HiLCoE Journal of Computer Science and Technology

December 2018

HiLCoE

School of Computer Science and Technology

HiLCoE Journal of Computer Science and Technology is published once a year by HiLCoE. Individual articles can be accessed and downloaded at http://www.hilcoe.net.

Address all correspondences to: *HiLCoE Journal of Computer Science and Technology*, P. O. Box 25304/1000, Addis Ababa, Ethiopia; e-mail: hjcst@hilcoe.net; Tel. +251 111 56 49 00/ +251 118 68 12 70

Papers for publication are welcome from researchers in Computer Science and Technology. Papers will be peer-reviewed by at least two reviewers who are active researchers in the respective area of each topic of the paper.

Acceptance Rate: A total of 28 papers were submitted for publication. After a rigorous reviewing process, 13 papers are accepted and 15 papers are rejected with an acceptance rate of 46.4%.

Primary Objectives:

- To serve as a forum for the advancement of computer science and technology.
- To contribute to a rapid information and knowledge exchange between researchers in computer science and technology.
- To serve as a forum for postgraduate students in computer science and technology to publish their thesis/research project results.
- To share best practices and case studies from practitioners in the industry that can be sources of new research ideas.

Contents

An Approach to Utilize Open Source ERP Framework: The Case of Ethiopian Governmental Financial Organizations

Amensisa Dereje
HiLCoE, Software Engineering Programme, Ethiopia
Information Network Security Agency, Ethiopia
amedereje@gmail.com

Abrehet Omer
HiLCoE, Ethiopia
abrehet@gmail.com

Abstract

This paper first attempts to find out and identify matters affecting Enterprise Resource Planning (ERP) adaptation in the Ethiopian context. Secondly, there are lots of Open Source ERP framework vendors, but which one to select is a difficult task. This paper tries to explore and provide ERP evaluation criterion, which will help in choosing an appropriate open source ERP framework. Accordingly, iDempiere ERP is selected as a best fit framework. Thirdly, the main contribution of this paper is high-level and low-level customization approaches or models which will have a significant impact to customize open source ERP framework in a simple way. Finally, one additional new component is added into the ERP framework which allows effortless system development and customization especially for organizations that have complex business logic.

Keywords: ERP Framework; Open Source; Customization Approach; Application Dictionary

1. Introduction

The power of ERP software is now universally seen as a necessity for any organization or business. The features and benefits of these programs can lead directly to increased efficiency and profit. However, organizations in different industries have different requirements for ERP software, which can be difficult to manage with traditional models [1]. To overcome this, in Ethiopia and most developing countries, governmental and non-governmental organizations are trying to develop an ERP framework and system from scratch. But still it is difficult to have a stable ERP system and it is challenging to make it fit with different organizational requirements with minimum customization cost. Some organizations are also trying to use proprietary ERP systems. However, this approach will also have its own drawback to implement in an organization. As a result, implementing ERP framework independently by various organizations and using proprietary ERP systems are not cost effective, ineffective utilization of manpower and make an organization stick to one specific vendor. In addition, local software vendors do not have the capacity to build it in the scale of the international standard. An ERP with acceptable international standard and reasonably low cost for a poor country like Ethiopia is highly beneficial.

This paper addresses questions such as a) Which Open Source ERP framework is best fit to an organization's business needs? b) What are the most challenging issues to adopt ERP framework for an organization to utilize? and c) What approaches are used to utilize or customize an Open Source ERP framework and system to incorporate an organization's unique requirements?

To answer the identified questions, potential challenges of utilizing Open Source ERP framework are identified and solutions are provided. The evaluation criteria are derived and different Open Source ERP frameworks are compared based on the evaluation criteria and the best fit open source ERP framework is selected, which is iDempiere. Lastly, we provide a customization model and add one component which can simplify customization effort and development time and cost.

The data obtained from the respondents is quantified in the research. Information is gathered from Ethiopian Governmental Financial

Organizations (EGFOs). Two separate questionnaires comprising various questions related to different factors in selecting of Open Source ERP solutions and to identify the challenges to adopt Open Source ERP framework are used. An interview has been conducted with EGFOs to understand the business requirement and the existing system challenges. The results obtained from respondents were analyzed.

2. Related Work

Different related literatures are reviewed in different areas. To identify the existing challenges in adopting Open Source ERP framework several failure reasons have been reported in the literature. In [2], one of the reasons for the failure of ERP is that organizations fail to reconcile between the requirements of their human and business systems and those of the new technological system. ERP failure has been also attributed to a plethora of reasons: the complexity of ERP systems and lack of ERP product knowledge [3]. According to [4] inappropriate project management, lack of executive commitment, unclear business objective, poor communications, resistance to change within the organization, and others are stated as failure factors to adopt open source ERP solution. Different researchers used different evaluation criteria to compare and select best fit ERP framework. The author in [5] used five evaluation criteria and its sub-criteria aid to compare selected open source ERP systems. The author in [6] used two criteria; the dimension and feature as keys to evaluate which one is the best open source ERP. To identify the possible open source ERP framework customization approach the author in [5] described ERP system permits low-level customization through coding and high-level customization through metadata editing. The high-level customization approach is possible through the abstraction of parts of function, structure and behavior from the ERP systems into metadata while the higher level of ERP customization is doing most customization (metadata editing) graphically and supported by tools.

3. The Proposed Solution

The main reason that Open Source ERP framework is used is because organizations own the system without any cost and its full source content. There is no lock-in or dependency on the vendor and they are free to create new interface shell outside of the core system to meet the business needs and line-up with the organization's goal.

Thus, this paper meets an ultimate goal by providing a customization model and one major component which aids to utilize ERP Framework for various organizations. It addressed the following issues: i) It identified the challenges to adopt or utilize an Open Source ERP framework in the Ethiopian Organizations; ii) Selects best fit Open Source ERP framework (iDempiere) by evaluating existing different Open Source ERP Frameworks; iii) It provided a Low-level approach and High-Level approach model by focusing on iDempiere Framework and accordingly we have added one new core component into the framework in order to simplify and minimize the customization or development process, implementation time and cost.

3.1 Challenges of Adopting Open Source

Most of the organizations believe that possessing and exercising the use of an ERP system has become an essential asset in order to be able to adapt to environmental changes and be competitive in today's market. However, organizations encounter many challenges and issues when adapting or implementing an ERP system.

In order to gather data related to the existing challenges of adopting open source ERP system in the Ethiopian context a questionnaire is designed and data analysis is performed and covers a sample population of four financial organizations. IT departments or developers are chosen in order to determine stakeholders' needs based on their experience and familiarity with the topic. Thus, the major factors or challenges that have hindered an ERP system to be adopted by EGFOs are shown in Table 1.

Table 1: Major Challenges of ERP System Utilization

Priority	*Factor*	*Number of Responses*	*(%)*
1	Lack of executive or top management commitment	75	73.5
2	Unable to get an approach to adopt/implement ERP system	71	69.6
3	Lack of knowledge of ERP product	64	62.7
4	Incompatible business process with ERP products	53	51.9
5	Lack of education about the system's advantages & functionalities	47	46.1
6	Lack of commitment, acceptance and readiness to deploy the system	38	37.3
7	Resistance to change within the organization	27	26.5

3.2 Evaluation and Selection of Open Source ERP Framework

There are many open source ERP software. Out of these, one of them will be identified and selected through different mechanisms. In the first mechanism, the leading Open Source ERP systems are identified by searching the Internet [1, 7].

In the second mechanism, after a preliminary review of different research works on each of the identified open source ERP systems and according to the organization's requirements, it became evident that only some of the ERP frameworks deserve being looked into in more details due to its competitive offering. These two mechanisms resulted in the selection of the three open source ERP systems for comparison. They are: Apache OFBiz, OpenERP, and iDempiere. In the third mechanism, the pre-selected or the three chosen open source ERP frameworks are evaluated through different criteria.

To evaluate and select one best fit Open Source ERP framework, we adopt six criteria that are documented by different researchers. The criteria are Functional fit, Flexibility, Support Availability, community, Maturity and Organization Size and sub-criteria that serve to compare the selected open source. We also added one basic need to cover the requirements and set point for the objective of the research which is "Plugin custom code/module" as presented in Table 2.

Table 2: Open Source Evaluation Criteria

No.	*Main Criterion*	*Sub-criteria*
1.	Functional Fit	Number of tables, Inventory & Warehouse, e-commerce, Accounting MRP & POS.
2.	Flexibility	Customization, Flexible Upgrades, Architecture, Security, Internationalization, User friendliness, Interface, Scalability, OS Independence, DB Independence & Programming Language.
3.	Support	Support Infrastructure, Training & Documentation
4.	Continuity	Project Structure, Community Activity, Transparency, Update Frequency & Other lock-in effect.
5.	Maturity	Development Status & Reference Site

In this paper, weight scoring method is used for evaluation. Thus, string values Yes, No, Above Average, Average and Below Average are used. The result is a single weighted score for each criterion and the evaluation criterion necessarily assigns numeric values to judgments. Accordingly, there is a margin being assigned between 0 and 10. Yes is assigned to an available feature and for the value of maximum which is 10. No is assigned if that specific feature is not available within the open source ERP and assigned the minimum value which is 0. Above Average is found in between Yes and Average and supports all features with certain limitations and it is closer to the value Yes which is 8. Average -

Between Above Average and Below Average - is when it supports half of the features and it is assigned a value of 5. Below Average is between the value score of Average and No and has a certain function with limited value and it is assigned a value of 3. A score to specific product per selection criteria is denoted as Score:

(Criteria, OSERP) = Count (Yes) X 10 + Count (Above Average) X 8 + Count (Average) X 5 + Count (Below Average) X 3 + Count (No) X 0

Table 3 shows the summary of evaluating the different open source ERP frameworks. For each product a score value is computed.

Table 3: Open Source ERP Frameworks Comparison Summaries

No.	*Evaluation Criteria*	*Open Source ERP Framework*			*Max Value*	*Criteria Weight*
		OFBiz	*OpenERP*	*iDempiere*		
1.	Functional Fit	160	160	180	180	10
2.	Flexibility	51	36	56	60	25
3.	Support Availability	18	15	16	20	10
4.	Maturity	20	18	15	20	10
5.	Continuity	28	28	28	30	10
6.	Organizational Size	40	40	40	40	5
7.	Plugin Custom Code/Module	26	34	48	50	30
Total Criteria Value Sum (Criteria i X Criteria Weight i)		*4,515*	*4,330*	*5,430*	*5,700*	*100*
Percentage (Total Criteria Sum X 100)/Max Value		*79.21*	*75.96*	*95.26*		

In the context of multiple-criteria decision-making, the result of selecting open source ERP systems can depend on the criteria used for the evaluation [8]. According to our criteria used for comparison and organizations requirement, iDempiere is found as a best fit Open Source ERP system than the other two (OpenERP and OFBiz).

3.3 Implementation of New Component and Design of Customization Model

iDempiere is selected with its highly configurable, customizable and extensible platform with high stability. It also has the benefit of being used with no changes and cost. Sometimes iDempiere framework does not perfectly suit organizational needs and may need some changes to parts of the source code. Some customizations are not possible to achieve through iDempiere's core Application Dictionary and we have to modify the source code for that.

The newly added component will minimize customization cost and development time. In addition, by using this component creating any iDempiere is not required (such as User Interface, M class, I class, X class, Model factory class and Component definition (XML file) class in minimum). Instead the new developed component will automatically generate every necessary menu, class and code from the pre-designed database and based on the information registered on this component. So the developer only considers the unique business logic to be written in the generated class.

Thus, with the new added component and customization model as shown in Figure 1, every necessary model class and code is automatically generated. So, the developer merely focuses on the unique business logic to be coded. Thus, the new developed components and customization model will make easy the customization and used to include any types of organizational business logic.

3.4 Prototype

The prototype development is the basis for EGFOs requirement which need business logic code. It is also developed based on the low-level customization model and the new added component into the iDempiere framework.

The window shown in Figure 2 will help to generate the Menu for the newly added modules. The following windows are generating the class with the required code based on the database tables and our configuration on the windows. After setting the required information on the specified field, one can generate by selecting the 'Generate CCM' toolbar.

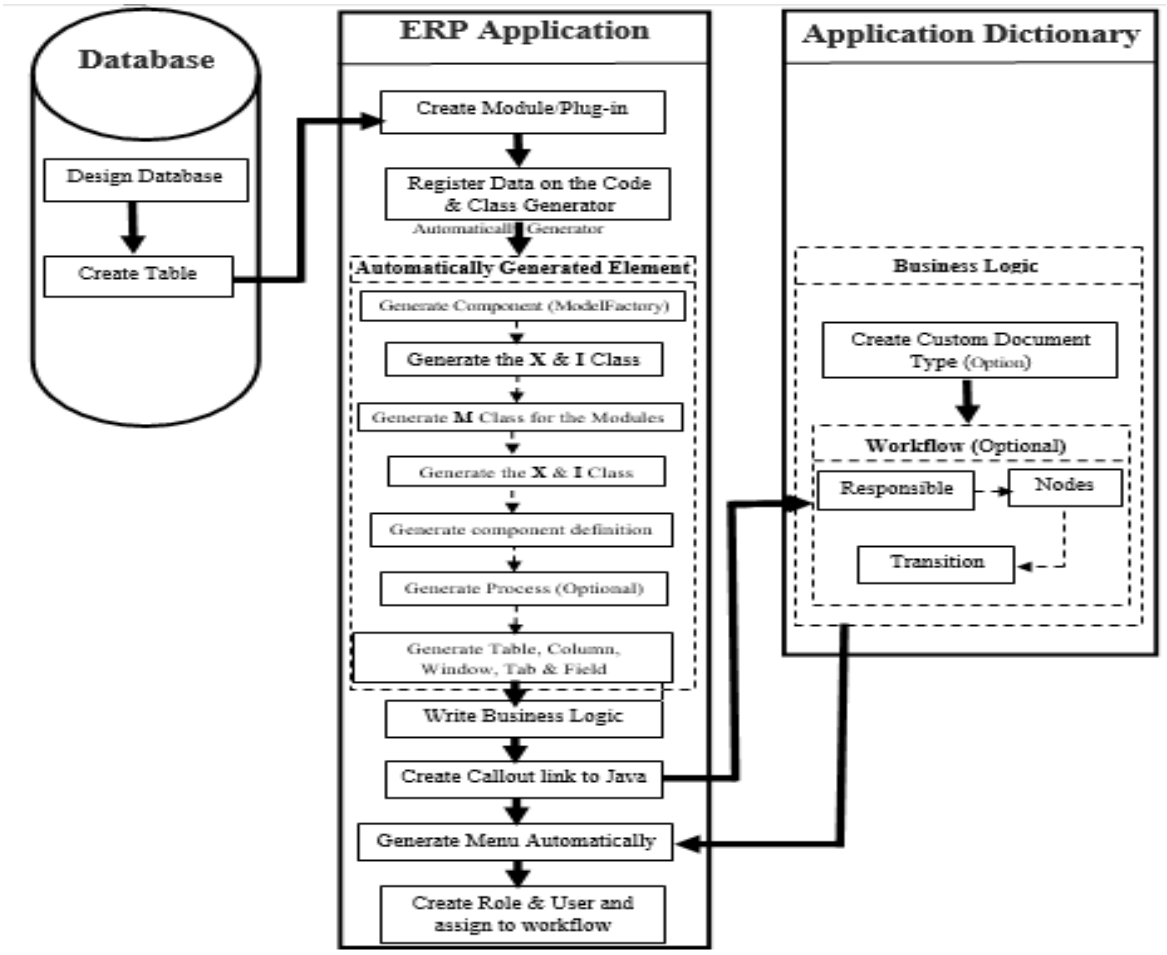

Figure 1: Low-level Customization Model to Add New Business Logic Using the Newly Developed Component

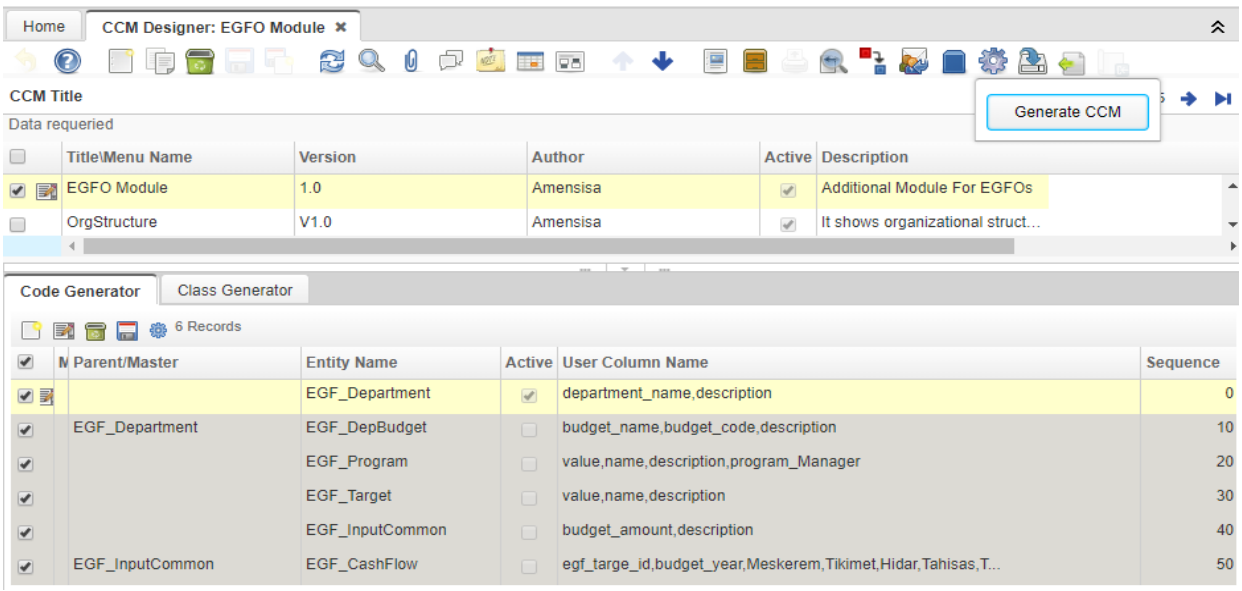

Figure 2: Class, Code and Menu Generator Window

Then the system or component automatically generates the required Tables and Columns, Window Tab and Fields, Classes with their codes and Menu as shown in Figures 3 and 4.

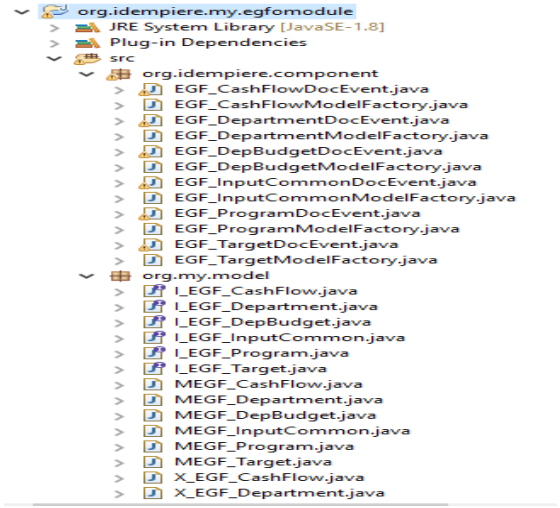

Figure 3: Dynamically Generated Class with its required Source Code

Figure 4: Automatically Generated 'EGFO' Menu

4. Conclusion and Future Work

The purpose of this paper was to provide high-level and low-level customization (utilization) approach or model for Open Source ERP framework and one new component is added into the ERP framework which makes more simplified and minimize the customization or development process and implementation time for developers. The supplementary aim was, since lots of Open Source ERP software vendors exist, selecting Open Source ERP framework is a difficult task. This paper formulated ERP evaluation criteria, which help in choosing the best fit open source ERP framework. In addition, this paper also revealed the challenges to adopt open source ERP framework in the case of EGFOs and provide a solution.

There are several lines of research arising from this work which should be pursued.

- Even though the customization approach or model is provided for most of iDempiere framework components, the customization model and the newly added component need some extension which can fit with customization needs based on the unique requirements of organizations.
- The current iDempiere framework works only for Java based application. By modifying the bottom layer architecture of the framework, it is possible to avail the framework for other programming languages such as .Net.

References

[1] Steve Floyd, "The Top Open Source ERP Software", http://www.axzm.com/top-open source-erp-software, Last accessed on July 18, 2014.

[2] Davenport, T. H., "Putting the Enterprise into the Enterprise System", Harvard Business Review, 1998.

[3] Chang, S., "ERP Life Cycle Implementation, Management and Support: Implications for Practice and Research", in Proceedings of The thirty-seventh Annual Hawaii International Conference on System Sciences, Hawaii, January 5-7, 2004.

[4] Loh, T. C. and Koh, S. C. L., "Critical Elements for a Successful ERP Implementation", SME. International Journal of Production Research, Vol. 42, pp. 3433-3455, 2004.

[5] Herzog T., "A Comparison of Open Source ERP Systems", Unpublished Master's Thesis, Vienna University of Technology, Austria, 2006.

[6] Raksha Jaiswal, "Evaluation of Open Source ERP for Small and Medium Scale Industries", Nagpur University, India, October 2014.

[7] Vittorio Gianni Fougatsaro, "A Study of Open Source ERP Systems", School of Management, Blekinge Institute of Technology, Sweden, 2009.

[8] Shouhong Wang and Hai Wang, "A Survey of Open Source Enterprise Resource Planning (ERP) Systems", March 2014.

Secure Notarization using Blockchain for the FDRE Document Authentication and Registration Agency (DARA)

Ashenafi Zena
HiLCoE, Computer Science Programme, Ethiopia
IMC Worldwide, Ethiopia Ltd.
ashenafizena@gmail.com

Mesfin Kifle
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
kiflemestir95@gmail.com

Abstract

Notarization is an important facet of civilization. It creates trust between interacting parties enabling them to do mutual exchanges. Financial services and civil service organizations in Ethiopia are among the consumers of notarized documents. These organizations resort to authenticating notarized documents against registered copies at the notary as a measure of security. Being subjected to such a lengthy security measure leads to customer dissatisfaction and delayed service delivery. Automating this key process has been slow until recently due to the difficulty of creating efficient and effective digital trust relationship between transacting parties as well as legal framework unavailability and incompatibility across jurisdictions. In this paper, we compare blockchain with similar technologies and adapt it for the purpose of notarization. The proposed solution is a notarization, registration and verification system based on an open source blockchain infrastructure. The proposed solution promises instantaneous and independent verification and immediate dissemination of revocation closing potential security gap while still protecting fraudulent activities as well as the brick and mortar notarization institution.

Keywords: Document Notarization; Document Security; Blockchain; Distributed Database

1. Introduction

Notarization, the action or process of notarizing a document, a record of this, as stamped and signed by a notary directly on a document, or in the form of a certificate [1] is ubiquitous. It plays a big part in the healthy functioning of a modern society. In all sovereign countries, there are always one or more entities constituted by law to perform the function. The Federal Democratic Republic of Ethiopia (FDRE) Document Authentication and Registration Agency (DARA), the primary focus of this paper, is established by the FDRE Proclamation No. 922/2015 [2].

The current predominant method of paper-based notarization requires that the notary physically sign and stamp or emboss the paper document being notarized [3]. "This has historically been a norm and was effective in achieving the goal of security-traditional notarizations were difficult to forge, modify or erase, because the signatures and seals are impressed or embossed into the very fiber of the paper, and they explicitly track back to a government-commissioned notary who can verify the facts related to the notarial act" [3]. However, technological advances are easing this difficulty thereby increasing the risk that forgery afflicts on the traditional signing of paper documents. Forensic science cannot guarantee that any given ink signature can be verified. Scientists can only offer an educated opinion as to whether the signature is authentic, and they can do so only under the right circumstances. This includes, for example, the availability of several good specimen signatures [4]. One possible approach to mitigate forgery is for the notary to keep a notarial journal and copy of notarized documents. Greenwood, in a review of electronic notarization [3] provides examples of how the mere existence of a notary journal served as a deterrent and a prosecutorial tool. By law, the Federal Democratic

Republic of Ethiopia requires the notary to keep a register (journal) and deposit copies of authenticated documents for itself [2]. When issues of authenticity arise, reauthentication is made against said deposits. Thus, it can be argued that the FDRE has installed a mechanism to deter forgery.

On the other end of the spectrum, financial services and civil service organizations in Ethiopia are among the consumers of notarized documents where the said documents are routinely presented for execution by agents like the Attorney-in-Fact. Such organizations are cognizant of potential economic loss to Principals that may arise from transactions involving high value assets like property and liquid assets. In practice, organizations resort to authenticating notarized documents against registered copies at the notary as a measure of security. This occurs at the expense of an expedited service delivery to the detriment of both the service rendered and the customer (Attorney-in-Fact). The authentication process conjures up the security-usability trade-off described in Saltzer and Schroeder's principle of psychological acceptability [5]. The principle says that a security mechanism should not make accessing a resource, or taking some other action, more difficult than it would be if the security mechanism were not present. In practice, this principle states that a security mechanism should add as little as possible to the difficulty of the human performing some action [6]. *In stark contrast, the Attorney-In-Fact, by being subjected to a lengthy security measure is denied an immediate response to execute its Representative Capacity.*

We hypothesize that through implementing automated, secure and easily verifiable notarization employing blockchain technology, we can significantly improve service delivery responsiveness of these organizations while at the same time guaranteeing the same or improved level of security as currently practiced. Thus, the following research questions are formulated to address the research.

- How secure is notarized document verification as currently practiced in financial and civil service providers of (how effective is the security) and how long does it take, on average, to verify notarized documents by such providers (how efficient is the security)?
- What are the desirable characteristics of a technology based notarization?
- Among potential technological solutions including the hypothesized blockchain, which is better suited to provide effective and efficient security?
- How can a notarization solution with the desired characteristics be built and evaluated?

Our primary research approach is design science- we design a secure notarization system with properties geared towards solving the efficiency problem of existing practice. The hypothesis is, blockchain, a distributed trustless consensus database, can deliver a secure, open and efficient solution to notarization. We have anecdotal understanding of how a manual notarial system can be improved by the application of technology. However, to make this understanding concrete, we employed a qualitative research design using interview as instrument to seek an understanding of different groups of users experience and perception about the state of affair in current notarization systems. Blockchain is an innovative technology with great promise. However, there are contending proposals that can potentially be applied to automate notarial process. As a consequence, we followed the qualitative study with a comparison of alternative technologies to select an appropriate technology from an array of alternatives based on desirable properties. The desirable properties are garnered from industry best practice documents, regulatory requirements as well as requirements gathered as a result of the interviews conducted.

The same desirable properties are used to drive the design and implementation of a proof-of-concept of the blockchain based notarization system. We used an open source blockchain infrastructure - Hyperleger Fabric [7] - as the backbone of the

prototype. Access to the blockchain is provided by a RESTful API provided by another Hyperledger project - Hyperledger Composer [8]. We designed a web application using Angular - a JavaScript client-side library - to implement the user access to the blockchain system and effect the notarization domain business logic in coordination with the RESTful API.

2. Related Work

In reviewing the body of research in digital notary, one can observe a rather concerted effort starting from the late 90s to apply the concepts of public key cryptography infrastructure (PKI) to the domain of long term preservation of digital assets using digital signature in a variety of ways. This is likely due to the maturity and wide spread application of PKI with the growth of the web in that era and the rather attractive notion of applying the technology to its natural use of enforcing authentication, non-repudiation (with certificate authority), and integrity which seems to fit nicely to the domain of notarization. In this section, we review related publications highlighting the problems they tackle and their limitations.

Publicized in the 2000 European Conference on Information Systems proceedings, the work in [9] aims at solving disputes between entities on events and/or actions that are of interest to the entities. Claim is made that commercial certification or notarization services are incomplete in that these services do not distinguish the creator from the sender nor store the data itself for a long time, nor prove the action (creation, sending, receiving) in the context of communicating entities. The paper's objective is to address this gap employing an empirical approach through the creation of a typical Trusted Third Party (TTP) service the authors named CYNOS- CYberNOtary System.

The proposed solution is a signature-based system and uses various security techniques to enable the timely certification of entities and fair and highly reliable delivery. First, the *existence notarization* service stores evidence proving the existence of a piece of data and its possessor. Second, the *delivery notarization* service stores the evidence of the delivery and the reception by the intended recipient as well as the existence of the data and the originator. *Notary token* CYNOS issues as the certification of facts containing the evidence stored in CYNOS after each notarization event and is accompanied with a signature of CYNOS. The solution relies on CA (Certification Authority), NA (Notary Authority) and RA (Registration Authority).

A paper presented at EuroPKI 2012 conference [10], proposes a new PKI model for long-term signatures based on X.509 and the real world of handwritten signatures. The problem is framed in the context of using long term keys to sign a document. The paper claims preserving information is useless if authenticity, integrity and datedness cannot be ensured in perpetuity. To address this, PKI has been commonly utilized. However, the paper posits PKI solutions suffer from two major problems. The first is the amount of protection data (which they define as timestamps, certificates and revocation data accumulated in the life time of the document) to store and verify is not optimal both storage wise and processing overhead. The second is the difficulty for verifiers of dealing with trust anchors in terms of unavailability of previously trusted entities. This can happen, for example, with changes to technology and the corresponding data model. The paper's objective is, therefore, focused in finding a solution to reducing the data protection burden and the chain of trust that needs to be verified in continuity.

