A text based drug query system for mobile phones

Akhil Langer*

Department of Computer Science,
University of Illinois at Urbana-Champaign,
201 N. Goodwin Avenue,
Urbana, IL 61801, USA
E-mail: alanger@illinois.edu
*Corresponding author

Rohit Banga

Master's Student at College of Computing,
Georgia Institute of Technology,
1038 McMillan Street NW,
Atlanta, GA 30332, USA
E-mail: rohit.banga@gatech.edu

Ankush Mittal

Department of Computer Science and Engineering,
Graphic Era University,
Dehradun 248 002, India
E-mail: dr.ankush.mittal@gmail.com

L.V. Subramaniam

Information Quality and Discovery,
IBM Research,
New Delhi 110 070, India
E-mail: lvsubram@ibm.co.in

Parikshit Sondhi

Department of Computer Science,
University of Illinois at Urbana-Champaign,
201 N. Goodwin Avenue,
Urbana, IL, USA
E-mail: sondhi1@illinois.edu

Abstract: Dissemination of medical information using mobile phones is still in a nascent stage because of their limited features – lack of penetration of mobile internet, small screen size etc. We present the design of a drug QA system that could be used for providing information about medicines over short message service (SMS). We begin with a survey of the drug information domain and classify the drug related queries into a set of predefined classes. Our system

uses several natural language processing tools coupled with machine learning classification techniques to process drug information related queries. We focus on developing a natural language interface allowing the user to be flexible in phrasing their queries and attain an accuracy of 81% in classifying the drug related questions. We conclude that it is feasible and cheap to deploy such a system to encourage the practice of evidence based medicine.

Keywords: mobile communication; text processing; question answering system; healthcare; drug; medicine; SMS; short message service; machine learning.

Reference to this paper should be made as follows: Langer, A., Banga, R., Mittal, A., Subramaniam, L.V. and Sondhi, P. (xxxx) 'A text based drug query system for mobile phones', *Int. J. Mobile Communications*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Akhil Langer is a PhD student in the Parallel Programming Laboratory, Department of Computer Science at University of Illinois at Urbana-Champaign. He received his BTech in Computer Science from the Indian Institute of Technology Roorkee in 2010. His research interests are high performance computing, parallel algorithms, stochastic optimisation and information retrieval. As an undergraduate at IIT Roorkee he worked under Dr. Ankush Mittal on a project titled *Providing Natural Language Interface to Health Care Services over SMS* which was awarded the Innovative Students Project Award by the Indian National Academy of Engineers.

Rohit Banga is a Master's student studying Computer Science at Georgia Institute of Technology. At Georgia Tech he has worked on Massive Scale Graph Analytics Algorithms and RNA secondary structure prediction. He obtained his Bachelor's degree in Computer Science from Indian Institute of Technology Roorkee in 2010. As an undergraduate at IIT Roorkee, he worked under Dr. Ankush Mittal on projects related to High Performance Computing and Text Processing for Question Answering systems. His undergraduate thesis was titled *Providing Natural Language Interface to Health Care Services over SMS* which was awarded the Innovative Students Project Award by the Indian National Academy of Engineers.

Ankush Mittal received the BTech in Computer Science and Engineering from the Indian Institute of Technology, Delhi in 1996 and Master's in 1998 respectively. He received PhD from Department of Electrical and Computer Engineering, The National University of Singapore in 2001. He was a Faculty Member in the Department of Computer Science, National University of Singapore for two years. He has also worked as an Associate professor at IIT-Roorkee for seven years. He has contributed more than 240 research papers in journals and conferences of high repute with significant impact factor, many of which are in IEEE transactions, Springer and Elsevier journals. He is also a co-author of a book on Bayesian Network technologies published by Information Science, USA. He has keen interest to facilitate publications, grants, and research aptitude of faculty and students. He is serving as Director (Research) at Graphic Era University, Dehradun.

L. Venkata Subramaniam manages the information processing and analytics group at IBM Research India. He received his PhD from IIT Delhi in 1999. His research focuses on unstructured information management, statistical natural language processing, noisy text analytics, text and data mining, information theory, speech and image processing. His work on data cleansing and entity resolution has been deployed on the field in scenarios involving

cleansing of millions of data records. He co-founded the AND (Analytics for Noisy Unstructured Text Data) workshop series and also co-chaired the first four workshops, 2007–2010. He was guest co-editor of two special issues on Noisy Text Analytics in the *International Journal of Document Analysis and Recognition* in 2007 and 2009.

Parikshit Sondhi received an Integrated Dual Degree (BS + MS) in Computer Science from Indian Institute of Technology Roorkee, India, in 2008. He is currently a Computer Science PhD candidate at the University of Illinois Urbana Champaign. His research interests include information retrieval and mining applications in the biomedical domain.

This paper is a revised and expanded version of a paper entitled 'Mobile medicine: providing drug related information through natural language queries via SMS' presented at *IEEE International Advance Computing Conference (IACC) 2009*, Patiala, India, 6–7 March, 2009.

1 Introduction

Many efforts have been made in providing drug related information to physicians as well as laypersons through the means of publishing drug information resources – printed, as well as in digital form (through CDs, free information sites on the internet, etc.), web interfaces where one can submit the queries for them to be answered by experts in the respective fields or one can visit drug information centres (DICs) to get his or her questions answered. DICs are primarily operated by medical schools or respective department/ministry of the government. These centres answer drug and drug therapy related questions both from professionals and general public via phone calls, email or fax. Additionally, numerous online drug information sites provide forums for discussing patient-specific problems/queries along with detailed factual information about drugs. These sites report tens of millions of unique visitors per month (Google, 2010).

