I.J. Modern Education and Computer Science, **2015, 9, 59-65**
Published Online September 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.09.08

Two Way Question Classification in Higher Education Domain

Vaishali Singh
Department of Computer Science, B.B. Ambedkar Unibersity, Lucknow-226025, India
Email: singh.vaishali05@gmail.com

Sanjay K. Dwivedi
Department of Computer Science, B.B. Ambedkar Unibersity, Lucknow-226025, India
Email: skd.200@yahoo.com

Abstract—Question classification plays vital role in Question Answering (QA) systems. The task of classifying a question to appropriate class is performed to predict the question type of the natural language question. In this paper, initially we have presented a brief overview of classification approaches adapted by different question answering systems so far and then propose a two-way question classification approach for higher education domain which not only identifies focus word and question class but also reduces answer search space within corpus comprise of question-answer pair, adding to the classification accuracy. For precise semantic interpretation of domain keywords, a domain specific dictionary is constructed which primarily have four domain word type. Classified features are built upon domain attributes in the form of constraints. The experiment proved the efficiency for restricted domain, even though we used quite simplistic approach.

Index Terms—Question answering system, question classification, question taxonomy, focus word, restricted domain, generic.

I. Introduction

The Question classification is one of the prime components of question analysis task. Although different type of Question Answering systems so far developed have different type of architecture, most of them follow a framework in which question classification has key role. Generally, a question given to a QA system, first preprocess through tokenizer and parser. Thereafter, the question is directed towards question classification component which in particular, used to assign labels for identifying expected answer type and question focus. For instance, the question "Who is the chairman of University Grant Commission?" implies expected answer type as the name of a 'person' and question focus as 'chairman'. Determination of answer type and question focus assist in confining the search scope for response. The task of question classification is followed by keyword extraction for further expansion and formulation of question into an appropriate query.

A. Classification Approach

A QA system has to classify question given to its interface into one of the predefined classes as identified by the system while developing question class taxonomy. So far the researchers have generally opted two methods for categorizing their question class. One of these methods is based on 'Wh' words while other relies on domain specific keywords.

Generic approach based on 'Wh' word: Such QA systems classify questions into semantic categories based on Wh word involved in the question. The question asked to a QA system is put into its respective class on the basis of question word e.g., who, when, what, where, how, why. Earlier, some of the significant researchers like Harabagiu et al. [3] and Singhal et al. [2] had defined their taxonomies following this approach. Riloff et al. [1] developed rules to classify questions into classes based on Wh words. However, in current scenario, most of the systems are built upon the taxonomy proposed by Li and Roth [4]. Li et al. has defined two layered taxonomy which contains 6 coarse grained classes (Abbreviation, Entity, Description, Human, Location and Numeric Value) and 50 fine grained classes. Most of the QA system use coarse grained category definition to identify appropriate class. However, it is obvious that a fine grained category definition is more beneficial in locating and verifying the plausible answer. The method of classifying questions on the basis of question word has been successfully applied to many open domain systems and to their counterparts, restricted domain systems as well.

Domain based approach: Such classifications primarily rely on domain specific keywords. Therefore, selection of appropriate keywords to represent different aspects of the domain is key to this categorization approach. Only those keywords are selected which have enough distinctive capability to represent different features of the given domain. No doubt, this approach is not applicable to open domain QA systems as it would result in exhaustive and ambiguous classification due to presence of so many keywords. Xia et al.[7] for Chinese cuisine domain, Athenikos et al.[8] for medical domain, Han et al.[9] for tourism domain, Dang et al.[11] for e-library system and Fu et al.[10] for music domain have

build their question class taxonomy on this approach. The question presented to these QA systems is categorized for particular class according to the features best represented by the keywords in the question. However, the span of domain for such QA systems is not very large. This approach is therefore quite successful while classifying questions at coarse level but not at fine level. QA systems implementing this approach actually, looks for most appropriate information resource and confine their search space only for that feature of the particular domain which is best represented by the keyword(s).

