Vol 5. No. 4, June 2012
African Journal of Computing & ICT
ISSN 2006-1781

www.ajocict.net

Evaluating SMS Parsing Using Automated Testing Software

[1]A.O. Adesina, K. K. Agbele, A.P. Abidoye & N.A. Azeez
Department of Computer Science, Faculty of Natural Science,
University of the Western Cape, Private Bag X17
Bellville, South Africa
inadesina@gmail.com, agbelek@yahoo.com, ademola.abidoye@gmail.com, nurayhn1@yahoo.ca

[1]Corresponding author

ABSTRACT
Mobile phones are ubiquitous with millions of users acquiring them every day for personal, business and social usage or communication. Its enormous pervasiveness has created a great advantage for its use as a technological tool applicable to overcome the challenges of information dissemination regarding burning issues, advertisement, and health related matters. Short message services (SMS), an integral functional part of cell phones, can be turned into a major tool for accessing databases of information on HIV/AIDS as appreciable percentage of the youth embrace the technology. The common features by the users of the unique language are the un-grammatical structure, convenience of spelling, homophony of words and alphanumeric mix up of the arrangement of words. This proves it to be difficult to serve as query in the search engine architecture. In this work SMS query was used for information accessing in Frequently Asked Question FAQ system under a specified medical domain. Finally, when the developed system was measured in terms of proximity to the answer retrieved remarkable results were observed..

Keywords: Short Message Service (SMS), SMS Parsing, SMS Normalization, Text Normalization, Keyword Extraction

1. INTRODUCTION

The introduction of telecommunications technologies in healthcare environment has led to an increased accessibility to healthcare providers, more efficient tasks and processes, and a higher overall quality of healthcare services [1-6]. However, these achievements are not without challenges, like medical errors [7-8], increasing cost of healthcare services and a partial coverage of healthcare services in rural and underdeveloped areas [9-10]. Mobile communication has come to ameliorate some of the challenges in the area of consultancy, treatment, drug prescription and administration, and laboratory tests results.

The mobile phone has transformed from a device for voice communication to advanced computing machine. Services offered on mobile phones range from advertisements, opinion polls, transaction alerts, Internet browsing, Internet banking to e-commerce. However, SMS stands out as the most frequently used service on any mobile phone; from the low to high end handsets, thus, many information based services are built around this technology, for example, translation of SMS text into different languages so as to communicate without considering the recipient preferred language of choice [11]. SMS has given rise to communication in a unique and continually evolving form of language.

African Journal of Computing & ICT Reference Format:
A. O.. Adesina, K. K. Agbele, A.P. Abidoye & N.A. Azeez (2012). Evaluating SMS Parsing Using Automated Testing Software. Afr J. of Comp & ICTs. Vol 5, No. 4. pp 53-62

SMS communication has its own specificities, where groups of users have their own patterns of writing, inventing new abbreviations, and using non-standard orthographic forms [12]. However, this nature of communication, with non-standard abbreviations, transliterations, phonetic substitution and omission presents difficulties in building information access systems, such as FAQ databases. In this research, the focus is on addressing how such peculiarities can be overcome in a manner that SMS can be used as a means of accessing information. The study considers the normalization and parsing of SMS text using both Medical and English dictionaries as database; having the awareness of the query on health related matters will be a combination of the two languages. The query is then used to make enquiry in the FAQ database for information searching and retrieval.

Therefore, the evaluation is carried out to ascertain the possibility of using SMS text to extract information from the database of FAQ system using AIDS/HIV medical domain. In order to achieve this, SMS was parsed and transformed into its original English spelling that can be used to query the search engine and retrieve information needs of the user. SMS was used as a tool for accessing information on FAQ system about HIV/AIDS from English and Medical dictionaries built mainly for the purpose of this study. In this research, the use of automated testing software (Quick Test Professional - QTP) is incorporated to assist in parsing of the computer-mediated language.

QTP is a Mercury Interactive powerful functional testing tool that is useful in testing Web objects, ActiveX controls and Visual Basic Control package written in Visual Basic scripting language. QTP

undergoes the processes of creating, running and analysing SMS input before being used as a query to generate results. The choice of QTP was for the application to work on web. Manual test of applications on the web have been accompanied by drawbacks. It is time-consuming and tedious, requiring heavy investment in human resources. Consequently, time constraints often make it impossible to manually test every feature thoroughly before the application is released. Automated testing with QTP addresses these problems by dramatically speeding up the testing process. Tests that check all aspects of application or Web site can be created and ran every time the site or application changes.

2. RELATED WORKS

There are different information transmission technologies in mobile phones nowadays which include SMS, Multimedia Messaging service (MMS), Voice and Video Technologies. The most common among them is the SMS because of its convenience, cost and the message could be relayed even in noisy places [13].

Some of the challenges in the usage of SMS as query in an information retrieval system are the variance of a singular word to various formats aggravating its use as a search engine. It is important that the disambiguate-nature of the language be worked on to give a term that can be used as a query for the searching. SMS has to be translated to the right English word which involves the technology referred to as SMS normalization and subsequently used for information accessing (i.e. parsing).