Many printed drug reference sources are easily available starting from handbooks to detailed guides for prescription and non-prescription drugs. These drug information sources however are either unknown or inaccessible to the semi-urban or rural population. Also, many of them such as the DICs and the printed books require delayed processing wherein one has to manually search through the available material for the required information. This information is also accessible via the internet but the internet is yet to become ubiquitous, particularly in developing countries. Drug related information can be useful to the general public in emergency situations when one has many medicines lying in his or her medicine-box but does not know which one to use in that situation. In such circumstances, the patient or anybody else in the vicinity of the patient would want to know various aspects of a given drug like its usage, dosage, side-effects, directions on how to take the drug, etc. In a developing country like India where doctor-to-patient ratio is very low (1 : 1700)[1] availability of such information becomes extremely important because then the situation can be tackled based on evidence without requiring to contact a physician who is generally not easily reachable. Some other situations in which such information can be useful are – when one wants to know the adverse interactions between two or more drugs that he or she is taking or the precautions to be taken while on a medication or whether a drug is contraindicated in

case of old age, pregnancy, etc. Availability of drug related information instantaneously in the absence of accessibility to the earlier mentioned resources makes it an important domain of interest.

Keeping the above points in context, mobile phones become a potential medium through which such information can be made available at the point of need. While research efforts like (Suh et al., 2012) are being invested in improving the quality of smartphone applications, most of the people in developing countries still do not use smartphones and depend on low-end mobile phones which do not have storage capabilities or internet connectivity, the drug resources available on the internet as well as digital drug databases still remain inaccessible to the common man. Short message service (SMS) however is available on almost every phone and is widely used being trivial to use and inexpensive. The use of text messaging or SMS has become ubiquitous and commonplace for recreational and business purposes. It has been estimated that 2 billion messages are sent daily across the world. In India cost of sending a SMS is very low, much less than Rs. 1 (US$ 0.02).

Certainly, SMS poses considerable restrictions too which challenge its usage for providing relevant information. 160 characters per SMS limit makes it infeasible for sending lengthy but relevant information. SMS being a text-only service makes it less interactive. These make SMS suitable for answers to factoid questions which are inherently suited for display on small screens. However the ubiquity of SMS based communication makes it a desirable channel for providing access to various services and we argue their use on technical grounds for encouraging evidence-based medicine.

The paper is organised into eight sections. Section 2 discusses the need of a Natural Language Drug QA system and thereby formulates the problem. Section 3 puts light on the related work. In Section 4, we present the analysis of the drug information domain and databases collection and organisation. Detailed description of the various components of the proposed system architecture is given in Section 5. In Section 6, we report the analysis of performance of various machine learning classifiers on the *DrugQuery* dataset. Finally, In Section 7, we conclude with some discussion, academic and managerial implications and future work.

2 Problem formulation

First of all, we consider the need of a natural language system for the drug information domain. By and large, masses are not acquainted with the jargon of the medical world like contraindications, interactions, indications, etc. These arcane terms are not expected from layman. Thus, the system should also be capable of recognising queries formulated in regular English terms. Minimalistic learning on the part of the user should be required. Below are two examples of questions asked in different manners but are inquiring for the same information:

Query with technical keyword: What are the interactions between Alcohol and Aspirin?

Same query in simple English: Can I drink alcohol while I am on Aspirin?

Query with technical keyword: What are the contraindications of Cetirizine?

Same query in simple English: Is Cetirizine advised in Hypersensitivity?

Our query dataset which we collected from online forums consists of several such examples. Table 7 contains more examples of queries posed in simple English, which are also the target queries for our QA system.

Secondly, it behoves us to consider the need of an automatic question answering like this. Setting up a call centre (semi-automatic) system, is inefficient as it will require a 24 h personnel engagement and from our experience of the railway enquiry system in India it is well known how ineffective such a system can be. Moreover, there is a need for an autonomous system that provides the mobile community access to the medical repository in parallel to the internet.

As a solution to these considerations we develop a fully automatic QA system that takes drug information related query in SMS text form as input, processes it and then tags it with an appropriate class using a machine learning classifier. Following this, it generates a SQL query to retrieve its answer from a comprehensive database and sends the answer back in SMS text format back to the user.

The main contributions of our paper are:

- an analysis of issues in building a drug information QA system on SMS and how to deal with them
- how to build a drug information related factoid QA system
- experiments and analysis of performance of various state-of-the art classifiers on the drug query dataset.

3 Prior work

Various services are provided to the customers via SMS namely, railway PNR enquiry, train schedules, stock quotes, etc. Google SMS (Schusteritsch et al., 2005) offers a number of search features like Weather, Sports, Movies, Flights, etc. for querying in a standard fixed format. Agarwal (2008) presents a model design for domain specific question answering on mobile which access information from sample database for automatic railway inquiry. He demonstrates through a railway enquiry system that how SMS which is easier, faster, reliable and available in all networks can be combined with a powerful method of artificial intelligence: question answering to facilitate information access on mobile phone. He concludes that performance of such a system mainly depends on the question typing rules, with precise question typing rules leading to higher accuracy.