Therefore, it is apparent that domain specific terminologies are quite helpful in limiting search scope (categorizing questions at coarse grained level) while question words help in identifying question and answer type at fine level. Therefore, in this paper, we are trying to propose a hierarchal integration of these two approaches for education domain which would be helpful in identifying appropriate question type, question focus and information resource (answer class) as well.

II. Two Way Classification Approach

In this paper, we performed two level of classification for higher education domain with the intention to refine focus word and question type at initial phase as proposed by Dwivedi et al. [15] and shrink the search space for required answer with the help of domain based terminologies at next phase. We first fetch the focus word and question type of the question for primary classification. In addition to focus words (which may also include main verb), we also extracted other keywords from the question along with the first auxiliary verb. Actually, the focus words, as name suggests are necessary to determine focus of the question but are not themselves sufficient to determine precise answer. Therefore, we need some additional keywords in the question which are necessary to frame context of the answer. After, extracting domain keywords, we would go for secondary classification. Secondary classification of the question is performed to identify answer domain inside the corpus as shown in Figure 1 which is followed by the query reformulation phase. The reformulated query is passed to document analysis phase.

A. Requirement

The idea of two-level question classification originates due to inherent architecture of our proposed QA system, which utilizes two type of information resource. Notion of two different types of resources is adapted depending on the nature of the asked questions. Katz et al. [5] had also worked on the idea of integrating web based and corpus based techniques for QA. The first information resource which we are incorporating in our proposed QA system is a corpus designed for higher education domain and another one is World Wide Web. The corpus is constructed to efficiently search for the answer of such questions which remain invariable or changes less frequently with time while Web is exploited to deal with frequently changing answers of the questions or extremely fresh questions. The structure of the QA corpus is organized according to the different sub-domains within the higher education domain. Whole of the QA corpus is divided into seven categories such as History, About, Admission and Scholarships, Academics, Examination, Events and Finance to provide efficient searching within the corpus. Therefore, secondary classification is required for mapping of asked question to the sub-domain of the QA corpus to extract answer efficiently. Here, we are going to adapt quite simple approach for secondary classification however, some of earlier works [6, 12, 13] suggests complex methods to search in large corpora.

B. Taxonomies

Two-way classification approach necessitates two taxonomies due to intrinsic requirements of individual approach. First level of classification focuses on determining question class and focus word for the asked question while second level of classification shrinks the search space for the answer within the corpus. Therefore, two taxonomies are defined here, one for generic classification based on Wh word and another one based on domain keywords.

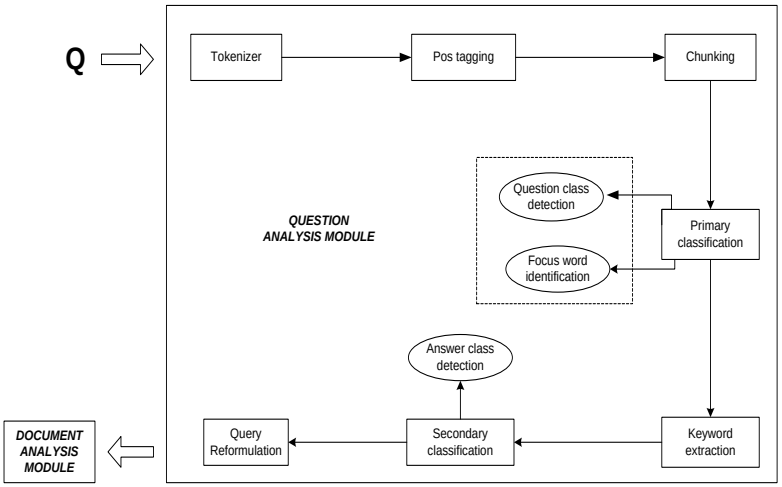

Fig.1. Question Analysis with Two Level Question Classification Approach

Table 1. Question Taxonomy for Generic Classification

Coarse classes	Fine classes
Abbr	Abbreviation , expression
Entity	Disease medicine, event, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word, caste, domicile, coarse, scholarship, fellowship, degree, subject area, job
Loc	City, country, state, other, website URL
Person	Individual, description
Number	code, count, distance, money, period, percent, speed, temp, size, weight, zip, phone number, grade, score, age
Org	Organization or institute, group or committee
Desig.	Designation
Description	Definition, description, manner, reason, criteria, syllabus, address
Time	Year, day, month, date, hour

