes to be followed in an organization are registered in a fixed collection of documents that only changes from time to time.

Developing a QAS for general questions is a very different job than developing a QAS for the closed domain [9]. General purpose algorithms that work well in the former generally do not perform well in the latter due to the different properties of the knowledge bases. In closed-domain QAS it is possible for the specialist to analyze and model in depth because the knowledge base is small and very specific, while this is not feasible in general bases such as the Web [4].

In any case, the first step in a QAS is the analysis and understanding of the question. In this phase of analysis, classifying the question by type or subject is of great value as this facilitates the recovery of more correct and better expressed answers. There are some general forms of question types such as the 6Wh [5] collection (what, who, when, where, which, how much) that can be useful, but classification by subject is desirable in many cases and is heavily dependent on QAS knowledge base and purpose [18]. A classification by type 6Wh is very useful in the syntactic generation phase of the response while a classification by subject is valuable in the information retrieval phase.

In this universe restricted to a closed collection of documents, the performance of the classification by subject is closely related to the modeling of the subject in classes. This work presents the approach adopted to model the subject by classes and to classify the questions of a closed-domain QAS (CD-QAS) in order to enhance a better accuracy in information retrieval. The experiments show that the approaches described for defining classes and classifying issues can be used successfully in QAS based on closed collections of documents.

After this Introduction, Section 2 reviews related work and Section 3 presents the context of the project and the methods used. Section 4 presents the results of two experiments and Section 5 the conclusion.

2. RELATED WORKS

Some representative works of the approaches used to classify questions in QAS are summarized in this section. A survey of additional methods and references can be found at [18].

Work by [9] develops the design of a QAS to answer queries from customers in the closed domain of the services of a telecommunications company. Its methodology consists of aggregating semantic information in the knowledge base through keywords and headwords to improve the information retrieval module. The project is developed manually and does not include question classification or expansion.

In [4] a question classification scheme in just three broad categories is proposed and evaluated. The classification procedure includes analyzing the syntactic structure of the question and thus

suggesting the syntactic pattern and types of expected answers. Thus, evaluating on a public dataset, the work shows that even with a small set of question categories, the algorithm is able to classify questions with competitive results with those of the state of the art. The work allows the authors, in the end, to suggest more general syntactic patterns for questions than the Wh types.

Grammar-based syntactic pattern recognition is also a technique used to categorize questions [13]. The approach consists of developing general grammatical rules and categories or, for specific domains, exploring the grammatical structure of the question. Classes are defined by standards formed by grammatical categories corresponding to terms in the text. Then a grammar is constructed to explore the structure of the question and to classify it. The idea is that a grammar allows integrating a priori domain information about each question category in the tagging and classification phases. The experimental results for the specific domains tested are promising.

This work [20] focuses on how to extract and select different features adapted to different types of questions. Based on this idea, the authors build a method using a feature selection algorithm to determine the appropriate features for different types of questions. Guided by this result, they design a new type of feature based on question patterns. Tests using the SVM classifier on the TREC dataset benchmark resulted in an accuracy of 95.2% and 91.6% for coarse and fine-grained datasets, respectively, which are better compared to the previous studies considered in the article.

Semantic analysis techniques were also used in the work [10] to enhance improvements in the classification of questions. The authors shows that, for this type of application, semantically enriched syntax trees that are structurally more oriented towards task semantics can substantially improve classification and information retrieval performance.

The approach followed in this work uses some of the ideas presented in this section. Specifically, semantic knowledge of the target closed domain is associated with keywords and question patterns to generate the proposed classification procedure.

3. METHODS

3.1 Context

The question classification task dealt with in this work is in the context of the development of closed domain QAS whose functional diagram is shown in Figure 1. At the macro level, the IR-based QAS is composed of three modules [1]: query processing module, processing documents (information retrieval) and response processing module. After receiving and pre-processing a question with the already standardized pipeline of removing stop words, stemming and standardization, the question processing module performs two relevant operations: question classification, object of this work, detailed in the next subsection, and question expansion or reformulation [6].

