Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Domain Specificity for Focused Retrieval in SMS based FAQ Retrieval

Jotsna Waghmare*, M. A. Potey
Department of Computer Engineering,
Savitribai Phule Pune University, India

Abstract— As information retrieval has become a most- appreciated part of everyone's life, it has turn into an interesting area of research that is how to make information retrieval system convenient. Nowadays, there are different resources through which users can access information such as internet, telephone lines, mobile phones, etc. Short Messaging Service (SMS) is a service popularly used in mobile phones to provide information access to people on the move. This has resulted in the growth of SMS based Question Answering (QA) services. However automatically handling SMS questions poses significant challenges due to the intrinsic noise in SMS questions. In this paper we proposed a SMS based FAQ (Frequently asked question) retrieval system in which user send a query question in SMS language through their mobile phones as text message and system retrieves the ranked answers from FAQ dataset and results are then sent back to user as SMS text on user's mobile phone. We validated the effectiveness of our approach on a real-life dataset.

Keywords— Short Message Service (SMS), Frequently Asked Question (FAQ), Question Answering (QA), and Information Retrieval (IR)

I. INTRODUCTION

In this era of Information and Communications Technology (ICT) everyone is looking for a better and an easy way to access information. Different resources are also available to retrieve the information, in that mobile phone has become a fast and convenient way to communicate over a network. The growing use of information technologies such as mobile devices have had a major social and technological impact such as the growing use of SMS, a communication system broadly used by cellular phone users. The length limitation of an SMS has led to create a sort of "sub-language" which includes a vocabulary of words, similar to that of the original natural language, but that regularly omits grammatical forms, punctuation marks and vowels. This leads to various lexicographical errors which create problem in analysis of query and providing corresponding correct results. An SMS query contains certain noisy terms; most of them are due to the limitation of characters allowed in an SMS and typographical errors by the sender. Therefore, retrieving the relevant queries for user query from FAQ dataset is a very challenging task.

Question answering is novel field of research in which we try to retrieve answers for different questions instead of systems of information extraction like search engines which usually retrieve appropriate text documents. Most of question answering techniques are applied to retrieve appropriate answers for queries typed in native natural language of that region. Typically, there are two types of question answering systems: (1) closed-domain question answering that deals with questions under a specific domain, and can be seen as an easier task on one hand as the NLP systems can exploit domain-specific knowledge frequently formalized in ontology but harder on the other as the information is not generally available in the public domain; and (2) open domain question answering that deals with questions about nearly everything, and can rely only on general ontology and world knowledge. On the other hand, as mentioned earlier these systems usually have much more data available in the public domain from which to extract the answer.

FAQ databases are standard ways to present domain specific information in the form of expert answers to user's questions. Each FAQ consists of a question and an answer, probably added with supplementary Meta data (e.g., keywords). A FAQ retrieval engine provides an interface to a FAQ database. Given a user query in natural language as input, it retrieves a ranked list of FAQs relevant to the query. FAQs are an efficient way of communicating domain specific information to the users. Unlike general purpose retrieval engines, FAQ retrieval engines have to address the lexical gap between the query and the usually short answer. FAQ retrieval can be considered half way between traditional document retrieval and question answering. Unlike in full blown QA, in FAQ retrieval the questions and the answers are already extracted. On the other hand, unlike in document retrieval, FAQ queries are usually questions and the answers are typically much shorter than documents. While FAQ retrieval can be approached using simple keyword matching, the performance of such systems will be strictly limited due to the lexical gap, a lack of overlap between the words that appear in a query and words from a FAQ pair.

The SMS based Question Answering service can be one of the easiest way to provide information access to the mobile users on move. This popularity and ease of use encourages the information providers to provide the access to information

through mobile phones. SMS Based FAQ Retrieval System is a question answering system in which user ask question in SMS language to system and system matches this question with FAQ database.

