SMSFR: SMS-Based FAQ Retrieval System

Partha Pakray,[1] Santanu Pal,[1] Soujanya Poria,[1]
Sivaji Bandyopadhyay,[1] Alexander Gelbukh[2]

[1] Computer Science and Engineering Department,
Jadavpur University, Kolkata, India
[2] Center for Computing Research,
National Polytechnic Institute, Mexico City, Mexico
parthapakray@gmail.com, santanupersonal1@gmail.com, soujanya.poria@gmail.com,
sbandyopadhyay@cse.jdvu.ac.in, www.gelbukh.com

Abstract. The paper describes an SMS-based FAQ retrieval system. The goal of this task 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 used in this paper contained FAQs in three languages: English, Hindi and Malayalam. The FAQs were from several domains, including railway enquiry, telecom, health and banking. We first checked the SMS using the Bing spell-checker. Then we used the unigram matching, bigram matching, and 1-skip bigram matching modules for monolingual FAQ retrieval. For cross-lingual system, we used the following three modules: an SMS-to-English query translation system, an English-to-Hindi translation system, and cross-lingual FAQ retrieval.

Keywords: domain-specific information retrieval, *n*-grams, spell checker, Bing translator.

1 Introduction

In spite of great advances of information retrieval systems and associated natural language processing technologies [1], domain-specific retrieval systems and retrieval systems used with special types of queries continue to represent a challenge for current technology and to be a topic of active research.

The Forum for Information Retrieval Evaluation (FIRE)[1] [2] is a forum for Information Retrieval evaluation that is traditionally mainly focused on Indian languages. After the success of FIRE 2008 and FIRE 2010, the main goal of the FIRE 2011 was to support the continuity of research on Information Retrieval, with the stress on non-English languages, specifically those used on the Indian sub-continent.

India, with its huge population, has a very high rate of using mobile phones, with service costs low enough for even very poor people to use the phones actively. Accordingly, FIRE 2011 included an SMS-based FAQ retrieval [3] task. The goal of this task was to find a question Q* from corpora of FAQs (Frequently Asked Questions) that best answers or matches a given SMS query S.

[1] http://www.isical.ac.in/~clia/

SMS queries, which are written in “SMS language” specific for this kind of communication and different from the usual grammatically correct text, tend to be noisy, because the users try and compress text by omitting letters, by using slang, etc., due to a restriction on the length of such messages (at most 160 characters are allowed for one SMS), the lack of screen space, which makes reading large amounts of text difficult, and other similar reasons.

The messages also frequently contain unintended typographical errors due to small size of keypads on mobile phones as also poor language skills of the users (which, in turn, are caused both by the availability of mobile services to the poorest strata of the population with low literacy and the fact that major languages used in India are not native languages for a large share of the population). The presence of such noise makes this task different and more challenging than traditional Question Answering (QA) retrieval tasks.

The SMS retrieval task consists of three sub-tasks.

Task 1: **Monolingual FAQ Retrieval:** In this sub-task, the SMS query and the FAQ corpus is in the same language. The goal in this task is to find the best matching question Q* from a monolingual collection of FAQs Q that matches an SMS query expressed in the same language as the FAQs. This task is most similar to the classical information retrieval task.

Task 2: **Cross-lingual FAQ Retrieval:** In this sub-task, the SMS query and the FAQ corpus are in different languages. That is, if the SMS query is in Language L1 (for example, Malayalam), then the FAQ corpus will be in a language L2 other than L1, for example, in English. Thus, the goal in this task is to find the best matching question Q* from the set of FAQs in language L2 while the SMS query is in a different language L1.

Task 3: Multi-lingual FAQ Retrieval: In this sub-task the SMS queries can be in multiple languages and these can match to FAQ collections in multiple languages. For example, SMS queries could be written in English, or in Hindi, or in Malayalam, and these queries could match FAQ collections of FAQs in all languages that participate in the task. That is, the goal is to find a pertinent FAQ item that can be in English, Hindi, or Malayalam.

In this paper we report the systems developed for the monolingual task, Task 1, and for the cross-lingual task, Task 2, while developing a system for Task 3 is the topic of our future work.

The paper is organized as follows. Section 2 briefly presents the datasets used in the work, along with examples and some statistics. Section 3 presents the block diagram of the architecture of our system developed for the monolingual task, Task 1. Section 4 presents the block diagram of the architecture of our system developed for the cross-lingual task, Task 2. Section 5 describes the experimental results obtained with these two systems. Finally, Section 6 concludes the paper and discusses the directions of our future work.