The second module is document processing where information retrieval (IR) takes place to build a response to the user. In this project, the knowledge base (KB) is composed of two sets of documents in XML format: text documents (KB1), which define the closed domain that is the object of the QAS and a duly validated question and answer base (KB2). A strategy for recovering passages (text passages) from KB1 or responses from KB2 that meet the user's information needs should be developed based on these two sources of information. In this context, we are faced with a problem of information retrieval in closed document collection [7].

Finally, the third module is the response processing. It receives a set of passages from the IR module, possibly ranked by some criterion and must merge this information with information received from the question analysis module to finally select and format one or more responses to the user.

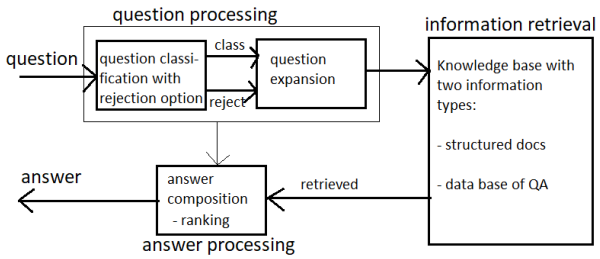

Fig. 1. Functional block diagram of the Closed Domain Question-Answer System (CD-QAS) being developed.

Note that the correct classification and expansion of the question is an essential and critical step that determines the success of the QAS functioning. It is critical because it is very difficult for subsequent modules to recover from an error in the classification or expansion of a question. Classification with reject option is being used to alleviate this problem [21]. Reject option is useful when the cost of misclassification is high and there is an alternative processing path. This must be considered in the design of the question expansion and the IR procedure. The next section presents the proposed classification procedure. This work does not address the problem of question expansion or reformulation, nor the alternative processing of the questions rejected by the classifier.

3.2 Question classification

Question classification is an instance of the short text classification problem [11]. The challenge in this task comes from the little semantic content contained in a short utterance. Figure 2 presents the general idea of the semantic model used for the knowledge base of the CD-QAS, for either the collection of documents and for the question-answer base, which is also considered, as a whole, a document.

The model is a semantic tree, inspired by [9], in which the root is a collection (col) consisting of documents (doc) with the internal units of the documents, such as chapters or sections, grouped coherently into semantic (cat) categories. A category can be formed by a set of classes with some affinity from the point of view of understanding of the users' domain or language. Categories, in fact, are broad classes. A class, in turn, is formed by groups of texts (groups), each dealing with a different subject, but which share some common property. Texts are the smallest units of information retrieval used in the model. The modeling procedure must take care that the segmentation in text units will not result in ambiguity of belonging to more than one group.

In this text, the initial knowledge base will be denoted KB, KBs will denote the semantically enhanced knowledge base composed of KBs1 (domain documents) and KBs2 (question-answers document). When developing a project, KB2 can be a set of questions-answers prepared by a domain expert for the purpose of the project, it can also be the use of a pre-existing FAQ (frequently asked questions) or a collection of QA carried out on the Web for the purpose.

The modeling procedure consists of segmenting documents into basic units of text, called passages, grouping them into mutually exclusive groups, organizing them in the semantic hierarchy and associating keywords (keyphrases) at each level of the semantic

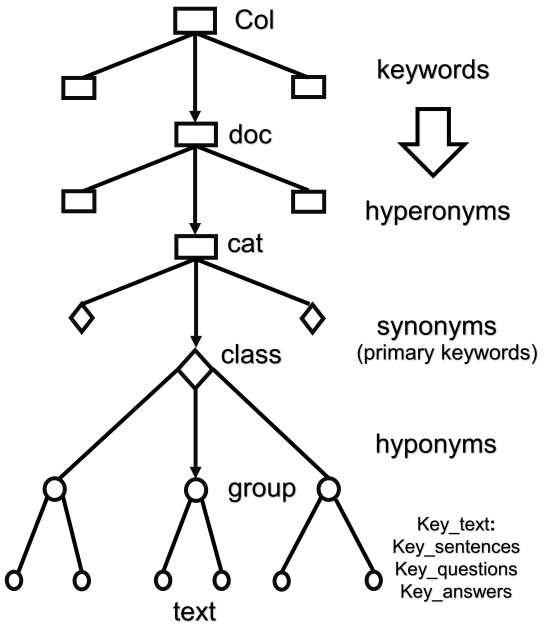

Fig. 2. Semantic model of the knowledge base used in the CD-QAS implementation.

tree. Then sets of classes and categories to cover KB1 and KB2 must be engineered by the expert to complete the model.