To the best of our knowledge, there has been no effort made specifically to provide drug related information through a question answering system, though a lot of related work has been done in query classification and information retrieval in medical domain. Considering the limitations posed by SMS a drug query classifier can significantly help in selecting only the relevant data and sending back to the user to solve his or her immediate need. In an attempt to recognise question type in a QA system for health (Cruchet et al., 2008) showed how SVM can be used to detect the medical type of the question and the type of answer expected. Their analysis was on the questions related to diseases which they classified into 11 types. However, they did not present any solution to obtain the answers for the queries and left it as a task of information retrieval from the web. Ely et al. (2002) created an evidence taxonomy to categorise questions that were

potentially answerable with evidence and concluded that only evidence questions are potential answerable with evidence using medical literature and other online medical resources. Yu and Sable (2005) developed approaches to automatically separate answerable medical questions from unanswerable ones. Yu et al. (2005) focused on the harder task of automatically classifying questions to the specific categories presented in the evidence taxonomy namely, clinical vs. non-clinical, general vs. specific, evidence vs. no evidence and intervention vs. no intervention. They studied use of various supervised machine learning systems like SVM, Naïve Bayes, etc. for binary classification of medical questions.

Millions of people use Google every day to search for information. There has been growing interest in providing access to applications, traditionally available on internet, on mobile devices using SMS as many users still have older handsets which do not support the features of mobile web browsing. Kopparapu et al. (2007) piloted with a major telecom operator in India to provide natural language mobile interface that gives user an unconstrained mode of asking for information from the yellow pages directory, 24 × 7. Their system takes care of spelling mistakes and SMS lingo. Research efforts have been invested in understanding the structure of texting language (Aw et al., 2006), (Acharyya et al., 2008). Adoption of mobile phones for content delivery has been studied in Priporas and Mylona (2008), Sun et al. (2009), Patrick (2011) and Chong et al. (2011). Specific studies related to adoption of mobile phones for healthcare related services can be found in Tounsi (2008) and Lin (2011).

4 Domain analysis

4.1 Query analysis

Pharmacists are the drug information experts. Pharmacists receive drug information from a variety of people, but the demographics of drug information requesters can generally be categorised into two main groups: healthcare providers and the lay public. Healthcare providers constitute physicians, nurses, occupational and physical therapists, medical technologists, social workers, psychologists and medical information specialists. The level of information complexity required by healthcare providers is likely to be high and requires systematic, effective and efficient search strategies for finding an answer as these questions are more targeted towards improving healthcare and patient outcomes. On the other hand, general public queries will be more straightforward – requiring searching of a structured database which contains comprehensive information about drugs. Such drug information questions can be broken down into well-defined classes (Table 1). Some of the important classes that need description are:[2]

- *Indication*: A condition which makes a particular treatment or procedure advisable e.g. CML (chronic myeloid leukaemia) is an indication for the use of Gleevec (imatinib mesylate).
- *Side-effects*: Problems that occur when treatment goes beyond the desired effects or problems that occur in addition to the desired therapeutic effect e.g., a haemorrhage from the use of too much anticoagulant (such as heparin) is a side effect caused by treatment going beyond the desired effect.

- *Contraindications*: A condition which makes a particular treatment or procedure inadvisable e.g., a baby with a fever should never be given aspirin because of the risk of Reye's syndrome.
- *Drug-drug interaction*: It is the phenomenon that occurs when the effects or pharmacokinetics of a drug are altered by prior administration of another drug. Drug interactions are not only unsafe, but also reduce medication effectiveness, produce undesirable side effects, and may cause death.

Hence, answering a query would require classification of the question into one of these classes and identification of the drug for which the information is being sought. Following this the information can be extracted from a well-structured database.

Table 1 Representative queries for each query class

Class	*Representative query*
General information	*What is Ambien?*
Indications	*What conditions or indications might Ambien treat?*
Side-effects	*What are the possible side-effects of Ambien?*
Interactions	*What other drugs affect Ambien?*
Contraindications	*When should I not take Ambien?*
Dosage	*What is the dosage of Ambien?*
Dose miss	*What happens if I miss a dose of Ambien?*
Overdose	*What happens if I overdose on Ambien?*
Things to avoid	*What should I avoid while taking Ambien?*
How to take (directions)	*How should I take Ambien?*
Substitutes	*What are the substitutes for Ambien available in the market?*
Onset of action	*How much time does it take for Ambien to come into action?*
Duration of action	*What is the duration of action for Ambien?*
Pregnancy	*Is it safe to take Ambien in Pregnancy?*
Old age	*Is it safe to take Ambien in Old Age?*
Breast feeding	*Is it safe to take Ambien during the duration of breast feeding?*

4.2 Database collection and organisation

From the vast pool of resources with varying amount and complexity of drug information, the task was to search for a digital database which is comprehensive as well as suitable for SMS domain i.e., while being within the limitations of SMS it should satisfy the needs of information seeker. Though none of these databases cater absolutely to the needs of the SMS domain, but in order to demonstrate the system we took MedClik's digital drug database, Alberta drug cards and crawled revolution health websites (Health, 2010) drug treatment pages to use them as our drug information resources. There is a significant amount of work that needs to be done in order to extract the answer (from these myriad resources) that meets the SMS size limitations and also matches with the health literacy level of the information seeker. In this work we focus on query analysis and use concise answers (may contain some technical words) from our

