SMS based FAQ Retrieval using Theme Matching Scheme

Deba P. Mandal and Saptaditya Maiti
Machine Intelligence Unit
Indian Statistical Institute, Kolkata, India
Email: {dpmandal,saptaditya}@isical.ac.in

***Abstract*—As a participant of FIRE 2012 monolingual English SMS based FAQ Retrieval Task, we proposed a theme matching scheme [1]. Once again the scheme is implemented for the same task in FIRE 2013 having different set of SMS and FAQ queries. An SMS text usually consists of certain noisy terms due to the limitations of characters allowed in an SMS, lack of screen space and unintended typographical errors; thereby retrieving the relevant FAQ query/queries corresponding to an SMS query is a very challenging task. We observed that depending on the parts of speech, different terms in a text carry different significance values in representing the theme of the text. The said scheme is mostly based on this observation. In order to recognize the noisy terms of an SMS, three approximate matching measures were derived in the retrieval procedure. Although the FAQ and test queries in FIRE 2013 consist of not only SMS text queries but also speech transcribed queries, our theme matching scheme is found to be very effective for the first category $i.e.$, SMS queries. The relevance set published for the speech transcribed query set is completely erroneous as there is absolutely no match between the test queries and the relevant FAQ queries. Accordingly, our scheme recognized the complete speech transcribed set as out of domain.**

***Index Terms*—SMS, FAQ retrieval, POS tagger, Theme matching, Approximate string matching**

I. Introduction

Frequently Asked Questions (FAQ) is a useful source of information about an organization and so, many organizations are now using the FAQ for providing answers to the user queries. Here a user sends a query over the SMS and gets the answer selected from FAQ, through SMS itself. However, automatic retrieval of a correct answer corresponding to an SMS query is a challenging task due to the inherent noise in the query. Because of the limitation of characters allowed in an SMS and the lack of screen space, the users tend to incorporate certain modifications in the texts. It is very common to use abbreviations and modified spellings (generally shortened) of terms in SMS and the texts are also ill formed grammatically. There are also some unintended typographical errors found in SMS. Therefore, before retrieving relevant queries (and subsequently answers) from FAQ against an SMS query the said modifications/errors are required to be corrected or handled properly.

As a part of FIRE 2012 monolingual English SMS based FAQ Retrieval Task, we proposed a theme matching scheme [1]. In that scheme, the theme of each of the FAQ queries is initially determined and then the extent (degree) of the matching of a pre-processed SMS query with a FAQ query is checked.

In case of theme matching is satisfactory, the remaining portions of FAQ query are matched with the unmatched portions of the SMS query. For the word/string matches, four different string similarity measures are applied sequentially. They are (*i*) complete match, (*ii*) partial match (cashless=cash less), (*iii*) soundex match and (*iv*) approximate match. Obviously, matching score differs with matching type ($i.e.$, similarity measure). A length normalization factor is adopted based on the lengths of the SMS and FAQ queries. All the FAQ queries having relevance scores above a predefined threshold value are decided as relevant to the SMS query. No relevant query will be returned for SMS query if either the FAQ query themes do not have satisfactory theme match or the final relevance scores of all the FAQ queries are below the threshold level.

In the FIRE 2013 monolingual English FAQ Retrieval Task, the FAQ and test queries consist of speech transcribed queries (generated using a speech transcription system) in addition to SMS queries. One can easily realize that the characteristics of noises/ambiguities in the SMS queries are quite different than that of in the speech transcribed queries. As mentioned earlier, noises in SMS text are due to use of abbreviations and modified spelling of terms and typographical errors and so, the terms may not be valid English words. In contrary, terms in speech transcribed queries are all valid English words but may not be appropriate ones due to the variations in pronunciations of the speakers. We therefore feel that the above two types of queries should not be put as a single task. It seems, there is a serious mistake in the relevance set determination procedure for the speech transcribed queries as there is absolutely no similarity between each of the test speech queries to its corresponding relevant FAQ query. The performance of the theme matching scheme on the SMS test queries of FIRE 2013 is found to be quite satisfactory as observed in FIRE 2012.

The article is organized as follows. In section II, a brief description of the theme matching scheme is presented. The implementation and the results of the method is described in section III. Finally section IV finds the conclusions.