Table 1. Dataset statistics for the SMS-based FAQ retrieval task.

FAQ	Language		Number of SMS queries in each task (in-domain / out-of-domain)		
			Monolingual task	Cross-lingual task	Multilingual task
7251	English	Training	701 / 370	291 /181	290 / 170
		Test	728 /2677	37 / 3368	724 / 2681
1994	Hindi	Training	181 / 49		183 / 47
		Test	200 / 124		200 / 124
681	Malayalam	Training	120 / 20		60 / 20
		Test	50 / 0		50 / 0

2 Datasets Used for SMS based FAQ Retrieval

In the "SMS based FAQ Retrieval" task[2] of FIRE 2011 the datasets used for training and testing of the systems were released. The training FAQ dataset and the training SMS datasets were released in three languages: English, Hindi, and Malayalam. Some statistical data on these datasets are shown in Table 1. In this table, the two figures separated by a slash stand for the number of in-domain items and out-of-domain items in each dataset.

A sample of XML formatted data for the training dataset is shown in Table 2.

3 System Architecture for Task 1: Monolingual

The architecture of the system for the monolingual SMS FAQ retrieval, Task 1, is shown as a block diagram in Figure 1. Our solution for this task is a rule-based system for ranking the candidate FAQ items. Various components of the system are: a pre-processing module, a unigram matching module, a bigram matching module, and the 1-skip bigram matching module.

In the following subsections we give a brief description of the modules used in the system.

3.1 Spellchecking

The SMS and FAQ statements both have various spelling errors. So, to achieve optimized and improved matched between them, both SMS and FAQ statements were passed through the spellchecker module. We used the Bing spellchecker for this pur-

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

Table 2. Sample XML-formatted training data.

FAQ

```
<FAQ>
<FAQID>ENG_CAREER_33</FAQID>
<DOMAIN>ENG_CAREER</DOMAIN>
<QUESTION>What is an effective resume?</QUESTION>
<ANSWER> An effective resume is one which makes your
phone ring or your email blink.
</ANSWER>
</FAQ>
```

SMS

```
<SMS>
<SMS_QUERY_ID>ENG_5</SMS_QUERY_ID>
<SMS_TEXT>are the carier conselling sessionss
confidensial</SMS_TEXT>
<MATCHES>
<ENGLISH>ENG_CAREER_43</ENGLISH>
<MALAYALAM>NONE</MALAYALAM>
<HINDI>NONE</HINDI>
</MATCHES>
</SMS>
```

pose, because it is free of charge and has acceptable quality as compared with other available options.

As Fig. 1 shows, we use the same Bing spellchecker module to process both the SMS dataset and the FAQ dataset, which guarantees that we correctly math coinciding words that are considered ungrammatical and changed by the spellchecker module: while both word are changed by the spellchecker module, they are changed in the same way and thus match normally.

Suppose SMS query is S and FAQ query is F. After spell-checking we obtain the changed form of the query S, which we denote S′, and the changed form of the FAQ statement F, which we denote F′.

3.2 Unigram Matching

After we have obtained the spell-checked statements S′ and F′, we fed them into the unigram matching module.

Suppose the text of the SMS statement S′ contains a word pattern that can be expressed as $<L_1, L_2, ..., L_n>$, and the text of the FAQ statement F′ contains the word pattern that can be expressed as $<W_1, W_2, ..., W_k>$. Then we searched for matching of each word of S′ in F′.

If a direct match occurred, then we do not search any further for the word L_i and pick up the next word from the list and search again for that word. We consider that a direct match occurred if there is some FAQ word W_i such that $W_i = L_j$ for some $i \le k$ and $j \le n$.