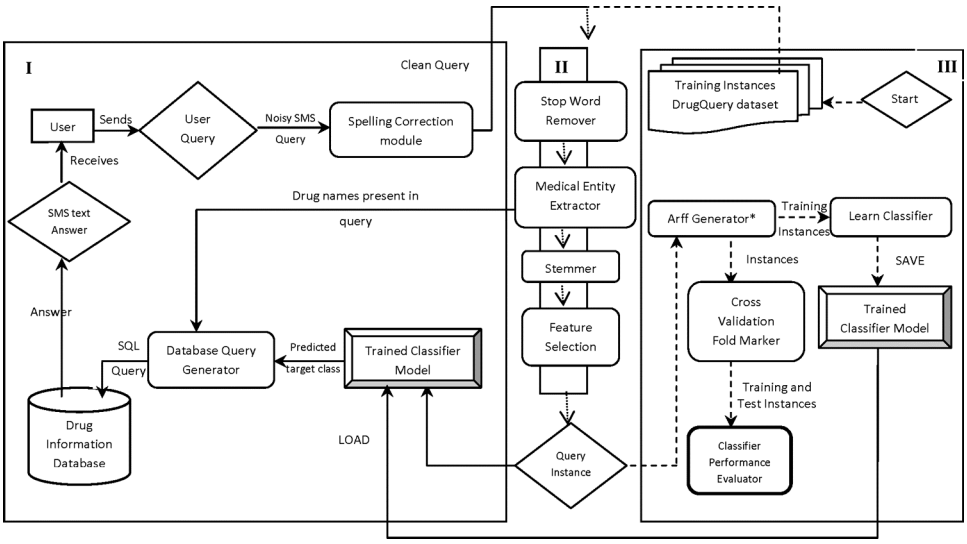
database that fit within a couple of SMSs. In the future we intend to work specifically on developing drug information resources having drug information that can be contained in 3–4 SMSs while serving the information needs of common man with low and moderate level health literacy.

Each medicine has two names associated with it, one is called the generic name/drug name which is given based on the salts present in it and other is the brand name which is given by the specific drug manufacturer. Therefore, corresponding to a given generic drug, the database consists of its' indications, side-effects, etc. and the brand-names available in the market.

5 Overall system architecture

Figure 1 gives the detailed system architecture. It comprises of three parts. Each of these parts has been described in the following three subsections.

Figure 1 System architecture



5.1 Query handler

This module receives the query from the user through SMS. Because the query is expected to contain spelling mistakes in drug names as a rule rather than exception, the query is first sent to the spelling correction module where the misspelled drug names as well as other words are replaced by their correct spellings.

5.1.1 Spelling correction module

This module receives the noisy query and corrects all the spelling errors using GSpell (2010). GSpell is a spelling suggestion tool that uses a mix of algorithms to retrieve close neighbours. The mix of algorithms includes Ngrams, metaphone, homophones and also uses common misspellings to figure out potential matches. Candidates are evaluated

by the Levenshtein (1966) edit distance, and similar ranked candidates are re-ordered by use of word based corpus frequencies.

GSpell outputs the following for a given input term:

Input term Suggestion Rank (from 0 to 1) Method used

The rank is the normalised ranking with a value between 0 and 1 where 1 is an exact match and 0 is no match at all. Table 2 gives an example of the output for the misspelled term Banadril.

Table 2 Gspell output for the misspelled term 'Banadril'

Input term	*Suggestion*	*Rank* (*from 0 to 1*)	*Method used*
banadril	benadryl	0.8	Metaphone
banadril	banal	0.3	LEFT_MOST
banadril	viadril	0.3	RIGHT_MOST
banadril	baytril	0.3	NGRAMS
banadril	benacil	0.3	NGRAMS

We have omitted other output terms as they are not relevant for our purpose. The dictionary used by Gspell consists of most of the drugs but their brand names which are manufacturer specific were not present in it. Thus we indexed a list of around 11,912 brand names (collected from our database) into the Gspell's dictionary and then calculated its efficiency (spelling correction ability). Methods to generate misspellings have been discussed in Damerau and Mays (1964). Damerau and Mays (1964) have shown that 80% of all spellings mistakes involve a single insertion, a single deletion, a single substitution, or a transposition of two letters in a word. The automatic error generator for our case created mistakes by applying randomly any of the two methods on the correct word. A sample of the misspelled words obtained after applying the methods:

Word: Acetaminophen

1 dropping a letter → Acetaminphen

2 adding a letter → Acet**k**aminophen

3 transposing adjacent letters → Acetam**ni**ophen

4 replacing a letter → Acetaminop**z**en

5 **performing two of 1–4** → cetam**ni**ophen,

Actaminophe, Aectamnophen, etc. **– one of these was used as misspelled version of Acetaminophen**

Two outcomes were measured:

- whether the correct suggestion was ranked number one
- whether the correct suggestion was found at all.

Table 3 shows the probability of success among the eight different search combinations for each of the two outcomes.

Table 3 Percent of misspelled words for which the correct word was found

Description	*Drugs*	*Brands*
Number of terms searched	744	11,912
Correct word ranked No. 1	81.3%	62.6%
Correct word found at all	87.2%	74.7%

The efficiency of GSpell for our domain is very encouraging and is higher than the general biomedical domain scenario for which efficiency has been calculated in Crowell et al. (2004). After spell-check, the query is passed through the query processor block (block II) which is delineated in the following section.

5.2 Query processor

This block passes the query through a number of query pre-processing modules and finally selects the features relevant for its classification. In other words, it outputs the query instance which is ready to be fed to the learned classifier model (described in block III). The classifier model tags the query with its predicted class. This target class, along with the drug name extracted from the query in medical entity extractor in block II, is used to generate the SQL query to retrieve answer from the drug information domain. Finally, the answer is sent to the user in the SMS text format.

5.2.1 Stop words remover

Some stop words were removed from the query in this stage. However, certain words that are generally considered as stop words played key roles in distinguishing one query class from the other. For instance, consider the queries in Table 4.

Table 4 Queries having common words as distinguishing features

Query	*Query class*
When should I not take Accutane?	Contraindications
What should I not take with Accutane?	Things to avoid
How should I take Accutane?	How to take (directions)

Keeping the above cases in mind, certain common words like what, when, how, not, etc. were not considered as stop words and were therefore used for query classification.