A procedure for the automatic generation of classes and categories has not yet been developed in this project. In closed domains this can be done ad hoc by a specialist. A good practice, which proved to be effective in the project, is to initially associate significant keywords at the class level and use hyperonyms and hyponyms of these at the category and group levels, respectively, always considering the vocabulary of the domain. In the current phase of the project, keywords and keyphrases are also added based on expert knowledge.

The proposed classification procedure is described in Algorithm 1. This algorithm essentially traverses the semantic tree in the bottom-up direction after finding $cos(p, pass())$, which is the cosine similarity of the TFIDF [12] representations of question p with each text passage $pass()$. It allows to consider weights ($\sum w_i = 1$) and costs ($\sum c_i = 1$) of classes and categories that can take into account knowledge of the domain or imbalance of the knowledge base [16]. It is a variant of KNN semantically weighted (weighted KNN) [2] and when all weights and costs are equal it executes simple KNN.

4. EXPERIMENTS

The method was evaluated in two experiments with knowledge bases from different sources. The first is a publicly available question-and-answer base about Covid-19. The second is a proprietary knowledge base of an authors project. The performance metrics used were Precision, Recall, F1 and Accuracy as they are classically defined [12]:

$$Prec = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 = \frac{2(Prec \times Recall)}{Prec + Recall},$$

Algorithm 1 Question classification procedure pseudo-code using semantic information.

Input: A question *p* for classification at the cat level, a semantic model as described in this section and the KBs Knowledge Base pre-processed.
Output: A question category, $cat(p)$.
$p \leftarrow preprocessing(p)$
for $i \in Cat$ **do**
  for $J \in Class(i)$ **do**
    for $k \in pass(j)$ **do**
      $sclass(i, j, p) =$
      $w_j \times argmax_k\{cos(p, pass(j, k))\}$
    end for
    $scat(i, p) = c_i \times argmax_j\{sclass(j)\}$
  end for
  $cat(p) = argmax_i\{scat(i)\}$
end for

and

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

,
where TP = true positive, TN = true negative, FP = false positive and FN = false negative.

In these experiments, the performance of the proposed algorithm, semantically weighted 1NN (swKNN) [2], is compared with Multinomial Naive Bayes (mNB) [12] as this was the one that presented the best result in preliminary tests with classifiers. Classification with reject option was not implemented in the experiments in this section.

Implementation: Processing was programmed in Python using resources from the numpy, NLTK [3], Sklearn [15] and GenSim [17] libraries, where convenient.

4.1 Covid-Q dataset

Covid-Q is an English language dataset of questions and answers about the Covid-19 virus and syndrome collected from several open pages such as Quora and Yahoo Answers [22]. These are questions issued by humans and answered by humans. The original file contains inconsistencies such as unanswered questions and also uncategorized question-answer pairs. After subtracting these two cases, the dataset used in this experiment has the characteristics listed in Table 1: 16 categories with each category subdivided into classes totaling 93 classes with 693 question-answer pairs. The experiment here used only the category level. For classification at this level, the author in [22] reports that the BERT [8] classifier achieves 58.1% accuracy when trained with 20 examples per category. This makes it a challenging dataset.

Note, however, that the categorization of question-answers does not mean that an answer actually meets what the question asks or that the recorded response to one question is not satisfactory to answer another question. The same question can be expressed in several ways. In this experiment, since the Covid-Q dataset has no texts, KB1 was taken as the question file and KB2 as the set of answers.