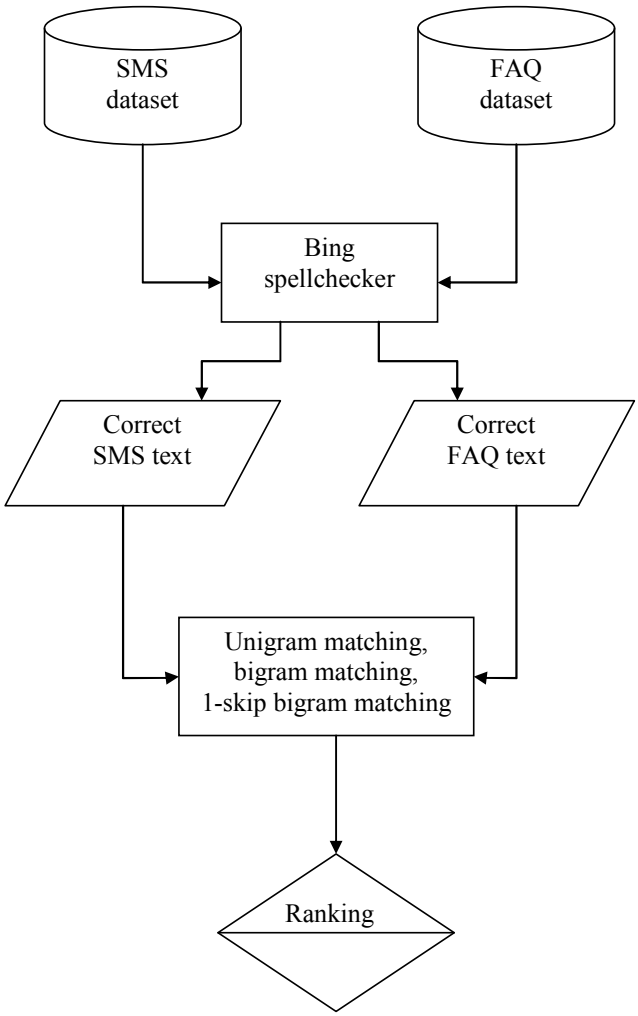

Fig. 1. Monolingual FAQ retrieval system architecture.

Now, if there is no direct match, then we looked up the word in WordNet 3.0 [4] and obtained its hyponyms, synonyms, etc., and searched each one of these words in the F′ list. If a match was found, then we passed on to next word.

Otherwise, we searched for an overlap between the synonym and hyponym list of the word L_j and the synonym and hyponym list of the word W_i. If in this case any matching words can be found, then we went on to the next word in the S′ list and store that word as a matched word. Otherwise we skipped the word and proceeded for next word.

The expression used as a score for the unigram matching module was:

$$\text{Unigram Score} = \frac{\text{Number of SMS word that match FAQ words}}{\text{Total number of words in SMS}}.$$

3.3 Bigram Matching

In this module, we aimed to find a match between two statements by considering the bigram occurrences of their words. We took the two consecutive words from the S′ list, represented as <L_i, L_{i+1}>, and similarly from the F′ list, <W_j, W_{j+1}>. If a match was found, then we went on to the next consecutive bigram.

Otherwise, we looked up in WordNet all hyponyms and synonyms for the words of the bigram <L_i, L_{i+1}>, and similarly we retrieved all hyponyms and synonyms for the words of the bigram <W_j, W_{j+1}>. Suppose we have a synonym list for L_i as (x_1, x_2, ..., x_n) and for L_{i+1} a list (y_1, y_2, …, y_k), and similarly for the pair <W_j, W_{j+1}>, obtaining lists (s_1, s_2, ..., s_n) and (t_1, t_2, …, t_k), correspondingly. If there was any matching words between the lists for the SMS bigram <L_i, L_{i+1}> and the list for the FAQ bigram <W_j, W_{j+1}>, then we considered that a match was found, and proceed to analyze the next bigram sequence.

The expression used as a score for the unigram matching module was as follows:

$$\text{Bigram Score} = \frac{\text{Number of bigram matching}}{\text{Total number of bigrams in SMS}}.$$

3.4 1-skip and Inverse Bigram Matching

For each pair of words <s_i, t_j> in the list S′ that is found in the inverse order <t_i, s_j> in F′, we applied various semantic rules, because such pairs can not be just rejected. The complete set of the rules is not given here. For example, if one of the two words in the pair was a verb, then we ignored the difference and considered the bigram sequence as a match.

If there was a negation word in one of the two texts between two given words, but the two words formed a bigram in the other text, then we just removed the negation word and, again, considered the bigram sequence as a match. That is, under some circumstances we considered a sequence of two words with one gap as a bigram.

3.3 Final Ranking

Finally, the overall scoring was calculated as the harmonic mean of the two particular scores:

$$\text{Total Score} = \frac{\text{Unigram Score} \times \text{Bigram Score}}{\text{Unigram Score} + \text{Bigram Score}}.$$

We presented to the user as the output the top five scores for a single SMS query.