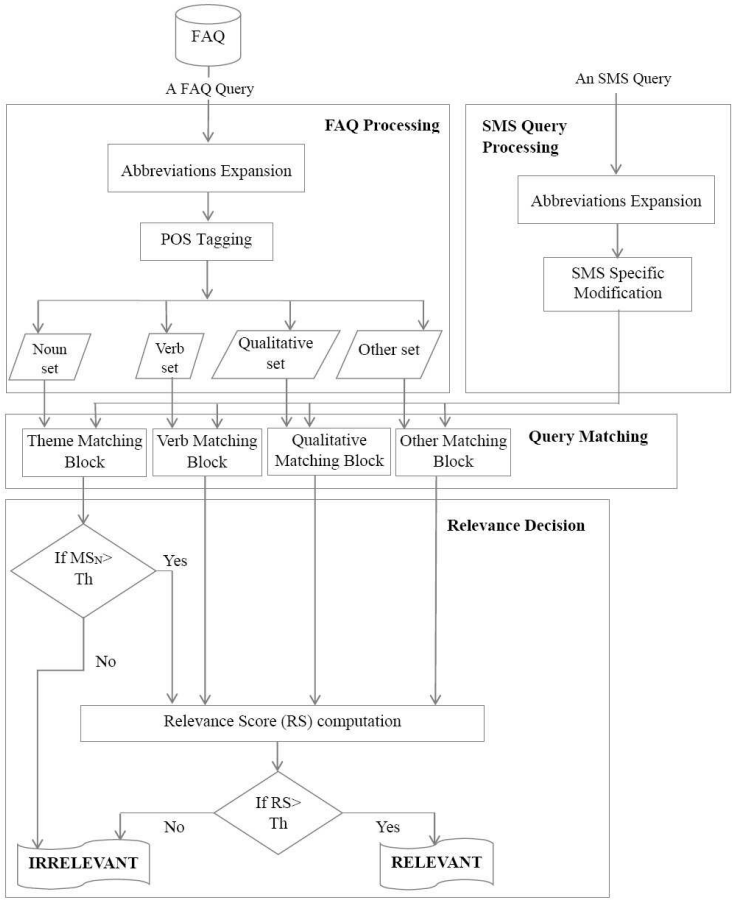

Fig. 1. Block diagram of the theme matching scheme [1]

II. Theme Matching Scheme

As a part of FIRE 2012 monolingual English SMS based FAQ Retrieval Task, we proposed a theme matching scheme [1]. A brief description of the scheme is provided here.

Fig. 1 shows the block diagram of the matching scheme between an SMS and a FAQ query. It consists of four sections namely, (*i*) FAQ Processing, (*ii*) Query Processing, (*iii*) Query Matching and (*iv*) Relevance Decision.

The FAQ Processing section is composed on Abbreviation Expansion and POS Tagging blocks. Expanded forms of some common abbreviations, if any, of the FAQ queries are expanded. Then POS tagging block determines the parts of speech of the terms of a query based on a POS tagger [2], [3]. Accordingly, the terms are decomposed into four sets of terms, namely, Noun, Verb, Qualitative term and Other term sets. That is, the FAQ processing section finds the said four term sets corresponding to each of the FAQ queries.

The SMS query processing section takes an SMS query as input. The function of the Abbreviation Expansion block is same as in the FAQ Processing section. That is, the section finds a modified query for the given SMS query. In addition, an SMS specific modification scheme is applied here.

The matching section consists of four sequential matching blocks, viz, theme, verb, qualitative and other term matching blocks which find the matching scores of the modified SMS query with the noun, verb, qualitative and other term sets of the FAQ query respectively.

The Relevance Decision block decides whether the FAQ query is relevant or not. Firstly, it checks the matching score (MS_N) of the noun/theme matching block and the Relevance Score Computation block is activated only if the said matching score is above a threshold value (Th). The matching scores of the four blocks of the Query Matching section are then averaged in the Relevance Score Computation block. The FAQ query is taken as relevant to the SMS query if the relevance score (RS) is above the threshold value (Th). That is, the FAQ query is considered to be relevant when both MS_N and RS are more than Th. If all the FAQ queries are found to be irrelevant, the relevance of the SMS query will also be null.

III. Implementation and Result

The theme matching scheme [1] is applied on the monolingual English task of FIRE 2013 SMS based FAQ Retrieval. In this section, the dataset, implementation of the theme matching scheme and experimental results are shown.