Parsing is the process of structuring a linear representation in accordance with a given grammar. For each grammar, there is generally an infinite number of linear representation or sentences that can be structured with it [14]. A finite-size grammar can supply structure to an infinite number of sentences. Parsing may involve reconstructing the production rules that indicate how the given string can be produced from the given grammar. Parsing can split a sequence of characters or values into smaller parts then used them for recognition of the characters or values that occur in a specific order.

However, there are different types of parsers- for example context free grammar (CFG) based on grammars and trees construct applying grammar productions rules. Bits of a tree are combined together to build up large trees in different variety and structure. Other types of parsing techniques are top-down and bottom-up models. The latter attempts to match the input to the grammar by considering the derivations of a grammar non-terminal, starting at the abstract level (the level of sentences) and breaking it down to the most concrete level (the level of words), as it traverses, it makes references to the production rules so as to know the right rules to apply. It commences at the goal symbol and move down to the leaf nodes. The former is the complete opposite of the top down which traverses from the leaf nodes to the goal symbols, asking and confirming what non-terminals the right hand elements reduce to [15].

In addition, the Example-based (EB) parsing approach [16] tries to analyse a word or sentence by finding the solution of a problem previously had. This will be most similar to the problem to be solved using a similar or closely related word or sentence as the input and output will be the same structure. With this the output can be re-used as part of the training data and thus enhancing the improvement and accuracy of the system. It learns how to interpret a new utterance by looking at some used examples and work on them. Other EB approach is the SMS parsing [17] that handles senders and process, analyse, parse and reformat the messages. An EB parser which uses statistics and learns to interpret new and recent utterances by keeping the track records of the way it was used in the past was proposed by Genereux [18].

In May 2000, EB parser for Chinese was introduced [19], this proved to be very reliable where a formal definition and derivation for its metric parsers' evaluation was confirmed. The rows of experiment that were collected to identify factors that support reliability of the parser are independent of the parsing approach. The parser retrieves trees from a tree bank via a fuzzy match of the sentence to be parsed as well as the terminal of all the trees in the tree bank.

The EB approach was adopted in this research as it applies to comparing the variants of SMS to its root words in its own native or original language. The researchers introduced some concluding or decision rules so as to determine the accuracy of the SMS being parsed in a situation of tying.

2.1 SMS Pre-Processing

Pre-processing of SMS words is a vital stage prior to its parsing as repeated words are pruned,(e.g. "yeeeeeeeessss!" can be pruned to "yes!" or "Its 222222222222222 much!!!!!!" can be pruned down to "Its 2 much!"), homophone words can be considered (i.e. gr8 can be considered as great), stop lists of words can be expunged (e.g. a, at, in, is, of, on, the, that, to etc) [20-21]. One important process common to SMS corpus after collection is that word that is not reasonable or sensibly in terms of information dissemination is discarded. It may be by reducing or totally eradicating the stop word i.e. removing most frequently used words that also exist in the web document by using a stop word dictionary or word stemming that is, reducing the occurrence of term frequency, which has similar meaning in the same document [22]. The pre-processing stage could also be in form of reassembling messages of more than 160 characters that were outside the capacity of the handset; this message may have split over to the next page. Other messages that may be removed in the pre-processing stage are commercial, graphical SMS, emoticons and any other outside the scope of this research [12]. The application of rules or protocols in research is also common in SMS corpus [12] as it restores normalcy or standardisation and preserves as much as possible the original messages and their meanings. Normalization is primarily to convert the selected SMS keywords to its right English word. With this, there is a better chance to use the query words for parsing, which can then return answer in a FAQ environment that the experiment is set up for.

The study objective is therefore to pre-process the chosen keywords for the SMS inputs that can lead to matching of question to relevant answer. Frequently asked questions were selected from

about ten different websites. These websites are concerned about the enlightenment, awareness, prevention, cure and education of the sexually transmitted disease to sexually active people. Keyword or phrases are selected from each question; which may be one word or combination of words to make the keyword or key-phrase. The selections are tagged and mapped with its SMS variants. Each question can stand with its frequently occurring words or phrases and use to query the database of carefully selected website questions that have been prepared in the database table. Average of 400 questions were selected from all the sites. Some of the questions extracted are as follows:

- What are antiretroviral drugs?
- What are the main routes of HIV transmission?
- Can I be infected if my partner does not have HIV?
- Can I get infected with HIV through biting?
- Does circumcision protect against HIV?
- What if I test positive?
- Where can I get tested for HIV?
- How do I prevent transmission of HIV?
- Why do people who are infected with HIV eventually die?
- Can I get HIV from a mosquito bite?
- How will I be sure that my future marriage partner is not infected with HIV?
- What is window period?
- How long does HIV virus lives outside the body?

SMS language expression comes in divergent forms. Getting a Machine Learning technique to be able to synchronize different ways has been a great challenge. There may be curiosity, as in Table 1, to the knowledge of some common questions like: What is HIV? What is AIDS? How is HIV passed on? How does the HIV test work? Where can I get tested? What is the window period? etc Considering these questions will create problem for its retrieval mechanism especially when they are all written in SMS language because a sample of ten SMS users mean ten different ways of writing the same question.