Their proposed solution to the problems is The Notarial Based Public Key Infrastructure (NBPKI). The notary in their solution is modeled against the real-world notaries which they refer to as Notarial Authorities (NA).

As with the current work, the goal is in finding an effective and efficient notarization solution using PKI. However, in the reviewed work, the NAs are modeled as trusted parties. This is just moving the known weak link of the Certificate Authority (CA) of current PKI systems to a new entity modeled after

the real-life notary which has no valid ramification. Even if a legal framework is coined to institute these digital NAs, the key decay problem tormenting the CA with time passage cannot be mitigated using legislative means alone.

A paper presented at the 12th annual ACM-SIAM symposium [11], proposes a method to notarize a paper document by computing a set of features from the page, inputs these features into a hash function, digitally signs the output, and then prints the resulting signature on the page. The notarized document is self-contained - verification does not require reference to online information.

The proposed solution is to have a digital scan of the document from which feature sets are extracted (glyphs or symbols) and hashed, employing public-private key cryptography, into a short string which can later be verified by a third party running the same algorithm as the authenticator on the same set of features of his/her scanned copy.

The authors' creative solution is to come up with a concept they named an assist channel. The value of the assist channel is to compensate and regulate for scan variations in a predictable manner so that any verifier can get the same hash compute result as the original authenticator provided the document has not been altered.

An article published in the Institute of Electrical Engineers of Japan journal [12], proposes storing document timestamp in the blockchain by encoding into Bitcoin addresses. The paper focused on the proliferation of data on the Internet without adequate ownership ascertaining mechanism for those who require it due to the nature of the digital data - creative assets that the owners would like to have copyright over. The authors believe the available solutions- timestamping, either by a central timestamping authority (TSA) or a scaled version of such service- suffer from a single point of failure and costly deployment respectively. The two further have a shared shortcoming of being easily exposed to security breaches. Thus, the paper's objective is to find a solution that has no single point of failure, economically scalable and tamper resistant.

The proposed solution is a method called decentralized trusted timestamping based on a blockchain. While they provide examples of services that provide timestamping on a distributed tamper free database they discount them on the basis of their data limitation. They all depend on Bitcoin's blockchain that put a 40-byte data constraint on extraneous data. Moreover, the 40-byte hash data cannot be reversed, which makes determining the creators and other related information difficult.

Thus, they propose a methodology for storing a maximum of N*20 bytes in the blockchain. For creators who prefer not to anonymously timestamp their creations, their proposed method can embed related information (e.g., the file name, the creator's name, and comments) and the hash of the digital data. The related information has the added benefit of being decrypted in plain text. The method creates transactions with N output addresses; therefore, one transaction can store at most N*20 bytes.

The reviewed work is closely aligned with the current research. The use of blockchain as an irrefutable secure storage of records and the result of the evaluation in terms of cost and effectiveness is what the current work aspires to establish.

A paper, published in the International Journal of Medical Informatics [13], deals with the long-term validation of the authenticity of electronic healthcare records (EHR). The authors' premise is that medical records should be preserved at least for the lifetime of a patient or even longer for research or other purposes. The challenge is that preserving electronic data for a lifetime is not a straightforward task, since ICTs are under intense development and are subject to continuous changes. The long-term preservation of EHR may be approached and analyzed from various perspectives, such as future data readability, storage media longevity and security.

The paper proposes a successive trust transition towards new entities, modern technologies, and

refreshed data to mitigate the key life-time problem. According to the paper, a future relying party will have to trust only the information provided by the last notary, in order to verify the validity of the initially signed EHR, thus eliminating any dependency on ceased entities, obsolete data, or weak old technologies.

3. The Proposed Solution

The proposed solution is a notarization, registration and verification system. We start by defining the functional and non-functional requirements of the system. Next, we describe a process for selecting a blockchain from a family of related chains based on system requirements. Finally, we construct an architecture and detail design of salient features of the notarization, registration and verification system. We arrived at this decision after a technology selection exercise. We made comparison of technologies we reviewed based on key attributes that a notary system should possess. The primary attributes for technology selection are the perceived effectiveness and efficiency of a notarization. We concluded that the proposed solution is a better fit to solve the problem at hand.

3.1 Functional Requirements

The prominent functional requirements are authenticate documents, authenticate and register copies of documents against their originals and register same, suspend document authentication and act in lieu of a Registry Authority. Figure 1 shows the proposed system functions.

3.2 Nonfunctional Requirements

The important quality attributes shape the structure and behavior of the overall system. We elucidate attributes, sourced from The Model Notary Act [14], that have relevance towards system building. These are technology neutrality, confidentiality, performance, integrity and nonrepudiation, capable of independent verification and digital signature source agnosticism.

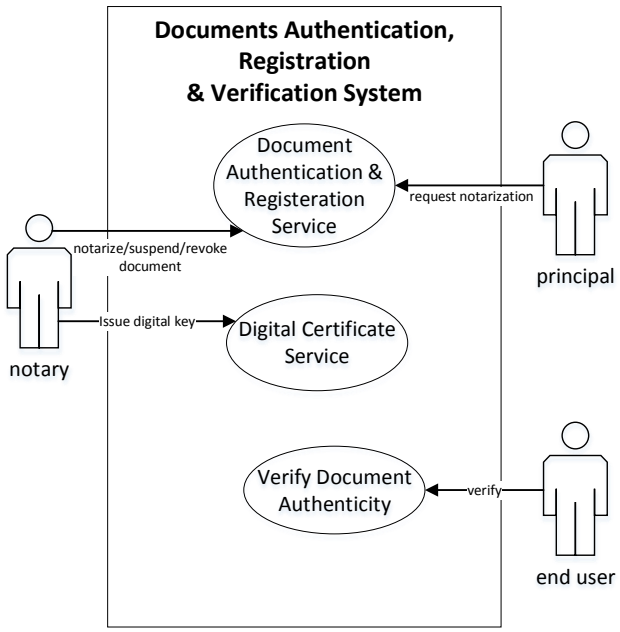

Figure 1: System Use Case

3.3 Blockchain Selection

"A major difficulty for architects designing applications based on blockchain is that the technology has many configurations and variants. We propose how to classify and compare blockchains and blockchain-based systems to assist with the design and assessment of their impact on software architectures" [15]. The selection process answers the following guiding questions: 1) Has trusted authority? 2) Can it be decentralized? 3) How to decentralize the authority? 4) Storage and computations on-chain vs. off-chain? 5) Need new blockchain? 6) What consensus protocol? 7) What incentive? Going through the guide, we selected A DARA controlled, decentralized by DARA branches and subordinated institutions, with optional off-chain document store, identity based consensus protocol (allowing only DARA and subordinates to write to the blockchain) dismissing the need for incentive.

3.4 Notary Application

The notary application is designed for the notary to complete notarial acts. It manages the notary's digital signature, it accepts user document for the notary's review, helps it generate digital keys for the principal and the attorney-in-fact as well as securely print it for later use, digitally sign the document using the principal's and its combined digital private

keys, digitally sign the document with the attorney-in-fact's public key and finally send the notarized document to the blockchain. The system architecture is shown in Figure 2.

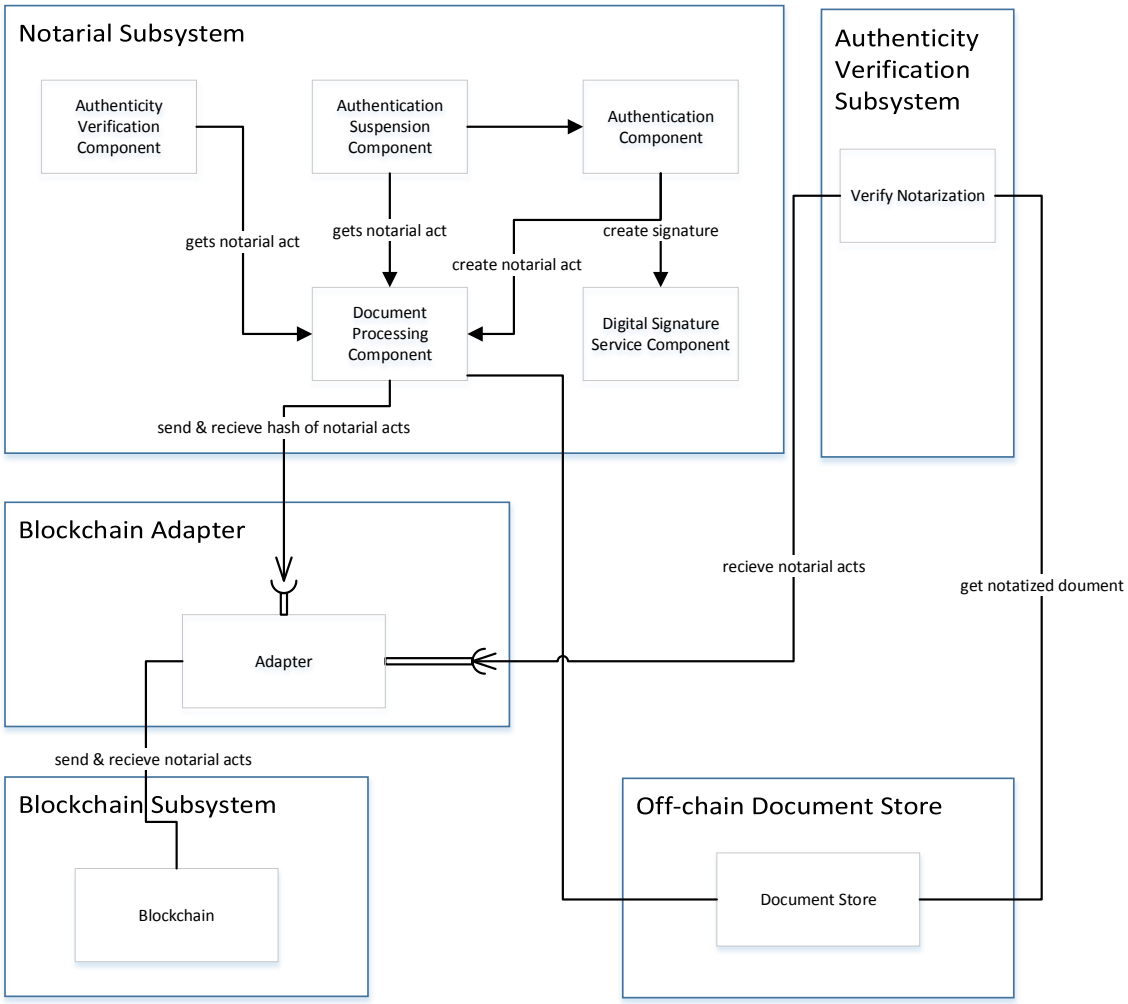

Figure 2: System Architecture

3.5 Verification Application

The verification application is designed for any interested party to independently verify a provided document from the off-chain document store that matches a hashed version in the blockchain. For this to happen, the attorney-in-fact needs to present the third party, a bank official for instance, with a raw document decrypted using his/her private key from the off-chain document store. The precondition is that at the time of notarization, the document is pushed to the document store encrypted using the attorney-in-fact's public key for the purpose of privacy (a copy can be made using the principal's public key for his/her reference as well). The document needs to be accompanied by metadata describing the notary and principal's public keys and a document reference. From there, the verification application computes and compares the hash with the one in the blockchain referenced by the same document reference. Thus, the physical architecture depicted in Figure 3 ensues.

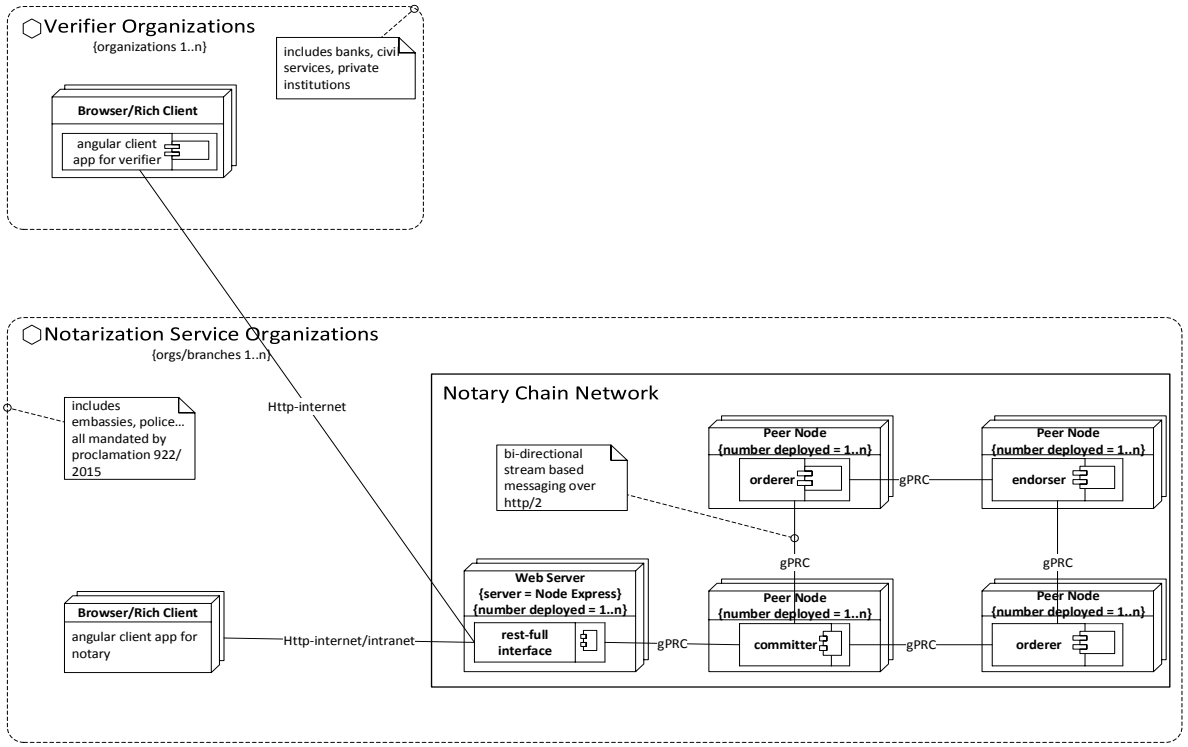

Figure 3: Physical Architecture

4. Conclusion and Future Work

We showed that recent development in distributed trust relationships with the advent of blockchain technology is a viable solution to a whole slew of problems - trusted third-party computation model, single point of failure and the related scalability and availability issues. We did so by designing a blockchain based solution to DARA.

The proposed solution promises instantaneous and independent verification and immediate dissemination of authentication revocation, closing potential security gap. It also affords service providers prompt reply to customers enhancing their brands and satisfaction for the customer.

Perhaps, the biggest hindrance to the practical implementation of a blockchain based notarization system is the absence of a conducive legal framework that enables the notary and individuals to execute digital signatures in leu of hand signatures and for these signatures to be legally binding. The foundation for this, The Ethiopian Electronic Signature Act (ESA), is already laid out but is still a draft under consideration. The speedy adoption of the act can enable the materialization of notarization and related domains to flourish using the blockchain as a secure distributed applications platform.

Most notarial acts need human intervention. They are performed in person and out of band - meaning they are not amicable for automation. However, other tasks like keeping specimen of signature, ascertaining the legality of documents; ascertaining the capacity, right and authority of persons, ascertaining title certificates and ascertaining legitimacy of documents source can be aided and become components of the blockchain system. This requires that other government institutions - land management and citizens management - also can be brought onboard the blockchain system. Property certificates and citizens' identity information are inputs for the notary to ascertain legality of documents; ascertaining the capacity, right and authority of persons, ascertaining title certificates and ascertain legitimacy of documents source. We believe this extension is a fruitful research arena to explore.

References

[1] "Definition/notarization," Oxford University Press, 2017, https://en.oxforddictionaries.com/

definition/notarization, Last accessed on 20/03/2017.

[2] Federal House of Peoples' Representative, Proclamation No. 922/2015, Authentication and Registration of Document, Addis Ababa: Federal Negarit Gazette, 2016.

[3] D. J. Greenwood, "Electronic Notarization," National Notary Association, Chatsworth, California, 2006.

[4] M. L. Closen, "Notaries Public - Lost in Cyberspace, or Key Business Professionals of the Future?," Journal of Computer & Information, Vol. 15, No. 4, pp. 703-758, 1997.

[5] J. H. Saltzer and M. D. Schroeder, "The Protection of Information in Computer Systems," Proceedings of the IEEE, Vol. 63, No. 9, pp. 1278-1308, 1975.

[6] M. Bishop, "Security and Usability: Designing Secure Systems that People Can Use", L. F. Cranor and S. Garfinkel, Eds., Sebastopol, CA: O'Reilly Media, Inc., 2005.

[7] Hyperledger, "About - Hyperledger," The Linux Foundation, https://www.hyperledger.org/about, Last accessed on 24/9/2017.

[8] Hyperledger, "Hyper Ledger Composer," https://github.com/hyperledger/composer, Last. accessed on 24/9/2017.

[9] S. Nakahara, "Electronic Notary System and its Certification Mechanism," in ECIS 2000 Proceedings, Vienna, 2000.

[10] M. Vigil, C. Moecke, R. F. Custódio, and M. Volkamer, "The Notary Based PKI: A Lightweight PKI for Long-Term Signatures on Documents," in Public Key Infrastructures, Services and Applications: 9th European Workshop, EuroPKI 2012, Pisa, Italy, 2013.

[11] M. Ruhi, M. Bern, and D. Goldberg, "Secure Notarization of Paper Text Documents," in SODA '01 Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms, Washington D.C., 2001.

[12] Y. Gao and H. Nobuhara, "A Decentralized Trusted Timestamping Based on Blockchain," IEEE Journal of Industry Applications, Vol. 6, No. 4, pp. 252-257, 2017.

[13] D. Lekkas and D. Gritzalis, "Long-term Verifiability of Healthcare Records Authenticity," International Journal of Medical Informatics, Vol. 76, No. 5-6, pp. 442-448, 2007.

[14] National Notary Association, "The Model Notary Act," 2010. https://www.nationalnotary.org/file%20library/nna/reference-library/2010_model_notary_act.pdf., Last accessed on 2/5/2017.

[15] X. Xu, I. Weber, M. Staples, L. Zhu, J. Bosch, L. Bass, C. Pautasso, and P. Rimba, "A Taxonomy of Blockchain-Based Systems for Architecture Design," in IEEE International Conference on Software Architecture (ICSA), Gothenburg, Sweden, 2017.

Large Vocabulary, Speaker Independent Continuous Speech Recognition for Ge`ez Language Using HMM

Assefa Mamo
HiLCoE, Computer Science Programme, Ethiopia
assefa.ma@yahoo.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

A speech recognition system implements the task of automatically transcribing speech into text. This paper is concerned with automatic continuous speech recognition using trainable systems. The aim of this work is to build acoustic models for Ge`ez language. Ge'ez is the classical language of Ethiopia within the Semitic language family. Today, Ge'ez remains only as the main language used in the liturgy of the Ethiopian Orthodox Tewahedo Church. In order to accomplish the aforementioned objectives, speech data is recorded at a sampling rate of 16 KHz and the recorded speech is converted into Mel Frequency Cepstral Coefficient (MFCC). The corpus of 400 utterances has been developed that is used for training. A set of 100 additional sentences and phrases is uttered using the same speakers participated in the training phase and another set of 100 sentences and phrases is recorded using 5 different speakers that did not participate in the training phase to test and evaluate the system. An attempt is also made to build bigram language model with task grammar. Then, the acoustic model is built using the Hidden Markov Model approach. Tied-state triphones have been considered as the basic sub-word units that are extended from context independent phonemes in this study. Finally, the performance of the speech recognizer has been evaluated for trained speakers and the result was 64.21% word level correctness, 62.05% word accuracy, and 29% sentence level correctness. Moreover, 57.72% recognition accuracy, 62.19% word level correctness and 24% sentence level correctness has been obtained when the system is tested with speakers who had no involvement in the training phase. Based on the experimental results, the Pause between words can also affect recognition accuracy whether the set time is too short or too long.

Keywords: Speech Recognition; HMM; HMM based Speech Recognition; Language Modeling

1. Introduction

Speech is one of the oldest and the most widely and effectively used means of communication for humans. It plays a great role especially in human-machine interaction. Automatic speech recognition (ASR) or computer speech recognition is the decoding of the information conveyed by a speech signal into a set of characters. The resulting characters can be used to perform various tasks such as controlling a machine, accessing a database, or producing a written form of the input speech. Speech recognition can be classified into various types depending on how capable it is. These classifications are:

- Discrete (Isolated) vs. Continuous;
- Speaker-dependent vs. Speaker-independent;
- Large vocabulary vs. Small vocabulary speech recognition;
- Read or spontaneous speech;
- Noisy or Noise free speech.

People have extensively studied speech and often tried to build machines to handle it in an automatic way. During recent years, automatic speech recognition has started to emerge as a pragmatic and useful technology for various applications. During the last two decades, a remarkable advancement has been achieved in the field of ASR technology.

The ultimate goal of ASR is to make a real time system to achieve 100% accuracy in recognizing all words which are independent of vocabulary, speaker and environment characteristics. But the research is not close enough to meet the human ability [1]. There have been many works in ASR systems for technologically favored languages for more than 60 years in the globe. Unfortunately, research in local language speech recognition started only during the last two decades. Over the past few years, several natural language processing applications such as speech-recognition and Text-To-Speech (TTS) have been developed for local languages such as Amharic and Tigrigna.

Ge`ez Language

Ge`ez is an ancient South Semitic language that originated in Eritrea and the northern region of Ethiopia in the Horn of Africa. It later became the official language of the Kingdom of Aksum and Ethiopian imperial court [10]. Today, Ge'ez remains only as the main language used in the liturgy of the Ethiopian Orthodox Tewahedo Church, the Eritrean Orthodox Tewahedo Church, the Ethiopian Catholic Church, the Eritrean Catholic Church, and the Beta Israel Jewish community [10]. It is also taught as a second language in the traditional schools of the Ethiopian Orthodox Church [11, 12].

Ge'ez is fairly massive in size, with its 182 syllographs. There are essentially 26 main syllographs, all consonants, while the rest are essentially those with additional strokes and modifications added on to the main forms to indicate a vowel sound associated with it or to make aural adjustments in the basic consonant sound [13]. With a combination of every twenty six original alphabet with the seven vowel sound alphabets, a matrix of 26×7 size is produced.

As each language has its own set of phonemes from which sounds of words are generated, the classification of Ge`ez language`s consonants and vowels differ in their set of phonemes. Ge`ez language has its own characterizing phonetic, phonological and morphological properties.

Unlike many of the modern Ethiopian Semitic languages, Ge'ez language has the two pharyngeal consonants [ʔ](ዐ), [ḥ] (ሐ). In addition, the phoneme | **ኈ** | **ጎ** is treated as a laryngeal by the phonotatic rule of Geez. **ኀ** and **ኈ** are used in orthography of the word Tigrinya, even though **ኈ** does not exist in the spoken language [14].

Ge`ez consonants may also be subjects to gemination. Germination is a widely employed inflectional and derivational process in Ge`ez. For instance *('häqqäl' (ሐቀል) countryman, farmer* from *'häql' (ሐቀል) field* and *'t.äbbäbt' (ጠበብት) wise men from 't.äbib' (ጠቢብ).* All Ge`ez consonants can geminate with the exception it appears of the laryngeal. Ge`ez can also be pronounced with slight variations by putting the stress on different parts of a word to produce different sounds thereby giving more than one meaning to one and the same word [15]. To give an example *"yenägg`era" (ይነ ግሩ) they [f. pl] speak vs "yenägger`a" (ይነ ግሩ) he speaks to her*. Sometimes words are compromised with a certain style of pronunciation: high level, low level, acute, grave accent and extra short [16].

2. Related Work

Amharic speech recognition of isolated Consonant-Vowel syllable was developed using Hidden Markov Model Tool Kit (HTK) based on HMM with recognition accuracy of 87.68% and 72.75% for speaker dependent and independent systems, respectively [2]. In this system about 41 CV syllables of Amharic language from eight people (4 male and 4 female) with the age range of 20 to 33 years were used.

Amharic speech recognition using sub-word units was also examined for the purpose of comparative analysis using phonemes, CV syllable and tied-state triphones as the basic sub-word. About 170 word vocabulary recorded from 15 speakers (8 male and 7 female) in the age range of 20 to 30 for both speaker dependent and speaker-independent models to obtain

performance for the training data using tied-state triphone models (91.46% vs. 77.33%) and for the testing data of practically speaking (77.87% vs.75.33%) [3].

Amharic speech input interface to command and control Microsoft Word was also designed using HTK based on HMM-based, speaker independent, small vocabulary, and isolated word recognizers [4]. Fifty command words were used from 26 people to develop the prototype system. The system was tested using 18 randomly selected command words and it performed accurately in 16 commands. The comparative study in [2] was tried to compare two different toolkits: HTK and the MSSTATE Toolkit from Mississippi State for small vocabulary speaker dependent continuous Amharic speech recognizers using 50 sentences with ten words (the digits). The recognition system developed using the HTK has 82.5% word recognition accuracy whereas the system developed using MSSTATE has 79% word recognition accuracy.

Amharic speech recognition system was developed to solve Out-Of-Vocabulary (OOV) words problem using morphemes as dictionary and language model units. For Small vocabulary (5k) system of morphemes as lexical and language modeling units was better than words and an absolute improvement (in word recognition accuracy) of 11.57% had been obtained [5]. In addition, there are researches done [6, 7, 8, 9] for speech recognition and Text-To-Speech (TTS) of Amharic and Tigrigna languages.

To the best of our knowledge, there is no published work on the development of speech recognition for Ge'ez. Since Ge'ez has its own features such as its characteristics of phonetic and stress, it is not possible to make use of an ASR developed for other languages. Therefore, the purpose of this study is to explore the possibilities of developing speech recognition for large vocabulary of Ge`ez language that helps anyone who uses Ge`ez to exploit the opportunities of information technology for the application of speech recognition. It also contributions to the automation activity of preserving and transferring of knowledge and heritage accumulated in Ge`ez language to the next generation. The knowledge might be accumulated in written form or transferred orally from past generations. The oral knowledge could be captured and documented in written formats. There are many elites in the church who can pass their knowledge in a spoken form, which can be documented in a written form using ASR. One of the ultimate goals of speech recognition is to create a human-like listening typewriter or automated secretary. In addition, ASR is used as part of Automated spoken language translation system and thus it can be used as one part to translate Ge`ez language to another language.

3. Data Collection and Preparation in HTK

Text and speech data should be prepared and made available before training the models and building the recognizer. This data is divided into 2 subsets; namely, training set and test set. The training set is used to train the recognizer while the test set is used to test the performance of the recognizer. All preparation and pre-processing steps are discussed in the subsequent sections.

3.1 Text Corpus

The set of speech recordings must be based on a proper written set of sentences and phrases. Therefore, it is crucial to create a high quality written (text) set of the sentences and phrases before recording them [17]. This work aims to produce Ge`ez phrases and sentences that are phonetically rich and balanced. To fulfil this aim, 250 sentences and phrases with a length ranging from 2 to 29 words are selected from prayer book (የቅዳሴ መፅሀፍ) and history book (የአጤ ልብነ ድንግል ታሪክ) written in Ge`ez . These selected sentences have been manually checked for grammatical, spelling, and abbreviation errors and have been manually corrected. Long sentences are shortened to have structurally simple sentences in order to ease readability and pronunciation.

3.2 Speech Corpus

Speech corpus is an important requirement for developing any ASR system. It is simply a set of recorded sentences. No readily available speech corpus exists so it is needed to be recorded. The database used for this work is a corpus with phonetically rich and balanced sentences and phrases used for training and testing the acoustic model. Baum-Welch algorithm is used to train the HMM models. This database is divided into training dataset and test dataset. The speech corpus was recorded by 11 male and 4 female Ge`ez speakers who represent a wide range of age group (20-50). The speech corpus was recorded using Audacity recorder software version 2.1.3 at a sampling rate of 16 KHz with a WAV audio file format. During the recording sessions, each speaker is asked to utter sequentially 20 different sentences (totally 200 different sentences). Additionally each speaker uttered the same 20 different sentences for training purpose. Thus, 400 utterances are recorded in total. A set of 100 additional sentences and phrases is uttered using the same speakers who participated in the training phase and another set of 100 sentences and phrases is recorded using 5 different speakers that did not participate in the training phase. AS Audacity is also audio editor software. Each utterance of the speaker is checked directly for errors and the duration of the pause between the words is also amended that affect the performance.

A total of 1.22h read speech corpus is collected with a minimum speech time of 1.21s, maximum speech time of 25.72s and an average speech time of 7.98s. The technical details of speech corpus are shown in Table 1.

