6

Towards Full Comprehension of Swahili Natural Language Statements for Database Querying

Lawrence Muchemi

Natural language access to databases is a research area shrouded by many unresolved issues. This paper presents a methodology of comprehending Swahili NL statements with an aim of forming corresponding SQL statements. It presents a Swahili grammar based information extraction approach which is thought of being generic enough to cover many Bantu languages. The proposed methodology uses overlapping layers which integrate lexical semantics and syntactic knowledge. The framework under which the proposed model works is also presented. Evaluation was done through simulation using field data on corresponding flowcharts. The results show a methodology that is promising.

1. Introduction

The quest for accessing information from databases using natural language has attracted researchers in natural language processing for many years. Among many reasons for the unsuccessful wide scale usage is erroneous choice of approaches where researchers concentrated mainly on traditional syntactic and semantic techniques [Muchemi and Narin'yan 2007]. Efforts have now shifted to interlingua approach [Luckhardt 1987 and Androutsopoulos 1995]. The problem requires syntactic and semantic knowledge contained in a natural language statement to be intelligently combined with database schema knowledge. For resource scarce languages the problem is acute because of the need to perform syntactic and semantic parsing before conversion algorithms are applied. In general database access problems should have a deeper understanding of meaning of terms within a sentence as opposed to deeper syntactic understanding. The successful solution to this problem will help in accessing huge data repositories within many organizations' and governments' databases by users who prefer use of natural language.

In this paper a methodology for comprehending Swahili queries is presented. The wider frame work for achieving the conversion to structured query language (SQL) is also explained. The scope for natural language (NL) comprehension reported in this paper is limited to the 'select' type of SQL statements without involving table joins. The approach used in this work borrows concepts from information extraction techniques as reported in Jurafsky and Martin [2003] and Kitani et al [1994] and transfer approach reported in Luckhardt [1984]. Like in information extraction, pattern search identifies terms and maps them to noun

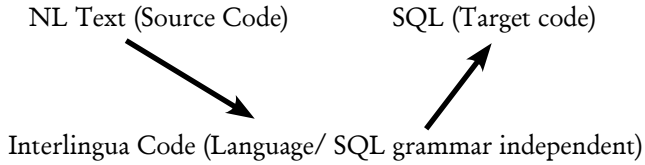
entities. A term is a word form used in communicative setting to represent a concept in a domain and may consist of one or more words [Sewangi 2001]. Templates are then used to map extracted pieces of information to structured semi-processed statements which are eventually converted to SQL code. Arrays or frames may be used to hold the entities and their meanings.

The rest of the paper first gives a synopsis of the general frameworks used for database access in natural language and shows criteria for the selection of approach. This is followed by an outline of a survey that investigates into Swahili natural language inputs. The conclusions are incorporated in the model presented.

2. General Frameworks for Data Base Access in Natural Language(Nl)

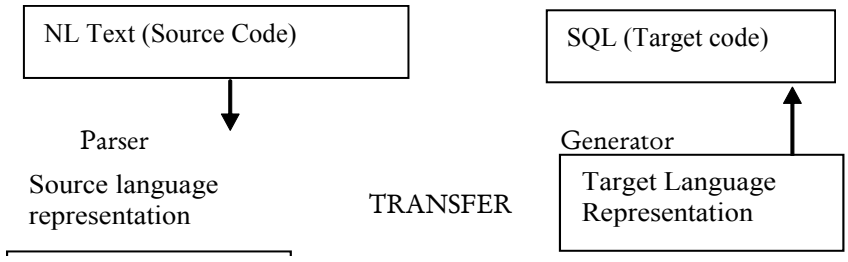
NL processing for generation of SQL statements has evolved from pattern matching systems to semantic systems and to a combination of semantics and syntactic processing [Androutsopoulos, 1996]. Perhaps what is attracting researchers to a great extent today is intermediate representation language, also referred to as interlingua systems [Jung and Lee 2002]. The two dominant approaches are direct interlingua approach and transfer models. The direct interlingua which may be roughly modeled as illustrated in figure 2.1 is attractive due to its simplicity. The assumption made in this approach is that natural language may be modeled into an interlingua. An SQL code generator would then be applied to produce the expected SQL codes.