Table1: Examples of some of the SMS Queries and their corresponding SMS representations

Query	SMS Representation
What is HIV?	*Wat is hiv, Wats hiv, Watz hiv, Wtz hiv Wht is hiv, Wat's hiv*
What is AIDS?	*Wat is aids, Wht is aids, Wots aids, Wt s aids, Wt's 8s*
How is HIV passed on?	*Auz hiv pasd on, Hows hiv psd on, Ao's HIV pssd on, Hws HIV pasd on*

This research focuses on the problem of using SMS communication to access information related to HIV/AIDS. It is proposed to provide access to carefully screened information on HIV/AIDS within the context of the Frequently Asked Question (FAQ) system. The answers to commonly asked questions reside in the database server and the SMS query will search and return the best answer. Information retrieval process begins by a user entering a query into the system. A query is an unknown request for user's information needs. The queries are entered into the retrieval system which translates the query into an understandable and meaningful form. This is made possible by use of the keywords/index terms. The keywords/index terms are then used for the string matching and all the relevant documents from the databases are retrieved. Typical information retrieval technique models are shown in Fig. 1 below

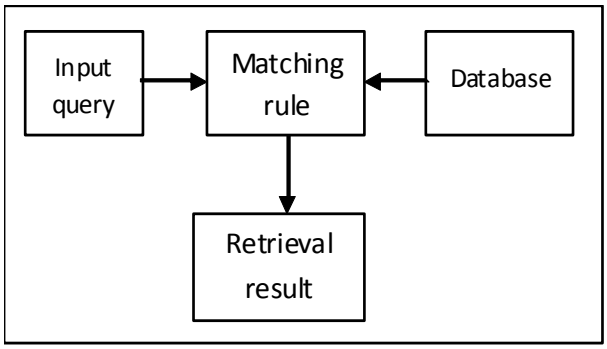


Fig. 1 General model for information retrieval system

However, automating SMS-based information search and retrieval poses significant challenges because of the inherent noise in its communication. The SMS messages were then analyzed, transcribed and classified with the aim of building a dictionary of SMS speech to English translation with reference to HIV/AIDS.

2.2 Advantages of Using SMS as a Communication Medium

There are other Computer Mediated Communication (CMC) languages like emails, SNS (Social Network Sites), IM (Instant Messaging) but the choice of SMS may be for the following reasons:

- It is cheaper to send/receive compared to the voice call [23].
- It is non-intrusive thus one can write or read an SMS in a meeting, bus, stadium, cinema hall etc and nobody hears you sending the message nor can one decipher what the incoming message is all about.
- It is persistent; it stays until the phone is checked or switched on.
- It enables direct conveyance of the message without interruption from the recipient. This ensures one way privacy i.e. one has time to compose and send a message unlike in normal conversation when the recipient interrupts or interferes with your statement.
- It offers a choice whether to reply, forward, or delete. Some phones now have delivery reports such that the sender is notified when the message is read or deleted. Thus a reply is expected but the good thing is that SMS gives one ample time to figure out the best possible reply.

- It can be saved for future reference unlike the spontaneous spoken word.
- It can be short, casual and precise.

3. METHODOLOGY

This research has proposed application of text messaging as a means of access to health care information. The preliminary stage of the experiment referred to as the corpus preparation/data collection stage has been executed. It involved the collection of SMS messages from a group of First year Computer Science students of the University of the Western Cape, South Africa totalling 150 and a set of 25 questions were administered. The SMS communication appeared in different forms because of flexibilities associated with this form of communication. A set of pre-formed questions related to HIV/AIDS were provided for all participants. The participants were then required to rewrite the same questions, assuming they were personally sending the same question via an SMS message transmission.

Three different methods were used in collecting question data sets: MXit platform, blue-tooth and hand-written of the SMS messages. MXit [24] is a free, instant messaging (IM) application developed in South Africa that runs on 3G General Packet Radio Service (GPRS/3G) mobile phones and on PCs. This technology allows the user to send and receive one-on-one text and multimedia messages to and from other users, as well as in general chat rooms. The HIV/AIDS FAQ were typed by the students and sent to the dedicated mobile device. There stand a risk of data corruption if the data are transferred through different media before being deposited into the server, but with this approach there was no need to copy the data through any other medium to the server. Blue -tooth became useful for some participants who did not subscribe to the previous technology. The information transfer in this Blue-tooth technology was as pure as in MXit. The hand -written technique involve the participants writing on the hard copy the way it will be written if they were to use a mobile set. This was a rigorous exercise as every other thing that was written will have to be keyed into the database. This method is prone to error during transcription unlike other methods. None of the methods adopted in data collection incurred any cost because of the peculiarity of our participants' environment. For all the methods used, a laptop was configured to serve as database server. It received all forms of text messages from the participants the way they all responded to the questionnaires.

4. SMS NORMALIZATION

This involves translating SMS-English to its original English root. The technique used in the translation services includes dictionary substitution without language model (LM). Other approaches are rarely employed in disambiguating between possible word substitution [25].
A noisy channel approach is a reasonable alternative to dictionary substitution in modeling the translation from SMS English to English. English SMS as a corruptible signal is sent across a noisy channel, in an attempt to translate to perfect English by using language and translation models. Using the models will be able to give the ability to disambiguate between ambiguous expansions of an SMS message. For example, in "whts 8s?" we need to disambiguate what each of the tokens means, specifically the 8; it could map to be a "figure" or "series of 8" or "AIDS". With machine translation (MT) there could be a better alternative for SMS English to English translation, known as SMS text normalization in the natural language processing (NLP) community.

