Data-Driven Methods
for SMS-Based FAQ Retrieval

Sanmitra Bhattacharya, Hung Tran, and Padmini Srinivasan
Department of Computer Science,
The University of Iowa,
14 MacLean Hall, Iowa City, IA 52246
{sanmitra-bhattacharya,hung-trv,padmini-srinivasan}@uiowa.edu

Abstract. SMS text messaging is one of the most popular data applications on mobile phones these days. Other than personal communication, text messaging can also be used for various purposes like bill payment, banking, inquiry, etc. However these messages are extremely noisy and contain misspellings, abbreviations, transliterations, etc. Keeping this in mind, FIRE 2011 introduced a new retrieval task called SMS-based FAQ retrieval in English, Hindi and Malayalam. Within-language and cross-language tasks were designed for this retrieval problem. As solutions we propose various data-driven retrieval techniques that includes noise reduction in the SMS queries and the FAQ corpora. Overall, we find that our methods work well for the retrieval experiments in the different languages. For English, the use of Google Spelling Suggestions and term expansion strategies improve retrieval scores. For Hindi and Malayalam retrieval experiments, we find that translation of queries and corpus to English increases retrieval accuracy.

1 Introduction

The mobile phone market has seen exponential growth since its inception in the mid-1970’s. This spark in growth can be attributed primarily to advances in cellular technology making mobile phones inexpensive as well as easy to use. Over 75% of the world population has access to mobile phones with China and India being the top contributors in its recent growth. The total number of mobile subscribers in India in May 2011 was over 840 million, up 233 million from May 2010. The most popular data applications on mobile phones is SMS text messaging. The number of SMS messages sent in 2010 was 6.9 trillion and for 2011 this number should reach over 8 trillion1. The wave of modernization and globalization in India has also resulted in an increased use of text messaging in various domains. SMS messaging is now used not only for personal communication but also for inquiry, advertising, marketing, polling, bill payment, banking, etc.

In line with these developments, FIRE 2011, in its third year, introduced a new retrieval problem called SMS-based FAQ retrieval. This problem addresses

1 http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats
finding question entries from a corpora of Frequently Asked Questions (FAQs) that best match a given SMS query. In a traditional question-answering (QA) retrieval task such as the TREC QA track (1999-2007), retrieval systems were required to retrieve answers (i.e. entries containing answers) rather than documents in response to questions. For example, the main task of the TREC 2007 QA track required participants to define a target by answering a series of factoid and list questions about the target. Participants also had to return any information not covered by the question series. For example, for a target such as “House of Chanel” participants are required to find answers to questions like “Who founded the House of Chanel?” (factoid) or “What museums have displayed Chanel clothing?” (list). The source of questions for the QA track were participants, assessors, FAQFinder system logs, MSNSearch logs, query logs of Microsoft and AOL, etc. The target corpus consisted of newswire documents, AQUAINT disks and blog data.

In contrast, the FIRE 2011 FAQ corpora consists of well-formed questions and answers in various domains like agriculture, career, general knowledge (GK), Indian Railways, telecommunication, etc. Here the queries are SMS messages inquiring about different topics of the FAQ corpus. In essence, participants were required to find FAQ entries that answer or match the SMS queries. One of the major challenges in this task is the structure of the SMS queries, or the lack thereof. SMS messages are generally noisy with misspellings, non-standard abbreviations, omission of words, etc. From a linguistic point of view SMS messages generally do not conform to proper grammatical usage and hence comprehension of such messages depend largely on context and intellectual effort. For example, the SMS message “wht is career counclng” highlights a few of these problems. Words like “wht” (for “what”) or “counclng” (for “counseling”) are non-standard abbreviations that require the expertise of experienced texters to decipher their meaning. Matching such ill-formed SMS messages to reasonably well-formed FAQs entries is the challenge.

2 Description of the Task

As already mentioned, the overall goal of the task is to match SMS queries to FAQs. Participants are permitted to use the information in both the questions and the answers in FAQs for matching the SMS queries.

A sample (abbreviated) XML-formatted FAQ topic is given below:

```xml
<FAQ>
<FAQID>ENG_CAREER_1</FAQID>
<DOMAIN>CAREER</DOMAIN>
<QUESTION>What is career counseling?</QUESTION>
<ANSWER>Career counseling is a process ...
</ANSWER>
</FAQ>
```