5.2.2 Medical entity extractor

We need to identify and mark named entities in the query, which in our domain are drug names, disease names and symptoms. To identify these entities in the query we do exhaustive search for every drug name, disease name and symptom in our medical database and replace them by generic terms like drug_name, disease_name and symptom_name respectively. All queries are expected to possess only one drug name except in the case of queries for drug-drug interaction where even two drug names can be

present. Now let us consider the following query on side-effects of Ambien to understand the generalisation procedure outlined in Figure 2:

Query: Can Ambien cause taste loss?

Generalised Query: Can drug_name cause symptom_name?

Now consider another query on side-effects of Ambien:

Can Ambien cause dizziness?

This query will also be generalised to the same form as above and hence classification of this new query containing previously unlearned medical terms becomes possible.

Figure 2 Pseudo code for medical entity extractor

```
//do this for all medical entities
for every drug in table drug_all
    if query contains drug
        replace drug by "drug_name" in query

save the drug for answer extraction after query classification
```

5.2.3 Stemmer

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not itself a valid root. To understand how stemming will be useful in our case consider the word 'indication' which can be used in following two ways in the query:

- What are the indications of Amoxicillin?
- Is Amoxicillin indicated for ear infection?

We used the famous Porter Stemming Algorithm (Porter, 1980) which has become the de-facto standard for English stemmers. The algorithm converts both the words 'indications' and 'indicated' to the word 'indic' as the base form for indications. Another example is of the basic word 'dose' which can also be used as 'dosage' in the query. Note that Stemming follows the medical entity extractor (Section 5.2.2) because otherwise it may convert medical entities in the query to unrecognisable forms.

5.3 Building the classifier

Classifying a query into one of the pre-defined classes delineated in Section 4 forms the main task of our paper. Manually writing heuristic rules for question classification requires tremendous amount of tedious work. In contrast, machine learning approach can automatically construct a question classifier (Zhang and Lee, 2003). However, a pre-requisite to training a model for classification is the availability of a large number of queries categorised manually in accordance with the class-set. The larger the training data set, more accurate the future predictions will be (Godbole, 2008).

We considered classification of questions to be a standard text classification task. Question classification problem can be solved quite accurately using a learning approach (Li and Roth, 2002). From amongst the various classifiers available we applied sequential minimal optimisation (SMO) algorithm (Platt, 1999) for supervised classification of the queries. SMO implements Platt's SMO algorithm for training a support vector classifier. Support vector machines (SVMs) is a binary classifier that learns a hyperplane in a feature space that acts as an optimal linear separator which separates (or nearly separates) a set of positive examples from a set of negative examples with the maximum possible margin (the margin is defined as the distance from the hyperplane to the closest of the positive and negative examples). SVMs have been widely tested to be one of the best machine-learning classifiers, and previous studies have shown that SVMs outperform other machine learning algorithms for open domain sentence classification (Zhang and Lee, 2003) and other text categorisation tasks (Sebastiani, 2002). Though SVMs are inherently two-class classifiers, there are various methods to do multi-class classification. In SMO, multi-class problems are solved using pair-wise classification (1-vs.-1).

Given the limited number of drug related information SMS queries available to us, the enormous size of the medical vocabulary as well as the huge number of drugs available today, we decided to generalise each of these queries in such a way that each query can be used for any drug or symptom. To achieve this, the medical entity extractor in block II replaces the medical entities in the query by their concepts.

SVMs perform classification by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories. In the linear case, the margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples. The formula for the output of a linear SVM is

$$u = \boldsymbol{w}.\boldsymbol{x} - b$$

where, w is the normal vector to the hyperplane and x is the input vector. The separating hyperplane is the plane $u = 0$. The nearest points lie on the planes $u = \pm 1$. By using geometry, we find the distance between these two hyperplanes is $2/\| w \|$, so we want to minimise $\| w \|$. Maximising margin can be expressed via the following optimisation problem

$$\text{Minimise over } (w, b) \| w \| \text{ such that } y_i(w.x_i - b) \geq 1 \text{ for every } i$$

where x_i is the ith training example, and y_i is the correct output of the SVM for the ith training example. The value y_i is +1 for the positive examples in a class and –1 for the negative examples.

6 Performance analysis

We collected an aggregate of 275 queries each of length less than 160 characters from university students studying engineering field and from a drug information forum at (Infonet, 2010). The literacy level of these users in the drug information was moderate. We would rate the literacy rate to be 4 on a scale of 10 where 10 is the highest level of expertise in the drug information domain. We consider these queries as representative

query set for the actual SMS queries that our SMS system will receive. Only the queries which were unambiguous and belonging only to one class were considered. The distribution of these queries across the 16 classes is as given in Table 5.

Table 5 Distribution of training instances amongst various query classes

(*a*)		(*b*)	
Side effects	73	Contraindications	6
Substitutes	58	Duration of action	6
General information	29	Dose miss	5
Things to avoid	23	Pregnancy	5
Interactions	19	How to take	4
Indications	17	Onset of action	4
Dosage	10	Old age	4
Overdose	6	Breast feeding	4

Various classifiers were trained over this dataset and the performance was obtained with different number of features used for classification. In our first experiment, we manually selected the attributes. Out of a list of 249 features generated from the processing steps described in Section 5.2, a list of 42 features was filtered out. These selected attributes represent the keywords for a drug query class e.g. *side-effects*, *risks* for side-effects class; *contraindication* for contraindication class; *dosage* for dosage class, etc. With just these attributes, the classifiers will basically act as rule-based classifiers looking only for these keywords for classification. However, as explained in Section 2, queries can contain regular English words, which when analysed independently cannot point to a target class. To handle these cases all the 249 attributes were used for classification in our second experiment. The results reported in Table 6 shows the increase in accuracy of the classifier in the second experiment as compared to the first one.