Table 1: Speech Corpus Technical Details

Criterion	*Training*	*Testing*
No. of utterances (sentences)	400	Two test, 100 for the same speaker in training, 100 for different speaker (untrained)
Number of Words	1248	316 (116 from training data and 200 out of training data)
Average No. of Words/Sentence	11	7
Min and Max No. of Words/Sentence	2 and 29	2 and 19
No. of Speakers	10	5 different speakers that is not participated in training phase, each 20 utterances (sentences), 10 speakers that is participated in training phase, each 10 utterances (sentences)
Speakers' Age (years)	20 – 50	20 – 50
Speakers' Gender	8 males and 2 females	11 males and 4 females
Sampling Rate (Hz)	16 kHz	16 kHz
No. of Bits	16	16
No. of Channels	1 (mono)	1 (mono)
Size of Raw Speech	102 MB	18.8 MB

3.3 Transcription Files

Many of the operations performed by HTK assume that the speech is divided into segments and each segment should have a name or a label. The set of labels associated with a speech file constitute a transcription. Since there is no hand labelled data to bootstrap a set of models, a flat-start scheme has been used instead. Phone level transcriptions are expected to be prepared out of the utterances. This is because, as the system is phoneme based and then a triphone, later during training examples of the phones are needed to adjust the parameters of these models. That is, to train a set of HMMs every file of training

data must have an associated phone level transcription [6, 18].

3.4 Parameters Conversion of Speech Data

In order to build a set of HMMs for acoustic modeling, a set of speech data files and their associated transcriptions are required. The speech data file is converted into an appropriate parametric format (or the appropriate acoustic feature format) which consists of the required phone or word labels.

In this experiment, HMMs were trained using Mel Frequency Cepstral Coefficients (MFCCs) which are derived from FFT-based log spectra, consisting of the length of the parameterized static vector (+13), the delta coefficients (+13) and the acceleration coefficients (+13) which add up to 39 acoustic features. In order to code the data, a configuration file (config) is created which specifies all of the conversion parameters. The following is the setting used in this experiment.

```
# Coding parameters
SOURCEFORMAT = WAV          # Gives the format of the speech files
TARGETKIND = MFCC_0_D_A     # Identifier of the coefficients to use
# Unit = 0.1 micro-second:
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0       # = 25 ms = length of a time frame
TARGETRATE = 100000.0       # = 10 ms = frame periodicity
NUMCEPS = 12                # Number of MFCC coeffs (here from c1 to c12)
USEHAMMING = T              # Use of Hamming function for windowing frames
PREEMCOEF = 0.97            # Pre-emphasis coefficient
NUMCHANS = 26               # Number of filterbank channels
CEPLIFTER = 22              # Length of cepstral liftering
ENORMALISE = F
```

3.5 The Pronunciation Dictionary

The pronunciation dictionary, i.e., the lexical model, is one of the most important blocks in the development of large vocabulary speaker independent recognition systems [6]. In this paper, Tied-state triphones have been considered as the basic sub-word units that are extended from context independent phonemes. In the process of building a dictionary, the first step is to create a sorted list of the required words. The desired word list used to build a dictionary is extracted automatically from the text corpus using a script file that comes with HTK. Pronunciation dictionary that contains 1,448 unique words is prepared by taking words from a text corpus consisting of 250 sentences and phrases: 1248 words for training and 316 words for decoding (116 from training words and 200 words out of training words).

3.6 Language Model (LM)

Language model is another important requirement for any ASR system. LMs are used to constrain search in a decoder by limiting the number of possible words that need to be considered at any one point in the search. The consequence is faster execution and higher accuracy [19]. The LM used in this paper is a statistical language model, i.e., a backed-off bigram language model and the task grammar. Creation of a language model consists of computing the word uni-gram counts, which are then converted into a task vocabulary with word frequencies, generating the bi-grams from the training text based on this vocabulary, and finally converting the n-grams into a binary format language model and standard ARPA format [20]. It is developed using two of the CMU-Cambridge

statistical language modeling tools: the HLstats and HBuild. HLstats is used to gather and display statistical information for the label files and HBuild command is used to build the network.

4. Acoustic Model Training

Training the model means estimating the HMM's parameters, which are the mean, the variance, and the transition probabilities based on the utterance structure and the extracted parameters (features). The basic operation of the HTK training tools involves reading in a set of one or more HMM definitions, and estimating the parameters of these definitions using the speech data. Training consists of applying an embedded version of the Baum-Welch algorithm.

The first step in HMM training is to define a prototype model (with all means and variances initialized to 0 and 1 respectively). The topology of the model consists of the start and end states and three emitting states, using single Gaussian density functions. Then the initial parameter is estimated for the HMM model. It computes the global mean and variance and set all of the Gaussians in a given HMM to have the same mean and variance. Once the initial monophone model set is available, the next task is re-estimating the model parameters, i.e., estimating the transition probabilities of the Context-Independent HMMs.

So far, a monophone Acoustic Model is created that is not looking at the 'context' of the monophone [21]. The last step of model building is to transform the monophone HMMs into context-dependent triphone HMMs. With a triphone acoustic model, it is essentially looking for a monophone in the "context" of other monophones, i.e., the one immediately before and the one immediately after (if they exist - it may be the beginning or end of the word). There are 1883 unique triphones extracted from the training transcripts.

4.1 Recognizer Evaluation

After acoustic model training is completed, it is ready to evaluate the performance of the models. In this experiment, performance is assessed from five different speakers that do not participate in the training phase and from ten speakers that participated in the training phase. The accuracy of speech recognition system has always been measured by comparing the system's final output with a manual transcription performed by an individual.

The triphone-based systems have been developed in this experiment. For the tied-state triphone models, performance is evaluated for different Sp (duration between words) and also by changing only two of its parameters - the word insertion penalty (-p) and the grammar scale factors (-s) using task grammar and bigram as language model.

Tables 2 and 3 show the word recognition rates (%) for different Sp (duration between words) in milliseconds. The Pause between words can affect recognition accuracy whether the set time is too short or too long. It is better to speak at a normal pace like reading the news (avoid speaking one word at a time) since this improves accuracy for continuous speech. The pause setting of 200 milliseconds has better accuracy in this experiment.

Table 2: Triphone Tested Using Task Grammar

Sp duration in milliseconds	*WORD: % Correct*	*SENT: % Correct*	*Accuracy*
350	47.58	23.00	29.75
250	50.70	27.00	46.84
200	64.21	29.00	62.05
150	62.48	23.00	60.75

Table 3: Triphone Tested Using Bigram Language Model

Sp duration in milliseconds	*WORD: % Correct*	*SENT: % Correct*	*Accuracy*
350	50.48	0.00	9.47
250	53.68	0.00	25.67
200	60.75	0.00	29.44
150	57.72	0.00	27.27

The remaining test is done using the speech data with Sp duration of 200 milliseconds.

Table 4 shows the word recognition rates (%) when the recognition tool, HVite, is executed again on the triphone models by changing only two of its parameters: the word insertion penalty (-p) and the

grammar scale factors (-s). These parameters can have a significant effect on recognition performance. The –p value was incremented from 0.0 earlier to 20.0 and the –s value was taken up to 50.0.

Table 4: Triphone Tested Using Task Grammar

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
M (-p=0.0, -s=5.0)	61.47	22.00	39.39
M (-p=0.0, -s=20.0)	63.35	29.00	61.47
M (-p=0.0, -s=25.0)	59.74	27.00	58.15
M (-p=0.0, -s=50.0)	36.65	13.00	36.51
M (-p=10.0, -s=20.0)	64.21	29.00	62.05
M (-p=20.0, -s=20.0)	64.21	29.00	61.62

All experiments above are performed without including QS command in the tree.hed file. *Tree.hed* contains the instructions regarding which contexts to examine for possible clustering.

QS - Question - is defined by the user and each QS command loads a single question and that question is defined by a set of contexts. Table 5 shows the word recognition rates (%) of the experiment with tree.hed containing Question that is generated manually to test the effects of *tying*.

Table 5: Triphone Tested with Tree.Hed Using Task Grammar

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
Performance	27.71	11.00	24.68

The possible reason for this difference could be lack of complete and exhaustive information about the rules of Ge`ez. It would have been useful to determine the rules of Ge`ez that decide which consonants may cluster together in which order, and which are prohibited. AS *tree.hed* is generated manually to test the effects of *tying*, the rules specified in the script file *tree.hed* are by no means exhaustive in this study. This requires the involvement of scholars with strong linguistics background.

HVite is also executed on the triphone models by changing only the values of RO - the outlier threshold and TB that clusters one specific set of states. These values can have a significant effect on recognition performance as shown in Table 6. The RO value is changed from 100 earlier to 200 and the TB value was incremented from 350 to 750.

Table 6: Triphone Tested with tree.hed with Different RO and TB Values Using Task Grammar

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
RO=100, TB=350	25.11	4.00	21.93
RO=200, TB=350	27.27	4.00	24.39
RO=200, TB=750	27.71	11.00	24.68

Finally, performance is done for untrained speakers. This test is performed using the speech data with Sp duration of 200 milliseconds as well as the -p value was 10 and the –s value was 20. Table 7 summarizes the recognition accuracy of speakers who do not participate in the training phase.

Table 7: Triphone Tested with Untrained Speakers

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
Performance	62.19	24.00	57.72

5. Conclusion

This paper was devoted towards the development and evaluation of speech recognition system for continuous speech Ge'ez language of multiple speakers. In addition to researches done previously by other researchers which have great contributions on related local language, this experiment shows a promising result to design an applicable system that recognizes Ge`ez language speech with more enhancements.

However, this work is not exhaustive due to some factors such as limitation of freely available Ge'ez speech corpus to adequately train such models. This system is also based on the statistical approach of HMM. But it is possible to design better speech recognition system based on a recently introduced artificial neural network (ANN) technique and a hybrid system that combines ANN with other state-of-the-art techniques to improve performance

References

[1] Speech Recognition, -whitepaper- ASR,: http://www.docsoft.com/resources/Studies/ Whitepapers/whitepaper- ASR.pdf, Last accessed on June, 2017.

[2] Solomon Teferra Abate, Martha Yifiru Tachbelie, and Wolfgang Menzel, "Amharic Speech Recognition: Past, Present and Future", Proceedings of the 16th International Conference of Ethiopian Studies, Trondheim 2009.

[3] Kinfe Tadesse, "Sub-Word Based Amharic Word Recognition: An Experiment Using Hidden Markov Model (HMM)", 2002.

[4] Solomon Tefera Abate, Wolfgang Menzel, and Bahiru Tefla, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Re-Cognition", In Proc of 9th Europ. Conf. on Speech Communication and Technology, Interspeech, Lisbon, Portugal, 2005.

[5] Solomon Teferra Abate, Martha Yifiru Tachbelie, and Wolfgang Menzel, "Morpheme-Based Automatic Speech Recognition for Morphologically Rich Language – Amaharic", Department of Informatics, University of Hamburg Germany, https://pdfs. semanticscholar.org/.../3483737c4f0053dacf2fd 5e1b.

[6] Hafte Abera Miruts, "Hidden Markov Model Based Tigrigna Speech Recognition", Unpublished Masters Thesis, Addis Ababa University, 2009.

[7] Mesfin Birile, "Synthetic Speech Trained - Large Vocabulary Amharic Speech Recognition System", Unpublished Masters Thesis, Addis Ababa University, 2008.

[8] Michael Melese, Laurent Besacier, and Million Meshesha, "Amharic Speech Recognition for Speech Translation", Addis Ababa, 2016.

[9] Alula Tafera, "A Generalized Approach to Amharic Text-To-Speech (TTS) Synthesis System", Unpublished Masters Thesis, Addis Ababa University, 2010.

[10] Ge`ez Language, – Wikipedia the free encyclopedia, https://www.en.wikipedia.org/ wiki/Ge%27ez, Last accessed on May, 2017.

[11] Wolf Leslau, "Comparative Dictionary of Ge‟ez (Classical Ethiopic): Ge‟ez-English/English-Ge‟ez with an index of the Semitic roots," Wiesbaden Harrassowitz, 1991.

[12] Mulusew Asratie, "The Syntax of Non-verbal Predication in Amharic and Geez", LOT, The Netherlands, http://www.lotschool.nl, 2014.

[13] Pilar Quezzaire-Belle, "The Comparative Origin and Usage of the Ge'ez Writing System of Ethiopia", 2001, https://www.scribd.com/ document/50818987/paper-gabriella.

[14] www.persee.fr/doc/ethio_0066-2127_1957_num_2_1_1264.

[15] https://books.google.com/books?isbn =1575060191, 1997.

[16] ትንሣኤ ግእዝ, አፍሮ የሕትመት ስራ ድርጅት, 2012.

[17] Mohammad Abushariah, Raja Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman Khalifa, "Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus", Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.

[18] Steve Young, Gunnar Evermann, and Dan Kershaw, "The HTK Book", Cambridge University, 2015.

[19] What-is-a-language-model, www.voxforge.org/ home/docs/faq/faq/, Last accessed on August, 2017.

[20] Speech recognition, – Wikipedia the free encyclopedia, https://en.wikipedia.org/wiki/ Speech_recognition, Last accessed on June, 2017.

[21] Create Acoustic Model Manually: Tutorial, 2017, www.voxforge.org, Last accessed on November, 2017.

Design and Implementation of Web Based E-Learning Management System: The case of HiLCoE School of Computer Science and Technology

Ephrem Abeselom
HiLCoE, Computer Science Programme, Ethiopia
ephabel@gmail.com

Workshet Lamenew
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
workshet@gmail.com

Abstract

Supporting the delivery of higher education with educational technologies is now becoming a common trend. As a result, many open source platforms emerged. Launching e-Learning system will be successful if prior investigation is made with issues related to need identification and environment analysis. HiLCoE was using Moodle for its e-Learning needs but the system stopped its operation when the College relocated from its Atlas campus. The objective of this paper is to assess factors leading to the failure of the e-Learning system with the view to design and develop state-of-the art e-Learning platform for the College.

To achieve the objective, interview, discussion and personal observation methods are used for data collection. A total of nineteen respondents were chosen in the process. An object oriented system development methodology and unified modeling language was used as analysis and design approach. The platform architecture was designed and the prototype was developed using HTML, PHP, and MySQL. Finally, the prototype was evaluated using user acceptance testing. Major functionalities are included in the designed e-Learning system including registration of a user as a student or instructor, approving or declining new users, posting news, notification, adding course, viewing notification and news are designed as well.

Among others, due to poor alignment to needs, poor implementation process, lack of management commitment and lack of support and technology, adoption of freely available e-Learning platforms might not be successful and systems will be much dependent while developed based on consideration of those factors and hence the system was developed with this rationale.

Keywords: E-Learning; Educational Technology; Moodle

1. Introduction

Learning is the act of acquiring new or modifying and reinforcing existing knowledge, behaviors, skills, values, or preferences which may lead to a potential change in synthesizing information, depth of the knowledge, attitude or behavior relative to the type and range of experience [1]. Exposure to learning experiences can be formal or informal. In the formal educational environment one can grasp knowledge by attending regular classes, distance learning or online learning as his/her convenience. In order to reach their students, institutions are now providing their teachings in various formats.

With the rapid increase of ICT infrastructures, every educational institution has the opportunity to make use of the Internet as a communication medium with the students [2]. For an effective and efficient access to learning materials, the concepts and methodologies of technology-based learning are increasing in importance with web based learning and becoming a crucial resource for institutions.

E-Learning is a learning that is enabled or supported by the use of digital tools and content. It is usually accessed via the Internet though other technologies are also used. It typically involves some form of interactivity, which may include online

interaction between the learner and the teacher or peers [3]. The operating environments for electronic learning can be rich, interactive, dynamic and customizable, connecting learners with an almost limitless wealth of information which in turn paves the way for new patterns of learning to emerge. Therefore, an increasing emphasis is given on information literacy, increased flexibility as to where, when and how people learn, and exploration of new ways in which learners can be empowered to structure and manage their own learning experiences.

Many e-Learning systems emerged as a result, Blackboard, Saba, Moodle, ATutor, Dokeos and Ilias to mention some. Among them Moodle is very popular and widely accepted across the globe.

Even though these systems are giants on the market, due to cultural, technological, infrastructural and environmental factors they might not be successful in all countries while adopting them. The challenge for instructors to develop and maintain an e-Learning course, security and technical issues are also additional headaches.

HiLCoE School of Computer Science & Technology is a leading institution and a center of excellence that adheres to computing technology and quality of education since August 1997 [4]. The school was using the open source platform Moodle for supporting its e-Learning offerings until it stopped working while they left Atlas campus. Since then it seems the emphasis given on supporting the delivery of learning using the fruits of e-Learning technologies are less.

Currently access to information is limited to only working hours, submit assignments and do related activities by being physically present at the school. Thus, the purpose of this paper is to analyze factors that lead to Moodle failure and based on the findings, analysis was made on how to enhance the current teaching-learning process of HiLCoE and designed state-of-the art web based e-Learning system.

2. Related Work

A case study to answer the question "why do e-Learning projects fail?" identified thirty three causes of failure [5], some of which are listed in Table 1.

Table 1: Factors Leading to Failures for e-Learning Projects

Factor	*Reason*
Poor Alignment to Needs	▪ Previous poor experiences of e-Learning that is boring or not relevant to individual and business needs.
Communication	▪ e-Learning project teams don't manage the expectations of business managers.
Lack of Implementation Skill	▪ Weak or ineffective ICT competence in the e-Learning project management. ▪ Content creation by teams who lack an understanding of learning processes and design issues. Content creation by teams who lack an understanding of the potential of IT and web resources. ▪ Errors in key administrative tasks which give a bad impression to learners, insufficient or inadequately trained staff or badly executed systems. ▪ Resistance from Trainers: Existing trainers do not integrate ICT into their educational practice and have difficulties adapting to new requirements of technology. ▪ Course content and delivery is too rigid, fails to give individual learners choice and flexibility and fails to respond to sudden business change; the syllabus is too rigid and too difficult to change. ▪ Not learning from others' experience (no benchmark comparison of success

Factor	*Reason*
	taken against other organizations elsewhere).
Poor Implementation Process	▪ Over reliance on synchronous events after the early stages of introduction. ▪ No pilot process for step-by-step development leading to large scale rollout. ▪ There is no in-built e-assessment, usage measurement, record keeping, evaluation recording, and links to competency measurement or business evaluation. ▪ Development and deployment too slow for the business who finds other solutions perceived to be faster.
Management Commitment	▪ Senior management do not sponsor projects and fail to encourage and lead adoption. ▪ Users of social networks and collaborative learning tools run out of enthusiasm and there is lack of senior organizational sponsorship. ▪ Changes in personnel at key stage of project development leading to lack of continuity/vision of the original goals. ▪ IT departments who fail to understand and manage the issues of web use, IT policy, location and support leading to access and firewall problems.
Scalability	▪ Early content implementations are heavily customized and insufficient attention paid to the future amendment and updating leading to rigid course content that is expensive to change.
Support	▪ Introduction can result in unplanned increases in workload of existing training staff; i.e., e-tutoring, content development, lesson preparation (virtual classrooms), learning to use learning technology tools or providing a help desk. ▪ There is no plan for the substantial human support services that are needed to blend e-learning into the mix of delivery options. Too much reliance is placed on e-learning as a stand-alone solution.
Technology	▪ Over reliance on technology tools alone with insufficient attention paid to getting the right blend of learning methods and learning tools.

Even though there are many e-Learning systems available in the market, content available for learning on the web is variable and thus the needs of content developers, educators and students can't be addressed that leads to a gap [6]. In order to bridge the gap, the authors developed an e-Learning system.

For identifying the gap between actual behaviors and desired outcomes they used surveys, direct and indirect observation, interviews and focus group discussions as tools.

They designed the system with available tools and tested the prototype and found that the systematic use of e-Learning as part of the instructional design process will improve the learning process.

3. The Proposed Solution

In this section, overview of the proposed system with its functional and non-functional requirements, analysis model using use case diagram, use case descriptions and sequence diagrams are presented. Requirement gathering and analysis is done through interview, discussion and personal observation. In the process of requirement elicitation, we interviewed two previous and current system administrators, discussed with five undergraduates, five each from postgraduate Computer Science and Software Engineering students and personal experiences on how the teaching learning process is directed. Various documents are revised as secondary data

source such as documents related to why e-Learning projects fail, advantages and limitations of e-Learning systems, preconditions in web based e-Learning development for best design and implementation of the system. As an analysis and design approach, object oriented methodology was chosen and UML was used because it makes it easier for adding new functionalities and modifications.

According to personal experience and the discussions made with school colleagues, HiLCoE is currently using

- E-mail for contacting instructors and submitting assignments.
- Physical presence at the Registrar to view course grades.
- Notice boards to post notifications, class schedules, call for papers, vacancies and announcements.

The use case diagram shown in Figure 1 was drawn from the functional requirements.

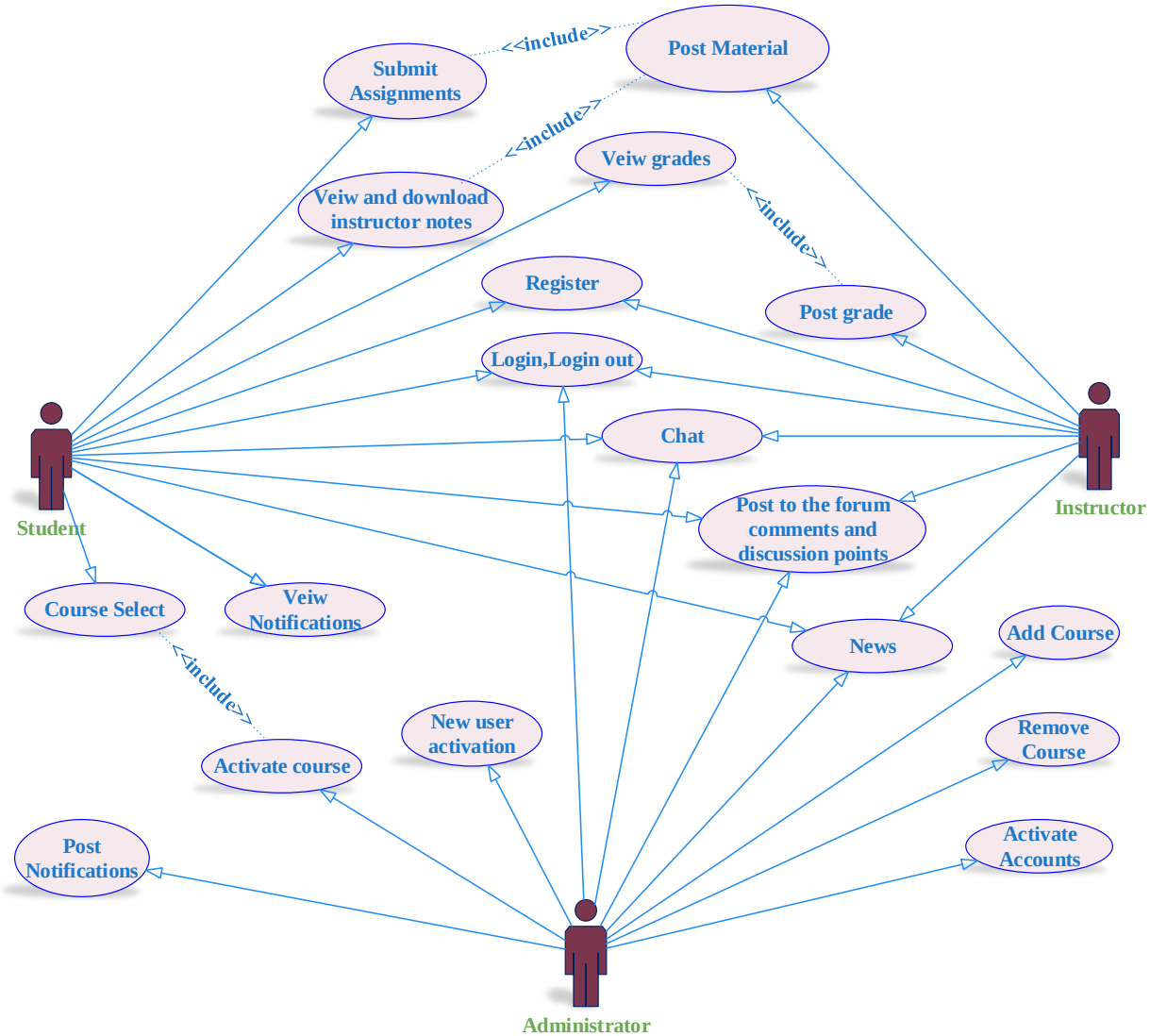

Figure 1: Use Case Diagram for the Proposed E-learning System

As shown in Figure 2, the architecture used in the e-Learning system is three tiered: presentation tier, logic/application tier and data tier. This type of architectural style is characterized by the functional decomposition of applications, service components and their distributed deployment which provides improved scalability, availability, manageability and resource utilization.

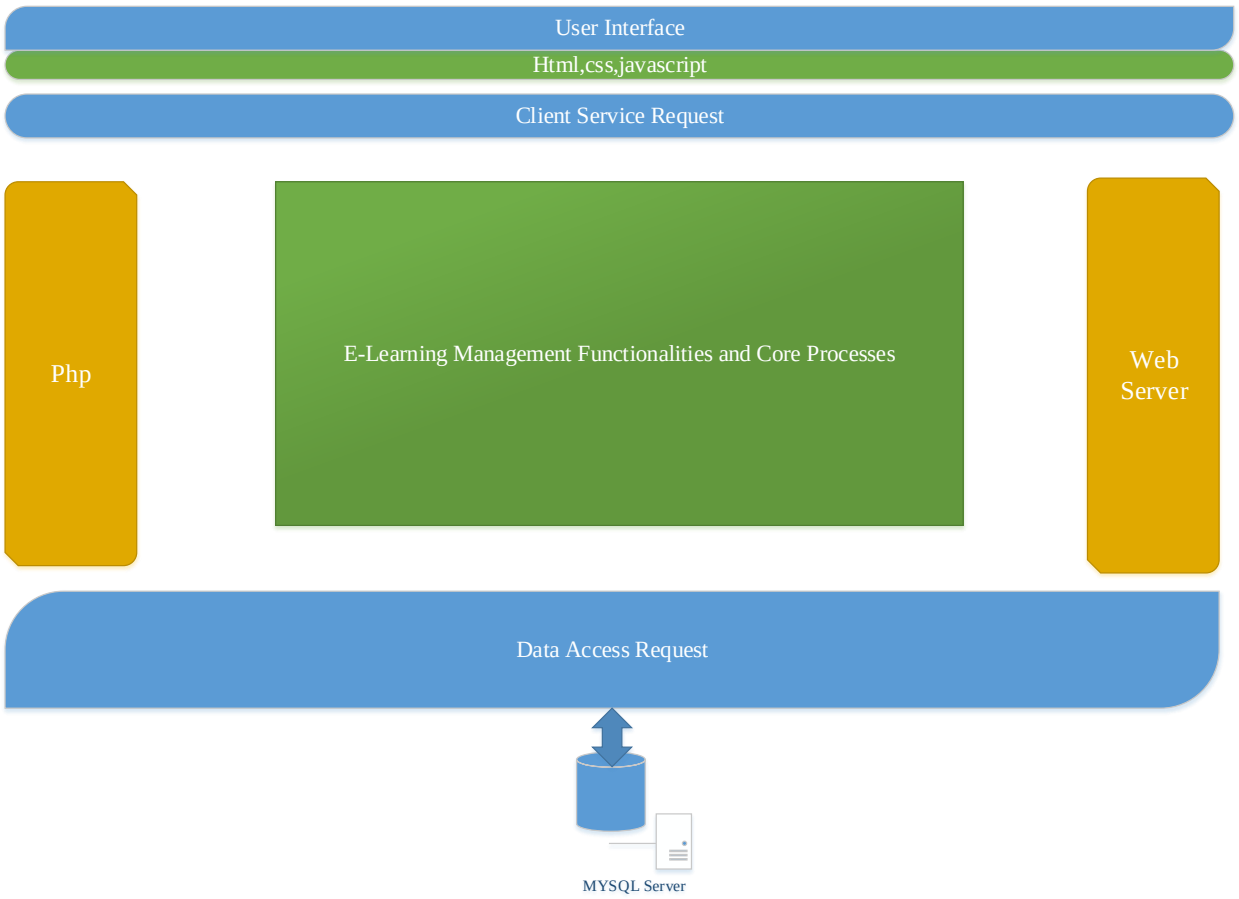

Figure 2: System Architecture

For the realization of the proposed solution, the use cases and sequence diagrams are converted to a source code and integrated in such a way that they work together for common purpose and to fill the gap between the analysis model and its realization. The home page of the prototype is shown in Figure 3.