Tables 2 and 3 show the results of the experiments for the classifiers mNB and sw1NN, respectively, when using different versions of the knowledge base: the original base before the semantic improvement (baseline - KB), only the base of the questions (KBs1) , only the answer base (KBs2), and with all the semantically improved knowledge base (KBs).

Discussion: 90 questions were used in the test in the leave-one-out modality. In summary, three main points can be highlighted in the comparison of these two tables. First, that the preprocessing of the Covid-Q dataset raised the accuracy from 51.8% published by the author to 68% and 70% respectively, in the base-

lines of Tables 2 and 3. Second, that the partial knowledge bases KBs1 and KBs2 are both necessary to obtain better results. And third, that semantic improvement is effective in improving classification results. Finally, comparing the last lines of the two tables, it is clear that the proposed algorithm is superior to the multinomial NB, which gave the best results among the tested standard algorithms. The precision gain was 5%.

Table 1. Information about the Covid-Q dataset used in the classification experiment.

id	cat	# class	# qa pairs
1	Individual Response	5	27
2	Symptoms	0	0
3	Nomenclature	4	22
4	Comparison	4	30
5	Speculation	1	43
6	Reporting	4	36
7	Societal Effects	3	101
8	Economic Effects	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	8	12
Total	16	93	693

Table 2. Question classification performance for the Multinomial NB (mNB) classifier as a function of the knowledge base used for the Covid-Q dataset.

knowledge base	Prec	Recall	F1	Acc
baseline (KB)	0.68	0.66	0.67	0.68
KBs1	0.79	0.64	0.71	0.72
KBs2	0.79	0.67	0.72	0.71
KBs	0.86	0.89	0.87	0.88

Table 3. Question classification performance for the semantically weighted KNN classifier (sw1NN) according to the knowledge base used for the Covid-Q dataset.

knowledge base	Prec	Recall	F1	Acc
baseline (KB)	0.76	0.65	0.70	0.70
KBs1	0.85	0.80	0.80	0.81
KBs2	0.80	0.74	0.77	0.76
KBs	0.91	0.89	0.90	0.90

4.2 rgMacc dataset

The second experiment is part of a Question-Answer System (QAS) project under development by the authors as described in Section 2. The document base is the normative regulation of the MACC (Academic Master in Computer Science - UECE), in brazilian portuguese language, and the QAS must answer questions from students and candidates during the selection process. The normative regulation was formatted as an XML file with each chapter, article, paragraph or caput being considered as a document (text passage) candidate for recovery in response to a question.

For the tests and for the formation of KB2, students and candidates were asked to prepare 200 free questions that were processed, labeled and answered for use in this classification module and in the other stages of the project. Table 4 shows the classes and the number of documents, including questions, answers or paragraphs of the normative regulations present in the knowledge base. Of these, 20% were set aside for testing, chosen at random, ensuring that you have at least one question from each class. The tests were performed in the leave-one-out mode.

Tables 5 and 6 show the results of the experiments for categories, that is, for binary classification (aca = academic, adm = administrative).

Discussion: The same conclusions found for the Covid-19 dataset were confirmed here. Of particular note is the fact that the gains in precision and accuracy were more than double those, reaching 13%.

Table 4. Information about the RgMacc dataset used in the classification experiment.

id	cat	class	# qa pairs	# texts
1	adm	Organization	14	8
2	adm	Collegiate	13	20
3	adm	Coordenation	16	21
4	adm	Committee	17	34
5	adm	teachers	3	1
6	aca	Selective process	13	24
7	aca	Course	9	16
8	aca	Curricular plan	20	32
9	aca	Assessment	7	9
10	aca	Conclusion	10	23
11	aca	Title	4	8
Total	2	11	126	196

Table 5. Question classification performance for the Multinomial NB (mNM) classifier as a function of the knowledge base used for the RgMacc dataset.

knowledge base	Prec	Recall	F1	Acc
baseline (KB)	0.56	0.83	0.67	0.67
KBs1	0.57	0.66	0.67	0.67
KBs2	0.57	0.67	0.67	0.67
KBs	0.75	0.67	0.71	0.67