To evaluate the performance of the classifiers we performed 10-fold cross validation on our training data and achieved best performance in the case of SMO classifier (parameter, $C = 1.0$ and $\varepsilon = 1.0\text{E} - 7$) with a recall of 79.9 and precision of 81% respectively.

Table 6 Performance of different classifiers on our training set

Classifier	*No. of features: 42*		*No. of features: 249*	
	Recall (%)	*Precision* (%)	*Recall* (%)	*Precision* (%)
SMO	62.8	71.8	79.9	81
Naive Bayes	62	72.8	76.6	78.1
J4.8 (decision tree)	61.7	73.8	71.5	71.3
Decision table	51.5	51.8	52.9	52.3

Decision tables simply generate rules, trying to make output of machine learning same as input. It does not give good results on new data set especially when the

test queries are not expected to have exactly the same features as in the training set. Decision trees also tend to over fit the training set and therefore do not perform very well on the test set.

With support vectors over fitting is unlikely to occur. In SVMs maximum margin hyperplane is relatively stable: it only moves if training instances are added or deleted that are support vectors. Over-fitting is caused by too much flexibility in the decision boundary. The support vectors are global representatives of the whole set of training points, and there are usually few of them, which gives little flexibility (Witten and Frank, 2011). Therefore support vectors perform better than others as the queries are expected to vary a lot and not many training instances were available. Since SVM works by pair-wise classification a total of 120 (16C2) binary classifiers were built. On analysis of the weights assigned to the features in each of these classifiers it was verified that the words like side-effects, contraindications, dosage got higher weights in comparison with other features.

7 Sample use cases

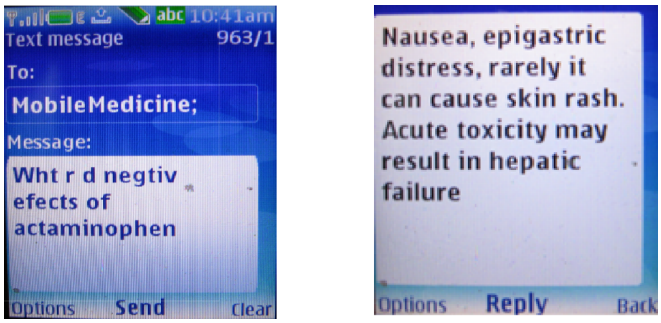
Snapshots demonstrating the actual usage of the drug QA system on SMS are given in Figure 3.

Table 7 gives some more examples of some expected queries and their corresponding answers.

Table 7 More SMS queries and their answers

Q: What is amvien used for
Ans: Short term treatment of lack of sleep(7-10 days)
Q: Wht r d conditions in wich acutance shud nt be taken
Ans: Acute inflammation of skin, acute eczema, inflammation of cheeks, nose, chin, abrasions, cuts
Q: Hw many times shud I tke ambien daily
Ans: Adults: 10 mg immediately before bedtime daily Elderly or weakened hepatic insufficiency: 5 mg before bed time
Q: Is Accutane advised in old age
Ans: Use with caution
Q: Are thr any risks of usin Zopidam
Ans: Symptoms of overdose include impairment of consciousness ranging from drowsiness to light coma
Q: give me sum brands same as adipen
Ans: Actedron Benzedrine Dexedrine

Figure 3 Snapshots demonstrating classification of misspelled SMS queries into their respective classes and their answers retrieved from our database (see online version for colours)



8 Discussion, implication and conclusion

SVMs has been successfully used in modelling complex, real-world problems such as text and image categorisation (Joachims, 1998), pattern recognition (Burges, 1998), bioinformatics (Byvatov and Schneider, 2003), etc. SVM has strong regularisation properties i.e., it can generalise the model to new data. SVM has been found to perform well on data sets with many features even if there are very few data instances to train the model. The accuracy of the SVM model for our drug queries application complies with results obtained in various other application domains. Some similar experiences can be found in Osuna et al. (1997), Zhang and Lee (2003), etc.

In this paper, we presented our novel approach of providing drug related information to people through natural language queries in a manner that meets their immediate need in a cost-effective manner. Through various sources (online drug reference sources in Appendix A) we identified the classes into which a drug query can be classified so that only the required information is extracted and presented to the information seeker. Using classical SVMs we achieved classification accuracy of 81% in this task. The biggest advantage of the system is that no installation of any kind is required on the user's mobile phone and the QA system can be quickly incorporated by a network service provider. Moreover, the system meets emergency needs in an inexpensive manner which otherwise would have required consulting a doctor.

The system as described in this paper requires minimal human involvement. The automated approach reduces response time for queries and at the same time offers a degree of flexibility to the user in terms of how they choose to formulate their query. This is clearly an advantage over DICs which require manned personnel to respond to user queries. The flexibility to ask natural language queries is advantageous to less privileged people who cannot afford to have access to internet enabled mobile phones. More importantly this system architecture can be extended to other domains as well. Even when mobile internet does gain popularity a system that interprets natural language queries potentially carrying spelling mistakes will still be useful. This system can also be used as an up-to-date reference tool by less experienced doctors practicing in villages where internet is not available.

From an academic perspective, development of such systems motivates development of new algorithms to increase the classification accuracy for multi-class classifier. If we extend the same system to multiple domains, better classification techniques need to be introduced to first classify a question to the domain it belongs to and then classify the class of question within the domain. This requires the use of multi-level classification techniques like hierarchical SVMs (Yu et al., 2003). One stage would be required to determine the domain of the query and another to determine the class of question within the domain.