The SMS based FAQ Retrieval System involved finding the best possible match from a given set of FAQs for a query written in texting language. Now the problem with texting language is that often there are misspellings, non-standard abbreviations, transliterations, phonetic substitutions and omissions making it difficult to build an accurate FAQ Retrieval System. In this paper we proposed SMS based FAQ retrieval system in which we focuses on generating or searching answers, in a FAQ database where the questions and answers are already provided by an expert. The goal is to identify the best matching question-answer pair for a given query. SMS query question given by user is first translated from SMS language into normal English language and then it is given to system as input query for finding ranked answers.

II. RELATED WORK

FAQ retrieval task consists of three tasks as (1) Monolingual FAQ Retrieval, (2) Cross lingual FAQ Retrieval, and (3) Multilingual FAQ Retrieval as shown in Fig. 1. In Monolingual sub-task SMS query and FAQ corpus are in same language and goal of the system is to find the best matching question Q* from a monolingual collection of FAQs Q. In cross lingual FAQ sub-task in which SMS query and FAQ corpus are in different languages and the system finds the best matching question Q* from a set of FAQs in language L2 while the SMS query is from different language L1. In multi-lingual FAQ sub-task SMS query can be from multiple languages and these can matches with FAQ collections from different languages. For example, SMS queries could be written in English, or Hindi, or Malayalam, and these queries could match FAQ collections of English or Hindi or Malayalam.

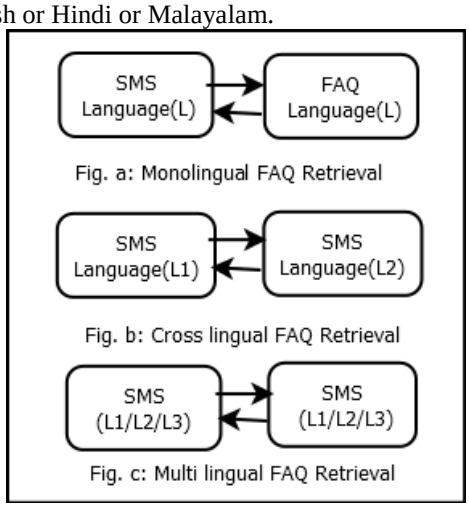

Fig. 1 FAQ Retrieval Tasks

Deirdre Hogan et al.[2] submitted system using monolingual FAQ retrieval in FIRE (Forum of Information Retrieval Evaluation) 2011, where they submitted their second system [3] using both Monolingual and cross lingual FAQ retrieval in FIRE 2012.

Kothari et al.[1] presented a FAQ-based question answering system over a SMS interface in which they handled the noise in a SMS query by framing the query similarity over FAQ questions as a combinatorial search problem, where the search space consists of grouping of all possible dictionary variations of tokens in the noisy query. They presented an efficient search algorithm that does not require any training data or SMS normalization and can handle semantic variations in question formulation. Their method is unsupervised and does not require aligned corpus or explicit SMS normalization to handle noise.

Deirdre Hogan et al.[2] from Dublin City University (DCU) has submitted a task on SMS-based FAQ Retrieval in the FIRE 2011 evaluation. The DCU team submitted three runs for the English monolingual task. The task consisted of retrieving the right answer to an incoming SMS question from an English FAQ, consisting of questions and answers on a mixture of different topics from career advice to popular Indian recipes. The in- coming queries were written in noisy SMS `text speak' and contained many misspellings, abbreviations and grammatical errors. Some SMS queries were out-of-domain and had no corresponding FAQ answer. Such queries needed to be identified and flagged as an out-of-domain (OOD) result before returning `NONE' as an answer string.