Basically, MT has been used on some works for SMS translations. SMS messages were normalized before being translated to Chinese language [20]. Dictionary substitution approach using frequency and bigram language model were compared to the use of MT under phrase-based machine translation. The result accounted for better results as MT boosts Bilingual Evaluation Unit (BLEU) scores for SMS English to English translation. BLEU is a metric for comparing the performance of the translation. [25]and [26] used 5000 SMS messages for their experiment from NUS corpus. The corpus has no translated SMS messages as a parallel English text, so they were produced by themselves. The study by Aiti [20], did not give report on how smiley faces were handled nor did it make mention of the translation of the punctuations. No standard tools, like MT tool, were used for the experiment but they had provided the tools and trained their n-gram language model on the English Gigaword corpus. There is no report of the test on NUS corpus and the use of the trigrams BLEU scores for the evaluations [20].

Spelling correction can stand better in solving the problem of SMS normalization. The only challenge is that the spelling correction will not consider the context of the language and cannot handle forms that are two words, for example shorthand form ("T42" being interpreted as "tea for two") to write its text may be very difficult. In homophony approach, the concept of speech recognition is essential in this kind of machine dialogue because the input is mostly full of errors as the user assumes that enough hints has been given before sending the text. Autocorrect, as part of several systems were designed to correct common typing or spelling errors automatically, changing two initial capitals, capitalizing both the first letter of a day and the first letter of a sentence. Several systems can analyse a text file for potential spelling errors, pointing out probable correct spellings. Spelling programs are expected to be a standard part of all text processing systems. Several companies, including Microsoft, have been working on two types of spelling programs: spelling checkers and spelling correctors. In the former, the input file of text find those words which are incorrect while the latter detects the misspelled words and tries to find the best substitute from the range of correct choices of words provided. This problem contains elements of pattern recognition and coding theory [26].

Error types are many hence there are different algorithms to solve each. The goal of this research is to resolve or reduce the orthographic nature of the SMS messages in text entry so as to form part of the method for SMS normalization; mostly word processor use string manipulation or matching technique to resolve the earlier work of finding and correcting errors brought about by specific input devices in specific contexts. For instance, the

concern may be to detect the potential of misspelled names of passengers for a specific airline flight, hotel reservation, population census registration, examination registration, and banking transaction as either the stored or inquiry name (or both) might be misspelled.

The approach of spelling correction used in this study has a way of making SMS normalization different from that which is used in Word Processing. It is an adaptive method, which submits corrections from the variance of the SMS inputs. In a situation where there is a tie, a certain rule system or decision can be applied to make the rightful selection within a specified domain so as to pick the likely English that will be used to interact with the database. Rather our system makes rapid automatic corrections since it has few seconds to normalize the input, update, and plan for the appropriate response from the database. It returns this selection into sentences and displays those sentences on the user's screen.

The parser accepts several replacements in order to build the database of many misspelled strings. The results of the algorithm in normalizing our SMS text make a selection that facilitate better results when our SMS query is sent into the database in a FAQ system. Information technique is adaptive as better function is returned for an answer. For the SMS normalization, the parser accepts several ways of texting and replaces the spelling corrector and selects the best by applying syntactic and context-dependent selection process.

5. GLOSSARY OF HIV/AIDS TERMINOLOGIES

According to [27], Dictionary is defined as a book that gives a list of the words of a language in alphabetical order; and explains what they mean, or gives a word for them in a foreign language e.g. a Spanish-English. It could be an electronic version where over 14,000 words were compiled and worked on. The dictionary version of Collins [28] are stored in the database for retrieval purpose.

This experiment uses two sources of database, English and Medical dictionaries. The latter is a collection of HIV related items that was compiled from 31 sources comprising of medical desk dictionary, Webster's encyclopaedia, white paper, research reports, journals, AIDS research paper. It opens up words that are commonly used in describing HIV-virus, epidermis, pathogenesis, treatment and their medical management in terms of therapy under related conditions. This indexed term contains up-to-date terms associated with HIV/AIDS with few technical terms omitted because of the enormity of its undertaking.

One of the unique things in the Medical dictionary is the irregularity of the word count unlike in English dictionary, 2 or 3 words or terminologies can be joined to describe a phenomenon e.g. cell membrane or central nervous system.

Our dictionary consists of 20,000 words and phrases from English and Medical terminologies and languages. These languages are mutually exclusive to each other hence; there is need to combine them for evaluation. Furthermore, there may be need to compliment medical issues with English word and vice versa.

5.2 The Structure of Medical Terminology

From Table 2, most of the terms used in Medical science can be broken into four word parts: roots, prefixes, suffixes and linking vowels like "o". The words could take their origin from Latin or Greek words, and they all stand for specific meanings. The table below shows some example:

An unknown term can often be understood if you know the meaning of common word parts. For example, the term pericarditis can be broken down into the root "card", which translates to "heart"; the prefix peri-, which translates to "surrounding"; and the suffix -itis, "inflammation." Pericarditis means an inflammation of the area surrounding the heart.