Fig. 2.1. Direct Interlingua Approach



This model has never been achieved in its pure form [Luckhardt, 1987]. Many systems use the transfer model. The transfer model, illustrated in figure 2.2 below uses two different types of intermediate code one closely resembling the source language while the other resembles the target language. This approach has experienced better success and has been used in several systems such as SUSY [Luckhardt, 1984] among others.

Fig. 2.2. Transfer Approach



Recent works in machine translation especially grammatical frameworks [Ranta 2004] have brought forth ways that inspire looking at the problem as a translation problem. However it remains to be established whether the heavy reliance on grammar formalism has a negative impact on the prospects of this approach. Rule-based machine translation relies heavily on grammar rules which are a great disadvantage when parsing languages where speakers put more emphasis on semantics as opposed to rules of grammar. This is supported by the fact that human communications and understanding is semantic driven as opposed to syntactic [Muchemi and Getao 2007].

This paper has adopted the transfer approach because of past experiences cited above. To understand pertinent issues for Swahili inputs, that any NL-SQL mapping system would have to address, a study was conducted and the results and analysis of collected data is contained in the sections here below.

3. Methodology

Swahili language is spoken by inhabitants of Eastern and central Africa and has over 100 million speakers. Only limited research work in computational linguistics has been done for Swahili and this brings about a challenge in availability of resources and relevant Swahili computational linguistics documentation. The methodology therefore involved collecting data from the field and analyzing it. The purpose was to identify patterns and other useful information that may be used in developing NL-SQL conversion algorithms. The survey methods, results and analysis are briefly presented in subsequent sections. The conclusions drawn from the initial review together with other reviewed techniques were used in developing the model presented in later sections of this paper.

A. Investigating Swahili NL inputs

Purposive sampling method as described in Mugenda and Mugenda [2003] was used in selecting the domain and respondents. Poultry farmers are likely beneficiaries of products modeled on findings of this research and were therefore selected. Fifty farmers in a selected district were given questionnaires. Each questionnaire had twenty five information request areas which required the respondent to pose questions to a system acting as a veterinary doctor. Approximately one thousand statements were studied and the following challenges and possible solutions were identified:

a) There is a challenge in distinguishing bona fide questions and mere statements of facts. Human beings can decipher meanings from intonations or by guessing. A system will however need a mechanism for determining whether an input is a question or a mere statement before proceeding.

For example: Na weza kuzuia kuhara? {Can stop diarrhea?}
From the analysis it was found that questions containing the following word categories would qualify as resolvable queries:

- Viwakilishi viulizi {special pronouns}such as yupi, upi, lipi, ipi,,/ kipi /kupi, wapi, {which/what/where} and their plurals.
- Vivumishi viulizi{ special adjectives} such as gani {which}, -pi, - ngapi{How many}
- Vielezi viulizi {special adverbs}such as lini {when}
- In addition when certain key verbs begin a statement, the solution is possible. For example Nipe{Give}, Orodhesha{list}, Futa{delete}, Ondoa{remove} etc

Hence in the preprocessing stage we would require presence of words in these categories to filter out genuine queries. In addition discourse processing would be necessary for integrating pieces of information. This can be done using existing models such as that described in Kitani et al [1994] among others.

b) Swahili speakers are geographically widely dispersed with varying first languages. This results in many local dialects affecting how Swahili speakers write Swahili text.

Examples of Swahili statements for the sentence "Give me that book"

1. Nipe kitabu hicho ..... . Standard Swahili (Kiugunja) dialect
2. Nifee gitafu hisho ..... Swahili text affected by Kikuyu dialect
3. Nipeako kitabu hicho ... Swahili text affected by Luhya dialect
4. Pea mimi gitavu hicho..... Swahili text affected by Kalenjin dialect