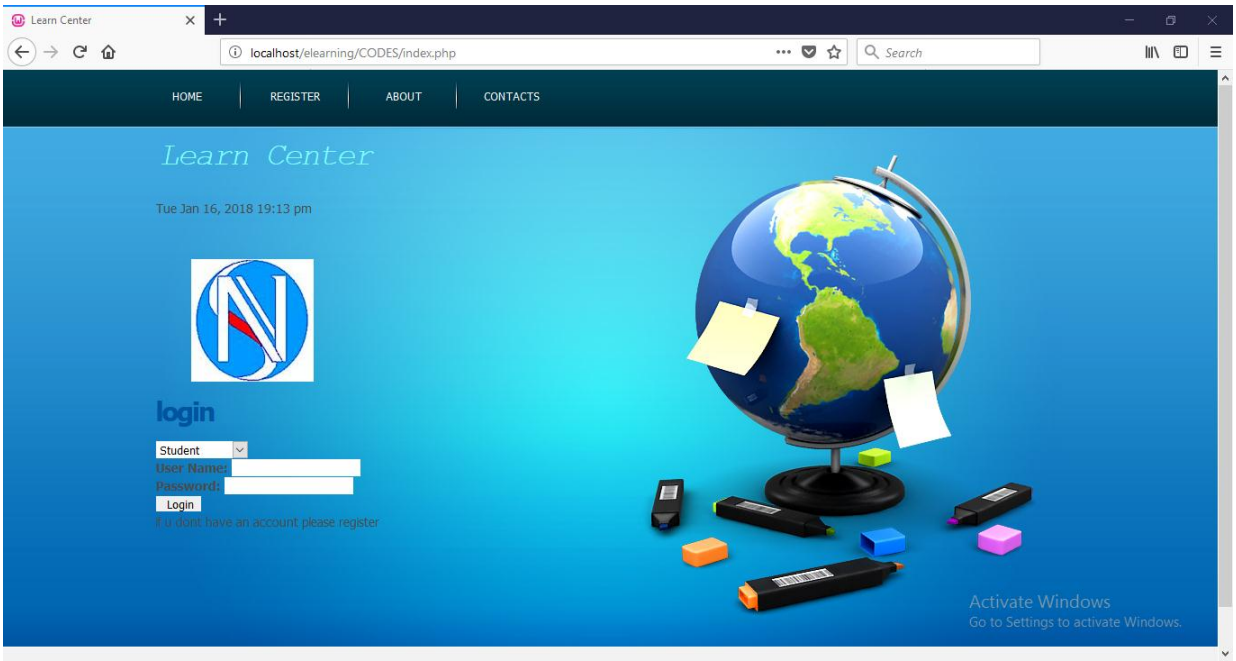

Figure 3: Home Page

The evaluation of usability is an important aspect of software design and development. In order to measure the success of the design, designers must evaluate the system. The evaluation of the e-Learning

platform used a method of questionnaires with 9 major usability criteria to evaluate different aspects of the prototype developed. The result is summarized in Table 2.

Table 2: Usability Testing Summaries for the Developed E-learning Management System

No.	*Question*	*Strongly agree*	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
1.	The steps required to install and run the e-Learning system on the web server are simple and easy	17	1		1	
2.	The e-learning system's user interface is attractive	19				
3.	Registration is facilitated easily	14	4		1	
4.	The e-Learning system provides easy navigation	18	1			
5.	Logging into the system is facilitated easily	18	1			
6.	If an error occurs, the application provides appropriate support to manage the error	15	2		2	
7.	The system is sufficient to fulfil the College's e-Learning need	13	4		2	
8.	The application is easy to use	17	2			
9.	The security of users is maintained properly	19				
Average Result		16.7	2.14		1.5	
Result in Percentage		87.72	8.78		3.5	

A questionnaire was distributed for a total of 19 respondents. The values of the responses were taken based on the Likert scale. According to the result of user interface evaluation as shown in Table 2, most of the respondents (87.72%) strongly agree that the system prototype has an interface that is easy to use, attractive and the security is maintained properly.

For the question regarding weather the developed system satisfies their need or not, 13 out of 19 respondents (68.43%) strongly agreed while 2 out of 19 respondents (10.53%) are not satisfied. This shows that still there are unseen needs which are not accessed with the requirement elicitation process. Regarding easiness of steps required to install, configure and run it on web servers, 17 out of 19 respondents (89.47 %) strongly agreed.

4. Conclusion and Future Work

Due to poor alignment to needs, poor implementation process, lack of management commitment and lack of support and technology, adoption of freely available e-Learning platforms might not be successful and leads to frustration on system users. Analysis of business requirements and operating environments must be done to launch systems. Moodle was adopted as an e-Learning platform for supporting the e-Learning needs of HiLCoE, but was not successfully implemented.

In this paper, we identified the failure factors for Moodle at HiLCoE and according to the analysis made, poor alignment to needs while choosing contents because of lack of prior need assessment and requirement elicitation, changes in system administrators and location of the College from Atlas campus led to its failure. Therefore, based on consideration of the failure factors, designing and developing a web based e-Learning management system can solve the problem so that the continual of the system which supports the delivery of education can be benchmarked.

The e-Learning management system was developed using PHP and runs on a broad variety of browsers. The usability test for the system was performed and the evaluation has shown the success of the web application.

Finally, the following recommendations are drawn:

- Additional need assessment using requirement elicitation and validation techniques must be done because 10.53% of the respondents are not satisfied with the developed system.
- We developed the system using a standard web service with the implementation of authentication and authorization only. Other security requirements like confidentiality, integrity, auditing and availability must be implemented in order to minimize risks on the school.
- We will identify additional gaps and factors that are prevalent in the area of e-Learning management system.
- Ease of use and other parameters are also research areas for further exploration.

References

[1] Learning-Wikipedia, "Learning", http://en. wikipedia.org/wiki/Learning, Last accessed on July 18, 2017.

[2] Waterhouse, S, "The Power of E-learning: The Essential Guide for Teaching in the Digital Age", Pearson Education Inc., Upper Saddle River, 2005.

[3] Waikato Research, "Introduction to e-Learning", White Paper, http://www. waikato.ac.nz/tdu/pdf/booklets/17eLearningIntro.pdf, Last accessed on April 04, 2017.

[4] HiLCoE School of Computer Science & Technology, Graduate Program Research Office, "M. Sc. Thesis/Project Guideline", Version TPG HiL11/2011, 2011.

[5] Towards Maturity Briefing, "Why Do E-Learning Projects Fail?", White Paper, 2010, https://www.aoc.co.uk/sites/default/files/Why_Elearning_Projects_Fail.pdf, Last accessed on August 10,2017.

[6] Noesgaard S. S. and Ørngreen R., "The Effectiveness of E-Learning: An Explorative and Integrative Review of the Definitions, Methodologies and Factors that Promote e-Learning Effectiveness", The Electronic Journal of e- Learning, Vol. 13, Issue 4, 2015, pp. 278-290.

Data Sharing System for Federal Government Institutions in Ethiopia

Fekadu Fikre
HiLCoE, Software Engineering Programme, Ethiopia
Information Network Security Agency, Ethiopia
fekadufkr@gmail.com

Fekade Getahun
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
fekade.getahun@aau.edu.et

Abstract

The need for sharing data between different governmental institutions in Ethiopia has been increasing along with the technology and economic growth of the country. Organizations are struggling to allow seamless and standardized data sharing and synchronization with the intention of data verification and avoid data redundancy. Ethiopia is currently well aware of its position and putting all rounded efforts to improve the services provided through the use of Information and Communication Technology (ICT). As a result, the present e-Government service initiatives are expected to share data using a standard data set.

This paper explores data sharing methods dedicated to handle foreign procurements that share common data exchange problem of federal government institutions. The proposed solution advocates the use of design science methodology. To do so, first the current efforts and technologies towards using central data sharing and integration of institution's internal systems using web service have been assessed.

The system is developed taking into account the state-of-the-art data sharing approaches and related technologies, identify the specific requirement of common data sharing system, propose and design an architecture for the Common Data Sharing System, and develop and implement a prototype to demonstrate the practicability of the designed system. Then information about existing systems and system requirements for data exchange is gathered. Based on the assessment and system requirement, a system is designed and a prototype is implemented. The system is evaluated by domain experts and the analysis shows acceptable result.

Keywords: Data Sharing; Foreign Procurement; Web Service; Common Data Sharing

1. Introduction

Technological advancement raised expectations for data sharing and data communication has brought the issue into sharper focus. With the growth of the digital economy, data sharing has become an essential business practice with different groups in the same organization, or among partners in larger platform endeavors, or even, as in growing common data movements, with the public. Data Sharing enables new insights from existing data, and lets organizations make full use of this core resource. Ensuring and improving data sharing between governmental organizations has become more important in recent years. Interaction of data between organizations improves outcomes in areas such as efficiency and low cost. Increased and meaningful data communication between sectors is likely to enhance performance.

Rapid data exchange, intense interaction and sharing of resources and information between different stakeholders become essential for enterprises in order to improve their productivity and services. ICT is crucial for the economic growth of a country. It is one important input for performing different tasks in several sectors including Municipalities, Banks, Insurances, Customs, Transits, Education, Health, Transportation, Justice, etc. As a result, information processing and delivery becomes an important part of an organization's overall operation [6].

ICT has dramatically altered the way companies manage their supply chain and has spawned a variety of new inter-organizational management approaches.

Tremendous cost and delay reductions from information sharing, and the ability to use advanced Information Technology (IT) to exploit superior expertise outside the boundaries of the firm have resulted in a number of virtually integrated government sectors and independent organizations that operate as a single vertically integrated firm. A common form of virtual integration results from sharing common information across organizations [5].

Nowadays IT is starting to spread all over Ethiopia and its significance is also becoming widely understood by the society. As a result, common data sharing has been a recurring focus of struggle within the community as improvements in IT and digital networks have expanded the ways data can be produced, disseminated, and used. But data disclosure also initiates research competitors [3].

Information management is everybody's business, the business of a government, as that of any ongoing organizational entity, involving strategic and long-term decision making involving goal setting, long term planning, resource allocation and monitoring and evaluation, and operational management and public service delivery. These business processes of the Government and its decision-making processes are based on the availability of the right information at the right time and at the right place. Empirical research in decision sciences and information management showed that the quality of management and decision-making increases proportionally with the availability of required information. Evidently, with better information and data availability, the uncertainty associated with decision making decreases [1].

It is known that Ethiopia is one of the countries with little ICT utilization. The government invested in the acquisition of ICT devices, encourages the use of ICT in the country, and most of the government offices invested in IT products acquisition. However, most of the organizations work in a manner similar to manual or semi-automated methods without coordination and sharing. Most of these government offices provide various services to citizens and yet keep their own copy of the same data and without shared by others.

In the current Ethiopian situation, data sharing among government offices is difficult and there is no way one can get data/information owned by another institution and validate the identity of the individual. The absence of data sharing environment forced organizations to request the same copy of data from service requester (i.e., citizen). One reason for this is that they work with little automation and not managed in an integrated way.

When organizational data are properly organized, preserved and well documented, and their accuracy, validity and integrity is controlled at all times, the result is high quality data, efficient, outputs based on solid evidence and results in saving of time and resources [2]. When it comes to sharing organizational data, good management is essential to ensure that data can be preserved and remain accessible in the long-term, so it can be re-used and understood by other organizations [2]. In addition, properly managed and preserved organizational data can be successfully used for future scientific and educational purposes, thus maximizing the investment made in generating the data and increasing the visibility of the organization [4]. The value of data to be preserved depends on the quality and efficiency of data management.

Nowadays in Ethiopia, federal government institutions are well aware of their present position and are putting all round efforts to improve their service provision through the use of ICT. Some organizations are engaged in the design and development of web-based applications to automate service delivery [10].

2. Related Work

Nowadays, technological advances have raised expectations for data sharing. Financial and Service exigencies have brought the issue into sharper focus. The use of technology has the potential to support data communication and consequently alleviating

some of the data communication challenges and barriers of time and distance between institution's experts that may face [14].

Data sharing and accessibility policy applies to the sharing of all community data and information collected, generated, and archived. In principle this policy is expected to strengthen the sector's capacity to effectively manage, curate, and analyze community data and publish research output as well as to foster balanced collaborations and partnerships with the scientific community [9, 13].

Several surveys have explored data sharing and withholding practices that are mostly framed from the perspective of data producers on the concerns or benefits of sharing. These studies explored where and how researchers are willing to share data, and what are the motivations and disincentives for sharing [8]. The results generally show minimal sharing practices and lack of incentives, which involve time, funding, labor, policies and standards, competition and ownership, conventions and discourses, and technical capabilities, as well as other limiting factors [4, 15].

Sharing of data can be done in several ways such as depositing in a common data repository and has many advantages [7, 8]. There are various benefits to diverse groups of users which include the following:

- *Avoiding duplication:* Sharing common data by different sectors results in significant cost savings in data collection and avoids data inconsistency.
- *Maximized integration:* By adopting common standards for the collection and transfer of data, integration of individual data sets may be feasible.
- *Maximizing data use:* Ready access to data will enable more extensive use of valuable public and institutional resource for the benefit of the data user and the federal government sectors at large.
- *Better decision-making:* In developing countries like Ethiopia, decision making by different government sectors has a great impact on the country's economy and lives of the citizens.
- *Equity of access:* A more open data transfer policy ensures better access to all authorized users in different sectors.
- *Beneficence:* Ensures maximum benefit to the community from which data and information is generated [9].

Web Services represent an architectural structure that allows communication between applications. A service can be invoked remotely or be used to employ a new service together with other services [11, 12]. Web Service applications are expected to be interoperable with other independent internal systems in each organization that are implemented in different platforms and programming languages [16, 17].

Web services are very important because they provide interoperability between loosely-coupled applications across distributed heterogeneous systems. A Web service is an interface that describes a collection of operations that are network-accessible through standardized XML messaging. These systems can include programs, objects, messages, or documents [17]. Web services are built on top of open standards such as TCP/IP, HTTP, Java, HTML and XML.

3. The Proposed Solution

3.1 Foreign Procurement Data Exchange

The exchange of foreign procurement data between the buyer organization and other concerned government institutions for further procurement process and their own decision making purpose is made through different ways, for example, using forms, formats, invoices, letters and telephone communication between the institutions.

The procurement data is recorded and collected from the buyer organization (e.g., Information Network Security Agency), Banks (National Bank of Ethiopia and Commercial Bank of Ethiopia), Insurance Company (Ethiopian Insurance Company) and Transistors (Ethiopian Revenue and Custom Authority and Ethiopian Shipping and Logistics Services Enterprise). The records are processed using letters and paper formats and usually kept on shelves and file cabinet boxes. Every foreign procurement data is recorded and stored on different record

formats and separate places like buyer's, banks, transistors and insurance institution.

i. Foreign Procurement Data Registration

The buyer institution registers the Foreign Procurement data. These are:

- The basic information about the supplier, and service providers information,
- Foreign Currency Exchange,
- Imported Goods and related information, and
- Claim recording and related information.

ii. Insurance Data Registration

The insurance company registers the insurance data relating to the foreign procurement services which includes the following:

- Insurance data registration for foreign procurement processing and
- Insurance for bank notification recording.

iii. Bank Services Data Registration

The banks register the bank data relating to the foreign procurement services which are:

- Foreign Currency Exchange approval information regarding foreign procurement and
- Letter of Credit (LC) recording information for foreign procurement processing.

iv. Transit Data Registration

The transistors register the transit data relating to the foreign procurement services. These are:

- Import shipping agreement information,
- The Container agreement information, and
- Bank Clearance and related information.

3.2 System Design Goals

The proposed data sharing system is designed based on the functional and non-functional requirements and analysis models. It describes the quality of the data sharing system that should be satisfied for the proposed solution.

a. Performance Criteria: indicate the speed and space requirements of the system.

b. *End User Criteria:* include qualities that are desirable from users' point of view.

Usability: The data sharing system should be designed to have an attractive and convenient interface for less experienced users, experienced users and system administrators. The system should be user friendly, and easy to learn and use.

c. *Dependability Criteria:* are used to determine how much effort should be spent in minimizing system crashes and their consequences.

Security: The important point in the design of the system is to ensure security because the system will only be reliable when confidential information can be transmitted securely. The system should not allow non-authorized users using a form based authentication.

d. *Maintenance Criteria:* determine how difficult it is to change the system after deployment.

Extensibility: The system has to be properly documented so as to facilitate the addition of new functionalities.

Portability: deals with how easy it is to port the system to different platforms. The system is developed using Java and it can run anywhere (i.e., it is platform independent).

The proposed data sharing system is designed to facilitate the sharing of common data and integrates organizations using their data at the national level having their own unique identification key. The system can allow registration and approval of data in procurement process and related data to be used and processed by other sectors across the nation as they request other organizations' data. In each institution, local data are accessed using web services. That is, when data is needed to be shared from the institution's internal database, it is accessed using the web service.

The architecture of a system describes its major components, their relationships (structures), and how they interact with each other. It involves set of significant decisions about the organization related to software development and each of these decisions

can have a considerable impact on quality, maintainability, performance, and the overall success of the final product. A centralized, Web-based data archive is the most suitable platform for the common data sharing systems. A centralized database is a database that is located, stored, and maintained in a single location. A major advantage of a centralized database is that data are prominently available in a uniform and readily searchable format.

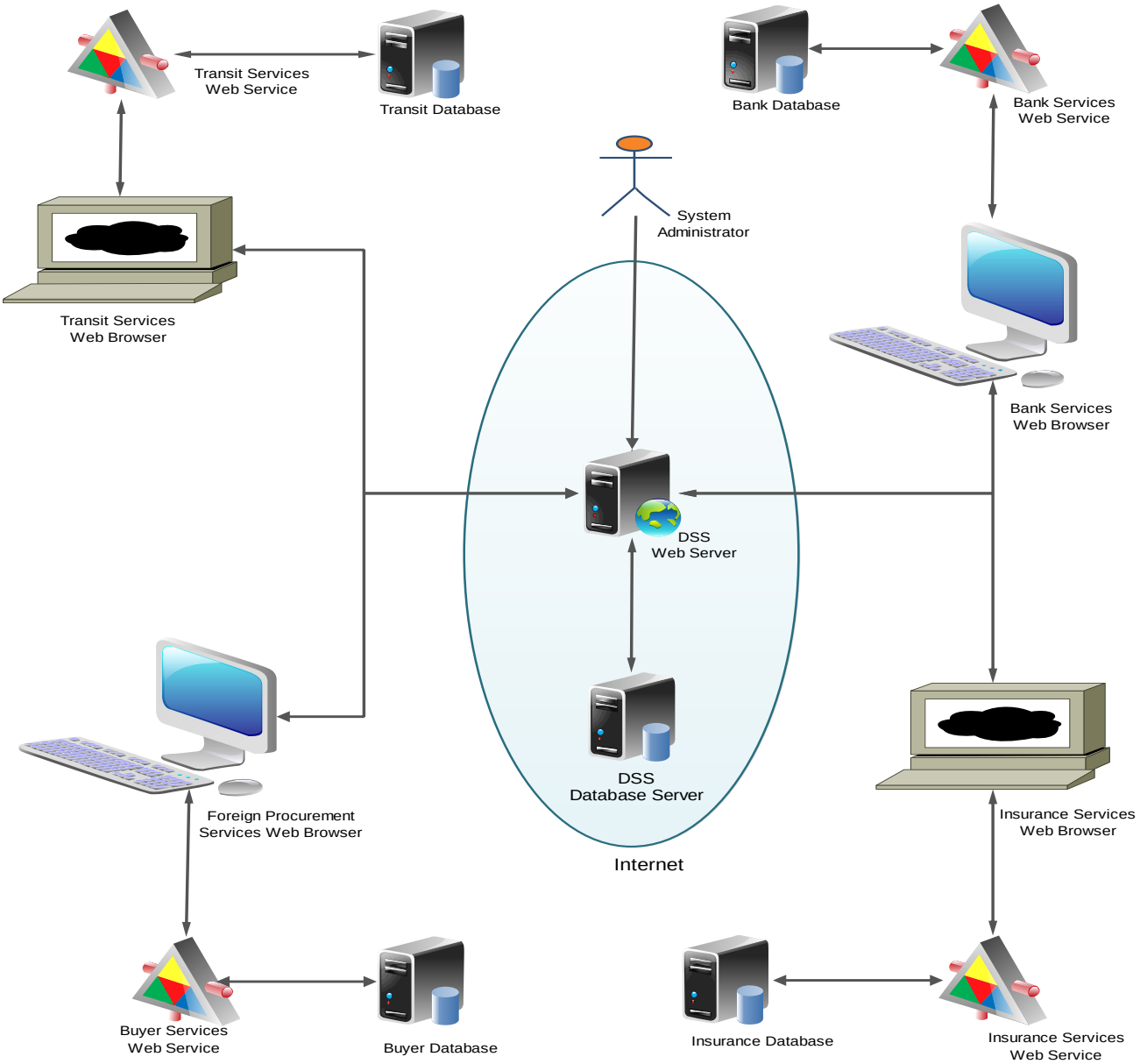

Figure 1: The Proposed Data Sharing System's Architecture

In data-centered architecture, the data is centralized and accessed by other components, which modify data. The main purpose of this style is to achieve integrality of data. Data-centered architecture consists of different components that communicate through shared data repositories. The components access a shared data structure and are relatively independent in that they interact only through the central data repository.

3.3 Architecture Components

The proposed Foreign Procurement Data Sharing System architecture has five components as shown in Figure 1. The description of each component is given below.

a. *Foreign Procurement Component:* enables users from the Buyer's institution to register foreign procurement records. It is responsible for adding

or registering all valid data about foreign procurement information regarding a particular organization's request.

b. *Insurance Services Component:* enables the Ethiopian Insurance Company users to register insurance related information, which is used as a prerequisite for banks information processing.

c. *Bank Services Component:* enables bank users (NBE & CBE) to register bank related information.

d. *Transit Services Component:* enables users to register the procured items transit related information. This component uses Insurance and LC information for further steps.

e. *Web Services Component:* is responsible to integrate the system's data with internal systems of each institution. It is used to access and fetch data that will be used by the common data sharing system in each of the Federal Government institutions from their internal database(s).

f. *Database component:* information related to foreign procurement data are stored in the central repository. The system uses MySQL database management system to create databases, tables, relations, views, and stored procedures.

4. Implementation

The system is required to assign each sequential procurement transaction Id (unique sequence number that will not be reused by any other government institution). The system needs to have a facility for authorized personnel to review the individual organization's information in order to check and approve foreign procurement transactions in federal government institutions.

4.1 Prototype of the System

A prototype was developed to test that the proposed solution meets the requirements specified and the design goals and quality attribute considerations set in the system architecture design. The programming language used to develop the system is Java with MySQL 5.1 database management system. User interfaces are developed in both English and Amharic languages.

i. User Login Page

System Admins and other system users can log in into the system using their privileges as shown in Figure 2.

Figure 2: User Login Page

ii. Letter of Credit Registration Page

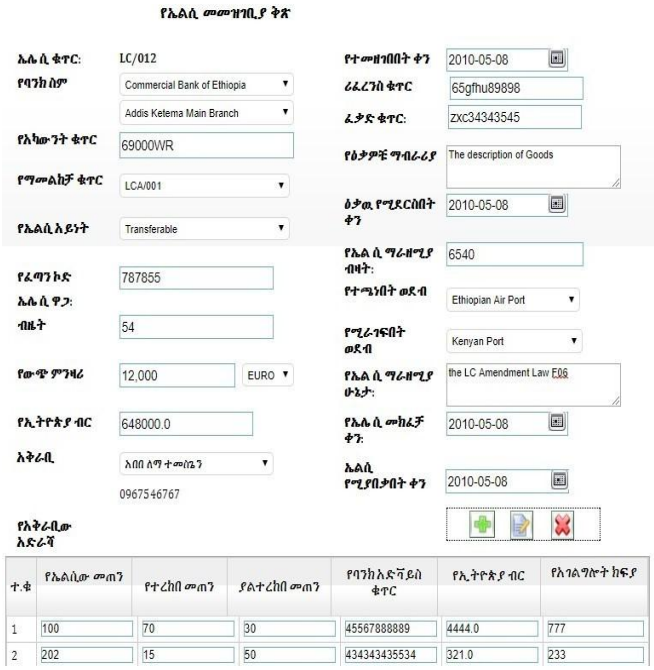

Figure 3: Letter of Credit Registration Page

Figure 3 is used to register Letter of Credit information by the buyer institution's officer and approved by the bank heads (NBE/CBE).

iii. Import Shipping Registration Page

The page shown in Figure 4 is used to register the Import Shipping and related information by the buyer's institution officer and approved by the concerned Transit (ERCA/ ESLSE) head.

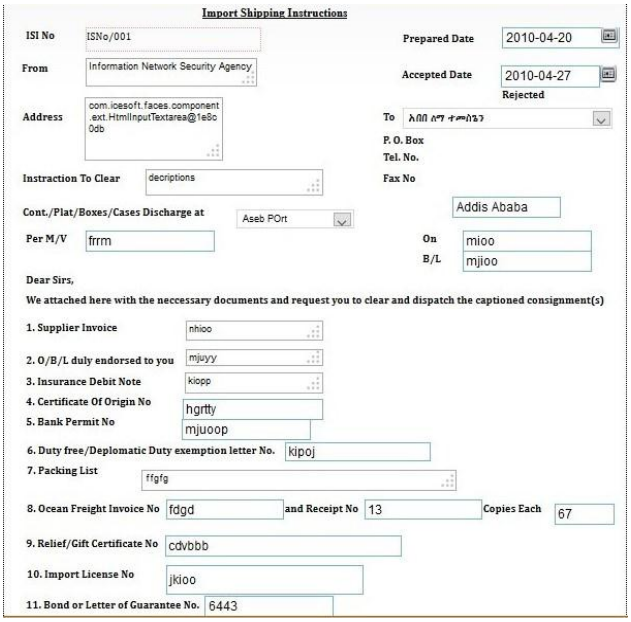

Figure 4: Import Shipping Registration Page

iv. Imported Goods Registration Page

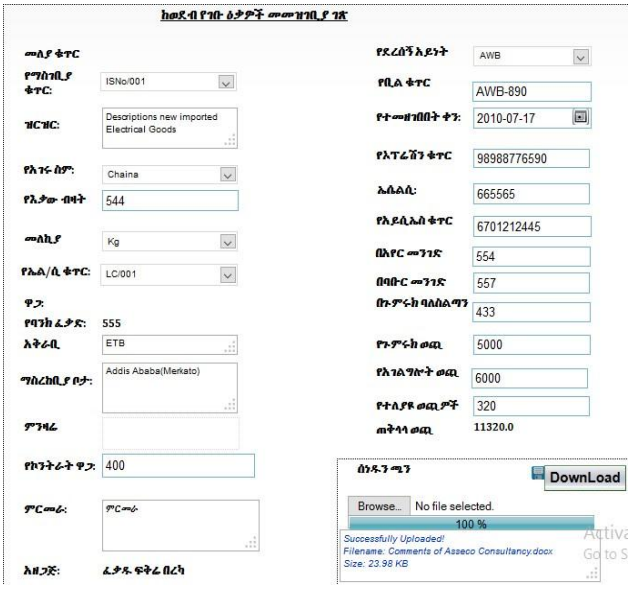

Figure 5: Imported Goods Registration Page

The page shown in Figure 5 is used to register the basic information related to the imported goods by the transit officer and approved and checked by the buyer's organization.

v. Insurance Registration Page

This page is used to register insurance information by the buyer institution officer and approved by the Insurance heads as shown in Figure 6.

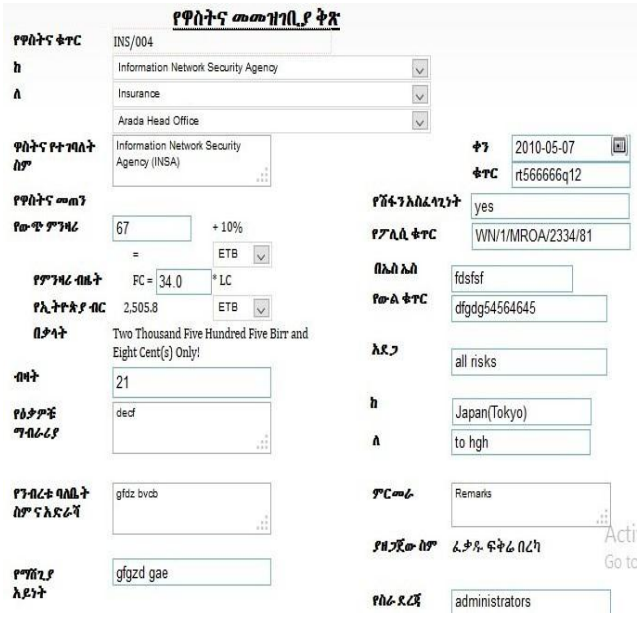

Figure 6: Insurance Registration Page

4.2 System Evaluation

The quality of the developed system is evaluated using known quality criteria in software engineering such as usability, impact of the system on users' decision, quality of information and usefulness and performance of the data sharing system.

For evaluation, a questionnaire was distributed to 8 employees working in different governmental institutions and associated with foreign procurement. Based on the evaluation, the system is usable with accuracy value of 92.2% which is promising.

5. Conclusion

In this paper the proposed centralized database has a substantial amount of advantages against other types of databases. Data integrity is maximized and data redundancy is minimized as the single storing place of all the data also implies that a given set of data only has one primary record. This aids in the maintaining of data as accurate and as consistent as possible and enhances data reliability. It is easier for use by the end user due to the simplicity of having a single database design.

Stakeholders can easily get and share timely information from such system and timely reports can be generated to be used and consumed for making decisions. Moreover, the system has been designed

and implemented with English and Amharic interfaces for maximizing the usability of the proposed system. As the end users evaluation analysis shows, the proposed system is successful in common data sharing and transforming the institution for better success.