Table 6. Question classification performance for the semantically weighted KNN classifier (sw1NN) as a function of the knowledge base used for the RgMacc dataset.

knowledge base	Prec	Recall	F1	Acc
baseline (KB)	0.56	0.83	0.67	0.67
KBs1	0.63	0.71	0.67	0.67
KBs2	0.63	0.71	0.67	0.67
KBs	0.88	0.78	0.82	0.80

5. CONCLUSION

This paper examined the question classification task for the development of a closed-domain CD-QAS when the knowledge base is composed of two types of information: some documents and a question-answer base. CD-QAS makes this a challenging task because the knowledge base is limited and there is not enough diversity of cases to train a classifier with wide coverage.

The proposed approach consisted of modeling the knowledge base as a semantic tree and associating keywords and key questions in the various subject categories both in the texts and in the question-answer pairs. The results on the datasets used in the tests showed that these semantic devices generated advances of more than 40% accuracy on the performance of the baseline method and 13% when the proposed sw1NN classifier is compared with multinomial NB.

In continuation of this work, it is intended, in the future, to expand the method to take into account syntactic properties of the question using grammars. These techniques have already been used successfully in other works [4, 13]. Another work in progress is a procedure supported by topical analysis (semi-automatic) to generate sets of classes and categories for small knowledge bases. In addition, the full potential of tuning the weights and costs of Algorithm 1 has not yet been explored.

The problem of segmenting documents written in natural language into passages that fall into mutually exclusive groups is not trivial. One of the future works is to program this CD-QAS so that groups can share passages without compromising the accuracy of the system.

Another promising step in the evolution of this project is to migrate all processing to the embeddings domain. Significant gains are expected at all stages provided by this semantically enriched representation.

6. REFERENCES

[1] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 373–383. Springer, 2020.

[2] Christopher G Atkeson, Andrew W Moore, and Stefan Schaal. Locally weighted learning. *Artificial Intelligence Review*, 1(11):11–73, 1997.

[3] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

[4] Payal Biswas, Aditi Sharan, and Rakesh Kumar. Question classification using syntactic and rule based approach. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1033–1038. IEEE, 2014.

[5] Eduardo G Cortes, Vinicius Woloszyn, and Dante AC Barone. When, where, who, what or why? a hybrid model to question answering systems. In *International Conference on Computational Processing of the Portuguese Language*, pages 136–146. Springer, 2018.

[6] 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.

[7] Fabiano Tavares da Silva and Jose Everardo Bessa Maia. Luppar: Information retrieval for closed text document collections. *International Journal of Applied Information Systems*, 12(28):1–6, March 2020.

[8] 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.

[9] 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.

[10] Ulf Hermjakob. Parsing and question classification for question answering. In *Proceedings of the workshop on Open-domain question answering-Volume 12*, pages 1–6. Association for Computational Linguistics, 2001.

[11] João Marcos Carvalho Lima and José Everardo Bessa Maia. A topical word embeddings for text classification. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pages 25–35. SBC, 2018.

[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. Question categorization and classification using grammar based approach. *Information Processing & Management*, 54(6):1228–1243, 2018.

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

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

[17] Radim Řehřek, Petr Sojka, et al. Gensimstatistical semantics in python. *Retrieved from genism. org*, 2011.

[18] Valtemir A Silva, Ig Ibert Bittencourt, and José C Maldonado. Automatic question classifiers: a systematic review. *IEEE Trans. on Learning Technologies*, 12(4):485–502, 2018.

[19] Irfandy Thalib, Indah Soesanti, et al. A review on question analysis, document retrieval and answer extraction method in question answering system. In *2020 International Conference on Smart Technology and Applications (ICoSTA)*, pages 1–5. IEEE, 2020.

[20] Nguyen Van-Tu and Le Anh-Cuong. Improving question classification by feature extraction and selection. *Indian Journal of Science and Technology*, 9(17):1–8, 2016.

[21] Andrew R Webb. *Statistical pattern recognition*. John Wiley & Sons, 2003.

[22] 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.