Question taxonomy based on generic classification: Following the two layered approach of Li and Roth [4] we have developed our own taxonomy according to the requirements of higher education domain as shown in Table 1. The proposed taxonomy for this domain contains 9 coarse grained classes and 63 fine grained classes. Aim of the proposed taxonomy is to assign a semantic class at the primary level of classification. However, even at the primary level of classification, our efforts are intended to identify fine grained classes along with the focus word.

Paper Question taxonomy based on education domain keywords: The target users of our education QA system are basically the students or individuals who are interested in knowing basic information regarding any university or college for higher education. Our system will provide the details regarding administration, infrastructure, admissions, available departments and courses, placements, scholarships, vacancies, conferences, workshops organized in various universities and colleges. We have defined question taxonomy according to the requirement of the user which our system is capable of answering.

We have divided education domain into eight sub domains as shown in Table 2. Class 1 deals with the historical facts i.e., the information which remain invariant with time, class 2 gives information about individual concerned with a designation in an organization, subject etc., class 3 questions are related to various events such as conferences, workshops, annual festivals etc. organized by the different educational organizations, class 4 investigates about queries related to finances i.e., scholarships, fee etc., class 5 interrogates admission, semester and entrance examination details, class 6 interrogates various teaching and non-teaching opportunities in organization, class 7 gives information about available courses, faculties, placement opportunities and research programs and class 8 is dedicated to the general queries related to infrastructure, location and administration of the universities and can be considered as our default class too.

Table 2. Domain Keywords Based Taxonomy with Examples

Answer Class	Class Features	Examples
Class1	History	Who was founder of XYZ university? When XYZ university was established?
Class2	WHO's who	Who is the Vice chancellor of XYZ university? Who is the UGC chairman?
Class3	Events	Is there any national conference in Feb 2014?
Class4	Finance & Scholarships	What is the application fee for Mass Communication course? How do I apply for merit-based scholarships?
Class5	Examination & Admission	When entrance exam result will be declared for XYZ course in ABC university? What are the eligibility criteria for admission in XYZ course of ABC university?
Class6	Establishment	How do I apply for post of Assistant professor in Computer science department of ABC university? What are the teaching and non-teaching vacancies available in XYZ university?
Class7	Academics	What courses are available in XYZ university?
Class8	About	How many hostels does XYZ university have?

C. Domain dictionary of Higher Education

Usually, in restricted domain, a user put a question which comprise of specific terminologies. However, for education domain QA system, the user will not only pose specific terminologies but also so many abbreviations as name of courses, designation, subjects etc. for ease. Also, the terminologies in higher education domain are somehow associated to each other [14]. Therefore, to use the domain knowledge efficiently and express the answer space distinctively, we have developed a dictionary specific to Higher education domain which includes domain word type as illustrated in Table 3.

III. METHOD

In the phase of primary classification, we identified the question class and focus words with the help of heuristic rules and pattern matching as proposed by Dwivedi et al. [15]. In addition to focus words which may also include main verb (if present in question), we also extracted keywords from the question along with the first auxiliary/modal verb.

Table 3. Domain Word Type

Domain word type	Example
Designation along their abbreviation	Assiatant Professor, VC
Course along their abbreviation	Bachelor of Technology, M. Pharm
Subject or Research area	Forensic Science & Criminolgy
Fellowships	RGNF, University Grant Commision Fellowship

A. Classification Constraint

In restricted domain, classification constraints can be effectively determined by using domain knowledge and pos tagging. First auxiliary/modal verb is fetched with the intention to identify such questions whose answer remain eternal and can be categorized as class 1 questions.