References

[1] R. Raina, A. Srinivasan, and Sumit Panda, "Development of National Data Set Master Plan", New Delhi, India, 2015.

[2] Veerle Van den Eynden, Louise Corti, Matthew Woollard, Libby Bishop, and Laurence Horton, "Managing and Sharing Data", UK Data Archive, University of Essex, 2011.

[3] Rebecca S. Eisenberg and Arti K. Rai, "Data Sharing in Genomics – Re-shaping Scientific Practice", 2006.

[4] Susan Brown Trinidad, Stephanie M. Fullerton, Julie M. Bares, Gail P. Jarvik, Eric B. Larson, and Wylie Burke, "Genomic Research and Wide Data Sharing: Views of Prospective Participants, Genetics in Medicine", Vol. 12, No. 8, August 2010.

[5] Matthew Brack and Tito Castillo, "Centre on Global Health Security, Data Sharing for Public Health Key Lessons from Other Sectors", April 2015.

[6] Hilda Tellioglu, "Using Electronic Document Repositories (EDR) for Collaboration", Vienna University of Technology, Austria, 2015.

[7] Mathieu d'Aquin, Alessandro Adamou, and Stefan Dietze, "Common Data Repository", First Version, 2013.

[8] Yi Shen Virgini, "Research Data Sharing and Reuse Practices of Academic Faculty Researchers", Polytechnic Institute and State University, November 2015.

[9] Kersa Demographic Surveillance and Health Research Center (KDS-HRC), "Data Sharing and Access Policy", Haramaya University, Ethiopia, October 2014.

[10] S. M. Afroz, Elezabeth Rani, and Indira Priyadarshini, "Web Application – A Study on Comparing Software Testing Tools"; International Journal of Computer Science and Telecommunications, Vol. 2, Issue 3, June 2011.

[11] Hans-Petter Halvorsen, "Web Services with Examples", University College of Southeast Norway, 2016.

[12] Erin Cavanaugh, "Web Services: Benefits, Challenges, and a Unique, Visual Development Solution", Altova, Inc., 2006.

[13] Irving Fisher Committee; "Data-sharing: Issues and Good Practices", Bank for International Settlements, Basel, Switzerland, 2015.

[14] Shawn W. Nicholson and Terrence B. Bennett "Data Sharing: Academic Libraries and the Scholarly Enterprise".

[15] Habtamu Sewnet Gelagay, "Geospatial Data Sharing Barriers across Organizations and the Possible Solution for Ethiopian Spatial Data Infrastructure Program (SDIP)", International Journal of Spatial Data Infrastructures Research, 2017, Vol. 12, 62-84.

[16] Hans-Petter Halvorsen, "Web Services with Examples", University College of Southeast Norway, 2016, Norway.

[17] Erin Cavanaugh, "Web Services: Benefits, Challenges, and a Unique, Visual Development Solution", Altova, Inc., 2006.

A Web Based Blood Donation System

Henok Kifle
HiLCoE, Computer Science Programme, Ethiopia
henokkifle01@gmail.com

Abrehet Omer
HiLCoE, Ethiopia
abrehet@gmail.com

Abstract

Currently the National Blood Bank of Ethiopia (NBB) has adopted semi-automated system which made blood donation difficult for manipulation and management of the service. Moreover, most blood banks work in isolation and are not integrated with other blood donation centers and health organizations which affect blood donation and blood transfusion services quality. Hence the aim of this paper is to design a web based blood donation management system which would improve the efficiency of blood collection and utilization. In doing so, three different web based technologies, namely server-side HTML web application, JS generation widgets (AJAX), and service-oriented single-page web apps (Web 2.0, HTML5 apps) were evaluated based on three points of view: software owner, software developer and end user. Hence for this research service-oriented single-page web apps (Web 2.0, HTML5 apps) is used. The reason is that it provides a communication that could involve either simple data passing or it could involve two or more services coordinating for some activity.

Keywords: Blood Donation System; Blood Donation Management; Blood Donation Service

1. Introduction

Life is a precious element from all the belongings we have [1] and it has been medically proven that no human being can survive without blood; it supplies all nutrients and oxygen in the human body [2]. Unlike Ethiopia, blood banks found in different hospitals take care of the entire donation process, including recruiting donors, collecting and screening of the donated blood and preparation and storage of blood [3]. Globally, over 81 million people donate blood annually, but only 45% of these are donated in developing countries, where 81% of the world's population live [4]. According to National Blood Bank, the annual national blood demand in Ethiopia is about 60,000-80,000 units. However, the collection at a national level is only 43% (of which 8.84% is collected from voluntary non-remunerated blood donors, 16.53% from mobile sessions and 76.64% from family replacement donors). Two key problems are the gaps between the supply and demand of a safe blood supply, and the serious safety concerns associated with inadequately screened blood.

A major constraint to solving these problems is the lack of strong infrastructure and systems to support the management of blood [5]. In addition, most blood banks work in isolation and are not integrated with concrete information system [3]. There is problem in keeping track of the actual amount of each and every blood type in the blood bank, which blood group is going to finish, etc. There is also no alert mechanism when the blood quantity is below its par level or when the blood in the bank has expired. Further, there is no automated way of reminding donors when the next donation time is expected [6].

Hence the question that this research would address is what kind of system helps to manage and effectively facilitate the blood donation service?

According to [8], qualitative methodology is tailored in system development research process. Since it consists of five stages, which are construct a conceptual framework, develop system architecture, analyze and design the system, build the (prototype) system, and observe and evaluate the system. Relevant data had been collected through primary and secondary data. The primary data collection methods deployed were interview and questionnaires, which are used to get original data directly from the respective donors and officials of the organizations.

The interview was conducted with the director of the center, lab technician and IT personnel. As a secondary data, review of the procedures and processes of blood donation and document used by the national blood bank have been used. This methodology enables us to explore the existing system of blood donation management system.

2. Related Work

A blood donation system based on mobile cloud computing is proposed in [3]. It is comprised of two main components: Cloud Computing (CC) component and Mobile Computing (MC) component. The main aim of this work was to design a framework for blood donation system using cloud computing and mobile computing technologies. Stakeholders use the blood donation system as an application installed on their smart phones to help them complete the blood donation process. The application helps the people receive notifications on urgent blood donation calls, know their eligibility to donate blood, search for the nearest blood center, and reserve a convenient appointment using temporal and/or spatial information.

The blood donation system proposed in [7] is designed to process as follows: two types of users are allowed in this system, the donor type and the patient type. For donor account, as input, the donor needs to enter the information needed for the patient to inquire necessary blood. Then the matcher decides to accept the donation or not using rule based knowledge. There are three main roles and three main processes. The three main roles are donor, patient and matcher. The three main processes are record the memberships of donors and patients, acquire the donor's purification blood and matching the patient with related donors. It needs main database for requirements like specific rules for donors. Main rules are divided into two classes in which patients can search on this page for their needs when they need blood.

Generally, the above related works provided an understanding of the different types of blood donation management systems. This also provided input regarding the provision of knowledge base in the integration of different stakeholders, web portal management, account management of stakeholders via the application and the extraction of knowledge using different data mining concepts and methodologies.

A location-based framework for mobile blood donation and consumption assessment using big data analytics [9] came up with the idea that blood banks usually suffer frequent shortage of certain blood groups at some locations. Hence, social networks have recently served as a quick channel for advertisements, looking for healthy individuals to donate blood for patients who urgently require blood transfusion at emergencies.

Other blood donation centers may have excess of the same required blood group, which it would eventually be wasted for expiry reasons. This is due to the fact that many blood banks work in isolation, having no integration with other health organizations that would dramatically affect the quality of BDT services.

The paper proposed a framework that uses large-scale time series regression analysis techniques to analyse blood demands and donations matching data. The proposed system also could use different data mining concepts to extract knowledge from the system as future work.

A knowledge based system for blood transfusion [10] was designed for the national blood bank of Ethiopia. The aim of the system was to acquire knowledge necessary in blood transfusion and designing knowledge based system that can provide advice to experts involved in blood transfusion. The system cross matches the compatibility of the patient's blood and the blood going to transfuse. The researcher used a rule based knowledge representation method to represent the relationship between facts and rules. The result shows that the system registers 83.3% complete knowledge of blood transfusion task. This system can be integrated as one

module in the whole blood bank information management system

This research come up with rule based knowledge representation to cross match patients' blood with donor's blood. Hence, it could be used as an input in extracting knowledge using different data mining concepts. After the implementation of the results, the research idea could be adopted to extract different kinds of knowledge.

Generally, the above related works provided an understanding of the different types of blood donation management systems. This also provided input regarding the provision of knowledge base in the integration of different stakeholders, web portal management, account management of stakeholders via the application and the extraction of knowledge using different data mining concepts and methodologies.

3. Requirement Analysis

Requirement analysis has been completed to assess the existing system in NBB and relevant data has been collected through primary and secondary data.

The primary data collection methods deployed were interview and questionnaires, which used to get original data directly from the respective donors and officials of the organization. The interview question was conducted with the director of the center, lab technician and IT personnel. It raises different issues that range from donor recruitment to blood distribution for hospitals and health centers. There were two types of questionnaires: for employees of the organization and for the blood donors. Half of quarterly regular donors fill the questionnaires provided for donors. Hence, out of 60 questionnaires 50 were correctly filled, and the result of this study was concluded based on those successfully returned.

As a secondary data, review of the procedures and processes of blood donation and document used by the national blood bank have been used. Since the data that have been collected through these methods are recent, they have been used in accordance with the primary data. This methodology would enable us to explore the existing system of blood donation management system.

According to the result of the interview with the professionals, the organization collects blood from different individuals and groups. This collection by the organization is planned based on the need by the health centers and other health organizations that need blood but this collection is not planned for each blood type.

It is possible to say that the current collection process that the bank follows is successful; it is possible to achieve more than 80% yearly but the bank has never achieved 100% collection.

In order to acquire blood from the national blood bank, the health centers sign memorandum of understanding (MoU). Once the bank and the centers agree on terms, it will be possible to request blood from the center with a stamped letter and assigned professional.

It is possible to know the amount remain in the stock of the blood bank but this record is found in hard copy kept in a file and health centers and organizations that need blood could call and check the amount of blood remaining in the bank which is error prone and tedious task. The national blood bank reports to the Ministry of Health (MoH). The reports are done monthly, quarterly, half-yearly, and yearly.

NBB currently has a relatively small network, which connects the data center with different departments with a maximum of 15 users.

For the national blood bank, it is possible to trace when a donor donates blood in the bag it will be given pack number, and after health centers receive the amount of blood they must return the feedback to the bank on how they used the blood taken from the bank.

Currently, the center uses a database application which is developed using SQL database management system. The center uses the database for the recording of donors' profile, blood distribution to health facilities, screening information and discarded

bloods information. In addition, the center generates reports using the information recorded in the database. In addition, the database does not have a front end, which could have made it easy to learn and operate. Thus the existing system is not user friendly.

In order to facilitate the blood donation and processing service, NBB uses various kinds of documents and forms. Those documents and forms currently used in the bank include memorandum of understanding (MoU), donor registration, blood distribution, blood discard and reporting formats. The documents used at the center helped us to know exactly what this documents are for about, what specific information included and what is left out (for example, the problem of not having a separate donor registration form and blood registry form, in addition the reporting format lacks blood specific (O, A, B, AB) output).

4. The Proposed Solution

The final output of this work is a web based blood donation system that is tailored to NBB and would serve as a medium to exchange information easily for the donors, health centers, NBB and the MoH. A prototype was also developed as shown in Figure 2.

Figure 1: The Developed Web Application

In addition, the system brings together different types of stakeholders including blood donors, blood banks, health organizations, and MoH staff through the web application. In the proposed architecture, as shown in Figure 2, there are five components as presented below.

1. *Registration Module:* has the following three services.
   - Donor Registration Service: handles donor registration; it enables them to reserve for donation and view their donation and reservation profile.
   - NBB Employee Registration Service: provides registration for the employees of NBB where the authorized organ to register the employees is MoH. The registration enables the employees to manage donor's reservation, hospitals and health centers request, view stock and manage the health centers account.
   - Health Centers Registration Service: provides health centers registration that enables them to request blood, view their requests and view request history.
2. *Request Submission Module*: has the following three services.
   - Reservation Request Service: enables donors to send reservation request, which will be managed by NBB.
   - Campaign Reservation Request Service: enables to broadcast campaign for donors, thus donors can register themselves as campaign donors.
   - Blood Request Service: enables health centers to send blood requests, which will be managed by NBB according to the blood stock.
3. *Broadcasting Module*: has the following two services.
   - Broadcast Low Par Value Blood Type: broadcasts the blood in stock with a lower par value.
   - Campaign Donation Management Service: enables the center to broadcast different campaigns which donors could directly sign up via the web application and donate blood according to the registered campaign.
4. *Request Response Management Module*: has the following two services.
   - Donation Management Service (Reserved, Campaign & Unreserved Donation): enables the management of donor's reservation. Thus donors reserved can easily donate blood. It is also responsible for the management of

campaign donation requests. In addition, donors who want to donate blood without reservation or walking donors can also be managed via this module.

- Stock Management Service: enables NBB employees to manage stock.

5 *Reporting Module:* has the following service.

- Reporting Service: enables to generate reports, including donor's reservation report, donor's donation, health centers request report, etc.

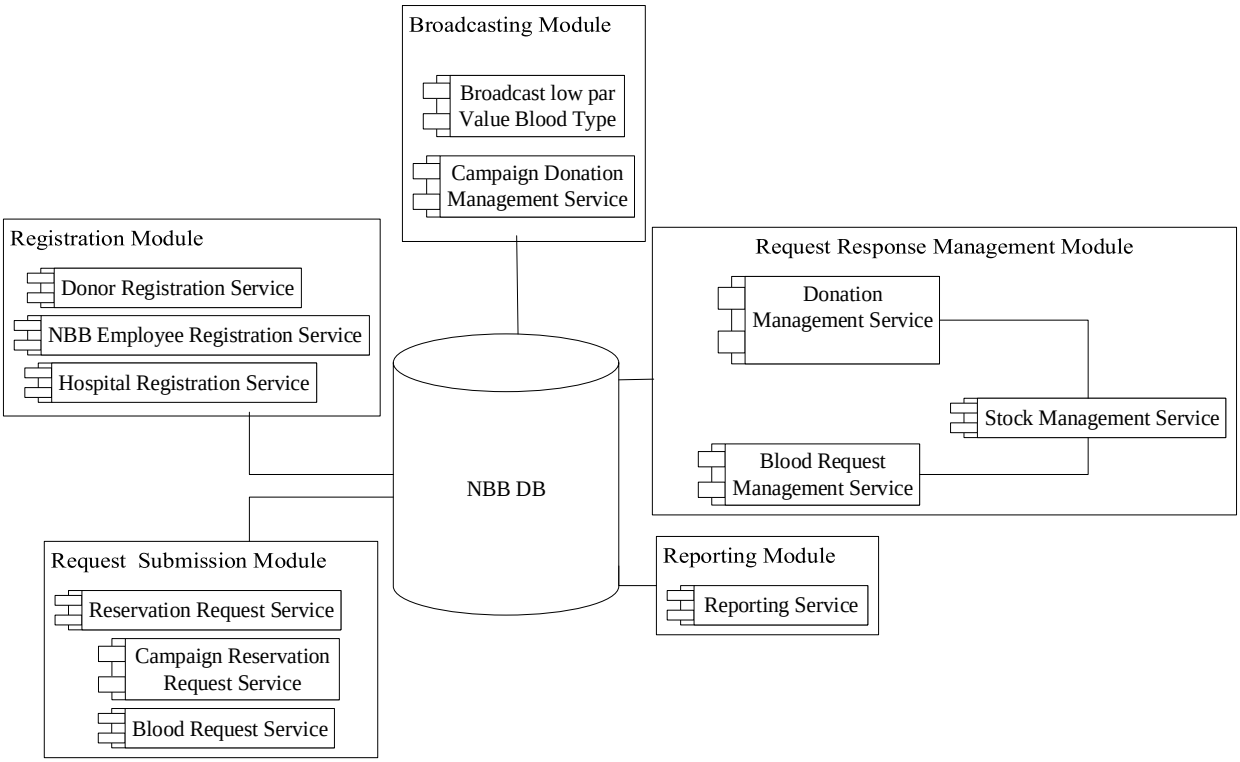

Figure 2: Blood Donation System Architecture

The evaluation of usability is an important aspect of software design and development and a questionnaire was distributed for a total of 19 respondents. According to the result of user interface evaluation most of the respondents, i.e., 84.89%, strongly agreed that the system prototype has an interface that is easy to use, attractive and the security is maintained properly.

5. Discussion

This research is carried out with the aim of proposing a web based blood donation management system and addressing the problem of direct control and management of the overall activities performed in the blood bank centers of MoH. Thus, web based blood donation mechanism was implemented for the stakeholders.

The prototype developed in this paper demonstrated how a service implementation facilitates the blood donation mechanism. Donors could register themselves for reserving a spot for donation via the web based application, which would also entertain viewing their donation profile. This process would facilitate the blood collection through minimized time and effort. Health centers could be registered by the NBB after agreeing with the terms and contract of the NBB. There after they must sign the MoU based on the terms of the national blood bank regarding the safety of the blood. It is also possible for them to view their blood request profile as well. NBB would manage donors and health centers to facilitate blood donation and processing mechanism efficiently and effectively. Since NBB reports to MoH, the system would be managed by MoH.

NBB currently has 24 branches where blood donation and distribution services are given. But, the

result in this work must be somehow modified with further analysis and minor modification of the existing systems on each branch to provide a comprehensive solution for sharing data sources. In developing the system prototype, requirements were acquired using in-depth interview, questionnaire and documents used at the center. The collected requirements were analyzed and designed using UML.

The research has provided an input regarding the provision knowledge base in the integration of different stakeholders. In comparison with other researches, this research has come up with three main web application architecture types. These are server-side HTML web application, JS generation widgets (AJAX), and service-oriented single-page web apps (Web 2.0, HTML5 apps) which each of them had been reviewed and service oriented single web apps (Web 2.0, HTML5 apps) was adopted. This was because it provides a communication that can involve either simple data passing or it could involve two or more services coordinating for some activity. In addition, Responsiveness/Usability, Performance and Conversion rate of the web application is high which will be demanded by the blood donation management service.

6. Conclusion and Future Work

This research is carried out with the aim of proposing a web based blood donation management system and addressing the problem of direct control and management of the overall activities performed in the blood bank centers for MoH. Thus, web based blood donation mechanism was implemented for the stakeholders. The prototype developed in this paper demonstrated how a service implementation facilitates the blood donation mechanism.

Donors could register themselves for reserving a spot for donation via the web based application, which would also entertain viewing their donation profile.

Figure 3: Donors Reservation

This process will facilitate the blood collection through minimized time and effort. Health centers could be registered by the NBB after agreeing in the terms and contract of the NBB. There after they must sign the MoU based on the terms of the national blood bank regarding the safety of the blood. In addition, it is also possible for them to view their blood request profile as well.

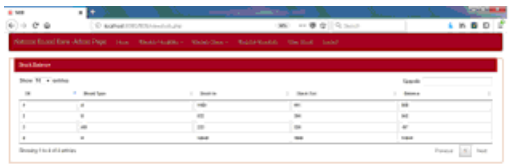

Figure 4: Manage Blood Stock

In order to implement our result, NBB and MoH should support the project in human resource and finance. In addition, the organizations should upgrade the hardware and network infrastructures of the main center and MoH. The regional health bureaus should support their respective blood bank centers with necessary human resource, hardware and network infrastructure for the implementation of this project.

It is recommended that interfacing issues like network infrastructure, security and particularly the data type that might be required by other stakeholders must be taken into consideration up on project initiation and planning.

Finally, the expansion of this system architecture is open for further work in accommodating other requirements like Android application based blood donation and consideration of other security aspects.

References

[1] Srcn Senanayake and Adai Gunesekara, “Designing an Information System Model for National Blood Bank of Sri Lanka”, In

Proceedings of the 8th International Research Conference, KDU, November 2015.

[2] Juma, Nazana, "Blood Bank Management Information System", University of Zambia, 2012.

[3] Almetwally M. Mostafa, "A Framework for a Smart Social Blood Donation System Based on Mobile Cloud Computing", Health Informatics-An International Journal (HIIJ) Vol. 3, No. 4, November 2014.

[4] World Health Organization and International Federation of Red Cross and Red Crescent Societies, "A Global Framework for Action. Towards 100% Voluntary Blood Donation", World Health Organization, 2010.

[5] World Health Organization, "Management of National Blood Programmes", In Proceeding of Three WHO Workshops (2007-2009), 2010.

[6] Abraham Zeleke, "Blood Transfusion Service in Ethiopia–from establishment up to date", Federal Ministry of Health special bulletin, 16th annual review meeting, 2014.

[7] SedaBa S., Giuliana Carello, Ettore Lanzarone, Zeynep Ocak, and Semih Yalçında, "Management of Blood Donation System", Literature Review and Research Perspectives Conference, January 2016.

[8] Orit Hazzan, Yael Dubinsky, Larisa Eidelman, Victoria Sakhnini, and Mariana Teif, "Qualitative Research in Computer Science Education", Department of Education in Technology and Science, March 2006.

[9] Ain Shams, "A Location-based Framework for Mobile Blood Donation and Consumption Assessment using Big Data Analytics", Asian Journal of Information Technology, January 2016.

[10] Guesh Dagnew, "Designing a Knowledge Based System for Blood Transfusion", Addis Ababa University, 2012.

Extractive Based Auto-Summarizer for Amharic News using K-Means Clustering

Jigsa Tesfaye
HiLCoE, Computer Science Programme, Ethiopia
jiks.tes@gmail.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

Today the amount of Amharic digital resource is growing rapidly. A number of news sources like fanabc.com, addisadmas.com and ethopianreporter.com are rolling out news daily. This makes it harder for us to efficiently read and extract relevant information from all these sources. On the bright side, the availability of these machine-readable Amharic news content creates the opportunity to easily accumulate statistical learning that could reveal important language patterns to develop unsupervised machine learning algorithms for the language. This paper mainly focuses on the design of an extractive automatic summarizer which uses statistical knowledge to intelligently determine the optimum extraction level and content using K-means unsupervised algorithm and it is tested specifically on Amharic news content. This approach was previously used in a similar way for an English language summarizer, but in this paper an improved selection process is provided which combines the restrictive feature of the conventionally known best-n method with the K-mean. Best-n is a method used to have control over the level of percentage of content extracted to produce the final summary. An evaluation was conducted to compare the two ways of sentence selection; the improved K-means method with the conventional best-*n*%. The K-means method showed a decrease in extraction percentage with an improved summary score of 58.1% which is a 5% increase from the conventional best *n*% selection method.

Keywords: Automatic Summarization; K-means; TF-IDF; Web Crawler; News Content

1. Introduction

Text summarization is the process of producing a shortened version of a given text retaining the most relevant information for the reader. Summarization can lessen the burden of information processing by condensing the original text into one which is more or less representative.

In this paper, an improved approach is introduced that uses K-means algorithm for selection of content in a language independent extractive summarizer. In previous researches, including such kind of language independent implementation researches done on Amharic, the prototype summarizers were designed to produce a summary with a strict level of extraction in percentage (usually 20% and 30%) or best *n*%; with *n* decided by the user. The best *n* approach ranks sentences based on their scores and selects the top *n* sentences. The problem with this approach is that when ranking the sentences, the score gap could be very critical for the selection process. Let us say we have twenty sentences in a document and the 6th and 7th ranked sentences have huge score gap and if the user assigns *n* as 35% then the inclusion of the last sentence has big negative influence on the accuracy of the summary.

K-means is a clustering algorithm that is used to model unsupervised data. In K-means, data objects are presented as points on a Cartesian plane. Starting with a random set of K reference points on n-dimensional space, the data points will be assigned to K clusters based on distance criterion. Centroids are special points that represent a cluster. These centroid points are selected with a calculated process that attempts to choose the points placed as far away from

each other as possible [1], setting as much distance among clusters as possible. The time spent when finding these ideal centroids is divided into iterations. The steps for the K-means algorithm are as follows:

1. On a single or multi-dimensional plane, the objects are positioned as points based on their features as a metric.
2. K points are selected from the objects as centers of clusters, also known as centroids.
3. Every data sample is assigned to a cluster with the closest centroid.
4. The mean of the data points is calculated for each cluster possibly choosing a new more representative centroid in the process.

The third and fourth steps are iterated until no changes happen in the centers of clusters. K is the number of clusters and should be defined ahead of time.

The content scoring algorithm used in this paper is Term frequency, inverse document frequency (TF-IDF). TF-IDF is a statistical method used to compute the significance of a word in a document by looking at its occurrence frequency in the document compared to a set of other similar documents [2]. Given the term ‘t’ in a document 'd', the TF value of ‘t’ is calculated as the frequency of ‘t’ in ‘d’ over the total number of all the terms found in ‘d’ as:

$$\mathbf{tf}(t,d) = f_{\downarrow}(t,d) / \Sigma_{\top}(t^{\intercal\prime} \in d)\, f_{\downarrow}(t^{\intercal\prime}, d)$$

The IDF of a word is assigned the inverse value of the number of documents it is found normalized by the total number of documents as:

$$\mathbf{idf}(t,D) = \frac{\log N}{|\{d \in D : t \in d\}|}$$

TF-IDF is a numeric weight given to a word which is a product of two formulas, TF and IDF.

$$\mathbf{tfidf}(t,d,D) = \mathbf{tf}(t,d) \cdot \mathbf{idf}(t,D)$$

The representation of sentences in the k-means clusters can be set with a numerical score value of an attribute or combined attributes [2]. Indicators such as term frequency and inverse document frequency can represent a relevance factor of a sentence based on frequency.

2. Related Work

Even though there weren’t as such similar summarization researches that specifically use TF-IDF and clustering on Amharic language, there are quite many papers that share most parts of the methodologies and algorithmic techniques with this paper [3, 4, 5]. The use of news articles and newspapers for developing and testing a summarizer is very common in the research community [6, 7]. The ease of finding large collection of news digital documents is a compelling reason that made researchers so interested to study the news domain.

The work in [1] is the closest in terms of the algorithms used and overall methodology. They used TF-IDF algorithm as a feature and K-means clustering for summarization. Sentences are scored using words’ TF-IDF sum to measure “goodness”. The difference in their approach comes down only to two decisions; the way k is chosen dynamically and the method used in cluster selection. Each sentence is represented as coordinates in a single dimensional Cartesian plane which will then be clustered using K-means. When selecting the sentences, they argue that the cluster with the maximum number of sentences should be selected. K, which represents the number of clusters to be created, is dynamically determined with the possibility of having k value more than two (as shown below) that was designed after running tests on multiple documents that calculate dynamic k.

If N>20 then k= N-4

Else k= N-20

where N is the total number of sentences in the document.

The authors argue that this algorithm will balance the amount of extraction with the size of the document. As implemented in this paper, their approach cluster sentences and in the end one cluster is chosen as a summary. The sentences are presented in the summary in their order found in the original document. But the work also had a limitation in a

way it produced an average of 43% extraction with the rate going up to 50% which only begs the question if 50% extraction really can be considered a summary.

This K-means and TF-IDF implementation is language independent. Past research implementation of Amharic summarizers incline towards a language independent, single document, unsupervised and extractive based approaches. There hasn't been as such many works done on the language with handful of researches only limited to the use of basic content indicators, topic modeling algorithm and Latent semantic analysis techniques. Almost all works used the best-n approach with different fixed extraction rates – 20%, 30% in [8], 30%, 25%, 20% in [9] and 100 words in [10].

3. The Proposed Solution

3.1. Prototype Design

The prototype is a program that has an intuitive user interface that takes input of an Amharic news article content in plain text format and a preferred level of content extraction in percentage to produce a summary output. The final summary consists of actual extracts of sentences that are deemed to have special relation and importance with the central topic of the writing. The summarizer is an unsupervised system that considers keywords or content words as summary indicators and uses frequency driven technique for scoring.