Table 2. The structures of Medical terminology

Roots	Prefixes	Suffixes
abdomin: abdomen	**endo**-:within	-**gen**: substance or agent that causes
acr: extremities; height	**hypo**-: below; deficient	-**itis**: inflammation
card: heart	**hyper**-: excessive	-**ologist**: one who studies and practices
home: sameness; unchanging	**peri**-: surrounding	-**plasty**: plastic or surgical repair
laryng: larynx	**tachy**-: fast; rapid	-**scopy**: visual examination
ot: ear	**mal**-: bad	-**stasis**: control; stop
path: disease		
vas: vessel; duct		

6. ARCHITECTURE OF AUTOMATED TESTING SOFTWARE

QTP is a regression and functional tool automation software used for enterprise quality assurance. The application supports keyword and scripting interface and features a graphical user interface. It uses the Visual Basic (VB) Scripting edition scripting language to specify a test procedure, and manipulate the objects and controls of the application under test. QTP facilitates creating tests on every application by recording operations as they are performed.

The tests are composed to become actions. Testing involves 3 main stages:

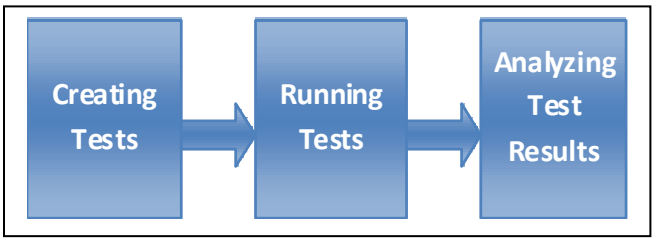


Fig. 2 QTP Testing stages

1. Creating Tests involves recording a session on the application, inserting checkpoints into every test that is made in the application. With this, the values and characteristics of the object or string are specified. The scope of the test is broadened by replacing the fixed values with parameters so as to check whether the application can perform the same operations with different data.
2. Running Tests is essential so as to check the application from line to line, checking the text strings, objects or table that are specified and lastly testing so as to debug the codes when they are identified with different commands.
3. Analyzing test results, results are viewed in the Tests Results window in form of a summary as well as detailed report and defects or errors that occur during a test run.

QTP is automatic testing software that supports keyword and scripting interfaces, file system operations, database testing, software application, web services and graphical user interfaces. This is a VB scripting language that is relevant in specifying a test procedure, manipulates and controls objects that are under tests. As parts of the strong weapons of QTP for the users, its ability to verify various aspects of an application, such as: the properties of an object, data within a table, records within a database, a bitmap image, or the text on an application screen and many more, motivates its choice as an application for this evaluation.

There is a strong capability for this automated testing software for exception handling; this scenario is to avoid running tests in case of an unexpected failure. During the experimentation, it was observed that if the application crashes, a message dialog box is expected to appear. With the compliments of Microsoft Excel, data can be uploaded to QTP for reusability by transferring to its data table. This happened in two ways - the Global data sheet and Action (local) data sheets. The test steps can read data from these data tables in order to drive variable data into the application under test, and verify the expected result.

7. RULES APPLIED FOR KEYWORDS SELECTION DURING SMS PARSING

1. SMS words are always shorter than the English counterparts.
2. First letter typed in by the user is assumed to be the first letter that starts the equivalent English word.
3. The arrangement of the characters in SMS is insignificant as it makes opportunity for the spelling errors that are acceptable in SMS words.
4. Homophonic SMS words, like 8, 4, 2, b, c, d, 9 assumed their corresponding interpretation of eight or ate or hate, for or four or fore, to or too, be, see, the, nine respectively.
5. One-letter words like a, b, c, d, e … may not be considered as either keyword or SMS word because they are considered as stop words.
6. In case there is a tie, words with lower possibility values will represent the parsed equivalence.

7.1 System Flowchart of the SMS Parsing

The input strings from the users are displayed and sent to the search engine to start the process of normalization. SMS normalization is a translational problem of converting an SMS language to the English language; it may involve the spelling correction, speech recognition, insertion of characters etc. The translation is being guided by some rules, as it is written in the session above and comparison is made with the English and Medical dictionaries that are part of the databases. For every SMS token parsed, there is a matching process of the input strings with the contents of the dictionary. If it is found in the dictionary, the tokens are parsed, but if not there will be need for the system to learn the word based on the positions of the characters on the token and the first letter that starts the token.