The study revealed that term structures in statements used by speakers from different language backgrounds remain constant with variations mainly in lexicon. The term structures are similar to those used in standard Swahili. This research therefore adopted the use of these standard structures. The patterns of standard Swahili terms are discussed fully in Wamitila [2006] and Kamusi-TUKI [2004]. A methodology for computationally identifying terms in a corpus or a given set of words is a challenge addressed in Sewangi, [2001]. This research adopts the methodology as presented but in addition proposes a pre-processing stage for handling lexical errors.

c) An observation from the survey shows that all attempts to access an information source are predominantly anchored on a key verb within the sentence. This verb carries the very essence of seeking interaction with the database. It is then paramount for any successful information extraction model for database to possess the ability to identify this verb.

d) During the analysis it was observed that it is possible to restrict most questions to six possible templates. This assists the system to easily identify key structural items for subsequent processing to SQL pseudo code. The identified templates have a relationship with SQL statements structures and are given below:

- Key verb + one Projection
- Key Verb + one Projection + one Condition

- Key Verb + one Projection + many Conditions
- Key Verb + many Projections
- Key Verb + many Projections + one Condition
- Key Verb + many Projections + many Conditions

The terms 'key verb' in the above structures refer to the main verb which forms the essence of the user seeking an interaction with the system. Usually this verb is a request e.g. Give, List etc. It is necessary to have a statement begin with this key verb so that we can easily pick out projections and conditions. In situations where the key verb is not explicitly stated or appears at the middle of a sentence, the model should assign an appropriate verb or rephrase the statement appropriately. An assumption here is that most statements can be rephrased and the original semantics maintained. The term 'projection' used in the templates above, imply nouns which can be mapped onto field names within a selected domain and database schema. 'Conditions' refer to restrictions on the output if desired.

e) Challenge in the use of unrestrained NL text as input

One major challenge with unrestrained text is that questions can be paraphrased in many different ways. In the example given above the same question could be reworded in many other ways not necessarily starting with the key verb 'give'. For example,

> *"Mwanafunzi mwenye alama ya juu zaidi ni nani?"..."The student having the highest grade is called who?"*

In such situations it is necessary to have a procedure for identifying the essence of interaction. Information contained within a sentence can be used to assign appropriate key verbs. For example 'ni nani' (who) in the above example indicates that a name is being sought, hence we assign a key verb and noun; Give Name.

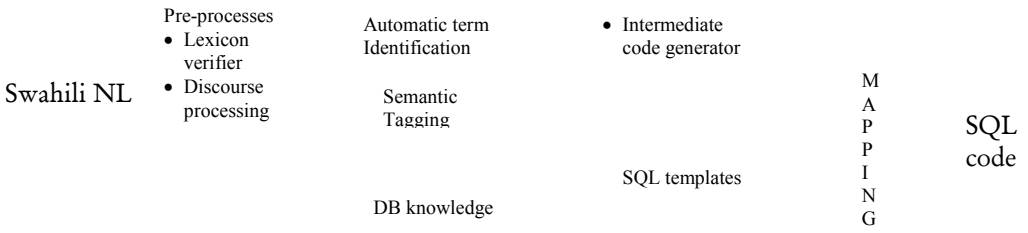
Nouns that would form the projection (Table name and column name) part of an SQL statement are then identified. This is followed by those that would form the condition part of the statement if present. Presence of some word categories signify that a condition is being spelt out. For example adjectives such as 'Mwenye, kwenye/penye, ambapo (whom, where, given)' signify a condition. Nouns coming after this adjective, form part of the condition. This procedure of reorganizing a statement so that it corresponds to one of the six identified templates would guarantee a solution. An algorithm for paraphrasing based on the above steps has so far been developed.

Model Architecture

The following is a brief description of the step by step processing proposed in the model. The input is unrestricted Swahili statement which undergoes pre-processing stage that verifies that the input is a genuine query, the lexicon is recognizable and there is no need for discourse processing. Terms are then generated and assembled into a suitable intermediate code. Generating intermediate code requires the use of

domain specific knowledge, such as semantics and templates proposed in section 3.1 above. The process proceeds by integrating this intermediate code with the specified database schema knowledge through mapping process and generates the SQL code. The process is illustrated in figure 3.1 here below.