Dataset: The task consists of 7251 FAQ queries representing certain domains including Railway Enquiry, Telecom,

Health, Banking, Career counselling etc. There are 9155 training queries and 540 test queries. In the training set, 8482 are SMS queries and 673 are speech transcribed queries; and in the test set, 299 are SMS and 241 are speech queries. Among the 299 test queries, 200 are In-domain ($i.e.$, relevant query/queries belong(s) to the FAQ) and 99 are Out-of-domain ($i.e.$, no FAQ query is relevant). Among the 241 speech queries, 192 and 49 are designated as In-domain and Out-of-domain queries respectively. It is asked to provide a maximum of 5 relevant FAQ queries (in order of preferences) for each of text queries.

A relevant query set, one for each of the In-domain test queries is also provided later $i.e.$, after the deadline of submission of run.

Our comments on the Speech transcribed queries: As we are informed, speech query texts were generated using a speech transcription system and simultaneously the relevant sets were also identified/selected. We would like to mention here that we observed absolutely no similarity/match between the test speech query set and the corresponding relevance FAQ query set $i.e.$, the adopted relevant set determination procedure is erroneous. This observation is also true for the training set as well.

Implementation: The theme matching scheme [1] has seven constants, one threshold value, Th [used in relevance decision block]; four significance factors I_N, I_V, I_Q and I_O [used in relevance decision block]; and two matching constants v_{pm} and v_{sm} [in query matching block]. For the present implementation of the system, we have taken the values of the constants as $Th = 0.3$; $I_N = 1$, $I_V = 0.8$, $I_Q = 0.5$, $I_O = 0$; $v_{pm} = 0.5$ and $v_{sm} = 0.8$.

Result: The result reported by the taskmaster is provided in Table I and we find the result quite encouraging which support the notion of theme match we applied in the proposed method. As mentioned earlier, the speech transcribed queries are very erroneous and all of these queries returned NULL as relevant in the FAQ. Hence, the results of the speech queries are not shown here.

TABLE I
EXPERIMENTAL RESULTS ON SMS QUERIES

Queries	In-Domain	Out of Domain	Total
No of queries	200	99	299
Correct	187 (0.93500)	97 (0.97980)	284 (0.94983)
MRR	–	–	0.97026

1) Analysis of result: As per the report of the taskmaster, the theme matching scheme made mistakes for fifteen text queries. The said text queries along with the relevant and retrieved (by our method) FAQ queries are furnished in Table II. The relevance scores of our retrieved queries are also shown in the table. The first thirteen queries are in-domain queries and the last two are out of domain queries. The first one was identified correctly (with relevance score 0.66301) by our method but it was wrongly reported as NULL while converting our result to the required format. There is a pair of relevant queries having identical text for each of the next six test queries (serial number $2-7$) and our scheme identified both of them (in alphabetic order) with the same relevance score. As the task management system considered only one of instead of the both in the pair our result becomes erroneous wrongly. Although, our system could identify the relevant queries corresponding to the next four test queries (serial number $8-11$), they were discarded from our final result having low relevance scores (< 0.3). On analysis, we found that the low scores are due to the adopted length normalization factor which actually performed well to restrict the out of domain queries to provide relevant FAQ queries. For the next two test queries (serial number $12-13$), our system identified the relevant queries in the second position from the top. One can observe certain similarity between our retrieved FAQ queries and the two out of domain queries (serial numbers 14 and 15).

Based on the aforestated analysis, we would like to claim that the output of the theme matching scheme is quite logical and appropriate for all the SMS based queries of FIRE 2013 FAQ retrieval using noisy queries task.

IV. CONCLUSIONS

An SMS query contains certain noisy terms, most of them are due to the limitation of characters allowed in an SMS, lack of screen space and unintended typographical errors by the sender. Therefore, retrieving the relevant queries from FAQ corresponding to the SMS query is a very challenging task.

A theme matching scheme for SMS based FAQ retrieval was proposed in FIRE 2012 [1]. It is again implemented for the same task of FIRE 2013. This scheme is based on matching of theme the FAQ queries with an SMS query. At first this scheme matched the theme of the FAQ query with the SMS query and then matched the rest of the query. Four string matching measures were adopted in the matching scheme. The method is implemented in monolingual English task in FIRE 2013 SMS-based FAQ Retrieval and it has performed significantly to retrieve the relevant queries from FAQ corresponding to an SMS query.

It is to be mentioned here that the given dataset consists of noisy SMS queries along with noisy speech transcribed queries. We found that the speech transcribed queries are too erroneous. Hence the proposed scheme is implemented only for the noisy SMS texts and the speech transcribed queries could not be verified in the present system.