In our question set, we analyzed our question answer corpus carefully and found that class 3, class 4, class 5 and class 7 necessarily exhibit some specific terminologies to uniquely determine their respective class while class 6 either rely on specific terminologies or occurrence of designation along with some other keywords. Therefore, we collected such kind of words and organized them into a keyword vocabulary for each specific class as the classification feature. Some of these representative keywords of the concerned classes are shown in Table 4. Keywords for each class are selected with utmost care but they need to be examined in a specific order with respect to a class to avoid any chance of the overlap as shown in Figure 3, for example, the question "*When entrance exam result will be declared for regular courses in JNU?"* may belong to both class 5 and class 7 according to matching keywords but order of similarity determination ensures that the question is associated to class 5 only.

Though, we have identified number of individual keywords for fsew classes but class 6 can be identified with individual keywords and co-occurrence of designation with some other additional keywords as well. These additional keywords are insufficient to be alone used as a classification feature but when integrated with some designation is capable enough to be used as predictive feature for the corresponding class. For instance, the keyword *vacancy*, when used with some designation of educational organization will correspond to class 6 of *establishment* otherwise will refer to class 7 (referring to placement related query) of *academics.* Consequently, after filtering through constraints of previous classes, course names or subject names can also be taken as characterizing feature of class 7 along with the specific keywords.

Table 4. Classification Constraint Corresponding to Answer Class

Answer class	Primary Constraint
Class 1	*Past verb phrase* (i.e., Main verb and its auxiliaries)
Class 2	*Primary class name, designation*
Class 3	*Keywords* e.g., conference, seminar, workshops, events, convocation
Class 4	*Keywords* e.g., fee, refund, payment, cost, scholarships, fellowship
Class 5	*Keywords* e.g., admission, entrance exam, exam, result, admit card, qualify
Class 6	*Keywords* e.g., teaching post, non-teaching post, OR *Designation combined with the keywords* e.g., recruitment procedure, pay scale, vacancy, job.
Class 7	Keywords e.g., schedule, academic syllabus, program, placements, courses, school, academic year, project, submission, stream, placements, semester etc. OR *Course name* OR *Subject name.*
Class 8	*About* (Default)

The classifying criteria for class 1 and class 2 has been chosen bit differently from rest of the classes. Class 1 is dedicated to history related questions or invariant past information. Therefore, instead of selecting specific terminologies as classification feature, we search for past verb phrase in the question, which specifically refers a word group that includes a main verb and its auxiliaries of past form. Class 2 is kept over the top of taxonomy after class 1 to answer very common questions usually asked by the user of who's who type for reducing system's complexity. The predictive feature of class 2 is recognized by the binding of primary class i.e., *person* and designation name. The default class for our classification task is class 8 which refers to general information regarding administration, infrastructure and many more. In summary, we implemented three strategies to extract classification feature: (1) Using keywords of question, (2) Using order of keyword similarity determination with respect to classes, (3) Using the past verb phrase (4) Using combination of primary class and domain attributes, (5) Using integration of keywords with domain attribute. The classifying features which are related to each answer class are listed in Table 4.

B. Classification Rule

According to the analysis of question instances and classified features, we summarized the formation of each kind of question and presented the classification rules in higher education domain.

R1: IF the question contains past verb phrase(main verb, auxiliaries or modal verb),
THEN the question must belong to the class 1.

R2: IF the question has primary class as WHO and keywords as designation,
THEN the question belong to class 2

R3: IF the question contains the keywords defined specifically for class 3
THEN the question belong to class 3

R4: IF the question contains the keywords defined for class 4
THEN the question belong to class 4

R5: IF the question contains the keywords defined for class 5
THEN the question belong to class 5

R6: IF the question contains the keywords defined specifically for class 6 or contain a designation instance along with the respective keywords,
THEN the question belong to class 6

R7: IF the question contains the keywords defined for class 7 or contain an instance of course or an instance of Subject,
THEN the question belong to class 7

R8: IF the question does not belong to any of the first seven classes,
THEN the question belong to class 8(default class)