4 System Architecture for Task 2: Cross-Lingual

The architecture of our system for Task 2 (cross-lingual task) is presented in Figure 2. Our system for this task has three major modules. The objective of the first module is

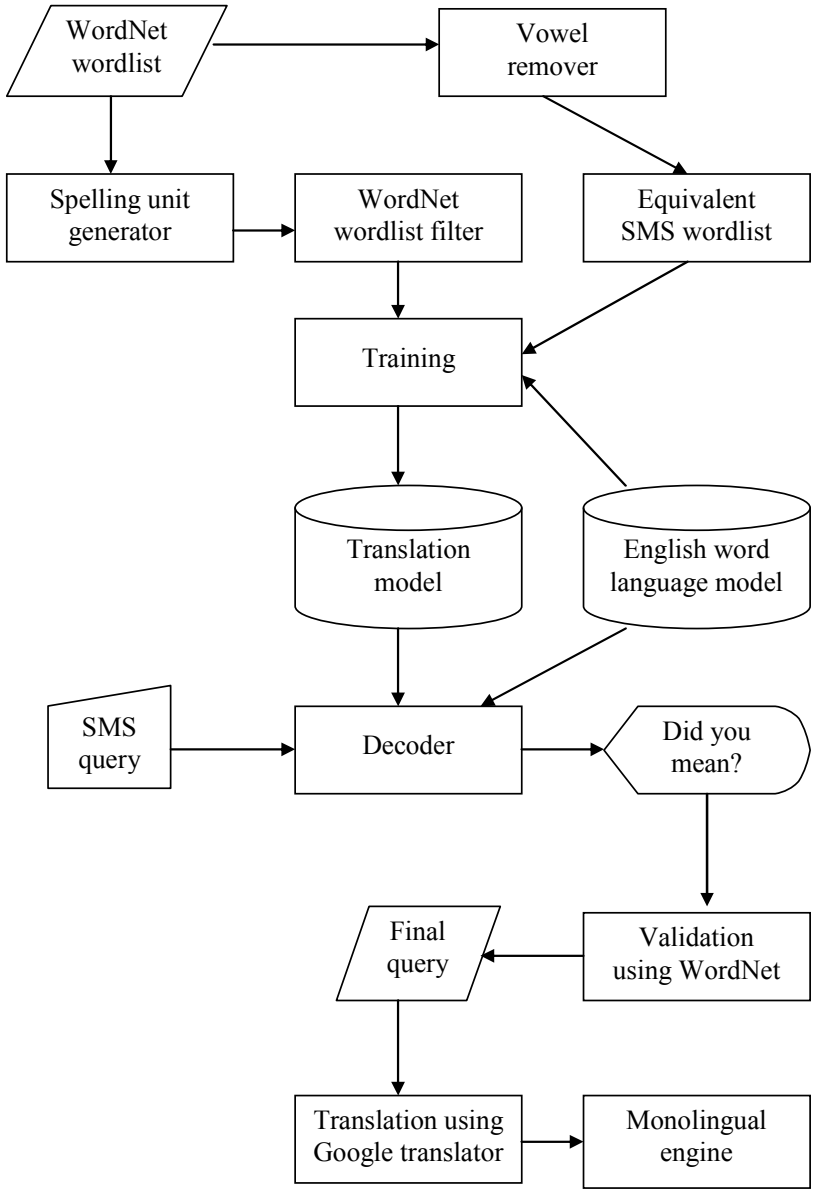

Figure 2: Cross-lingual FAQ retrieval system Architecture.

to compose an SMS query to the monolingual English query translation model. An English correction module corrects the output of the SMS-English query translation system. Finally, the system retrieves the answer from the FAQ dataset by using a cross lingual FAQ retrieval system.

4.1 SMS to English Query Translation Module

In this module, initially, we have all English words known to the system listed in a file. These words have been collected from the English WordNet 3.0 resource. The English wordlist file plays an important role in the development of parallel example-based Hindi SMS—English translation system.

Next, the system trained and decoded the SMS query using the English query translation. This module has the following sub-modules:

SMS-English parallel corpus We have created a database of English words, which were collected from the WordNet database: verbs, nouns, adverbs, and adjectives. We removed the vowels from each word: e.g., a word "Translation" became "Trnsltn". This is important because vowels are often omitted in SMS texts. We stored such words stripped from the vowels along with the original word and the word separated into spelling units, e.g., "tra-ns-la-tion".

Training and decoding We have developed an SMS-English word translation model by using a statistical tool: Moses toolkit. The translation model generates English words from SMS words. Given an input of as an SMS text, the system generates the corresponding English word-by-word translation. After decoding SMS query in this way, we form the English query corresponding to this SMS text.

Cleaning The generated English query is corrected by a spell checker, as discussed in Section 2. For further cleaning the English query, we validated each word against the WordNet resource. If the word is not present in WordNet, then we extracted the nearest candidates of that unknown word. Letters of the unknown word were matched against the nearest candidate word. The matching procedure followed the letter sequence matching.