In FIRE 2012, Johannes Leveling et al.[3] from Dublin City University (DCU) has submitted a second task on SMS-based FAQ Retrieval. For FIRE 2012, they submitted runs for the monolingual English and Hindi and the cross lingual English to Hindi subtasks. Compared to our experiments for FIRE 2011, system implemented in FIRE 2012 was simplified by using a single retrieval engine (instead of three) and using a single approach for detecting out-of-domain queries (instead of three). In this approach, the SMS queries are first transformed into a normalized, corrected form. The normalized queries are submitted to a retrieval engine to obtain a ranked list of FAQ results. A classifier trained on features extracted from the training data then determines which queries are out-of-domain and which are not. For cross lingual English to Hindi experiments, they trained a statistical machine translation system for Hindi to English translation to translate the full Hindi FAQ documents into English. The retrieval then works on the corrected English input and retrieves results from the translated Hindi FAQ documents.

An automated question answering system was designed by authors of [7]. Mukul Rawat et al.[8] have presented "Improved version of SMS Based FAQ Retrieval System". Their work is extension of the system described in [7]. They have mainly added three improvements to the previous system. They are: 1. Proximity score, 2. Length score and, 3. An answer matching system. The experiments show that the accuracy of the system gives the accuracy of current state- of-the-art system. They demonstrate the effectiveness of our approach by considering many real-life FAQ datasets from different domains (e.g. Agriculture, Bank, Health, Insurance and Telecom etc.).

Partha Pakray et al.[9] presented a SMS-based FAQ retrieval system where the goal of the system is to find a question Q* from corpora of FAQs (Frequently Asked Questions) that best answers or matches the SMS query S. The test corpus contained FAQs in three languages: English, Hindi and Malayalam. The FAQs were from several domains, including railway enquiry, telecom, health and banking. They first checked the SMS using the Bing spell-checker. Then they used the unigram matching, bigram matching, and 1-skip bigram matching modules for monolingual FAQ retrieval. Freely available Google translator is used for translation because Google translator has better performance than the Bing translator.

Johannes Leveling [4] investigated the effects of stopword removal in different phases of a system for SMS-based FAQ retrieval. He did Experiments on the FIRE 2011 monolingual English data. The FAQ system includes several steps, comprising normalization and correction of SMS, retrieval of FAQs comprising answers using the BM25 retrieval model, and recognition of out-of-domain queries based on a k nearest-neighbor classifier. Together retrieval and OOD detection are tested with different stopword lists. Results show that 1) retrieval performance is highest when stopwords are not removed and decreases when longer stopword lists are employed, 2) OOD detection accuracy decreases when trained on features collected during retrieval using no stopwords, 3) A combination of retrieval using no stopwords and OOD detection trained using the SMART stopwords yields the best results: 75.1% in-domain queries are answered correctly and 85.6% OOD queries are detected correctly.

III. IMPLEMENTATION DETAILS

A. System Architecture

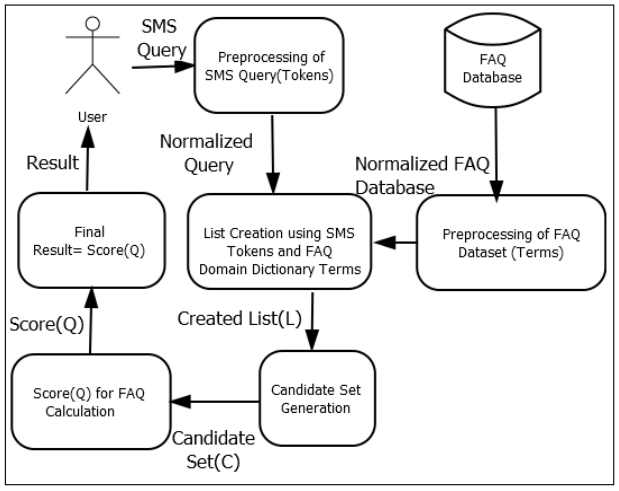

Fig. 2 Proposed System Architecture for SMS Based FAQ Retrieval System

Fig. 2 shows the proposed system architecture for SMS Based FAQ Retrieval System. In proposed system, user gives query in SMS language (SMS Query) which is then pre-processed into normal English language query. Pre-processing of SMS query contain 3 steps as (a) Training, (b) Learning and (c) Classification. Naïve Bayes Classifier is used for classification. For each token in SMS query class is predicted and clean SMS text is retrieved. System views the SMS as a sequence of tokens and each question in the FAQ corpus are viewed as a list of terms. The goal is to find a question from the FAQ corpus that matches best with the SMS query and return the answer of the selected question as a response of the input query.