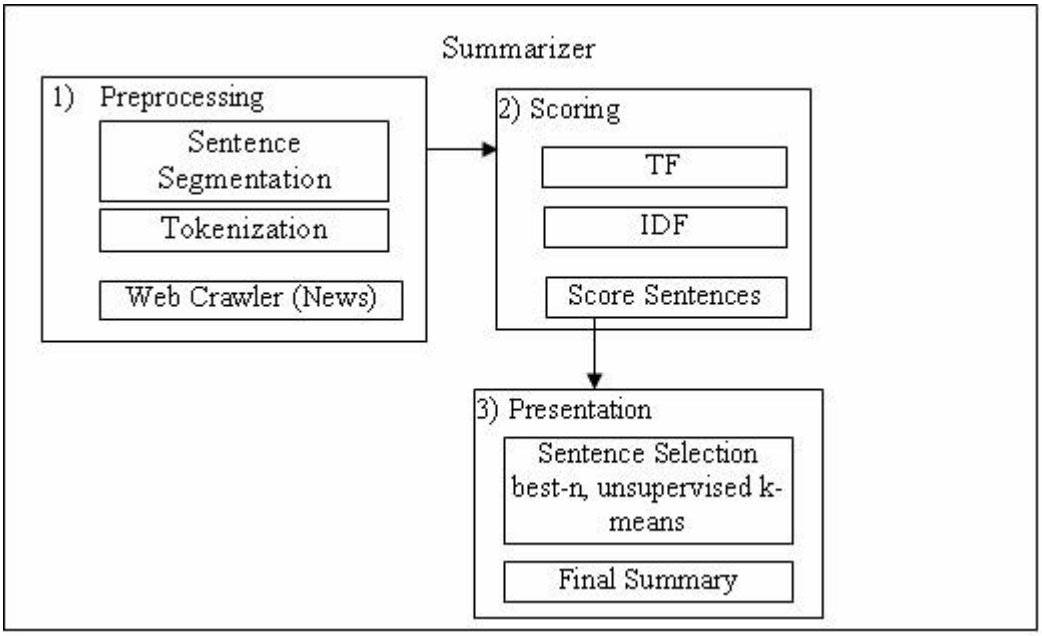

Figure 1: The Architecture of the Prototype System

The summary extraction process programmed in the prototype has distinct stages from preprocessing, intermediate representation, scoring, up to selection and finally summary presentation. The most important stage is scoring because it guides all the other stages. The scoring is done using frequency driven topic terms identification algorithm TF-IDF. The final stage of summary selection can be programmed to be determined using two options, the K-means or pure best *n* methods. Clustering is an unsupervised technique used in this particular case to select the most outstanding sentences from the input document. K-means is the clustering algorithm used to achieve this based on the TF-IDF score representation of the sentences. As mentioned before, this paper aims to evaluate an alternative K-means implementation approach for determining the summary extraction percentage threshold level automatically by identifying the outstanding cluster of sentences using their centroid distance measure. In a previous work in [1] the level of percentage of extraction was not limited when using a pure K-

means selection. In our case the K-means is processed in increased iteration depending on the extraction level set by the user which makes it a hybrid approach with the Best-n. The prototype should also support the conventional sentence selection method in the form of the best *n*% where the user specifies a rigid percentage level in which case the top ranked sentences are automatically added in the summary. These two selection methods are discussed more in Section 3.3. The architecture of the prototype system is shown in Figure 1.

3.2. Data Collection

The prototype is tested on Amharic language specifically the news domain. Amharic is still under resourced that hindered the development of language processing applications. Consequently, it is necessary to explore ways to efficiently populate resources. For languages like Amharic where most of the resources available are in unstructured text, in the form of news articles and messages, the design of unsupervised systems could be the way forward at the moment [11].

Collection of large amount of online news data can be very useful for identifying keywords that can in turn be used to produce a short summary of news articles [12]. The main content scoring algorithm used in this research is TF-IDF. In order to calculate the IDF of words there needs to be a considerably large amount of corpus. A document frequency is the number of documents a term is found in a corpus. The collection of documents from which the word document frequency knowledge is gathered should be similar to the document on which the system algorithm applies, specially for domain specific tasks like this [13]. Therefore, the corpus data is gathered from a large collection of news articles. There are commonly found terms in news articles such as "ዘገባ", "ቃለመጠይቅ", "ገለጹ". The use of this news corpus allows us to identify such words that are often used in the news domain. The scores of these terms should be lowered depending on their IDF value. IDF was designed for the purpose of achieving a frequency driven weighting score for terms. As its name suggests IDF holds the inverse value of the number of documents a term was found to appear in.

An automated web crawler program developed specifically for this research, adopted from utopiaio'saScraper implementation (found at github.com/utopiaio/eKeyboard), was used to help build the news corpus. This crawler automatically goes through hundreds of news web pages on the Internet to grab contents of news articles. So the document frequency information of each appearing word is then collected and shared to the summarizer application.

3.3. Sentence Selection Stage

The implementation of the K-means from this paper's perspective begins with a bag of words approach [14] where each word from each sentence is gathered and viewed collectively. The sentences are then represented by their scores in a single dimensional Cartesian plane where score is calculated as:

$$\text{Score}(S) = \sum_t \text{TF-IDF}_t / |S|$$

where S is sentence, |S| is sentence length and t are terms in the sentence.

These coordinates are used as input for unsupervised K-means clustering algorithm. As pointed out in [15], one of the proposed methods is to take points whose attribute values are average of the n-objects from the dataset. Therefore, the initial points are the 'k' sentences that are found in the middle of the score ranking. Finally, the cluster with the highest centroid value is chosen as the appropriate summary. The rationale behind this centroid-based summarization is that the centroid has a mean value representing the cluster. Therefore, a centroid with higher value signifies the sentences in the cluster have bigger relevance.

Algorithm 1: Best-N

```
Input: Percentage of extraction
  preferred by the user, list of
  sentences and their scores
Begin
  Rank the sentences based on their
    score
```

```
    Take the best n% sentences and
      present them to the user in the
      order they were presented in the
      original document
End
```

Algorithm 2: K-Means

```
Input: Maximum percentage of
  extraction preferred by the user,
  list of sentences and their scores
Begin
  Set k initial value to 2
  Rank the sentences
  While (percentage of selected
    sentences > restriction set by
    the user & k <= 10) {
    Choose the most average k
      sentences as initial points
      based on their rankings
    Cluster the sentences using k-
      means
    Choose the cluster with the
      highest centroid value as a
      possible summary
    k++
  }
  Present the sentences to the user
    in the order they were presented
    in the original document
End
```

There are two types of extractions, K-means and pure Best-n%. If the selection type is Best-n% (see Algorithm 1), the user selects percentage level option from 5 to 95% and the top n% of sentences directly qualify to the summary. If it is set as K-means selection (see Algorithm 2), then the sentences first get clustered based on their scores and the outstanding cluster of sentences with high centroid get selected. The user can set a restriction as the maximum level of extraction for the K-means. The K value depends on the restriction set by the user, for example, the user could set the restriction at 30% or less.

4. Discussion

For the evaluation, this paper presented results from experiments done to test the accuracy of the prototype from two perspectives. The first one tests the precision and recall measures of the conventional Best-n implementation. The second one checks summaries produced by K-means where clustering technique is used to determine varying level of extraction with a restriction of 30%. The results show a better F-measure score for the K-means by 5%. This is because the K-means reduces the extraction level by taking out irrelevant sentences from the summary increasing the recall score. The average extraction rate decreased by 6.75% when using K-means which is good because it is a shorter summary with better accuracy. But the precision is lowered by about 5%. This was because the best-n summaries taken were allowed higher percentage extraction. The system overall retained a 58.1% F-measure score from an evaluation done on four articles by comparing summaries received from four people.

When compared to other researches on Amharic language, the document set used for testing are long with an average length of 40 sentences and 1075 words and also had less extraction percentage average of 18%. The alternative K-means approach presented in this paper tries to balance the extraction level better when compared with the approach taken in [1] where they tested pure K-means for extraction and had summary results more than 40% in most cases, average of 43.2% and a maximum of 50%. The test cases in this evaluation show average of 18%, 25% lower. To support the content scoring, the crawler was successfully used to mine frequency data of 145,650 words from three well-known news sources.

5. Conclusion

This main focus was on the design of selection process for auto summarizers. It was attempted to provide an optional and improved implementation of the K-means as a selection method judging from a previous research in [1]. It showed how clustering algorithms like K-means can be integrated with TF-IDF to produce a more unsupervised summary. The experiments in this paper showed an approach that integrates K-means to produce a restrictive summary less than or equal to 30%. K-means is applied iteratively to determine k (number of clusters) and

take the highest centroid valued cluster as a summary. We conclude that this method improves the accuracy of the best-n by also solving the high percentage summary problem observed in [1].

The experiments focused on the comparison of the two summary selection approaches in the form of the newly suggested K-means unsupervised selection and the conventionally known best-n method. The results of the experiments show 5% increase in accuracy with the K-means cluster based selection method. The new approach described in this paper showed a more appropriate level of extraction unlike that in [1] with the average extraction level of 18%.

To counter the scarcity of resource in Amharic language this research developed a dedicated web crawler application and presented an approach on how to integrate this application as a subordinate service to compliment the summarizer and ensure a scalable and adaptable approach. This in turn showcased a method on how a useful corpus can be prepared automatically from Amharic news websites.

References

[1] Agrawal, A. and Gupta, U., "Extraction Based Approach for Text Summarization Using K-Means Clustering", International Journal of Scientific and Research Publications, 4 (11), 2014.

[2] Munot, N. and Govilkar, S. S., "Comparative Study of Text summarization Methods", International Journal of Computer Applications, 102 (12), 2014.

[3] Seki, Y., "Sentence Extraction by TF-IDF and Position Weighting from Newspaper Articles", The Third NTCIR Workshop, National Institute of Informatics, 2002.

[4] Christian, H., Agus, M. P., and Suhartono, D., "Single Document Automatic Text Summarization Using Term Frequency-Inverse Document Frequency (TF-IDF)", 2016.

[5] Radev, D. R., Fan, W., and Zhang, Z., "WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System", In NAACL Workshop on Automatic Summarization, Pittsburg, 2001.

[6] Jassem, K. and Pawluczuk, Ł., "Automatic Summarization of Polish News Articles by Sentence Selection", In proceedings of the Federated Conference on Computer Science and Information Systems, 2015, pp. 337-341.

[7] Zadbuke, A., Pimenta, S., Padwal, D. and Wangikar, V., "Automatic Summarization of News Articles using TextRank", Special Issue on 3rd International Conference on Electronics & Computing Technologies, 2016, pp. 124 - 127.

[8] Melese Tamiru, "Automatic Amharic Text Summarization Using Latent Semantic Analysis", Unpublished Masters Thesis, Addis Ababa University, 2009.

[9] Eyob Delele, "Topic-based Amharic Text Summarization", Unpublished Masters Thesis, Addis Ababa University, 2011.

[10] Mattias, Gesesse Argaw, "Efficient Language Independent Text Summarization Using Graph Based Approach", Unpublished Masters Thesis, Addis Ababa University, 2015.

[11] Bekele, Worku Agajyelew, "Information Extraction from Amharic language Text: Knowledge-poor Approach", Unpublished Masters Thesis, Addis Ababa University, 2015.

[12] Gupta, V. and Lehal, G. S., "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, 1 (1), pp. 60-76, 2009.

[13] Shams, R. and Elsayed, A., "A Corpus-based Evaluation of Lexical Components of a Domain-Specific Text to Knowledge Mapping Prototype", IEEE 11th International Conference on Computer and Information Technology, 2008.

[14] Gupta, V. and Lehal, G. S., "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, 2 (3), 2010, pp. 258-267.

[15] Baswade, A. M. and Nalwade, P. S., "Selection of Initial Centroids for K-Means Algorithm", International Journal of Computer Science and Mobile Computing, 2 (7), 2013, pp. 161-164.

Part of Speech Tagger for Hadiyyisa Language

Kacha Desta
HiLCoE, Computer Science Programme, Ethiopia
okoted@yahoo.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

Part of Speech (POS) Tagging is the process of assigning the correct part-of-speech to each word in a text that best suits the definition of the word as well as the context of the sentence in which it is used. A part-of-speech tagger is used in many language technology applications as a first step in more complex NLP applications such as grammar checking, machine translation, information extraction and information retrieval. There are different approaches to the problem of POS Tagging. In this paper, we have used two approaches: Rule based and TnT approach. We have compared the performance of these techniques for tagging using combination with backoff taggers for Hadiyyisa Language. These approaches use supervised POS Tagging. It requires a large amount of annotated training corpus to build the tagging model and perform acceptable tagging of a new text. At the initial stage of POS Tagging of Hadiyyisa Language, we have very limited annotated corpus, i.e., 1280 manually tagged corpus of Hadyyisa sentences tagged with 32 identified tag sets and 10 basic tag sets. The 10 basic tag sets are derived from the 32 identified tag sets. The corpus was divided into 80% for training and 20% for testing. The research attempted to investigate a technique that maximizes the performance with this limited language resource. By experiment, the best configuration was investigated using TnT with Affix unknown word taggers, Rule-based with Affix initial state tagger, Rule based tagger when the initial state tagger with backoff in the sequence of TrigramTagger, BigramTagger, UnigramTagger, AffixTagger, and TnT-unknown word handling with backoff in the sequence of TrigramTagger, BigramTagger, UnigramTagger, AffixTagger. These approaches have accuracy of 66.64%, 64.65%, 72.54%, and 73.06% with 32 derived tag sets tagged corpus and 80.34%, 72.67%, 87.25%, and 89.03% with 10 basic tag sets tagged corpus respectively. Therefore, better performance is gained by the basic tag set tagged corpus with the tagger TnT-unknown word handling with backoff in the sequence of TrigramTagger, BigramTagger, UnigramTagger, AffixTagger taggers than the other approaches.

Keywords: Part-Of-Speech tagging; TnT; Rule based Tagger; Natural Language Processing

1. Introduction

Natural Language Processing (NLP) is a field of computer science, artificial intelligence and computational linguistics, which is concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human-computer interaction [1]. NLP is important in machine translation, fighting spam, information extraction, summarization, question answering, speech-recognition, etc.

There are different types of languages in the world where the basics in every language are words. Words are the building blocks of a language and word classes (Parts of Speech) in languages are the building blocks of sentences.

A word class (in other terms Part of Speech) is a set of words that display the same formal properties, especially their inflections, i.e., the differing forms one word takes to signal differing grammatical roles, e.g., boy-singular vs. boys-plural and distribution. Commonly recognized categories of parts of speech include: adverb, adjective, article, conjunction, noun, preposition, pronoun, and verb. There are other additional categories along the way, for example,

articles are generally taken to be a subset of a larger category called determiners [2].

Knowing all the parts of speech can help in advancing and understanding of the language because humans recognize patterns. Grammar is essentially a system of patterns applied to words that organizes them in a particular way. Without a grammatical system, we wouldn't be able to communicate with one another through reading, speaking, and writing [3].

POS tagging in particular is the act of assigning each word in a sentence to a tag that describes how that word is used in the sentence. That means POS tagging decides whether a given word used in a sentence is a noun, adjective, verb, etc. As Pla and Molina [4] noted, one of the most well-known disambiguation problems is POS tagging. A POS tagger attempts to assign the corresponding POS tag to each word in a sentence, taking into account the context in which this word appears [1].

In Ethiopia, part-of-speech tagger is developed for different languages and it is a demanding research area for other languages. Up to now, to the best of our knowledge, there are no attempts to investigate the suitable approaches of part of speech tagging for Hadiyyisa language.

Hadiyyisa is a language of the Hadiya people in Ethiopia. Most Hadiyyisa speakers live in the Southern Nations, Nationalities, and People's Region in the Hadiya Zone, the nearby zones and Special Woredas. Hadiyyisa uses Latin-based alphabets and it is an official language of Hadiya Zone and also it has been used as medium of instruction in primary schools, as a subject at junior secondary schools found in the zone and as field of study at Hossana Teachers Training College and Wachamo University and medium of transmission of local FM Radio programs.

Hadiyyisa has 27 letters including letters borrowed from other languages. Out of these, 23 are consonants phonemes and ten of them are both long and short consonants. Like other related languages, the basic word order in Hadiyyisa sentence is Subject Object Verb (SOV).

2. Related Work

In this section, different works on part of speech tagging in different languages are reviewed and they are selected based on their use of POS approaches and the language used.

Afaan Oromo part of speech tagger using HMM approach was developed by Getachew Mamo [7]. In this work, the researcher used HMM approach which is one of the most common mechanisms under stochastic approach. For training and testing purpose, 159 sentences were collected with a total of 1621 words to make the corpus balanced and used 17 tag sets.

In the tagging process, the tagger assigns word classes to a given Afaan Oromo text with two main phases. In the first phase, the tagger trains on the training data in order to compute and store both lexical and transitional probability of training data. In the second phase, the tagger accepts untagged Afaan Oromo text and is tokenized into words. Then the tagger assigns the correct POS tag for each token. This is achieved using unigram and bigram model of the Viterbi algorithm by taking the stored information during the first phase. The author tested the performance of the tagger using tenfold cross validation mechanism. As a result, 87.58% and 91.97% accuracy for unigram and bigram model respectively were achieved.

Another work on Part of speech tagging for Tigrigna using hybrid model was done by Teklay Gebregzabiher [6]. In this work a hybrid approach, which is the combination of rule based and HMM tagger, is designed. The author preferred the hybrid tagger because it performs better than the individual ones. A corpus with a total of 26, 000 words from different Tigrigna news broadcasting agencies was collected. The corpus prepared for this work was only news domain. Thirty six tag sets were identified and used in annotating the corpus for training the

HMM and rule based taggers. A supervised learning approach is used.

The author divided the corpus into training set and testing set for testing and training purposes. The training set consists of 75% of the corpus (20,000 words) and the testing set consists of 25% of the corpus (6,000 words). Different experiments are conducted for the three types of taggers namely the HMM tagger, the rule-based tagger and Hybrid tagger. Accordingly, 89.13%, 91.8%, and 95.88% performances are obtained for HMM, rule-based and hybrid taggers, respectively. It is concluded that the hybrid tagger for Tigrigna performs better than HMM tagger and rule-based tagger used independently.

3. The Proposed Solution

3.1 Data Collection and Corpus Preparation

A corpus can be defined as a systematic collection of naturally occurring texts, may be written or spoken [9]. For this work there is no readymade and balanced corpus developed before, so raw text was collected for corpus development from different datasets like Grade 5, 6 and 7 Hadiyyisa student textbooks, text transliterated from newsletters, University Hadiyyisa language teaching modules and Hadiyyisa Proverbs [10]. A total of 1280 sentences were collected and manually tagged with the 32 derived tag sets and another corpus is derived from the 32 tag sets tagged corpus to 10 basic tag sets tagged corpus.

3.2 Model Selection

We experimented on two kinds of POS taggers and compared their performance with different baseline taggers and combination with sequential backoff taggers in different sequences. The taggers used in this work are a transformational tagger (Brill algorithm) and a Hidden Markov Model (Tag 'n' Trigrams algorithm) approach.

TnT, the short form of *Trigrams 'n' Tags*, is a very efficient statistical part of speech tagger based on second order Markov models that is trainable on different languages and virtually any tag set. The component for parameter generation trains on tagged corpora. The system incorporates several methods of smoothing and handling unknown words [5].

The rule based approach, as its name indicates, relies on rules which are either handcrafted or machine learned. The rules are the important elements for annotating words in the rule based approach and require only small amount of training data and useful for limited domain.

3.3 Natural Language Toolkit and Python

The selected models are tested using the infrastructure of natural language processing which is NLTK using Python programming language [8].

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces and it has lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, semantic reasoning, and wrappers for industrial-strength NLP libraries [8]. It is an open source Python library for Natural Language Processing.

Python is a simple but powerful programming language with excellent functionality for processing linguistic data [8, 11]. It has efficient and high level data structure with simple but effective approach to object-oriented programming [8]. In this work Python 2.7.11 is used.

3.4 The Algorithms used with TnT and Brill Taggers

One class of sequential tagger is a regular expression. By using a regular expression, instead of looking for the exact word, we can define a regular expression that follows some patterns, and at the same time we can define the corresponding tag for the given expressions. For Hadiyyisa text, we developed the code shown below for some of the most common patterns to assign the different parts of speech. The designed regular expression helps to tag words using the TnT and Brill tagger as backoff tagger with other baseline taggers as unknown word handling and initial state tagger respectively. The

code below shows the Regular expiration for 32 derived tag sets for tagging the corpus.

```
RegexpTagger([
  (r'(.*o|.*oo|.*i|.*yo|.*ee|.*aa|.*
    kkoo|.*kkok)$', 'V'),
  (r'(.*nne|.*nnee)$', 'NPREP'),
  (r'(ee*|ka*)$', 'PROND'),
  (r'(an|An||I|i||Ki|ki)$', 'PRON'),
  (r'.*K$', 'ADJ'),
  (r'^[0-9]+(.[0-9]+)?$', 'CARDN'),
  (r'.*', 'N')])
```

The AffixTagger class is another ContextTagger subclass, but this time the context is either the prefix or the suffix of a word. This means the AffixTagger class is able to learn tags based on fixed-length substrings of the beginning or ending of a word. One can combine multiple affix taggers in a backoff chain if there is an interest to learn the tagging on multiple character length affixes [11]. We use the three suffix AffixTagger classes learning on -1, -2 and -3 character suffixes for Hadiyysa text based on the behavior of the language and used the tagger as base line combination as backoff tagger with other taggers:

```
from nltk.tag import AffixTagger
suf1_tagger = AffixTagger(dtrain,
  affix_length=-1)
suf2_tagger = AffixTagger(dtrain,
  affix_length=-2,
  backoff=suf1_tagger)
suf3_tagger = AffixTagger(dtrain,
  affix_length=-3,
  backoff=suf2_tagger)
backoff=suf3_tagger
```

3.5 Combining Taggers

A straight forward way to improve tagging is to combine different taggers sequentially, hoping to take advantage of the fact that different taggers are good at tagging different constructions. Combining classifiers in the context of POS tagging has mainly been used to increase tagging accuracy.

Backoff tagging is one of the core features of Sequential Backoff Tagger. It allows chaining taggers together so that if one tagger doesn't know how to tag a word, it can pass the word on to the next backoff tagger. If that one can't do it, it can pass the word on to the next backoff tagger, and so on until there are no backoff taggers left to check [8]. In this work, as a method to increasing performance we combined and evaluated Uni, Bi and Trigram taggers as sequential backoff tagging.

3.6 Evaluation

After developing the tagged corpus, even though our corpus is very small compared to the others like brown corpus in NLTK, the performance of TnT and Rule based taggers were evaluated using the corpus 80/20 percent for training and testing with different initial state tagger for Brill and different unknown word handling tagger for TnT tagger and combination of taggers by sequential backoff tagging in two different corpuses (i.e., tagged with the basic and the identified full tag set corpus). The reason behind considering basic tags is to determine how much the result differs between tests using basic data and tests making use of full data.

4. Results and Discussion

One of the main purposes of running tests of the taggers was to determine how accurate each tagging tool is, i.e., how much of the testing set input it tags correctly. Accuracy results for all taggers are presented below.

4.1 Results of Baseline Taggers

Before evaluating the performance of the TnT and Brill Tagger, we evaluated baseline taggers with their default options to see the performance on the two tag set tagged corpus and use for the combination (i.e., for initial and unknown word handling and combination as sequential backoff tagging) for the proposed taggers. Table 1 shows the results of the taggers on the two corpuses tagged with basic and derived tag sets.

Table 1: Baseline Tagger Performance

Corpus tagged with	*Default (%)*	*Affix (%)*	*Regex (%)*	*Unigram (%)*	*Bigram (%)*	*Trigram (%)*
Basic tag set	32.58	58.10	51.89	53.84	4.18	3.27
Derived tag set	31.85	50.51	33.45	49.56	3.40	2.71

As shown in Table 1, Affix Tagger performs better than the other taggers on the two corpuses when applied alone on Hadiyyisa corpus. But trigram and bigram taggers give lower results in accuracy whereas the Unigram and Regex taggers results are higher than Bi- and Tri-gram taggers.

The bigram and Trigram taggers manage to tag every word in a sentence during training, but does badly on an unseen sentence. As soon as it encounters a new word, it is unable to assign a tag. It cannot tag the following word even if it was seen during training, simply because it never saw it during training with none tag on the previous word. Consequently, the tagger fails to tag the rest of the sentence.

4.2 Results of TnT-HMM and Brill Tagger

To evaluate the performance of the taggers, 6 different experiments are conducted with different unknown word handling (smoothing) for TnT and initial state taggers for Brill taggers namely Default, Affix, Regular Expiration, Unigram, Bigram and Trigram tagger on two corpuses (i.e., basic and derived tag set tagged corpus). The default tagger assigns Noun (N) as a default word class based on frequency of individual tag within the training set. While in case of Affix we use -1, -2, and -3 suffix letters and unigram tagger assigns the most likely tag for a given word based on the frequency of individual tag within the training set. Table 2 shows the different experiments and their corresponding performance of the taggers. When the rule based tagger uses default tagger as initial state tagger, it assigns tags to every single word, even for words that have never been encountered before. From the result in Table 2, TnT with Affix unknown word handling tagger on basic tagset tagged corpus performs better than the others.

Table 2: TnT and Brill Tagger Performance

Tagger	*TnT Taggers with*		*Brill Taggers with*	
	Basic tag set Tagged Corpus %	*Derived tag set Tagged Corpus %*	*Basic tag set Tagged Corpus %*	*Derived tag set Tagged Corpus %*
Default	67.24	57.80	61.25	49.48
Affix	80.34	66.64	72.67	64.65
Regex	72.54	65.77	60.64	53.96
Unigram	54.01	50.39	53.92	49.69
Bigram	54.00	50.30	51.811	51.25
Trigram	54.01	50.37	50.99	50.38

4.3 Results of Combination Taggers with Backoff Tagging

There are different types of taggers used as the sequential backoff tagger such as Regexp Tagger, Affix Tagger, Unigram tagger, Bigram tagger and Trigram tagger. The backoff taggers are implemented in the Brill‘s and TnT taggers to detect and tag the word which is not tagged by the preceding tagger and the unknown words [8]. When unigram tagger assigns the most likely tag for a given word, the bigram and trigram taggers assign tags to words based on the preceding one and two words with their corresponding tags respectively [12].

Table 3: Combination of Taggers with Backoff Tagging Rule Based and TnT Tgagger

Tagger	*Accuracy %*	
Corpus tagged with	*Basic tag set*	*Derived tag set*
Brill-Initial state with Backoff Taggers		
• TrigramTagger, BigramTagger, UnigramTagger, RegexpTagger	72.37	65.25
• TrigramTagger, BigramTagger, UnigramTagger, AffixTagger	87.25	72.54
TnT -Unknown word Handling with Backoff Tagger		
• TrigramTagger, BigramTagger, UnigramTagger, RegexpTagger	72.54	65.78
• TrigramTagger, BigramTagger, UnigramTagger, AffixTagger	89.03	73.06

On the experiments from combinations of the above taggers (Regexp, Affix, Unigram, Bigram and Trigram taggers), comparatively good performance is gained by the TnT tagger with unknown word handling with backoff in the sequence of TrigramTagger, BigramTagger, UnigramTagger, AffixTagger and Rule based tagger when the initial state tagger with backoff in the sequence of TrigramTagger, BigramTagger, UnigramTagger, AffixTagger annotator respectively 73.06% and 72.54% on derived tag set and 87.25% and 89.03% on basic tag set tagged corpus.

5. Conclusion and Future Work

Part of speech tagging is a very important step in most advanced language technology systems. It is a nontrivial problem due to ambiguous words and unknown words, i.e., words are not in the lexicon. POS tagging is harder for some languages than others. A straight forward way to improve tagging is to combine different taggers, hoping to take advantage of the fact that different taggers are good at tagging different constructions. This work selected a Rule based, TnT and combination by backoff chain taggers at sentence level for Hadiyyisa language.

Corpora are important for many types of linguistic research including part of speech tagging. We used a corpus with a total of 1280 sentences which is collected from different genres. There are several standard tag sets used in corpora and in POS tagging experiments. This work used 32 part of speech tags which are identified as a tag set for annotating a raw text and reduced version with 10 basic tag sets and used two different corpuses which are tagged with 10 and 32 tag sets.

The tagged corpus is divided into training and testing set. The training set comprises of 80% of the corpus and the reaming 20% of the corpus is used as a testing set.

In this work NLTK and Python are used in the experimentation of Hadiyyisa part of speech taggers. To test the performance of the taggers different experiments are conducted for the two types of taggers namely the Rule-based and TnT and also combination with backoff on the two different (basic and identified derived tag set) tagged corpus. As a result, 64.65%, 66.64%, 72.54% and 73.06% performances are obtained for Rule-based with Affix Tagger initial state tagger, TnT with Affix unknown word taggers, and Rule based tagger when the initial state tagger is backoff in the sequence of TrigramTagger, BigramTagger, UnigramTagger, AffixTagger, and TnT-unknown word handling with backoff in the sequence of TrigramTagger, BigramTagger, UnigramTagger, AffixTagger, respectively on identified 32 tag sets and 80.34%, 72.67%, 87.25% and 89.03% results were obtained for the basic 10 tag set tagged corpora respectively. From the experiment the performance using basic tag set tagged corpus performs better than the full identified tag set tagged corpus and also higher performance is gained by combining taggers for

initial state tagging and unknown word smoothing at sentence level for Brill and TnT taggers respectively.

For the future, we would like to extend the work in the following areas.

- Preparation of a balanced corpus that contains texts which represent different genres like Bible, Newspapers, Fiction, Textbooks, Parliamentary Reports, etc.
- Comparative study of different approaches HMM based, Rule-based, and Artificial Neural Network based taggers for Hadiyyisa language with more training and testing data.

References

[1] Daniel Jurafsky and James Martin, "Speech and Language Processing: An Introduction to Natural Speech Recognition", Prentice-Hall, Inc, New Jersey, 2000.

[2] Kim Ballard, "The Frameworks of English: Introducing Language Structures", 3rd ed. Palgrave Macmillan, 2013.

[3] Simon and Schuster, "Handbook for Writers; Psychology Encyclopaedia: Pattern Recognition", Lynn Quitman Troyka, 2002,

[4] Ferran Pla and Antonio Molina, "Natural Language Engineering: Improving Part of-Speech Tagging using Lexicalized HMMs", Cambridge University Press, UK, 2004.