4.2 English—Hindi Translation System and Final Retrieval

Finally, the generated English query was translated into the Hindi query using freely available online Google translator. With this, the cross-lingual SMS query was obtained. Then, the monolingual FAQ retrieval procedure of Hindi SMS queries described in Section 2 was used.

5 Experimental Results

We tested our SMS-based FAQ retrieval on one run for Monolingual (English), two runs for Monolingual (Hindi), and one run Cross-lingual (Hindi) of the FIRE 2011 competition.

We used the Mean Reciprocal Rank (MRR) as one of the evaluation measures for our system. MRR is non-zero if, for the in-domain queries, a correct match was found

in any of their top 5 answers. If the correct answer was returned as the top answer, then the score was the highest for that query; if it was returned as the 5th answer then the score was the lowest.

The evaluation scores obtained by our SMS-based FAQ retrieval are shown in Table 3.

Table 3. Evaluation Score for SMS-based FAQ retrieval.

Run	Descriptions	Statistics
Monolingual (English)	# of In-domain Queries	704
	# of Out of Domain Queries	2701
	In Domain Correct	29 / 704 = 0.0412
	Out of Domain Correct	0 / 2701 = 0
	Mean Reciprocal Rank (MRR)	0.0538
Monolingual (Hindi), run 1	# of In-domain Queries	200
	# of Out of Domain Queries	124
	In Domain Correct	0 / 200 = 0
	Out of Domain Correct	119 / 124 = 0.960
	Mean Reciprocal Rank (MRR)	0.0
Monolingual (Hindi), run 2	# of In-domain Queries	200
	# of Out of Domain Queries	124
	In Domain Correct	36 / 200 = 0.180
	Out of Domain Correct	0 / 124 = 0
	Mean Reciprocal Rank (MRR)	0.235
Cross-lingual	# of In-domain Queries	37
	# of Out of Domain Queries	3368
	In Domain correct	2 / 37 = 0.0541
	Out of Domain correct:	40 / 3368 = 0.0119
	Mean Reciprocal Rank (MRR)	0.0541

As it can be observed from Table 3, the system gives promising results, which are above a random baseline, though much lower than what would be required for practical use of the system.

We believe that such low results can be explained by the noisy nature of both SMS queries (which are usually written in a language very different from the norm) and the text of the FAQ answers.

6 Conclusions and Future Work

In this paper we have described our approaches for monolingual and cross-lingual SMS-based FAQ retrieval. For spell checking, we used the freely available Bing spellchecker. For translation, we used the freely available Google translator because Google translator has better performance than the Bing translator.

In our future work we plan to develop our own machine translation system optimized for this kind of tasks. In addition, we plan to explore the use of softer semantic

and syntactic [5] similarity measures, such as those that employ argument structure information [6] and information about synonyms [7].

Acknowledgements. The work was partially supported by the Governments of India and Mexico under the CONACYT-DST India (CONACYT 122030) project "Answer Validation through Textual Entailment", the Government of Mexico under the CONACYT 50206-H project and SIP-IPN 20121823 project through Instituto Politécnico Nacional, and the Seventh Framework Programme of European Union, project 269180 "Web Information Quality Evaluation Initiative (WIQ-EI)".

References

1. Ledeneva, Y., Sidorov G.: Recent Advances in Computational Linguistics. Informatica. International Journal of Computing and Informatics, 34:3–18 (2010).
2. FIRE 2011: Third Workshop of the Forum for Information Retrieval Evaluation, 2–4 December, IIT Bombay.
3. Contractor, D., Mittal, A., Padmanabhan D.S., Subramaniam L.V.: SMS-based FAQRetrieval, FIRE 2011: Third Workshop of the Forum for Information Retrieval Evaluation, 2–4 December, IIT Bombay.
4. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.
5. Sidorov G., Herrera-de-la-Cruz J.A., Galicia-Haro, S., Posadas-Durán, J.P., Chanona-Hernandez, L.: Heuristic Algorithm for Extraction of Facts using Relational Model and Syntactic Data. Lecture Notes in Artificial Intelligence 7094, 2011, pp. 328–337.
6. Castro-Sánchez, N.A., Sidorov, G.: Analysis of Definitions of Verbs in an Explanatory Dictionary for Automatic Extraction of Actants based on Detection of Patterns. Lecture Notes in Computer Science 6177, 2010, pp 233–239.
7. Castro-Sánchez, N.A., Sidorov, G.: Automatic Acquisition of Synonyms of Verbs from an Explanatory Dictionary using Hyponym and Hyperonym Relations. Lecture Notes in Computer Science 6718:322–331, 2011.