Figure 3.1. The Transfer Approach Frame Work for Swahili



Steps in the generation of SQL scripts

Preprocessing

To illustrate the processes of each stage of the above model, we consider a sample statement:

"Nipe jina la mwanafunzi mwenye gredi ya juu zaidi? "....."**Give me the name of the student with the highest grade?"**

The model accepts the input as a string delivered from an interface and ensures that key words are recognizable. If not recognizable, the user is prompted to clarify. Preprocessing also involves verifying whether a statement is a resolvable query. The above statement begins with the word 'give', hence the statement is a resolvable query using the criteria described in section 3.1. If the statement contains pronouns and co-referential words, these are resolved at this stage. The output of this stage is a stream of verified words. The words are labeled to indicate relative position within the sentence. This forms the input of the term identification stage.

Automatic Term Identification

Term identification is the process of locating possible term candidates in a domain specific text. This can be done manually or automatically with the help of a computer. Automatic implementation involves term-patterns matching with words in the corpus or text. The model described here proposes application of automatic term identification algorithms such as those proposed in Sewangi [2001] at this stage. A tool such as the Swahili shallow syntactic parser described in Arvi [1999] may applied in identifying word categories.

Examples of term-patterns obtained through such algorithms would be:

N(noun)	Example ..... Jina
V(Verb)	Example ..... Nipe
V+N	Example ..... Nipe jina
N+gen connective+N	Example ..... Gredi ya juu

There are over 88 possible term patterns identified in Sewangi [2001] and therefore this large number of patterns would be difficult to fully present in this paper. However these patterns are used in identifying domain terms within the model.

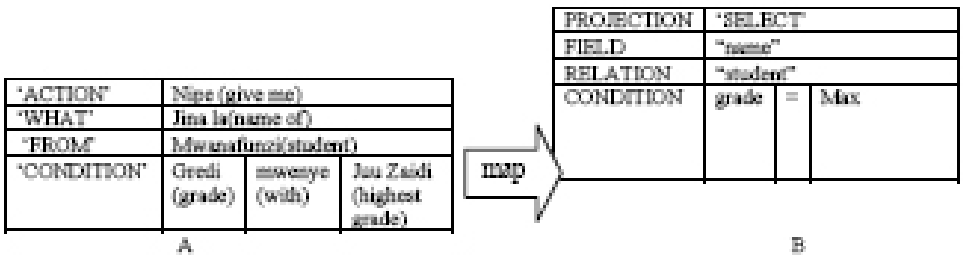
Semantic Tagging

The terms identified in the preceding stages are tagged with relevant semantic tags. These include terms referring to table names, column names, conditions etc. Information from the database schema and the specific domain is used at this stage for providing the meanings. For example, '*jina la mwanafunzi' (name of student)* gives an indication that column name is 'name', while table name is 'student'. Knowledge representation can be achieved through the use of frames or arrays.

Intermediate Code Generation and SQL Mapping

In the generation of intermediate code, we store the identified and semantically tagged terms in the slots of a frame-based structure shown in fig 3.2 (A) below. This can be viewed as an implementation of expectation driven processing procedure discussed in Turban et al. [2006]. Semantic tagging assists in the placement of terms to their most likely positions within the frame. It is important that all words in the original statement are used in the frame. The frame appears as shown here below:

Fig 3.2 Mapping Process



The semantically tagged terms are initially fit into a frame which represents source language representation. From a selection of SQL templates, the model selects the most appropriate template and maps the given information as shown in fig 3.2 (B) above. The generated table has a structure closer to SQL form and hence it can be viewed as a representation of the target language. This is followed by generation of the appropriate SQL code.

4. Discussions

As described, the methodology proposed here is an integration of many independent researches such as discourse processing found in Kitani [1994], automatic term identification found in Sewangi [2001], expectation-driven reasoning in frame-

structures [Turban et al. 2006]among others. The methodology also proposes new approaches in query verification as well as paraphrasing algorithm.