Generally, SMS string is bound to have misspellings and other distortions, which needed to be taken care of while performing the match. There is a pre-processing stage in which the system develops a domain dictionary containing all the terms that are present in the FAQ corpus. For each term t in the dictionary and each token S_i in the SMS query, similarity measure is defined $\alpha(t, S_i)$ that measures how closely the term t matches the SMS token S_i. They said the term t was a variant of S_i, if $\alpha(t, S_i) > 0$.

Apache Lucene is used to indexing FAQ database into terms. Indexed FAQ terms are looked up with SMS query tokens using similarity measure and list of such terms is maintained. This step is referred as List Creation. FAQs containing the terms present in the list are retrieved; all such terms are called Candidate set. For each FAQ in Candidate set, score is calculated and highest scored FAQs are returned to user. Four Different classifiers (Naïve Bayes, Fuzzy Lattice Reasoning (FLR), IB1 and Random Forest) from Weka (Waikato Environment for Knowledge Analysis) Library are used in pre-processing steps. Naïve Bayes classifier gives best results in classification among all classifiers. We build training dataset which is used in pre-processing SMS text query for classification. Correct class is predicted for each SMS token and feature set is generated for training dataset.

B. Steps in SMS Based FAQ Retrieval System

1. Pre-processing on SMS Query: 1. Training, 2. Learning, 3. Classification
2. Formation of Domain dictionary and Synonyms dictionary
3. List Creation (L): For each token S_i in SMS Query Find Similarity measure with term t.
4. Candidate Set(C) Generation
5. For each Q_i in C find Similarity Score
6. Highest Score FAQs are returns to User.
7. End

Following Formula is used to calculate Similarity Score as

Similarity_Score (Q) = $\sum_{i=1}^{n} \max_{t \in Q \text{ and } t \sim s_i} \omega(t, s_i)$ (2)

Where

$$\omega(t, s_i) = \alpha(t, s_i) * idf(t) \quad (3)$$

Where

Idf (t) is inverse document frequency of term t.

$\omega(t, s_i)$ is weight function by combining the similarity measure and idf of the term t.

C. Experimental Setup and Dataset:

The standard configuration required to build the system is using Java framework (jdk 7.0) on windows platform. Java Netbeans IDE 7.2 is used as a developer tool. Apache Lucene search engine is used for indexing FAQ database into terms. Any standard machine is capable of running this application.

In proposed system, FIRE 2013 English FAQ training dataset is taken from Forum of Information Retrieval Evaluation (FIRE) TREC website [12] which is in XML format. Dataset is available in different languages such as English, Hindi, and Malayalam from which we used English dataset containing 9524 FAQs (Frequently Asked Questions) from 16 different Domains (Agriculture, Banking, Career, General Knowledge, Health, Insurance, Online railway reservation, Sports, Telecom, Tourism, Loan, Internal Devices, Personality Development, Recipes, Visa, Web etc.)

IV. RESULTS

A. Result Set

Different classifiers (Naïve Bayes, FLR, IB1 and Random Forest) are used for classification of SMS query. Compared to all other classifiers, Naïve Bayes classifier gives best results in classification. In Table I, Statistical analysis of training dataset using different classifiers are given. Naïve Bayes classifier gives best result by classifiying 52 correct instances out of 87. For implementation we have taken 87 instances in our training dataset for classification but we can add instances as many as we want in training dataset. Accuracy (as shown in Table II) and all statistic measures (as shown in Table I) varies based on training data which is given as input for classification.