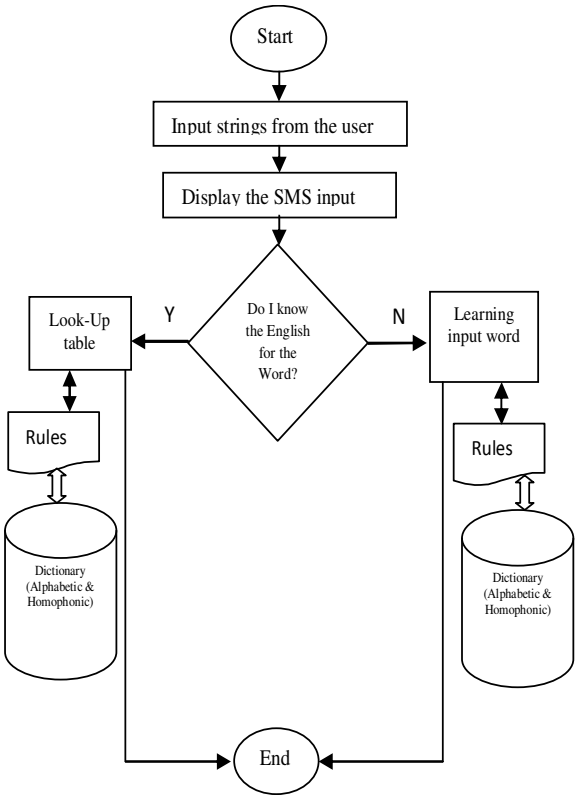


Fig. 3 Flowchart indicating the algorithm used in SMS parsing

The algorithms to decide the concluding rule in case there is a tie of the words that are parsed are shown below. It consists of six steps:

Algorithm for concluding rule

i. *get decimal values of each character that constitute the SMS word;*
ii. *get the values from the concatenation of the USER INPUT;*
iii. *get the values from the concatenation of the value in the database;*
iv. *calculate the possibility of the various SMS words given English word applying the Bayesian concept;*

Bayes' concepts

Let Pr be a probability measure on a probability space (Ω, Σ, Pr). Let Pr(A|B) denote the conditional probability of A given B.

Let $Pr(A) > 0$ *and* $Pr(B) > 0$.

Then:

$$Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$$

v. *if not available in the database*
system can learn and add to database
or
go to other search engine
vi. *then ask if it is alright to be added as the profile*

8. RESULTS

At the end QTP generates a test result. Using XML schema, the test result indicates whether a text parsed or not, shows error messages, and may provide supporting information that allows users to determine the underlying cause of a failure for possible correction.

The experiment of SMS query parsing was carried out on "What are antiretroviral drugs". Three words What, antiretroviral and drugs represents the keywords going by the judgment of some Linguists because of the relevance and importance of the words, "are" was identified as a stop word and was eliminated. Data was collected from about 400 First year Statistics students of the University of the Western Cape and the tables (Tables 3, 4, and 5) below represent the proportion of students that shows how the three terms were written.

Table 3. SMS Forms and Number

For *drugs:*

SMS forms	Number
drgs	185
dgs	65
drs	40
drg	85

Table 4. SMS forms and Number

For *What:*

SMS forms	Number
Wht	182
wat	70
wot	45
Wt	62

Table 5. SMS forms and Number

For *antiretroviral:*

SMS forms	Number
ARV	85
antirtrvrl	50
antirvrl	27
antrvral	45
antirtrvrl	38
antirviral	18
aretrval	15
antitroral	17
anietovl	15
antiroval	15
anttrol	56
anvral	25
aoviral	23
antirrol	14
antivirl	25
antirtroral	26
antroviral	35
antroviral	25
antioviral	14
antiviral	35
Others	12

This is the basis for our results and how the parsing was achieved. Using the algorithm before, we worked on the highest SMS form for each keyword identified and they were all parsed according to Fig 4 (i - iii) below. For instance "*wht*" a short code for "*what* " generated 21 words, "ARV" an SMS code for "Antiretroviral" generated 8 and for "drugs:" "drgs" a short-form of "drugs" generated 3 words from the corpus used. Further works need to be done on these results so as to reduce the number. The smaller the number the more accurate the results tend to be able to retrieve the answer or get better translation or interpretation to the SMS query.

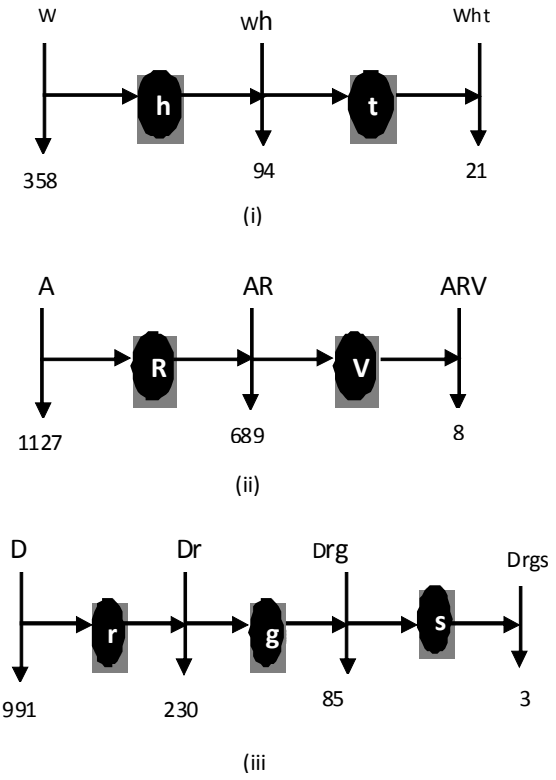


Fig. 4 (i- iii) Results of the Parsing of the keywords

Taking the N-grams concept and linking the individual unigram to represent the keywords in the query will give the likely translation of the short codes. As it were, this is work in progress. A n-gram is a contiguous sequence of n items from a given sequence of text or speech to identify and index word phrases [29]