Fig.2. Rules for Question Classification to Reduce Answer Search Space

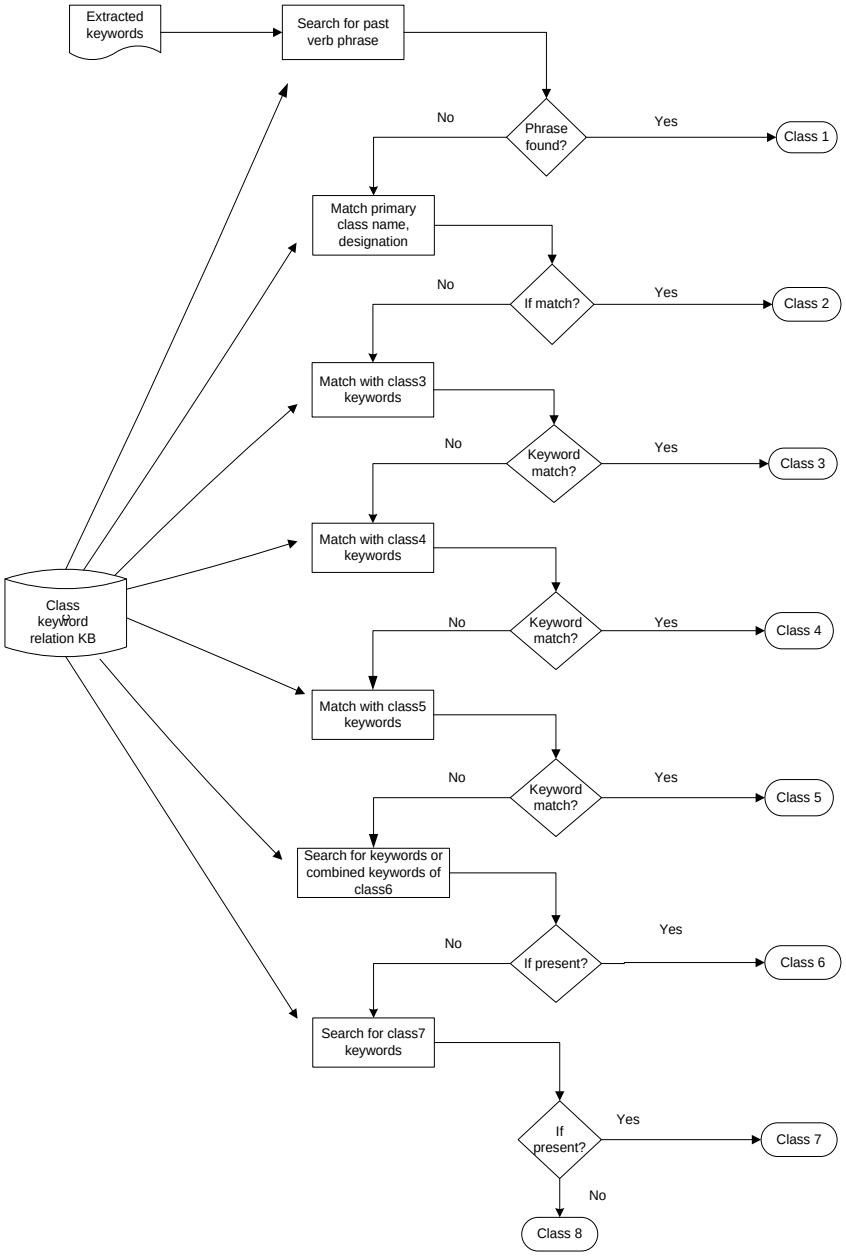

Fig.3. Flowchart for Answer Source Identification

Our rules classifying education domain questions and thus reducing answer search space can be are illustrated as Figure 2. These rules constitute a hierarchal classifier as matching filtering algorithm. If certain rules are fulfilled in user's question, we classify the question into corresponding question class otherwise continue to next level for matching. We follow the algorithm step by step.

IV. Experiment

We have used 254 questions of higher education domain as testing dataset for evaluating the performance of our question classification approach. These questions cover each type in the taxonomy. The distribution of the questions belonging to each class is shown in Figure 4.