TABLE I STATISTICAL ANALYSIS OF TRAINING DATASET USING DIFFERENT CLASSIFIERS

Classifier	Correctly classified instances (out of 87)	Incorrectly classified instances	Kappa	Mean Absolute Error	Root mean squared error	Relative absolute error	Root relative squared error
Naïve Bayes	52	35	0.5923	0.0209	0.1016	82.27%	89.85%
FLR	4	83	0.0188	0.0245	0.1564	96.14%	138.27%
IB1	5	82	0.0433	0.0242	0.1555	94.98%	137.44%
Random Forest	4	83	0.0254	0.246	0.1172	96.49%	103.62%

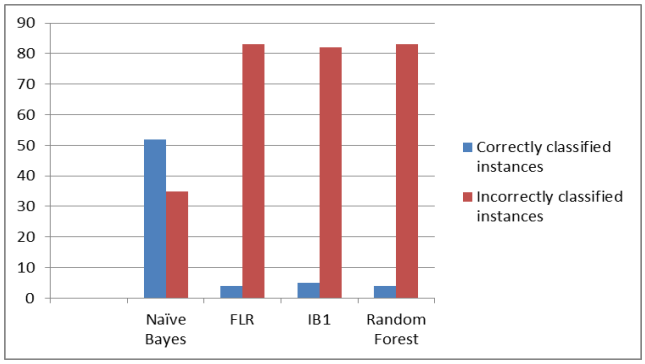

Fig. 3 Graph for Instance classification for different Classifiers

Correctly and incorrectly classified instances are then mapped into graph shown in fig. 3 for four different classifiers.

TABLE III ACCURACY OF CLASSES IN TERMS OF DIFFERENT STATISTICS

Sr. No.	Term Classes	TP rate	FP Rate	Precision	Recall	F – Measure	ROC Area
1.	Are	1	0.012	0.5	1	0.667	0.994
2.	What	1	0.012	0.5	1	0.667	0.994
3.	Who	1	0	1	1	1	1
4.	Your	0	0	0	0	0	0.994
5.	Best	1	0	1	1	1	1
6.	For	1	0.012	0.5	1	0.667	0.994
7.	World	1	0.012	0.5	1	0.667	0.994
8.	infected	1	0	1	1	1	1
9.	What_is	1	0.012	0.5	1	0.667	0.994
10.	Where	1	0.012	0.5	1	0.667	0.994
11.	developing	1	0.012	0.5	1	0.667	0.994
12.	relationship	0	0	0	0	0	0.994
13.	infection	0	0	0	0	0	0.994
14.	classifier	1	0	1	1	1	1
15.	begun	1	0.012	0.5	1	0.667	0.994
16.	Should	1	0	1	1	1	1
17.	People	1	0	1	1	1	1
18.	Become	1	0	1	1	1	1
19.	Between	1	0.012	0.5	1	0.667	0.994
20.	Control	1	0	1	1	1	1

In Table II accuracy of classes in terms of different statistics are given, which is then mapped into graph as shown in fig. 4. Each class is accurately classified using Nave Bayes Classifier. In Table II, accuracy of only 20 classes is given but we can add as many classes as we want in our file for classification.