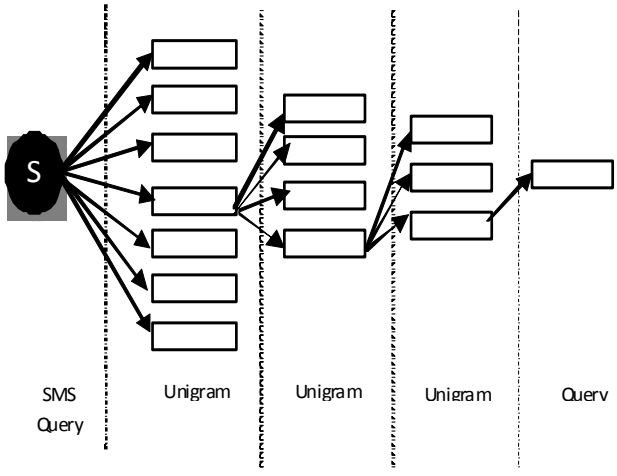


Fig. 5 Representation of the results

The result is represented in form of combinatorial search work, where the character sequence of the short-code is used to determine the keywords.

9. CONCLUSION

Firstly, there is an on-going work in the experiment; as arrangement is being made to apply further pre-processing techniques like stemming and clustering. These methods will not only reduce the number of query that will be parsed, but will also enhance the selection of the keywords so as to prune down the numbers of words that will be parsed by the end of the processing. At the end it is expected that the retrieval of information using SMS language will be improved remarkably especially when the technique is applied to a specific domain.

Secondly, more rules in terms of grammatical specification will be introduced in other to formulate an automatic keywords extractor using a *tf-idf* or other algorithmic keyword extraction methods.

Expectation at the end of this research is to enable the adoption of metric of Recall and Rejection to know the overall system performance of the experiment when each query is considered.

Accuracy of the SMS parsing is better noticed when the number of parses becomes more than three with respect to the keywords that was parsed.

10. FUTURE WORKS

The implementation of the project is still on. Further research works will be on the identification of the keywords from different grammar and medical contexts. The aim is to be able to have an overall parser that is 100% so as to enable users query to retrieve results regardless of the representation of the SMS lingo.

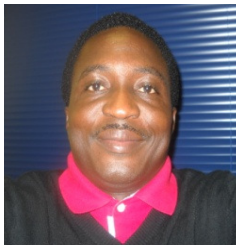
REFERENCES

1. Kern, S. E., and Jaron, D. (2003) Healthcare Technology, economics and policy: an evolving balance, IEEE Eng Med Biol Mag 22, 16-19.
2. Wells, P. N. T. (2003) Can technology truly reduce healthcare costs, IEEE Eng Med Biol Mag, 20-25.
3. Lin, J. C. (1999) Applying telecommunications technology to health-care delivery, IEEE Eng Med Biol Mag, 28-31.
4. Zhang, J., Stahl, J. N., Huang, H. K., Zhou, X., Lou, S. L., and Song, K. S. (2000) Real-time teleconsultation with high resolution and large-volume medical for collaborative healthcare, IEEE Trans Inf Technol Biomed 4, 178-185.
5. Holle, R., and Zahlmann, G. (1999) Evaluation of telemedical services, IEEE Trans Inf Technol Biomed 3, 84-91.
6. Lee, R. G., Shen, H. S., Lin, C. C., Chang, K. C., and Chen, J. H. (2000) Home telecare system using cable television plants- an experimental field trial., IEEE Trans Inf Technol Biomed 4, 37-44.
7. Kohn, L. T., Corrigan, J. M., and Donaldson, M. S., (Eds.) (1999) To err is human: building a safer health system.
8. Hayward, R. A., and Hofer, T. P. (2001) Estimating Hospital Deaths Due to Medical Errors The Journal of the American Medical Association (JAMA) 286, 415-420.
9. Singh, M. P. (2002) Treating healthcare, IEEE Internet Computing, 4-5.
10. Parsloe, C. (2003) World apart? Healthcare technologies for longlife disease management, IEEE Eng Med Biol Mag, 53-56.
11. Samanta, S. K., Achilleos, A., Moiron, S. R. F., Woods, J., and Ghanbari, M. (2010) Automatic language translation for mobile SMS, International Journal of Information Communication Technologies and Human Development (IJICTHD) 2, 43-58.
12. Cédrick, F., and Sébastien, P. (2010) A Translated Corpus of 30,000 French SMS, In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 770-779 Uppsala, Sweden
13. Tagg, C. (2009) A Corpus Linguistics Study of SMS Text Messaging, In Department of English, p 402, School of English, Drama and American and Canadian Studies, Birmingham.
14. Masizana-Katongo, A., and Ama-Njoku, T. (2011) Example-Based Parsing Solution for a HIV and AIDS FAQ System, International Journal of Research and Reviews in Wireless Communications (IJRRWC) 1, 59-65.
15. Anderson G., Asare S.D., Ayalew Y., Garg D., Gopolang B., Masizana-Katongo A., Mogotlhwane O., Mpoeleng D., and Nyongesa H.O. (2007) Towards a Bilingual SMS Parser for HIV/AIDS Information Retrieval in Botswana, In Proceedings of the second IEEE/ACM International Conference of Information and Communication Technologies and Development (ICTD), pp 329-333, Bangalore, India.
16. Copestake, A., Corbett, P., Murray-Rust, P., Rupp, C. J., Siddharthan, A., Teufe, S., and Waldron, B. (2006) An Architecture for Language Processing for Scientic Texts, In Proceedings of the UK e-Science Programme All Hands Meeting (AHM2006), Nottingham, UK.
17. Kobel, M. (2005) Parsing by Example, Institut fur Informatik und angewandte Mathematik.
18. Genereux, M. (2002) An example-based Semantic Parser for National language, In Proceedings of EMCSR 2002, Vienna, Austria.
19. Streiter, O. (2000) Reliability in Example-Based Parsing, In Workshop TAG+5, Paris France.
20. Aw, A., Zhang, M., Xiao, J., and Su, J. (2006) A phrase-based statistical model for SMS text normalization, In Proceedings of COLING-ACL, pp 33-40.