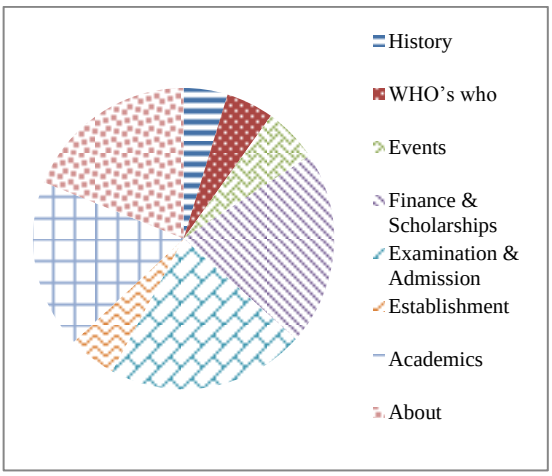

Fig.4. Distribution of Questions among Classes

Table 5. Result of Proposed Classification Approach

Answer class	Questions	Correctly classified	Incorrect Classification	Accuracy
Class1	12	11	1	91.67
Class2	13	12	1	92.30
Class3	15	13	2	86.67
Class4	53	49	4	92.45
Class5	54	47	7	87.03
Class6	13	11	2	84.61
Class7	46	40	6	86.95
Class8	48	41	7	85.41
Total	**254**	**224**	**30**	**88.39**

We conducted our experiment for evaluating answer class detection ability of our approach with in the higher education domain. Performance is evaluated by accuracy for a specific class c, defined as:

$$Accuracy = \frac{\#\ of\ correctly\ classified\ instances}{\#\ of\ total\ instances}$$

The overall accuracy shown by our system is 88.39 which will be quite effective in locating answer sub domain within the corpus.

V. Discussion

As we can observe from the Table 5 that our simple domain keywords based classification approach achieves satisfactory performance, but the accuracy of class 6 and class 8 is relatively low. The reason behind lower accuracy of class 6 is due to the fact that vacancy related questions may belong to class 6 if asking about jobs in educational institution but may also belong to class7 if it is about some placement information. Also, our algorithm flow ensures that the algorithm will search and match the classified feature step by step from the class 1 to class8. Therefore, after all the filtration and matching, the class 8 belongs to rest of the questions which actually only defined for the administration and infrastructure related information.

We expressed the classifying features with domain keywords, so when the fresh keyword occur or there are missing keywords or too general words in a question, the corresponding class cannot be detected appropriately and question by default goes to class 7. Among 30 wrong instances of classification, 11 questions are incorrect due to too general terminologies, 2 questions are missing in essential keywords while 2 questions posses fresh keywords.

Furthermore, the lower accuracy of class 3 that belongs to event related information is due to the name of such events in acronym form. It is an exhaustive task to gather prior information regarding acronym of all event name that are going to be happened in future and hence according to classifying feature, the question will be classified to wrong class. The primary reason behind erroneous classification instances of class 5 is owing to the variety of questions belonging to the academics in higher education, which make it difficult to design perfect classification mechanism.

From Table 5, it can be observed that both class 1 and class 2 have shown good accuracy while class 4 is most promising one among all classes. The success of our classification approach heavily depends on the selected keywords for the concerned classes. Being restricted within a specific sub domain, these classes found most unambiguous set of keywords which clearly represent their concerned category. Also, the primary class constraint worked as good predictive feature for class 2. The algorithm flow is another prime factor which adds to the performance of our approach in the associated classes.

In summary, the instances which are incorrectly classified are either result of too general keywords, missing keywords or fresh keywords in the questions. Our mechanism fails to classify few other questions because of misinterpretation of keywords belonging to another class. However, the adopted classification approach has shown accuracy of 88.39 in identifying the answer sub-domain despite being simplistic.

VI. Conclusion And Future Scope

This paper proposed a two way classification approach for higher education domain. The aim is to improve accuracy for extracting answers to question from the corpus with in a specific domain. The approach actually aids in identifying answer resource within the QA corpus. The questions involved in education domain relatively posses fewer number of question patterns which can be easily represented with the help of our approach employing a little bit of human effort. In the proposed paper, we show that constraints based on only domain keywords perform satisfactory well for answer sub domain identification. In future, we want to exploit semantics and semantic knowledge resources such as WordNet to identify appropriate context of the question having too general or missing keywords.

