Retrieval of FAQs based on SMS-input through phonetic equivalence

Akhil Lohchab Amanpreet Singh Ashish Mahajan Ritu Agarwal

Delhi Technological University Delhi Technological University Delhi Technological University Asst. Prof., Delhi Technological University

Abstract : Short Messaging Service(SMS) has become really popular in the present day and presents a unique opportunity to make automated query response reach a wider audience.Public information systems such as passenger query systems and patient query systems in a hospital can prove to be quite useful when implemented this way. In this paper, we have addressed the problem of mapping the user queries on government portals in the form of SMSes to their equivalent plain text frequently asked questions (FAQs) stored in the database by generating phonetic equivalents of both the queries and the FAQs. Lucene indexer has been used to index the FAQs and the score for a document is determined by primarily by the number of tokens in the SMS query that have a high similarity score with an FAQ. Experiments show high success rate on the new unseen SMSes.

1. Introduction

The number of mobile users is growing at an amazing rate. In India alone a few million subscribers are added each month with the total subscriber base now crossing 370 million. The anytime anywhere access provided by mobile networks and portability of handsets coupled with the strong human urge to quickly find answers has fueled the growth of information based services on mobile devices. These services can be simple advertisements, polls, alerts or complex applications such as browsing, search and e-commerce. The latest mobile devices come equipped with high resolution screen space, inbuilt web browsers and full message keypads, however a majority of the users still use cheaper models that have limited screen space and basic keypad. On such devices, SMS is the only mode of text communication.[1]

Language usage over SMSes differs significantly from the language of communication from person-to-person. The characteristic features of the mobile devices have brought about the unique ways of communication, notably in its usage is the restriction in the number of characters/words engaged by the user to communicate because of the memory space, pad structures of the design and writing skills of the query and bandwidth issue . Its construction is based on convenience of spelling and homophony of the wordings. Regardless of the range of the handsets (low, medium or high end).This nature of texting language makes it difficult to build automated question answering systems around SMS technology. This is true even for questions whose answers are well documented in a FAQ database.

In this paper we present a SMS query-based FAQ retrieval system that allows the query to be input in the SMS texting language. Our goal is to accept the queries and respond by finding the most appropriate FAQ from a corpus of FAQs.

2. Our Approach

Our initial idea was to use term correction and find the closest words to SMS-tokens based on edit distance from the FAQ-list and then use these corrected SMS tokens to calculate a similarity score with the FAQs. After a few preliminary experiments, it was observed that certain words rarely had any effect on the list or the ranks of the FAQs retrieved. Hence, a stop word list was added that removes all the words within the file from the SMSes as well as the FAQs. However, this approach did not cater well to the 'Speech_queries' and a different direction was sought, namely phonetic equivalents where the Metaphone library was used to generate phonetic equivalents.

We indexed all FAQs using Lucene. Lucene is a public domain search utility that builds an inverted index on the given document set. Whenever a search is made on some keywords, this index is searched.

Every FAQ document was tokenized and phonetic key was generated for each token using the Metaphone library with a maximum possible length of key. All SMS queries were also tokenized and converted to phonetic equivalent tokens.

Metaphone uses a set of rules to code a token by using the 16 consonant symbols 0BFHJKLMNPRSTWXY, where '0' represents "th" (as an ASCII approximation of Θ), 'X' represents "sh" or "ch", and the others represent their usual English pronunciations.

Once, the FAQs have been tokenized, stop words removed, the phonetic equivalents of the tokens are generated and committed to the index. A similar procedure allowed the phonetic equivalents of the tokenized SMS query to be generated. The index was then searched with each token of the SMS query, and a score for each document with respect to the SMS token is calculated. A list of top scoring documents is then compiled and a maximum of 5 documents are selected. Selection of the documents is filtered by a threshold function.

3. Results

Text-Only queries:

No. of In-domain Queries	200
No. of Out of Domain Queries	99
In Domain correct	179/200 (0.895)
Out of Domain correct	58/99(0.58)
Total Score	0.7926421

Mean Reciprocal Rank (MRR): 0.95936835

Overall :

No. of In-domain Queries	392
No. of Out of Domain queries	148
In Domain Correct	180/392 (0.4591)
Out of Domain correct	83/148 (0.5608108)
Total Score	0.48703703

Mean Reciprocal Rank (MRR): 0.6298412

4. Future Work and Conclusion

Given the problem of retrieving an FAQ based on SMS query, we obtained a good accuracy on the text queries and although our attempt at mapping the 'Speech_queries' fared poorly, one can consider trying other approaches. In particular, using a synonym lookup dictionary to improve the matching of SMS tokens to FAQ tokens is something one can try.

5. References

1. SMS based Interface for FAQ Retrieval. Govind Kothari. IBM India Research Lab gokothar@in.ibm.com. Sumit Negi. IBM India Research Lab sumitneg@in.ibm.

2. A Query-Based SMS Translation in Information Access System. Ademola O. Adesina, Kehinde K. Agbele, Nureni A. Azeez, Ademola P. Abidoye, International Journal of Soft Computing and Engineering (IJSCE) 2011.

3. E. Prochasson, Christian Viard-Gaudin, Emmanuel Morin. 2007. Language Models for Handwritten Short Message Services, In Proceedings of the 9th International Conference on Document Analysis and Recognition.

4. E. Sneiders. 1999. Automated FAQ Answering: Continued Experience with Shallow Language Understanding, Question Answering Systems. Papers from the 1999 AAAI Fall Symposium. Technical ReportFS-99−02, November 5−7, North Falmouth, Massachusetts,

5. Dr. Hadeel Showket Al-Obaidy, "Building Ontology Web Retrieval System Using Data Mining," ed.