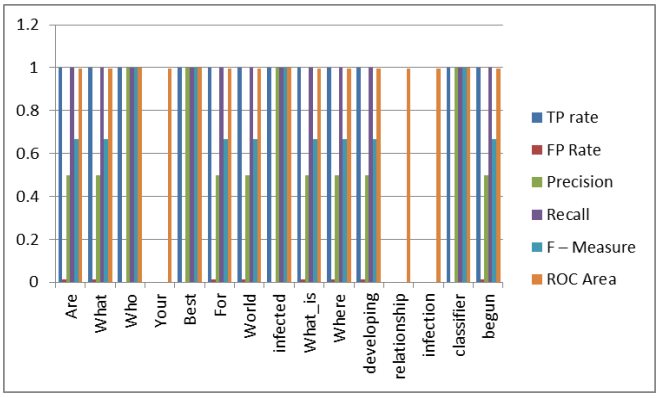

Fig. 4. Graph for Accuracy of Classes in terms of different Statistics

TABLE IIIII EXAMPLE OF RANKED RESULT GIVEN BY SYSTEM WITH HIT SCORE

User Query: wht s hiv?
Normalized Query: What is HIV?
Ranked Result from FAQ Dataset with Hit score is given below:
Q1: What is HIV? **A:** HIV or human immunodeficiency virus is a virus that is known to result in AIDS. This virus… **Score**: 4.32652
Q2: What is the test FOR AIDS AND HIV? **A:** The test is a test for the ANTIBODIES to HIV it is NOT a test for HIV and it is NOT a test for AIDS. If a person tests HIV positive… **Score**: 3.461216
Q3: What is the difference between AIDS and HIV? **A:** HIV is the virus that causes the disease AIDS. AIDS is…

Score: 3.028564
Q4: What work is Action Aid doing on HIV? **A:** action Aid advocates that people should have a life of dignity in… **Score:** 2.595912
Q5: What efforts are being made to integrate HIV/AIDS/STD prevention and control activities into primary health care? **A:** Integration into primary health care is a priority because it is… Score: 1.730608

Table III has given an example of how system works when user gives a query. When user gives a query e.g. "wht s hiv?" as shown in above table I, it is pre-processed and normalized into normal query. And that query terms are then matched with the FAQ dataset and 5 top Ranked results are retrieved. Hit score of each answer is calculated which is shown as in Table III and mapped into graph in fig. 5 shown below. Hit score is nothing but how many time that particular answer hits for user query question. It varies based on how many time answer hits for user query.

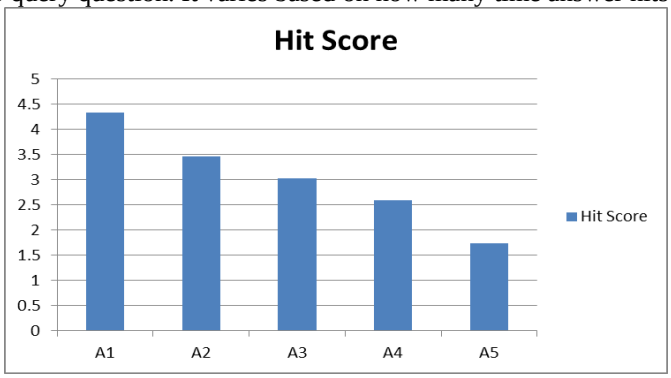

Fig. 5 Graph for Ranked Answers over Hit Score

In fig. 5 graph of ranked answers over hit score is plotted. Higher ranked answer is having the high hit score which is then decreases with respect to rank.

V. CONCLUSIONS

SMS is really popular short message-based communication service among mobile phone users. In this paper, SMS based FAQ retrieval system is presented in which user asks query questions through users mobile phone in SMS texting language and ranked answers for the user query are then retrieved from FAQ dataset and given back to user as text message to their mobile phones. As SMS noise is a challenging problem in SMS language, classification technique is used in pre-processing of SMS query text which gives good matched result to the user. If system is used for particular domain it will give more accurateness in result. As a future work, calculating proximity score at run time using user search history will improve efficiency for the system.

ACKNOWLEDGMENT

We would like to thank the publishers, researchers and teachers for their guidance. We would also thank the college authority for providing the required infrastructure and support. Last but not the least we would like to extend a heartfelt gratitude to friends and family members for their support.

REFERENCES

[1] Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarthy, L. Venkata Subramaniam *"SMS based interface for FAQ retrieval."* Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009.

[2] Deirdre Hogan, Johannes Leveling, Hongyi Wang, Paul Ferguson, and Cathal Gurrin *"DCU@ FIRE 2011: SMS-based FAQ retrieval."* 3rd Workshop of the Forum for Information Retrieval Evaluation, FIRE. 2011.