References

[1] Riloff E and Thelen M., A Rule-based Question Answering System for Reading Comprehension Tests. In ANLP /NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, Vol. 6, pp. 13-19, 2000.

[2] Singhal, A., Abney, S., Bacchiani, M., Collins, M., Hindle, D. & Pereira, F., In Proceedings of the 8th Text Retrieval Conference, NIST, 2000.

[3] Harabagiu, S. M., Moldovan, D. I., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R. C., Giriju, Rus, V. & Morarescu, P. FALCON: Boosting Knowledge for Answer Engines. In Proceedings of the 9th Text Retrieval Conference, NIST, Vol. 9, pp. 479-488, 2000.

[4] Li, X. & Roth, D., Learning question classifiers. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), pp.556-562, 2002.

[5] Katz, B., Lin, J. J., Loreto, D., Hildebrandt, W., Bilotti, M. W., Felshin, S., Fernandes, F., Marton, G., Mora, F., Integrating Web-based and Corpus-based Techniques for Question Answering. In TREC, pp. 426-435, 2003.

[6] Kaisser, M., and Becker, T., Question answering by searching large corpora with linguistic methods. In Proceedings of the 13th Text REtrieval Conference (TREC 2004), 2004.

[7] Xia, L., Teng, Z. & Ren, F., An Integrated Approach for Question classification in Chinese Cuisine Question Answering System, Second International Symposium on Universal Communication, 2008.

[8] Athenikos, S. J., Han, H., & Brooks, A. D., Semantic analysis and classification of medical questions for a logic-based medical question-answering system, In IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp. 111-112, 2008.

[9] Han, L., Yu, Z. T., Qiu, Y. X., Meng, X. Y., Guo, J. Y., & Si, S. T. Research on passage retrieval using domain knowledge in Chinese question answering system. In IEEE International Conference on Machine Learning and Cybernetics, Vol. 5, pp. 2603-2606, 2008.

[10] Fu, J., Xu, J., & Jia, K., Domain ontology based automatic question answering. In IEEE International Conference on Computer Engineering and Technology, Vol. 2, pp. 346-349, 2009.

[11] Dang, N. T. & Tuyen D.T.T., Document Retrieval Based on Question Answering System. In IEEE Second International Conference Information and Computing Science, Vol. 1, pp. 183-186, 2009.

[12] Fakhr, M. S., & Abadeh, M. S., AISQA-An Artificial Immune Question Answering System. International Journal of Modern Education and Computer Science, Vol. 4(3), pp. 28-34, 2012.

[13] Arai, K., & Handayani, A. N., Question Answering for Collaborative Learning with Answer Quality Predictor. International Journal of Modern Education and Computer Science, 5(5), pp. 12-17, 2013.

[14] Mashat, A. F., Fouad, M. M., Philip, S. Y., & Gharib, T. F., Discovery of Association Rules from University Admission System Data. International Journal of Modern Education and Computer Science, 5(4), pp. 1-7, 2013.

[15] Dwivedi S.K. & Singh V., Integrated Question Classification based on Rules and Pattern Matching. In International Conference on Information and Communication Technology for Competative Strategies, 2014.

Authors' profiles

Vaishali Singh is Research Scholar at Department of Computer Science in B.B. Ambedkar University, Lucknow, India. She has received her M.C.A. Degree in the year 2010 from UP Technical University. Her research interest includes Information Retrieval and Question Answering Systems. She has published some of the research papers in international conferences and journals.

Prof. S.K. Dwivedi is Professor and Head at Department of Computer Science in B.B. Ambedkar University, Lucknow, India. He has received his Ph.D. Degree from Banasthali Vidyapeeth in area of Web Mining in the year 2006. His research interest includes Web content Mining, Semantic Web, Search Engine performance evaluation, Machine translation, Information Retrieval etc. He has published many of the valuable research papers in various national and international Journals of repute.