[5] Brants, Thorsten, "TnT – A Statistical Part-of-Speech Tagger", In Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP 2000), Seattle, WA, 2000.

[6] Teklay Gebregzabiher, "Part of Speech Tagger for Tigrigna Language" Unpublished Masters Thesis, Addis Ababa University, 2010.

[7] Getachew Mamo, "Part-of-Speech Tagging for Afaan Oromo Language", Unpublished Masters Thesis, Addis Ababa University, 2009.

[8] Steven Bird, Ewan Klein, and Edward Loper. "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit", O'Reilly Media, 1st ed, Cambridge, 2009.

[9] Nadja Nesselhauf, "Corpus Linguistics: A Practical Introduction", 2011.

[10] Getahun Waaxummo, "Hadiyy Heessechchaa Kobillishsha", Nigd Mattemiy Dirijjit, undated.

[11] Python Programming Language: http://www. python.org, Last Accessed on Aug. 8, 2013.

[12] Jurafky. D. and Martin. H. James, "Speech and Language Processing, An introduction to Natural Language Processing, Computational Linguistics and speech recognition", 2nd ed. New Jersey, Prentice Hall, 2006.

Grapheme Based Tigrinya Speech Synthesis using Statistical Parametric Speech Synthesis

Luel Negasi
HiLCoE, Computer Science Programme, Ethiopia
luelnegasi@gmail.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

In this paper a model-based speech synthesis prototype for Tigrinya language idiom is developed in an integrated speech synthesis framework (Festival speech synthesis system).While the frontend of the framework is Grapheme based synthesizer, the backend is CLUSTERGEN Synthesizer which is an instance of statistical parametric speech synthesis. The under resourced linguistic nature of the language was the main reason to choose this framework. 249 Tigrinya graphemes were considered as phonemes independently; irrespective of its 32 phonological phonemes. For this study, 800 previously prepared sentences and rerecorded again in a recommended way is used as corpus. Amendments and additions to the adopted methodology was done. The whole prototype synthesis development was done automatically. A tenfold threshold method was used for training and testing of the prototype. The synthesized speech was Android deployable prototype. This synthesized speech resulted in a score of 5.82 using Mel Cepstral Distortion (which is built-in objective measurement metric); while subjective evaluation resulted in 4.5 and 4.3 out of 5 score, naturalness and intelligibility of the synthesized speech respectively. Both evaluations were interpreted as the synthesized speech was almost the same as natural human speech.

Keywords: Speech Synthesis; Statistical Speech Synthesis; Grapheme; Spoken Language

1. Introduction

Speech synthesis is playing different roles in today's modern human activities like aid for the disabled and in the telecommunication sector [11]. Its task is converting written text to speech. Naturalness and intelligibility are the main quality measurements of a synthesized speech. Meanwhile text to speech consists two main sub parts known as natural language processing and digital processing that are also termed as frontend and backend systems respectively [11].

Frontend systems are used for processing linguistic part of a language that is going to be synthesized. Many of its tasks relay on the nature of a language. In general, it consists of activities such as text processing, converting text to sound and prosody modeling.

In the past few decades, different synthesizing methods (backends) have been implemented. Nowadays the popular method is statistical parametric speech synthesis which is model based [4, 11, 16].

Tigrinya is a Semitic language spoken mainly in Eritrea and northern part of Ethiopia (Tigray). It is phonetic to mean that what is written is what it pronounces [6, 12, 13]. Meanwhile it is under resourced language in its linguistic part that can be used as an input for speech synthesis mainly for the frontend part.

Very few attempts in synthesizing Tigrinya had been done such as those by Tesfay Yihdego [15], Agazi Kiflu [7] and Lemlem Hagos [5]. All these attempts are based on different types of concatenative synthesis techniques. They used different transliteration mechanisms but all were tried

to transliterate the Unicode based orthography to Latin alphabets in order to get pronunciation dictionaries using these Latin alphabets. This way of finding pronunciation dictionaries adds extra process though it is possible to convert its Unicode letters directly to International Phonetic Alphabet (IPA). Moreover, all attempts were based mainly on the phonological phonemes of the language. In addition to that the backends they used were not model based techniques. The frontend and backend systems used to synthesize Tigrinya language so far lack integrity.

In this paper 249 graphemes of the language were used as means of finding pronunciation. This is because it is easy if the graphemes are considered as phonemes since Tigrinya writing system is phonetic. Grapheme based Synthesizer which is integrated in Festival/Festvox framework was used as frontend.

The backend system used was also model based statistical parametric speech synthesis method known as CLUSTERGEN Synthesizer.

1.1 Tigrinya Language and its Writing System

Tigrigna is the official language of Tigray region of Ethiopia and is the language most widely spoken in this region. It is also the national language of Eritrea which is the neighboring country of Ethiopia. Hence it is a medium of instruction of primary schools in the aforementioned areas and other Tigrigna speaking people [6, 12, 13].

Tigrigna's script has phonetic nature that has 249 characters, known as fidels or Abugida, orthographically. This orthographic representation of the language is organized into seven orders borrowed from Geez [13]. These seven orders are grouped by concatenating consonant and vowel. The following are the first two rows of seven ordered fidels.

Hä, hu, hi, ha, he, hə, ho / ሀ፡ ሁ፡ ሂ፡ ሃ፡ ሄ፡ ህ፡ ሆ

Lä, lu, li, la, le, lə, lo / ለ፡ ሉ፡ ሊ፡ ላ፡ ሌ፡ ል፡ ሎ

Table 1: Sample Seven Orders of Tigrinya Graphemes (Fidels)

	ä	U	i	a	e	(ə)	O
H	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
Xl	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
M	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ

Text to speech systems have two main parts known as natural language processing and digital signal processing.

1.2 Natural Language Processing (NLP)

This is a module that involves text analysis, phonetic analysis and prosodic generation, and is responsible to make the text ready to be used by the synthesis module [2].

An instance of NLP module is Grapheme based synthesizer that is integrated in the Festvox voice building tool. It works for Unicode based orthographic languages using Unitrantransliteration tool [10]. It is more efficient for languages that have phonetic natured writing systems such as Spanish and Indic. Tigrinya is one of those languages that have direct correspondence between its letters and sounds of those letters.

1.3 Digital Signal Processing (DSP)

Digital signal processing is concerned with application of algorithms and models to generate synthetic speech from text manipulated by the NLP module [8, 12]. This is a backend part of speech synthesis also termed as speech synthesis method.

Text to Speech System Methods can be generally grouped into three main categories known as parametric speech synthesis, concatenative speech synthesis and statistical parametric speech synthesis according to their behavior of conversion of phone sequence to speech waveform [9].

Parametric synthesis systems (SPSS) parameters prediction coefficients are extracted from speech signal for each phone unit. Speech parameters are derived manually.

Concatenative synthesis method synthesizes based on the concatenation of prerecorded speech segments. This technique alleviates the drawback of parametric synthesis method which is manual derivation of parameters. However, it is limited to one speaker and one voice and usually requires more memory capacity than the other synthesizers [3, 12].

The third backend method, which is statistical parametric synthesis, is described as speech synthesis method that enables generating an average of some set of similarly sounding speech segments using parametric models [4]. This method differs from the traditional parametric method by using machine learning techniques to learn the parameters and any associated rules of co-articulation and prosody from speech data. It also differs from concatenative method since it does not use prerecorded speech segments. Instead it uses parameters that are extracted from speech signal [4]. Generally, SPSS is model based but the other two are not. SPSS has two distinct phases termed as training and synthesis phases.

One instance of SPSS is CLUSTERGEN Synthesis where its framework is integrated in Festival Speech Synthesis System [1]. It is a method for training models and using these models at synthesis time. The training requires well recorded utterances, and text transcriptions of what has been said.

2. Related Work

Sitaram *et al*. [10] developed Universal Grapheme-based Speech Synthesis. They tested it on 12 languages such as English and German from resourced languages, and Tamil and Thai from under resourced languages. The results were remarkable, especially for languages that have one to one mapping between their orthography and phoneme. They used CLUSTERGEN's Mel Cepstral Distortion (MCD) objective measure for their synthesized speeches. The lowest (that has good quality of synthesized speech) was 4.30 which is German while the highest was 6.71 which is Ojibwe.

Tesfay Yihdego [15] developed a prototype for Tigrigna Language using TTS System that is Diphone based concatenative speech synthesis. Diphones are used as the basic concatenation units to synthesize sample Tigrigna texts and also Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) is used as a technique to generate the synthetic speech.

Unit selection based Text-to-Speech (TTS) system for Tigrinya using the Festival framework and practical applications of it was done by Agazi Kiflu [7]. The result was generating speech from the prepared corpus via selecting the needed units. Another work by Lemlem Hagos [5] developed Tigrigna Text to Speech Synthesizer. A prototype is developed using Festival Speech Synthesis Framework.

All the reviewed works are concatenative speech synthesis method based that is advantageous in high quality speech. But due to its consumption of enormous costs and time for constructing corpora and is not straightforward to synthesize diverse speakers' such as emotions, styles and the like, it becomes less favourable for under resourced languages and implementation in different application areas [11, 14]. In addition, they all transliterated Tigrinya letters to Latin alphabets and then to IPA instead of directly to IPA which adds extra process.

3. The Proposed Solution

The proposed solution of the gap shown in the previous attempts of synthesizing Tigrinya was aimed to the transliteration from Geez character to Latin Alphabet by making it directly to IPA and changing the backend to model based synthesizer which is CLUSTERGEN. The other solution to the gap was considering Tigrinya writing system as main base for synthesis instead of its phonological phonemes (32 phonemes). The proposed architectural view is shown in Figure 1 to indicate how the whole synthesis process looks like.

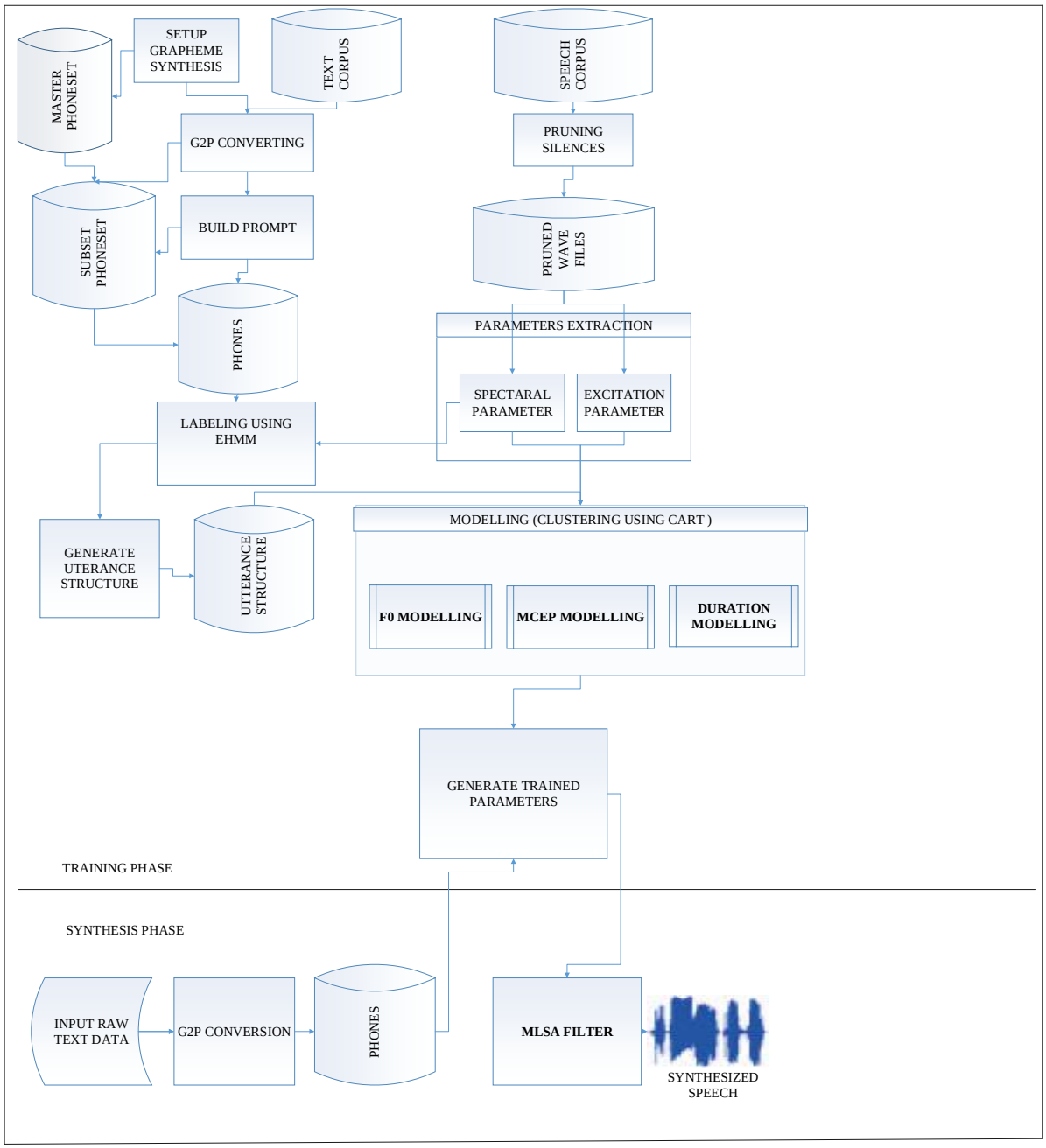

Figure 1: The Proposed Tigrinya General Speech Synthesis System Architecture

Eight hundred sentences that were previously prepared but recorded again in professional studio by native experienced female journalists was used as corpus.

Different software toolkits from recording to synthesis were used. Dell Laptop Core i7 that has 4 Gigabyte RAM and dual operating system, 32 bit Windows 10 and 32 bit Ubuntu 14.04 LTS was used in order to run this software. Wired microphone that had earphone was used for recording purpose while PRAAT was used as recording tool.

The list of software used, in which all are free, are: PRAAT, speech_tools-2.5.0-current.tar.gz, festival-2.5.0-current.tar.gz, festlex_CMU.tar.gz, festlex_POSLEX.tar.gz, festvox_kallpc16k.tar.gz, festvox-2.7.3-current.tar.gz and Speech Signal Processing Toolkit (SPTK-3.6.tar.gz). These software, except PRAAT, which was installed on Windows, were installed on Ubuntu.

4. Discussion

Preprocessing activities such as removing byte order and marker order from UTF-8 based Tigrinya

characters and preparing the corpus was done initially. The whole synthesis process was done using an incremental development process where first 200 utterances (sentences) were synthesized and in the next doubling of the first utterances were added, till it reached 800.

From the prepared text corpus phoneset was derived using the Grapheme based synthesizer frontend system. However, the phoneset was not full enough since it ignored 14 pharyngeal consonants. Tigrinya's 14 consonant characters which are pharyngeal voiceless fricatives (ሐ, ሑ, ሒ, ሓ, ሔ, ሕ) and pharyngeal voiced fricatives (ዐ, ዑ, ዒ, ዓ, ዕ, ዖ) were added to make the synthesis system efficient and effective. A phone '&' derived automatically also caused a problem, in the next steep, while labelling. This was due to the occurrence of this phone almost in every word of the language. This phone is changed to AMP and automatic labelling using Embed Hidden Markov labeler (EHMM) was done successfully. This was the beginning of training of Tigrinya speech synthesis. Spectral and excitation parameters were extracted using SPTK from the parallel speech corpus. These extracted parameters are used to develop three models named Spectral (F0) model, Excitation (MCEP) model and Duration model using Classification And Regression Tree (CART) via SPTK. Afterwards, these models are combined into one to use for synthesis phase. The text corpus was separated into training and testing, based on tenfold threshold. In the synthesis phase from the test dataset phones were derived automatically and these phones were aligned acoustic features from the combined trained model. Finally, synthesized speech was enabled using Mel Log Spectrum Approximation (MLSA) vocoder.

For the synthesized speech an objective measure, MCD, was generated automatically.

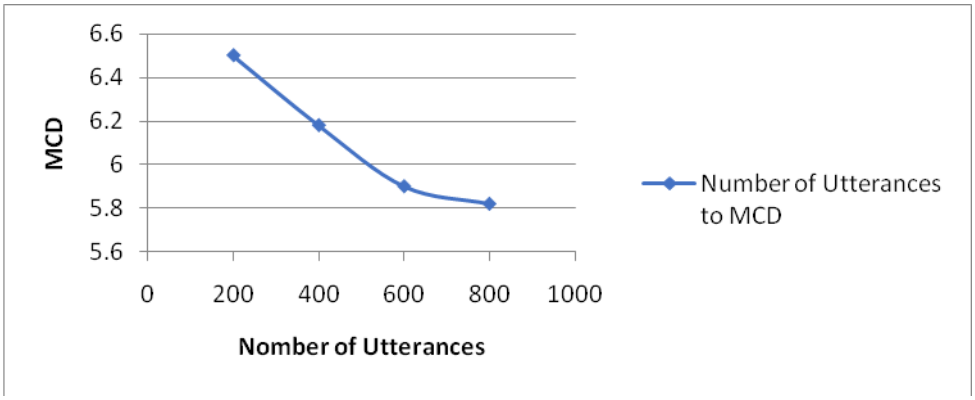

Figure 2: Experimental Results (for Objective Evaluation)

As shown in Figure 2, MCD of final 800 sentences was 5.82 which can be interpreted as the synthesized speech almost as the same as natural human speech sound in its naturalness and intelligibility.

Subjective evaluation using mean opinion score (MOS) that was done by 8 native speakers for randomly selected 10 sentences also resulted the same to MCD when interpreted. These evaluators were 4 males and 4 females aged between 18 and 64, which were diploma to master educational levels. Naturalness was evaluated as 4.8 while intelligibility was 4.3 out of 5.

5. Conclusion and Future Work

Lack of efficiency and effectiveness of synthesizing Tigrinya language was identified as a gap. Based on the identified gap, we developed a model-based speech synthesis prototype which is integrated into one comprehensive framework. The main reason for selecting this framework was the under resourced linguistic nature of the language. The Unicode encoded characters (orthography) of Tigrinya were considered as phonemes irrespective of the phonology of the language. The text corpus was existing. However, recording (speech corpus)

was done professionally in a recommended way for speech synthesis.

The whole end to end synthesis was done automatically using Grapheme based synthesis and CLUSTERGEN on Festival speech synthesis system. The synthesis methodology used on this study is an adoption of a well-accepted system, that had got recognition from other phonetic nature languages and for the sake of this study we made some additions and amendments.

Direct transcribing Tigrinya letters to IPA, that alleviated transliteration of the characters from Tigrinya letters to Latin that had been used in earlier works was tested and enabled; additions of 14 pharyngeal letters to the system derived Tigrinya phoneset was made; modification of one phone that frequently occurs in almost all words of Tigrinya that is used in the case of vowel insertion was changed from '&' to 'AMP'. A prototype that could possibly exploitable flite voice which could run on any Android device was developed.

Generally, an efficient, effective and model based speech synthesis prototype is developed (in Festival speech synthesis system) that has remarkable natural and intelligible speech. The test and evaluation results of the prototype have shown promising possibility of developing model-based Tigrinya speech synthesizer.

As future work, adding gemination handler module, using phonetically balanced and prosodically rich corpus and adding interpolation for a variety of speaking styles will be done for better result.

References

[1] Black, A. W., "CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling", In Ninth International Conference on Spoken Language Processing, 2006.

[2] Black, A. W. and Lenzo, K. A., "Building Synthetic Voices", Language Technologies Institute, Carnegie Mellon University, 2014.

[3] Black, A. W. and Taylor, P. A., "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", 1997.

[4] Black, A. W., Zen, H., and Tokuda, K., "Statistical Parametric Speech Synthesis", In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, 2007.

[5] Lemlem Hagos, "Developing Tigrigna Text to Speech Synthesizer", Unpublished PhD Dissertation, Addis Ababa University, 2015.

[6] Kievit, D. and Kievit, S., "Differential Object Marking in Tigrinya", Journal of African Languages and Linguistics, 2009.

[7] Agazi Kiflu and Tibebe Beshah, "Unit Selection Based Text-to-Speech Synthesizer for Tigrinya Language", HiLCoE Journal of Computer Science and Technology,Vol 1. No. 1, 2013.

[8] Prahallad, K., "Automatic Building of Synthetic Voices from Audio Books, Unpublished PhD Dissertation, Carnegie Mellon University, 2010.

[9] Shinoda, K. and Watanabe, T., "MDL-based Context-dependent Subword Modeling for Speech Recognition, 2001.

[10] Sitaram, S., Parlikar, A., Anumanchipalli, G. K., and Black, A. W., Universal Grapheme-based Speech Synthesis", In the Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[11] Taylor, P., "Text-to-Speech Synthesis", Cambridge University Press, 2009.

[12] Tesfay Tewolde, "A Modern Grammar of Tigrinya", Rome, Italy, 2002.

[13] Tigrigna Ethnologue, www.ethnologue/ language/tir, Last Accessed on 27 November, 2016.

[14] Yamagishi, J., "An Introduction to HMM-Based Speech Synthesis", Technical Report, 2006.

[15] Tesfay Yihdego, "Diphone Based Text-to-Speech Synthesis System for Tigrigna", Unpublished Masters Thesis, Addis Ababa University, 2004.

[16] Zen, H., Tokuda, K., and Black, A. W., "Statistical Parametric Speech Synthesis", Speech Communication, 51(11), pp. 1039-1064, 2009.

Developing Loan Advisor Model for Customer Loyalty using Data Mining

Teshome Bulo
HiLCoE, Computer Science Programme, Ethiopia
World Vision Ethiopia (WVE)
teshebb@gmail.com

Temtim Assefa
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
temtimasefa@yahoo.com

Abstract

Customer relationship management is vital in financial institutions for their successful operations. To develop a good customer relationship, it is required to understand customer characteristics by analysing historical customer data. Financial organizations capture massive data which isn't feasible to do by traditional methods. Wasasa is not exceptional to this problem. Even though the main goal of microfinance is to provide financial service to alleviate poverty, it is also necessary to analyse data and generate new knowledge so as to satisfy its customer needs and consequently its future existence.

The purpose of this paper is to use data mining to analyse customer data and predict customer's behaviour. In this paper, an attempt is made to apply different classification algorithms namely Naïve Bayes, KNN, Rule induction (PART), and Decision Tree. As methodology, we used CRISP-DM methodology and WEKA software to implement the selected algorithms. Different experiments were conducted repeatedly by adjusting different parameters until the required output is reached.

There were twelve experiments conducted using four selected algorithms. According to the result of the experiments, KNN, PART, and J48 achieve an accuracy of 99.96%, 99.93% and 99.91% respectively and ROC area of 1.0 that is equal for all algorithms. The rules are generated using the PART algorithm because it achieved slightly better output than the J48 algorithm.

Based on the rules generated using PART rule induction, a prototype system was developed for domain experts. The system will help Wasasa's employees to understand their customers' need and provide better services. As customers become more satisfied with the services they will remain loyal to Wasasa.

Keywords: Microfinance; Loyalty; Data Mining

1. Introduction

The establishment of microfinance has brought many benefits to low income citizens which have direct relationship on the country's development. Even though microfinance is playing a great role in reducing poverty, recently the explosively growing, widely available, and gigantic body of data becomes an issue in all business firms. Analyzing these data easily through known traditional methods has lots of problems. Thus business firms started searching for powerful and versatile tools that automatically extract valuable information from these data and convert such data into organized and understandable knowledge. Microfinances were one of the business firms which were looking for this solution that support them to analyze their customer information during decision making. The popular and fast-growing technology to solve such problems is data mining or knowledge discovery.

Data mining uses a variety of data analysis tools to discover patterns and relationships in the data which may be used to make valid predictions [1].

Data mining is a component of business intelligence that will extract non-trivial knowledge and hidden patterns from large amount of data which enables a business firm to make the right decision.

The derived knowledge must be new, not obvious, relevant and can be applied in the field where this knowledge has been obtained. It is also the process of extracting useful information from raw data. Data mining is applicable in all areas of business including manufacturing, financial, governmental and health institutions to discover new knowledge. Out of these application areas this research is conducted on financial institutions [2, 4].

Data mining presents an opportunity to increase significantly the rate at which the volume of data can be turned into useful information [5]. The application of data mining is essential in any organization to make important decisions that might be crucial for the well-being of the whole business [3]. Data mining has different algorithms from which a suitable algorithm can be selected to generate meaningful patterns. Hence, data mining is recently becoming a popular tool for analysing big data and reveals relevant information for decision making.

Even though different attributes are in use to define loyalty within the domain area, the definition given to loyalty is not consistent and vary from company to company. The same is true for Wasasa and loyalty is defined from different perspectives in Wasasa; such as based on saving, repayment schedule, gender and others. This shows that within the same company there is no proper definition assigned to loyalty.

Accordingly, Wasasa is facing the same challenges in relation to customer loyalty definition. One of the challenges is that the institution is fully dependent on the credit officer ability and judgments to assign label to its customers which makes the company vulnerable because of the subjective definition assigned to loyalty. Moreover, apart from the operational manual there is no standard set to define loyalty within the company during loan appraisal that all branches follow.

2. Related Work

The work in [6] developed a model on a selected MFI that segments and predicts customers based on historical data. The author follows the CRISP-DM process model and employed k-means clustering to segment the dataset into different clusters and J48 classification algorithm was used to develop the final model. One of the limitations is that clustering is based on the distance between each point and cluster which takes a longer time to calculate when the dataset size becomes large. The other problem is that selecting the initial K is also challenging and the author used only decision tree algorithm to develop the final model with dataset of 13,057 records and 6 attributes for the experiment.

Simret Solomon [3] conducted a research on predicting customer loyalty using data mining techniques and built a model that predicts customer loyalty with a dataset of 6,447 records and 10 attributes. The researcher used three classification algorithms: Rule Based ZeroR, Decision Tree (J48), and Bayes (Naïve Bayes) and three test option training set, 10 fold and split option to build a model. Finally, the researcher selected J48 decision tree algorithm with 10 fold cross validation among the other algorithms which produces an accuracy of 97.8% for customer loyalty. The researcher concluded that the MFI can use this model during pre-loan disbursement to identify default customers and minimize risks credit management.

Wagari Dibaba [7] conducted a research in the application of data mining in microfinance using a dataset of 9,550 records and 13 attributes. The objective of the research was building customer classification model for better customer relationship management. The author followed CRISP-DM methodology to conduct the research and used a classification decision tree algorithm to develop the final model. There were six experiments conducted using the same algorithm with varying size and other parameter settings. Finally the algorithm that produced an accuracy of 78.54% was selected as best out of the six experiments conducted.

Sara Worku [8] conducted a research in the Addis Credit and Saving Institution. The aim of the research was to develop a predictive model that

investigates credit risk in financial institutions using data mining techniques. The author followed CRISP-DM methodology and uses WEKA tool to develop the model. The data was extracted from four branches of the institution within the same geographical location and contain 4000 records and 7 attributes. There was a class distribution imbalance in dataset so that the author used SMOTE technique to rebalance the minority class distribution. Three classification algorithms were employed to develop the model, namely decision tree, J48, Naïve Bayes and PART rule induction. Cross validation test method was used to validate the model. Finally, the author selected PART rule induction that produces optimal accuracy of 99.83% among the nine experiments conducted for the rebalanced dataset with seven attributes. The author concluded that the result shows that it is possible to extract useful pattern through data mining algorithm that help to make accurate decision for credit risk assessment.

3. The Proposed Solution

The following steps are followed to achieve the objectives stated in this paper.

- Reviewing previous works within the area to understand the domain and attributes selection method, identify gaps and how to employ CRISP-DM methodology that includes business understanding, data understanding, data preparation, evaluation, and deployment.
- Preparing the data in the format that is appropriate for the algorithm selected.
- Importing the prepared dataset to the selected tools and conduct the experiments with 4 algorithms.
- Generating the rule based on the optimal result and develop the prototype
- Present the developed prototype to domain experts and senior management for evaluation.

The main activities are discussed below:

Business understanding: is a process where the core business flows are understood. To understand the business process various documents were reviewed that include loan processing procedure, repayment methods and operational policy.

Data understanding: initially the data were collected from the central database of Wasasa Microfinance. In addition interview and discussion were conducted with selected domain experts. Since the data were extracted for all branches separately, the first task was consolidating these data into one Excel sheet. The data set contains a total of 31,996 records and 44 attributes.

Data preprocessing: is a task that is applied on the dataset to have the dataset a suitable format appropriate for the data mining techniques. There were different steps involved in data preprocessing such as data reduction, attribute selection, data cleaning, data integration, and data transformation [9].

Data Reduction: to enhance the mining process of large data, the reduction of the data size should be performed without jeopardizing the mining results [10]. Attributes that are less useful for the purpose are removed.

Attribute selection: is performed based on the purpose of the research and also the mining tasks applied. Thus, we selected 12 attributes for the final experiments.

Data cleaning: is a task to clean the data by filling missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Data integration: having large amount of redundant data may slow down or confuse the knowledge discovery process [9].

Data Transformation and Discretization: is the last preprocessing step where data is transformed into a suitable form for data mining tasks that increases the mining process efficiency.