[3] Johannes Leveling *"DCU@ FIRE 2012: Monolingual and Cross lingual SMS-based FAQ Retrieval."* (2012): 37-39.

[4] Johannes Leveling "On the effect of stop word removal for SMS-Based FAQ retrieval." Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2012. 128-139.

[5] Danish Contractor, Govind Kothari, Tanveer A. Faruquie, L. Venkata Subramaniam, Sumit Negi *"Handling noisy queries in cross language faq retrieval."* Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.

[6] Thuma, Edwin, Simon Rogers, and Iadh Ounis. *"Detecting Missing Content Queries in an SMS-Based HIV/AIDS FAQ Retrieval System."* Advances in Information Retrieval. Springer International Publishing, 2014. 247-259.

[7] Sneiders, Eriks. "Automated question answering using question templates that cover the conceptual model of the database." Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2002. 235-239.

[8] Anwar D. Shaikh, Mukul Jain, Mukul Rawat, Rajiv Ratn Shah, and Manoj Kumar *"Improving Accuracy of SMS Based FAQ Retrieval System."* Multilingual Information Access in South Asian Languages. Springer Berlin Heidelberg, 2013. 142-156.

[9] Partha Pakray, Santanu Pal, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander Gelbukh *"Smsfr: Sms-based faq retrieval system."* Advances in Computational Intelligence. Springer Berlin Heidelberg, 2013. 36-45.

[10] Pinto, David, Darnes Vilarino, and Yuridiana Alem. *"The Soundex Phonetic Algorithm Revisited for SMS-based Information Retrieval."* Department of computer science, Mexico.

[11] Contractor, Danish, Tanveer A. Faruquie, and L. Venkata Subramanian. *"Unsupervised cleansing of noisy text."* Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.

[12] Danish Contractor, L. Venkata Subramanian, P. Deepak, and Ankush Mittal. Text retrieval using SMS queries: Datasets and overview of FIRE 2011 track on sms-based FAQ retrieval. In Multilingual Information Access in South Asian Languages- Second International Workshop, FIRE 2010, Gandhinagar, India, February 19-21, 2010 and Third International Workshop, FIRE 2011, Bombay, India, December 2-4, 2011, Revised Selected Papers, pages 86-99, 2011.

[13] Hogan, Deirdre, et al. *"SMS Normalization, Retrieval and Out-of-Domain Detection Approaches for SMS-Based FAQ Retrieval."* Multilingual Information Access in South Asian Languages. Springer Berlin Heidelberg, 2013. 184-196.

[14] Mogadala, Aditya, Rambhoopal Kothwal, and Vasudeva Varma. *"Language modeling approach to retrieval for SMS and FAQ matching."* Multilingual Information Access in South Asian Languages. Springer Berlin Heidelberg, 2013. 119-130.

[15] Pinto, David, Darnes Vilarino, and Yuridiana Alem. "*The Soundex Phonetic Algorithm Revisited for SMS-based Information Retrieval.*" Department of computer science, Mexico.

[16] Chung-Hsien Wu, Jui-Feng Yeh, and Ming-Jun Chen. *Domain-specific faq retrieval using independent aspects.* ACM Transactions on Asian Language Information Processing (TALIP), 4(1):1-17, 2005.

[17] MUKUL RAWAT. *Improving SMS Based FAQ Retrieval Using Proximity and Length Score*. PhD thesis, DELHI TECHNOLOGICAL UNIVERSITY, 2012.

[18] Jotsna Waghmare and Mrs MA Potey. "Survey of SMS Based FAQ Retrieval Systems.", International Journal of Engineering and Computer Science, 2319-7242 Volume 4 Issue 2 February 2015, Page No. 10259-10263.

[19] Valentin Jijkoun and Maarten de Rijke. *Retrieving answers from frequently asked questions pages on the web*. In Proceedings of the 14th ACM international conference on Information and knowledge management, pages 76-83. ACM, 2005.

[20] http://www.isical.ac.in/~fire/faq-retrieval/faq-retrieval.html