Extensive user testing is necessary to eventually make the system available for users. We have already conducted preliminary user testing of related systems which utilise NLP techniques for processing SMS queries targeted on the healthcare domain (Langer and Banga, 2010; Banga et al., 2010; Langer et al., 2010). *Find a Doctor* (Langer and Banga, 2010) processed free form SMS queries to extract doctor name, address and the health problem to find a closest doctor for the health problem. We developed another system (Langer et al., 2010) which performed a search over Yahoo! Answers in health domain to find close matching answers to related questions for an SMS question with potential short forms typically used in SMS domain. We received encouraging feedback from the people who tested these systems because of their flexibility. We learnt that incorporating flexibility in a SMS based system makes it very interesting both from the user and developer points of view. We believe ours is a preliminary system that leverages well established algorithms from NLP and machine learning to reduce the workload of DICs and also sets an example for other factoid based systems.

In this paper we have not considered thoroughly the intricacies of SMS texting language (besides handling common spelling mistakes). Incorporating language models and noisy query analysis can further improve the user experience in such a system by relaxing the rules on the types of questions handled. Larger training set that can be obtained once the system is taken into pilot stage with a network service provider will help us better understand the user behaviour and the kind of queries expected through SMS. Based on this analysis the system can be adapted to handle wider class of query. Current available drug databases provide us with more general information e.g., it will give all the side-effects of a drug even if information on a particular side-effect is asked. Either developing a database having detailed description can be done or searching the online drug reference sources for matching FAQs can be done to retrieve specific information. This will enable the system to answer queries like What is the side-effect of <drug_name> on <organ>?, Does <drug_name> cause <symptom>?, etc. A functional system will require efforts in the right direction to ensure adoption of service by the majority of population. In their study, Lee et al. (2012) discuss the effects of perceived enjoyment for services adoption.

References

Acharyya, S., Negi, S., Subramaniam, L. and Roy, S. (2008) *Unsupervised Learning of Multilingual Short Message Service (SMS) Dialect from Noisy Examples*, ACM, Singapore.

Agarwal, A. (2008) 'Using domain specific question answering technique for automatic railways inquiry on mobile phone', *Fifth International Conference on Information Technology: New Generations, 2008. ITNG 2008*, 7–9 April, Las Vegas, NV, pp.1111–1116.

Aw, A., Zhang, M., Xiao, J. and Su, J. (2006) *A Phrase-based Statistical Model for SMS Text Normalization*, ACM, Sydney, pp.33–40.

Banga, R., Langer, A., Mittal, A. and Sondhi, P. (2010) *SMS Based Natural Language Interface for Locating HealthCare Service Providers*, Telecom Regulatory Authority of India, India, April.

Burges, C. (1998) 'A tutorial on support vector machines for pattern recognition', *Data Mining and Knowledge Discovery*, Vol. 2, pp.121–167.

Byvatov, E. and Schneider, G. (2003) 'Support vector machine applications in bioinformatics', *Applied Bioinformatics*, Vol. 2, p.67.

Chong, J-L., Chong, A.Y-L., Ooi, K-B. and Lin, B. (2011) 'An empirical analysis of the adoption of M-learning in Malaysia', *International Journal of Unications*, Vol. 9, No. 1, pp.1–18.

Crowell, J., Zeng, Q., Ngo, L. and Lacroix, E. (2004) 'A frequency-based technique to improve the spelling suggestion rank in medical queries', *J. Am. Med. Inform. Assoc.*, Vol. 11, No. 3, May–June, pp.179–185.

Cruchet, S., Gaudinat, A. and Boyer, C. (2008) 'Supervised approach to recognize question type in a QA system for health', *Stud. Health Technol. Inform.*, Vol. 136, pp.407–412.

Damerau, F. and Mays, E. (1964) 'A technique for computer detection and correction of spelling errors', *Communications of the ACM*, Vol. 7, No. 3, pp.171–176.

Ely, J.W., Osheroff, J.A., Ebell, M.H., Chambliss, M.L., Vinson, D.C., Stevermer, J.J. and Pifer, E.A. (2002) 'Obstacles to answering doctor's questions about patient care with evidence: qualitative study', *British Medical Journal*, Vol. 324, pp.710–713.

Godbole, S. (2008) *Text Classiffcation, Business Intelligence and Interactivity: Automating C-Sat Analysis for Services Industry*, 24–27 August, ACM, Las Vegas, Nevada, USA.

Google (2010) http://trends.google.com/websites?q=webmd.com

GSpell (2010) Url: http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/gSpell/current/GSpell.html

Health, R. (2010) *Revolution Health Website*, http://www.revolutionhealth.com/

Infonet, D. (2010) *Drug Infonet Website*, http://druginfonet.com/index.php?pageID=drug_faq.htm

Joachims, T. (1998) 'Text categorization with support vector machines: learning with many relevant features', *Machine Learning: ECML-98*, Vol. 1398, pp.137–142.

Kopparapu, S., Srivastav, A. and Pande, A. (2007) 'SMS based natural Language Interface to yellow pages directory', *Mobility '07 Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology*, 10–12 Septemeber, Singapore, ACM, pp.558–563.

Langer, A. and Banga, R. (2010) *Providing Natural Language Interface to Health Care Services Over SMS*, Project Report Submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering, IIT Roorkee, India.

Langer, A., Banga, R., Mittal, A. and Subramaniam, L.V. (2010) 'Variant search and syntactic tree similarity based approach to retrieve matching questions for SMS queries', *AND '10 Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, ACM, New York, NY, USA, pp.67–72.