21. Kothari, G., Negi, S., Faruquie, T. A., Chakaravarthy, V. T., and Subramaniam, L. V. (2009) SMS based Interface for FAQ Retrieval, In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp 852- 860, Suntec Singapore.
22. Patel, D., and Bhatnagar, M. (2011) Mobile SMS Classification An Application of Text Classification, International Journal of Soft Computing and Engineering (IJSCE) 1, 47-49.
23. Back, L., and Makela, K. (2012) Mobile phone messaging in health care - where are we now?, Information technology & software engineering 2.
24. As cited by en.wikipedia.org/wiki/MXit.
25. Raghunathan, K., and Krawczyx, S. Investigating SMS Text Normalization using Statistical Machine Translation.
26. Peterson, J. L. (1980) Computer Programs for Detecting and Correcting Spelling Errors, Communication of the ACM 23, 676-687.
27. Hornby, A. S. (2006) In Oxford Advanced Learner's Dictionary of Current English, In Special Price Edition (7th, Ed.) Oxfrod University Press.
28. http://www.collinslanguage.com/wordlist.aspx.
29. Huston, S., Moffat, A., and Croft, W. B. (2011) Efficient Indexing of Repeated n-Grams, In WSDM'11, Hong Kong, China.

Authors' Briefs

Ademola Olusola ADESINA, a lecturer of Computer Science in Lagos State University (LASU), obtained his First Degree in Computer Science from Ogun State University (now Olabisi Onabanjo University), Ago-Iwoye and Masters Degree in Computer Science from the University of Ibadan, Nigeria. He is presently a doctoral student at the University of the Western Cape, Department of Computer Science (Machine Learning and Intelligent Systems Research Group), Cape Town, South Africa. His research interests are in Mobile computing, Text Processing, Agent Technology, Information Retrieval, Web Search and Mobile Security. Email: inadesina@gmail.com



Kehinde Kayode AGBELE is a Lecturer at Department of Mathematical Sciences (Computer Science Option), EKSU, Ado-Ekiti, Nigeria. AGBELE received B.Sc (Hons) degree in Computer Science from Ondo State University (now Ekiti State University), Ado in 1997, and M.Tech degree in Computer Science from the Federal University of Technology, Akure, in 2005. Currently, AGBELE is a Doctoral Research student at University of the Western Cape, Computer Science Department (Machine Learning and Intelligent Systems Research Group), Cape Town, South Africa. His research interests include Information Retrieval, Data Mining, Text Mining, Web Search Engine, Agent Technology, Pattern Classification and Clustering, Ubiquitous Healthcare, ICTs & Applications. He can be reached by phone on +27789345755 and through E-mail at agbelek@yahoo.com.

Abidoye Ademola Philip received B.Tech degree from Federal University of Technology Akure, Ondo State, Nigeria in 2001. He went further for his master's degree (M.Sc.) in Computer Science from University of Ibadan, Oyo State, Nigeria and completed it in 2006. He is presently a PhD student in Computer Science at University of the Western Cape, Cape Town South Africa. He is working as a Lecturer in the Department of Computer Science, Lagos State University, Ojo Nigeria. He has attended both local and international conferences, written many papers published in reputable international journals. His research areas include wireless sensor network, energy optimization, security, and mobile health. He can be contacted through Email: ademaola.abidoye@gmail.com

Azeez Nureni Ayofe graduates with B.Sc. (Hons) degree in Computer Science with Second Class (Hons.) Upper Division, from the Federal University of Technology, Akure (FUTA), Ondo State, Nigeria in 2004. He proceeded in 2006 to the University of Ibadan, Oyo State, Nigeria, after completing his National Youth Service Corps (NYSC) programme, for his Masters Degree programme in Computer Science which he successfully completed in 2008. He is currently a PhD research student in Computer Science at the University of the Western Cape, South Africa. He was a lecturer in the Departments of Computer Science of Crescent University, Abeokuta, Ogun State, Nigeria and the Fountain University, Osogbo, Osun State, Nigeria between 2008 - 2010. His areas of research include security and privacy; Grid and Cloud computing, knowledge representations and Computer Education & Applications. He can be contacted on +277 3 899 1735; nurayhn@yahoo.ca; and 3008814@uwc.ac.za.