Research for this work is on-going. The algorithms for paraphrasing and mapping are complete and were initially tested. Randomly selected sample of 50 questions was used to give an indication of level of success. The statements were applied to flow charts based on the algorithms and the initial results show that up to 60% of these questions yielded the expected SQL queries. Long statements cannot be effectively handled by the proposed algorithm and this is still a challenge. Due to the heavy reliance on automatic term generation which relies on up to 88 patterns, there is over generation of terms leading to inefficiencies. Machine learning may improve the efficiency of the model by for instance storing successful cases and some research in this direction will be undertaken. Though not entirely successful, the initial results serve as a good motivation for further research.

5. Conclusions

This paper has demonstrated a methodology of converting Swahili NL statements to SQL code. It has illustrated the conceptual framework and detailed steps of how this can be achieved. The method is envisaged to be robust enough to handle varied usage and dialects among Swahili speakers. This has been a concept demonstration and practical evaluations would be required. However, test runs on flow charts yield high levels of successful conversion rates of up to 60%. Further work is required to refine the algorithms for better success rates.

References

ANDROUTSOPOULOS, I., RITCHIE, G., AND THANISCH, P. 1995. Natural Language Interfaces to Databases - An Introduction. *SiteSeer*- Scientific Literature Digital Library. Penn State University, USA.

ANDROUTSOPOULOS, I. 1996. A Principled Framework for Constructing Natural Language Interfaces to Temporal Databases. *PhD Thesis*. University of Edinburgh

ARVI, H. 1999. Swahili Language Manager- SALAMA. *Nordic Journal of African Studies* Vol. 8(2), 139-157.

JUNG, H., AND LEE, G. 2002. Multilingual question answering with high portability on relational databases. *In Proceedings of the 2002 conference on multilingual summarization and question answering. Association for Computational Linguistics, Morristown, NJ, USA. Vol.19, 1–8.*

JURAFSKY, D., AND MARTIN, J. 2003. *Readings in Speech and Language Processing*. Pearson Education, Singapore, India.

KAMUSI-TUKI. 2004. Kamusi ya Kiswahili Sanifu. Taasisi ya Uchunguzi Wa Kiswahili, Dar es salaam 2ND Ed. Oxford Press, Nairobi, Kenya.

KITANI T., ERIGUCHI Y., AND HARA M. 1994. Pattern Matching and Discourse Processing in Information Extraction from Japanese Text. *Journal of Artificial Intelligence Research*. 2(1994), 89-110.

LUCKHARDT, H. 1987. *Readings in Der Transfer in der maschinellen Sprachübersetzung*, Tübingen, Niemeyer.

LUCKHARDT, H. (1984). *Readings in* Erste Überlegungen zur Verwendung des Sublanguage-Konzepts in SUSY. *Multilingua*(1984) 3-3.

MUCHEMI, L., AND NARIN'YANI, A. 2007. Semantic Based NL Front-end For Db Access: Framework Adaptation For Swahili Language. In *Proceedings of the 1st International Conference in Computer Science and Informatics*, Nairobi, Kenya, Feb. 2007, UoN-ISBN 9966-7284-0-6, Nairobi, Kenya 151-156.

MUCHEMI, L. AND GETAO, K. 2007. Enhancing Citizen-government Communication through Natural Language Querying. . In *Proceedings of the 1st International Conference in Computer Science and Informatics* ,Nairobi, Kenya, Feb. 2007, UoN-ISBN 9966-7284-0-6, Nairobi, Kenya 161-167.

MUGENDA, A., AND MUGENDA, O. 2003. *Readings in Research Methods: Quantitative and Qualitative Approaches*. African Centre for Technology Studies, Nairobi, Kenya

RANTA, A. 2004. Grammatical Framework: A type Theoretical Grammatical Formalism. *Journal of Functional Programming* 14(2):145-189.

SEWANGI, S. 2001 Computer- Assisted Extraction of Phrases in Specific Domains- The Case of Kiswahili. PhD Thesis, University of Helsinki Finland

TURBAN, E., ARONSON, J. AND LIANG, T. 2006. Readings in Decision Support Systems and Intelligent Systems. 7th Ed. Prentice-Hall . New Delhi, India.

WAMITILA, K. 2006. Chemchemi Ya Marudio 2nd Ed., Vide Muwa, Nairobi, Kenya.