Lee, S., Noh, M-J. and Kim, B.G. (2012) 'An integrated adoption model for mobile services', *International Journal of Mobile Communications*, Vol. 10, No. 4, pp.405–426.

Levenshtein, V.I. (1966) 'Binary codes capable of correcting deletions, insertions and reversals', *Soviet Physics Doklady*, Vol. 10, pp.707–710.

Li, X. and Roth, D. (2002) 'Learning question classifiers', *Proceedings of the 19th International Conference on Computational Linguistics*, Vol. 1, Association for Computational Linguistics, pp.1–7.

Lin, S.P. (2011) 'Determinants of adoption of mobile healthcare service', *International Journal of Mobile Communications*, Vol. 9, No. 3, pp.298–315.

Osuna, E., Freund, R. and Girosit, F. (1997) 'Training support vector machines: an application to face detection', 1997 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997. Proceedings*, 17–19 June, IEEE, San Juan, pp.130–136.

Patrick, P.L. (2011) 'Content relevance and delivery time of SMS advertising', *International Journal of Mobile Communications*, Vol. 9, No. 1, pp.19–38.

Platt, J.C. (1999) 'Fast training of support vector machines using sequential minimal optimization', *Advances in Kernel Methods*, MIT Press, pp.185–208.

Porter, M. (1980) 'An algorithm for suffix stripping', *Program*, Vol. 14, pp.130–137.

Priporas, C-V. and Mylona, I. (2008) 'Mobile services: potentiality of short message service as new business communication tool in attracting consumers', *International Journal of Mobile Communications*, Vol. 6, No. 4, pp.456–466.

Schusteritsch, R., Rao, S. and Rodden, K. (2005) *Mobile Search with Text Messages: Designing the User Experience for Google SMS*, 2–7 April, ACM, Portland, Oregon, pp.1777–1780.

Sebastiani, F. (2002) 'Machine learning in automated text categorization', *ACM Computing Surveys (CSUR)*, Vol. 34, No. 1, March, pp.1–47.

Suh, Y., Lee, H. and Park, Y. (2012) 'Analysis and visualisation of structure of smartphone application', *International Journal of Mobile Communications*, Vol. 10, No. 1, pp.1–20.

Sun, J., Koong, K.S. and Poole, M.S. (2009) 'Critical success factors for context-aware mobile communication systems', *International Journal of Mobile Communications*, Vol. 7, No. 3, pp.290–307.

Tounsi, M. (2008) 'A bluetooth intelligent e-healthcare system: analysis and design issues', *International Journal of Mobile Communications*, Vol. 6, No. 6, pp.683–695.

Witten, I.H. and Frank, E. (2011) *Data Mining: Practical Machine Learning Tools and Equipments*, 2nd ed., Morgan Kaufmann Publishers.

Yu, H. and Sable, C. (2005) 'Being Erlang Shen: identifying answerable questions', *Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasoning for Answering Questions*.

Yu, H., Sable, C. and Zhu, H. (2005) 'Classifying medical questions based on an evidence taxonomy', *Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains*.

Yu, H., Yang, J. and Han, J. (2003) 'Classifying large data sets using SVMs with hierarchical clusters', *KDD '03 Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, USA, pp.306–315.

Zhang, D. and Lee, W. (2003) 'Question classification using support vector machines', *SIGIR '03 Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 28 July–1 August, ACM, New York, NY, USA, pp.26–32.

Notes

[1]http://www.mciindia.org/tools/announcement/MCI_booklet.pdf

[2]Taken from www.medterms.com

Bibliography

Bouvy, M.L., Meijboom, R.H.B., van Berkel, J. and de Roos-Huisman, C.M. (2002) 'Patients' drug-information needs: a brief view on questions asked by telephone and on the internet', *Pharmacy World and Science*, Vol. 24, No. 2, pp.43–45.

Choudhury, M., Saraf, R. and Jain, V. (2007) 'Investigation and modelling of the structure of texting language', *IJDAR*, Vol. 10, Nos. 3–4, pp.157–174.

Emmanuel, P., Christian, V. and Emmanuel, M. (2007) 'Language models for handwritten short message services', *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, Vol. 1, pp.83–87.

Even-Zohar, Y. and Roth, D. (2001) 'A sequential model for multi class classification', *EMNLP-2001, the SIGDAT Conference on Empirical Methods in Natural Language Processing*, pp.10–19.

Fontelo, P., Liu, F., Muin, M., Tolentino, H. and Ackerman, M. (2006) 'Txt2MEDLINE: text-messaging access to MEDLINE/PubMed', *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, Vol. 2006, p.259.

Kothari, G., Negi, S., Faruquie, T.A., Chakaravarthy, V.T. and Subramaniam, L.V. (2009) 'SMS based interface for FAQ retrieval', *ACL '09 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*, Singapore, pp.852–860.

Langer, A., Kumar, B., Mittal, A. and Subramaniam, L.V. (2009) 'Mobile medicine: providing drug related information through natural language queries via SMS', *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pp.546–551.

Weka (2010) *Weka*, http://www.cs.waikato.ac.nz/~ml/weka/index.html

Yu, H., Lee, M. and Kaufman, D. (2007) 'Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians', *Journal of Biomedical Informatics*, Vol. 40, No. 3, June, pp.236–251.

Appendix A

Some of the drug reference sources available for getting information about drugs including their therapeutic usage, adverse-effects, toxicity, dosage, etc.

Online drug reference sources

1 Revolutionhealth.com

2 Mayoclinic.com

3 Medclik.com

4 RxList.com

5 WebMD.com

6 Druginfonet.com

7 Alberta.com

8 Drugs.com

9 Drugdigest.com