REFERENCES

[1] Working Notes of FIRE 2012

[2] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[3] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

TABLE II
ANALYSIS OF EXPERIMENTAL RESULTS

Sl. No	Test Query Id with Text	Relevant FAQ Id with text	Retrieved FAQ Id with text	Relevance score
1	ENGLISH_NR_422: from whre i get admission in mechnical enginering dploma?	ENG_CAREER_669: from where i get admission in mechanical engineering diploma?	NULL	0.66301
2	ENGLISH_NR_442: Will IRCTC moble wrk on CDMA mble conections like TTSL, TTML, Reliance, MTNL Garuda, BSNL WLL?	ENG_INDIAN_RAILWAYS_335: Will IRCTC mobile work on CDMA mobile connections like TTSL, TTML, Reliance, MTNL Garuda, BSNL WLL?	ENG_INDIAN_RAILWAYS_61: Will IRCTC mobile work on CDMA mobile connections like TTSL, TTML, Reliance, MTNL Garuda, BSNL WLL?	0.75189
3	ENGLISH_NR_462: How shld customers go about geting these services from the prvt sector?	ENG_SPORTS_442: How should customers go about getting these services from the private sector?	ENG_AGRICULTURE_217: How should customers go about getting these services from the private sector?	0.8625
4	ENGLISH_NR_487: organisations authorised to collect LIC premiums through their net banking/phone banking facility?	ENG_INSURANCE_464: Which organisations are authorised to collect LIC premiums through their net banking/phone banking facility?	ENG_INSURANCE_1012: Which organisations are authorised to collect LIC premiums through their net banking/phone banking facility?	0.51664
5	ENGLISH_NR_494: It looks as if you are in fact degrading NWS services, in violation of the Cramer Report. Please explain?	ENG_SPORTS_468: Last year response to the Cramer Report determined that there would be no degrading of NWS services along the Eastern slopes of the Cascade Mountains in both Washington and Oregon and in the Sacramento Valley. This looks as if you are in fact degrading NWS services, in violation of the Cramer Report. Please explain?	ENG_AGRICULTURE_226: Last year response to the Cramer Report determined that there would be no degrading of NWS services along the Eastern slopes of the Cascade Mountains in both Washington and Oregon and in the Sacramento Valley. This looks as if you are in fact degrading NWS services, in violation of the Cramer Report. Please explain?	0.38485
6	ENGLISH_NR_558: Where 2 find info about registering a pesticide?	ENG_AGRICULTURE_7: Where can I find information about registering a pesticide?	ENG_AGRICULTURE_195: Where can I find information about registering a pesticide?	0.46089
7	ENGLISH_NR_575: benefits of paying your LIC premiums through net banking/phone banking?	ENG_INSURANCE_461: What are the benefits of paying your LIC premiums through net banking/phone banking?	ENG_INSURANCE_1009: What are the benefits of paying your LIC premiums through net banking/phone banking?	0.60810
8	ENGLISH_NR_508: Services of ILL cost?	ENG_LOAN_190: How much do you charge for your services of ILL?	NULL	0.2
9	ENGLISH_NR_555: time it takes to get a visa to India?	ENG_VISA_134: How long does it take to get a visa to India?	NULL	0.22063
10	ENGLISH_NR_567: Haplogroup?	ENG_HEALTH_178: What is a haplogroup?	NULL	0.25
11	ENGLISH_NR_597: After diaphragm spasms?	ENG_SPORTS_151: What's most likely to occur when your diaphragm goes into spasms?	NULL	0.16667
12	ENGLISH_NR_510: Final weight of pure gold confirmed?	ENG_BANK_369 : How can I be sure about the final weight of pure gold?	ENG_GK_686: How many carats is pure gold?	0.35476 / 0.41425
13	ENGLISH_NR_544: IS majors jobs?	ENG_CAREER_205: Where do IS majors look for jobs?	ENG_CAREER_204: What jobs do IS majors get?	0.42857 / 0.5
14	ENGLISH_QUERY_2013_30236: How can I check on the status o my application?	NULL	ENG_BANK_644: How can I enquire about the status of my application?	0.35714
15	ENGLISH_QUERY_2013_30256: How many factory farms are dere in the uk?	NULL	ENG_AGRICULTURE_95: How many farmers are there in the India?	0.33333