After preprocessing, 26,517 records were made ready for the experimentation where 96.28% are

loyal, 2.99% non-loyal, and 0.73% spurious loyalty. But the dataset contains imbalance class distribution within the consolidated data that is 3% of the two classes (spurious and non-loyal) while loyal class is 97% in the dataset which needs rebalancing. A dataset is said to be imbalanced if one class (called the majority or negative class) vastly outnumbers the other (called the minority or positive class). This is a common problem in classification data mining [11]. To rebalance this class distribution there are various techniques available such as Random under Sampling (RUS), Random over Sampling (ROS), and Synthetic Minority Over-sampling Technique (SMOTE). For the purpose of this paper, Synthetic Minority over-sampling Technique (SMOTE) was selected and finally the rebalanced dataset is used for experimentation.

Once the dataset becomes ready and converted into the format supported by the selected tool, that is WEKA 3.8.1, and importing and checking the data is error free, the experimentation follows. There are different algorithms available to develop a predictive model. Among those four algorithms are selected to achieve the objective of this paper. The model developed is evaluated using different metrics to select the optimum result out of the twelve experiments. The metrics used were the one that is available and provided within WEKA tool, and different interview questions designed and conducted with the domain experts. Based on these finally the optimum result is selected and rules generated with the algorithm used to develop the prototype that is easily understandable and used by the business user developed using vb.net programming language.

4. Discussion

To develop the predictive model using data mining, classification algorithms such as Bayes Naïve net, K-Nearest Neighbor (KNN-IBK), PART Rule Induction and Decision Tree (J48) are used. The final cleaned and rebalanced dataset with a total of 75,523 records was given to WEKA tool version 3.8.1 and twelve experiments were conducted using these four selected algorithms. Those experiments are summarized and presented in Table 1.

Table 1: Summary of Experimental Results

No.	*Algorithm*	*Experiments*	*Attributes*	*Evaluative*	*Accuracy*	*ROC*
1.	Naïve Bayes	Experiment 1	12	training set	97.21	0.999
		Experiment 2	12	10 folds	97.19	0.999
		Experiment 3	12	70/30	96.79	0.998
2.	K-nearest Neighbor	Experiment 1	12	training set	99.98	1.000
		Experiment 2	12	10 folds	99.79	1.000
		Experiment 3	12	70/30	99.67	1.000
3	PART	Experiment 1	12	training set	99.97	1.000
		Experiment 2	12	10 folds	99.96	1.000
		Experiment 3	12	70/30	99.94	1.000
4	J48	Experiment 1	12	training set	99.97	1.000
		Experiment 2	12	10 folds	99.95	1.000
		Experiment 2	12	70/30	99.95	1.000

Among the twelve experiments conducted, the model built with algorithm KNN-IBK with training set test method produced the highest accuracy which is 99.98% and ROC area 1.0. Even though the model achieved the highest result, it doesn't have an option to generate rule so that the next model that produced

the highest results was selected to generate the rule that was PART rule induction having accuracy of 99.97% and ROC area of 1.0. The rule generated using this algorithm (PART) is used to develop the prototype shown in Figure 1.

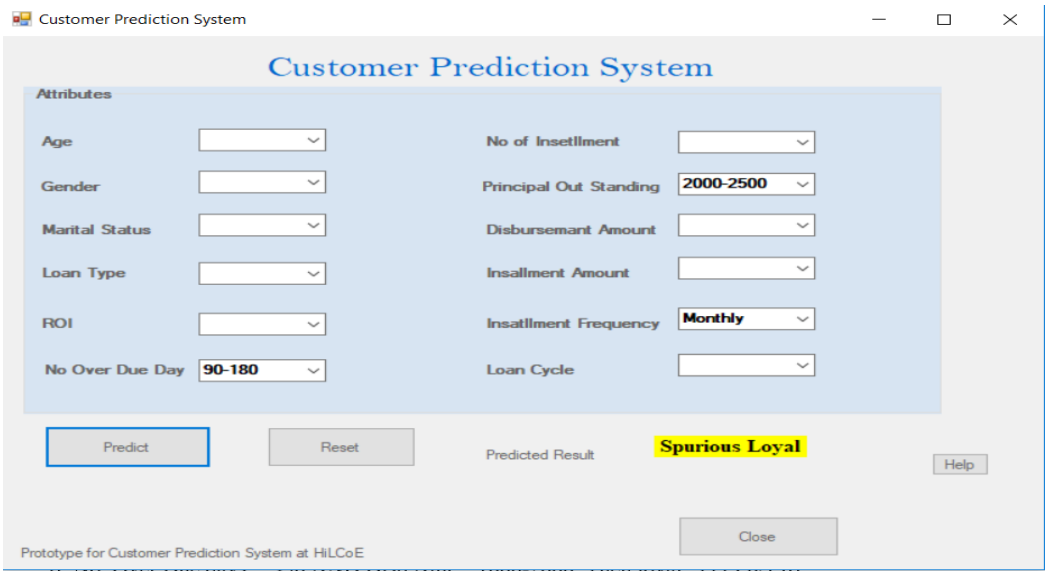

Figure 1: Customer Prediction System Prototype

The model produced better result when compared to a similar previous work [3]. The model developed in this paper is better than that in [3] because: it achieved highest accuracy that approaches to perfection, the model categorized customers into three classes, and there was no model developed previously to allow MFIs to check their customers after disbursement.

5. Conclusion and Future Work

Customer loyalty is vital to any business firm especially in financial institutions like Wasasa. From the dataset prepared for experimentation spurious and non-loyal customers accounted for 3%. Though it looks less when compared with other researches conducted, it has significant impacts on financial institutions business. Analyzing and extracting new knowledge through manual methods from customer records are time consuming and difficult as the size of data keeps increasing from day to day. Not only analysing but also discovering a pattern and relationship will become difficult. Use of the state of the art data analysis technology such as data mining increases organizations' learning capability from their data and solve their problems in a novel way.

From a total of 31,996 records with 44 attributes, 26,517 (82.87%) records with complete data were validated and preprocessed. Among all attributes, 12 relevant attributes were selected to build the model.

The CRISP-DM standard process model and WEKA tool were adopted to achieve the research objective. The classification algorithms Bayes Naïve net, KNN, PART and J48 were employed to develop the model. From experiments conducted on four algorithms, KNN and PART with accuracy of 99.97% and 99.91% respectively were selected to customer loyalty model development. Even though KNN achieved optimum result, it couldn't generate rule. Hence, PART algorithm with better accuracy result next to KNN was used to generate the rules.

Finally, the predictive model that enables the MFI for decision making and loan appraisal is developed.

Future research includes the following:

- Conducting study on developing a prototype integrating to a live database and generating different reports that will support decision making.
- To study further by incorporating additional attributes like income and saving products to widening the study area.

References

[1] Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery", 3rd edition, 1999.

[2] A. J. Hamid and T. M. Ahmed, "Developing Prediction Model of Loan Risk in Banks using Data Mining", University Khartoum, March 2016.

[3] Simret Solomon, "Predicting Customer Loyalty Using Data Mining Techniques," Unpublished Masters Thesis, HiLCoE School of Computer Science and Technology, Addis Ababa, January 2012.

[4] Anteneh Fentahun, "Mining Road Traffic Accident Data for Predicting Accident Severity to Improve Public Health – Role of Driver and Road Factors: The Case of Addis Ababa", Addis Ababa, 2011.

[5] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining", Daimler Chrysler Research & Technology, 1999.

[6] Belachew Reganie, "Application of Data Mining Techniques for Customer Segmentation and Prediction: The Case of Busaa Gonofa MFI," Unpublished Masters Thesis, Addis Ababa University, 2013.

[7] Wakgari Dibaba, "Application of Data Mining Techniques for Effective Customer Relationship Management of Microfinance: The Case of Wisdom Microfinance," Unpublished Masters Thesis, Addis Ababa University, 2010.

[8] Sara Worku, "Application of Data Mining Technology for Credit Risk Assessment in Addis Credit and Saving Institution", Unpublished Masters Thesis, Addis Ababa University, June 2016.

[9] Women in Microfinance and its Social Effects on the Community: A Case Study in Hue, Vietnam, 2016.

[10] V. K. Tripathi, "Microfinance - Evolution, and Microfinance-Growth, of India," International Journal of Development Research, Vol. 4, No. 5, pp. 1133-1153, May 2014.

[11] T. R. Hoens and N. V. Chawla, "Imbalanced Datasets: From Sampling to Classifiers", John Wiley & Sons, Inc., USA, 2013.

Developing a Decision Support System using Fuzzy Logic for Sugarcane Supply Cycle Management

Tinsae Berhanu
HiLCoE, Software Engineering Programme, Ethiopia
ztinsae.berhanu@gmail.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

Supply of sugarcane at the right time in the right amount for years to come is pivotal if it is to achieve maximum productivity of commercial sugar from sugar factories. Sugarcane supply cycle is a process that will stabilize the temporal productivity of sugar from sugar factories by supplying the right amount of sugarcane at the right time from the plantations that are owned by the factories.

This planning system is particularly important in the pioneering sugar factories of Ethiopia as the average cane maturity age in most cases elapses beyond a year. An automated sugarcane supply cycle management system, which is the subject of this paper, will eliminate the hideous drudgery of cane supply planning currently in practice and could facilitate the subject matter experts to review multiple sugarcane supply cycle alternatives in a short period of time. This proposed intelligent software application will generate seven sugarcane supply cycle scenarios; one conventional and six in fast track planning methods.

In addition, some of the features of the estimation process are subjective variables where integrating fuzzy logic to the automated system will make the analysis of supply sugarcane management more objective and intelligent. This paper, therefore, tries to come up with a software application that will generate multiple sugarcane supply cycles in two categories, viz., conventional method and fast-track method with the aid of fuzzy logic.

The proposed prototype is satisfactory as per subject matter experts. This is tested by letting them interact with the prototype and generate various sugarcane supply cycles using both conventional and fast track methods. After generating the supply cycles they were able to manage the printable report with little learning curve to the system's user interface.

Keywords: Fuzzy Logic; Sugarcane Supply Cycle; System Automation; Decision Support System

1. Introduction

Sugarcane is the major raw material used by sugar factories in sugar production. Unlike many other countries involved in the manufacturing of sugar, the Ethiopian sugar industry adopted the nucleus estate sugarcane growing modality whereby both the sugar factory and the sugarcane plantation operate under single management structure [1].

Under nucleus estate cane development modality, the sugar mill is required to develop its own integrated cane plantation that is sufficient to meet the annual uninterrupted cane crushing capacity of the mill [1]. With the help of carefully designed cane supply planning, the sugarcane to be supplied to the mill on a daily basis is required to meet the desired optimum variables in terms of age, varietal composition, cuttings, etc.

Planning the cane cycle program and its management has persisted as major challenge in all sugarcane growing countries in general and in Ethiopian sugar factories in particular. This is attributed to the inclusion of a number of variables in the planning process that are governed by natural phenomena which are beyond the direct control of a

human. It is to be noted that sugar factories are established in different project areas with variable environmental settings that have direct influence on these factors.

In Ethiopian sugar factories, the cane cycle planning process and its management is carried out manually by subject matter experts, reported to be very few in number as the knowledge is gained through long years of work experience that is passed from one person to the other rather than formal training [1]. The planning process is highly drudgery as the work involves huge iteration work, and, as a matter of fact, highly disliked by most staff involved in the process [1].

Currently, the process of developing sugarcane supply cycle is done using MS Excel application. The problem with this process is that it costs a lot of time and energy to come up with the proper supply cycle due to a lot of variables to be considered in the process making it highly prone to error. In addition, usually, one sugarcane supply cycle is not sufficient for plantations.

The aim of this paper is, therefore, to develop a robust software that can assist subject matter experts to develop alternative scenarios in cane supply planning and enable them to select the optimum scenario for implementation. The outcome from this work will be of use for managing the cane supply plan.

Two distinct phases are identified while formulating cane supply plan for a factory [1]. These are (i) initial development (establishment) phase plan and (ii) steady state operation phase plan. With respect to development rate, basically two alternative systems are available [1], namely:

- *Conventional system*: which dictates the development rate to be according to the annual rotation area. Annual rotation area is area to be planted regularly every year during steady state operation phase.
- *Fast track system*: where the allowed development rate is well beyond the annual rotation area, and naturally a number of scenarios meeting this criterion could be formulated. This condition makes it very difficult to formulate the possible supply cycles even with a computer system.

The controlling point will be how closely the long-term cane supply plan in terms of the quantity of cane delivered to the factory on daily, monthly and annual basis matches with the required design capacity of the factory [1]. However, among the possible scenarios, identifying the optimum development rate for the establishment phase planting has been a problem for long and requires human assistance in addition to an automated system that helps to investigate between the various available scenarios within a limited period of time.

Fuzzy logic, in this regard, can deal with information arising from computational perception and cognition, that is, uncertain, imprecise, vague, partially true, or without sharp boundaries [3]. Fuzzy logic allows for the inclusion of vague human assessments in computing problems, where it also provides an effective means for conflict resolution of multiple criteria and better assessment of options [3]. Therefore, the use of fuzzy logic in sugarcane supply cycle planning automation to facilitate decision-making is a wise endeavor.

This paper tries to develop a cane supply planning software prototype that has the capacity to forecast the various cane supply cycle options with change of project parameters in accordance to the characteristics of the project location. Hence, the following research questions are formulated.

- How to design and implement a software to manipulate sugarcane supply planning data?
- How could fuzzy logic model be incorporated in the development of the proposed system?

2. Related Work

The biggest challenge in developing a sugar cane supply cycle plan is the complexity resulted from handful of variables. These, ever unstable variables, make the supply cycle plan difficult to be suitable for

acceptable operation. Sugarcane supply cycles in the world differ in their nature mainly due to the fact that they can be greatly affected by managerial decisions as much as the natural phenomena.

As the requirements elicited from the subject matter experts show, quite a number of inputs that are constant parameters to a specific factory are required to formulate the cane supply planning [2]. These parameters are mainly determined with the help of an in-depth study and analysis of the natural resource base of the specific area where the envisaged sugar project is located. Then, the results of the resource base will be used to formulate detail project report in which all project parameters are determined.

As argued in [2], the elements of the natural resource base that need to be studied are:

- The land resource of the project area – soil type, land suitability and slope class;
- Climate - long years records of the major climatic variables (temperature, rainfall, potential evapo-transpiration, wind run, relative humidity, length of sun shine hours, etc); and
- Hydrology of the area.

It is to be expected that the characteristics of the natural resource base is highly influenced by geographic locations thus exhibits spatial variability. Table 1 shows the list of project parameters determined by detail study and inputs to be used in formulating cane supply planning and management which are used in the proposed system.

Table 1: Expected Inputs

No.	*Input/Parameter*	*Unit*	*Source*
1.	Factory milling capacity	Ton cane per day (TCD)	Detail project report (DPR)
2.	Annual milling days	days	Resource base study
3.	Cropping pattern	Number of ratoons	Resource base study
4.	Harvest/maturity age	months	Detail project report
5.	Average Cane yield	Ton/ha	Detail project report
6.	Commercial cane area (CC)	ha	Detail project report
7.	Seed cane area (SC)	ha	Detail project report
8.	Initial seed cane area	ha	Detail project report
9.	Total plantation area	ha	Detail project report
10.	Length of one occupation period	year	Detail project report
11.	Harvest proportion between 1RS and 2RS cane	%	Detail project report

Two step planning is involved in preparing this initial stage plan, namely “the cropping pattern plan” and “plantation development and cane production projection plan” [6]. The input for preparing this plan is mainly generated from Resource Base Study Report and Detail Project Report carried out at the time of project feasibility study.

The choice of a well suited cropping pattern is of primordial importance to ensure a regular cane and sugar production, a constant juice quality and a seamless management [4, 5]. In the current trend of producing sugarcane supply cycle, as discussed above, the following terminologies are directly used in the formulation of cane supply cycle:

- *Conventional*: in short, this method is dependent on annual rotation area. For a specific environment where the plantation is placed, there is only one kind of conventional sugarcane supply cycle method. Using the ratio provided by the user on initial seed cane and commercial cane the system will provide the amount of area that will be planted starting from the first year until plantation capacity is met. Just before the end of virgin land

planting, depending on the average month of cutting variety, it starts replanting to avoid down time of the factory. According to the planted cane amount, harvesting starts at the third year when the commercial cane is mature enough to produce the expected amount of sugar. Every year, that amount of developed area is determined by Annual rotation area. Depending on the plan on how many ratoon the cane is supposed to last, the first cut will be the first ratoon for next year which will vary in the yield expected. Thus, there will be a variation in the cutting amount.

- *Fast track*: unlike the conventional method, this method can be varied according to the business statues, resource capability and feasibility against the environmental demands. In this method, everything is similar to the procedure used in conventional method except that the annual developed area is not determined by annual rotation area instead it is affected by external factors. Fast track method usually varies by increasing the annually developed area by some percentage. This variations achieved in the process are represented as Fast-Track-One, Fast-Track-Two, Fast-Track-Three, etc.

The main difference between the two methods is that conventional method uses annual rotation area as annual developed area and fast track method uses different external factors of annual developed area to guide the generation of sugarcane supply cycle.

The differences between various fast track methods depend on the business needs of stakeholders. Which means, a faster factory mill has to meet its designed capacity, annual developed area will be annual rotation area plus some percent of the annual rotation area. Usually, adding more than 30% of annual rotation area is not feasible according to the subject matter experts [2].

A research done on Vietnamese farmers decision-making simulation by using fuzzy logic is the most related work found for review. The objective of this paper [2] is to explore the effect of farmers' family motivation on farm diversification and integration using a fuzzy logic model simulating their decision making. Even though the research uses fuzzy logic and agricultural decision-making simulation, it is very difficult to directly relate it with sugarcane supply cycle.

Another work on decision support system [7] presents a detailed framework, conceptual design and prototype of a decision support system which can be used to aid business executives in deciding the appropriate e-business models to use. The design is based on the ideas and concepts of fuzzy logic theory and fuzzy expert systems. The proposed system solves important challenges such as the use of linguistic terms to capture the executives' assessments of the key business measures.

These related works enlighten us the possibility and success of applying fuzzy logic in business area field with much uncertainty and subjective judgments.

3. The Proposed Solution

3.1 Architectural Design

The architectural design (shown in Figure 1) is the high level description of the software application developed. A layered architecture is a very clean architecture for applications that tend to be messy during the development process. It is an appealing option because of the ability it presents to decouple sophisticated procedures in a way the development team can test different related modules without disrupting other modules.

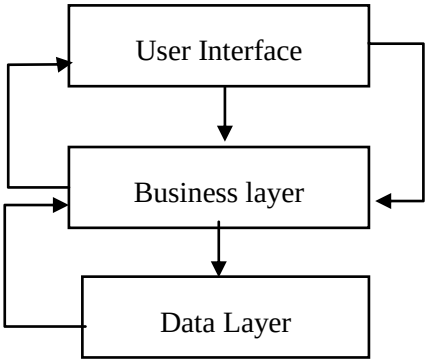

Figure 1: Architectural Design

3.2 Fuzzy Logic Model

Fakhry [7] proposed a decision support framework based on Hayes and Finnegan and the results of Fuzzy Logic Theory and systems by Zadeh in 1965 [8]. Fuzzy logic joins language and human intelligence together using the mathematics of fuzzy membership functions and provides a formal framework to represent and reason with vague, uncertain, and linguistic terms. Using this understanding of fuzzy logic model, the proposed flow of function actions are summarized in Table 4 and described in detail as follows.

a. Linguistic Variables and Terms

In this paper, fuzzy logic model has been used for the generation of fast track method sugarcane supply cycle. Subsequently, developing a linguistics resembling to the terms the professionals have been using to communicate is important for the interface in generating supply cycle. We have come up with these terminologies in consultation with the domain experts, which are presented in Table 2.

Table 2: Linguistics

Supply cycle	*level*
Fast track one	Low
Fast track one	High
Fast track two	Low
Fast track two	High
Fast track three	Low
Fast track three	High

b. Membership Functions

Membership function deals with the association of the linguistics terms with crisp range values. To come up with this numeric values, the subject matter experts have been consulted.

Table 3: Membership Function

Linguistics	*% of annual rotation area*
Fast track one, low	0<x<5
Fast track one, high	5<x<13
Fast track two, low	13<x<15
Fast track two, high	15<x<23
Fast track three, low	23<x<25
Fast track three, high	25<x<30

c. Rule Base

1. If fast track one and low then 5% of annual rotation area is added to annual rotation area.
2. If fast track one and high then 13% of annual rotation area is added to annual rotation area.
3. If fast track two and low then 15% of annual rotation area is added to annual rotation area.
4. If fast track two and high then 23% of annual rotation area is added to annual rotation area.
5. If fast track three and low then 25% of annual rotation area is added to annual rotation area.
6. If fast track three and high then 30% of annual rotation area is added to annual rotation area.

Using the membership functions (parameters and rules) shown in Table 3, the generation of fast track development method part will be executed using the Java code and will be able to generate sugarcane supply cycle.

d. Defuzzification

Defuzzifiaction is the process of using the numeric values and producing the final supply cycle result. Hence, after the process passes the fuzzy rule, it will generate a new fast track method sugarcane supply cycle with the newly acquired value of annual developed area.

Table 4: Functions of the Fuzzy Logic Model

Fuzzyfication	*Fuzzy rule*	*Defuzzification*
▪ Collects linguistic values ▪ Changes into numeric value	▪ Applies fuzzy rule and sends value	▪ Generates sugarcane supply cycle according to the value passed by the fuzzy rule

3.3 Sample User Interface

The user interface appears just after a person is logged in as a professional. It helps the user to fill prerequisite data that will be used to generate sugarcane supply cycle.

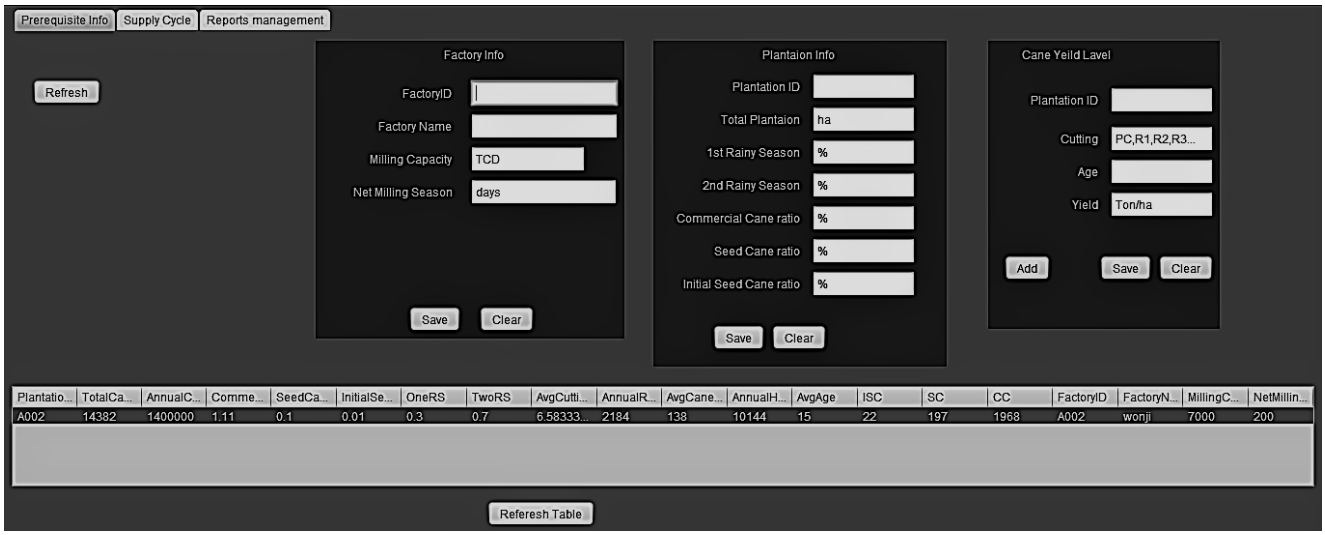

Figure 2: Prerequisite Info Page

Figure 3 shows the page that is used to generate the sugarcane supply cycle and tabular report.

This report is generated based on the data that was provided previously using the screen shown in Figure 2. It also provides the capability to choose between the two types of methods that are used to generate a sugarcane supply cycle. It also provides the option to choose the level of fast track method.

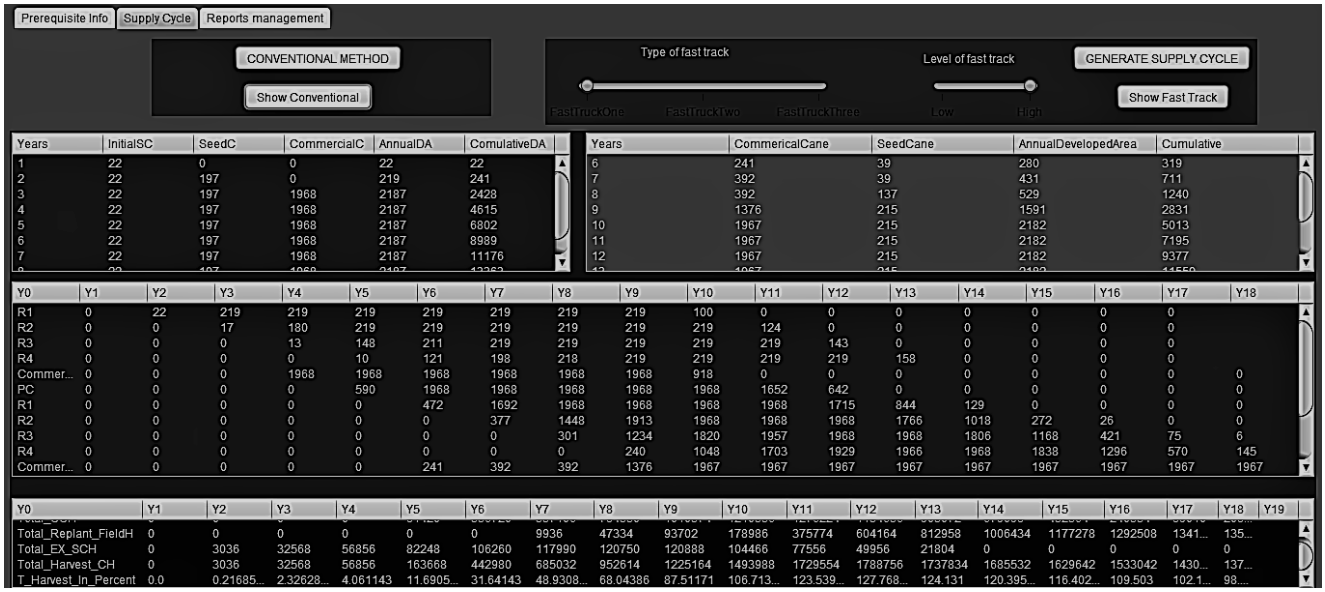

Figure 3: Supply Cycle Management

4. Discussion

Fuzzy logic has been most useful in generating fast track method sugarcane supply cycle. It has been used to determine the different types of fast track supply cycles with specific plantation area information but varying the annual developed area of plantation. In the process of conducting this research, the intent was to incorporate fuzzy logic to the system in place so that it could be more intelligent and reduce the hideous drudgery the subject matter experts go through. It is found that most current professionals who are familiar with the concept of sugarcane supply cycle use manual and complicated MS Excel file to do their job. This is the main

motivation to design and develop such software which employs fuzzy logic to address the challenge of uncertainty.

The prototype can generate sugarcane supply cycle and has an integrated fuzzy logic for the generation of a fast track method supply cycle generation. As it was mentioned above, there are two main parts of cane supply cycle management, establishment state and steady state management. From this two cycle managements, we have used establishment part of the management in this research.. In order to come up with the cane supply cycle there are two distinct methods, conventional and fast track methods. We have developed the prototype to accommodate both methods and report a printable result. While generating the supply cycle the system calculates the yearly cutting schedule by comparing with the given annual yield value, the yearly cutting schedules are generated according to the cutting yield level, average age of PC + Ratoon, design capacity of the mill, etc.

Evidently, fuzzy logic model applicability is for the fast track method cane supply cycle. In the generation of a fast track method the user only has to select the type of fast track method (Fast track one, two, or three) and the ceiling of the selected fast track method. The fuzzy logic model will convert this human understandable linguistic but not computable values of information to computable forms and generate the suitable cycle.

5. Conclusion and Future Work

In order to construct a feasible linguistics for the fuzzy logic as well as to come up with the crisp values of those linguistic values, subject matter experts' opinion was pivotal. Nevertheless the testers of the prototype were those people who provided the information needed. The results from the tests conducted were promising. Testing is not a one-time event; it had to be an iterative process to validate the usability of the prototype. In the testing process, there were two domain experts involved who were closely working in this research. After using the application in the presence of the developer, their perception towards the system was found to be satisfactory. Although the prototype can generate seven cycles, they were very much interested in the application of fuzzy logic. Especially in the process of finding the most optimum cane supply cycle and widening the scope of the application are the most essential ideas that could improve the acceptability of the system. Subsequently, usability of the prototype could be improved by making it more responsive to the user's action.

Concluding from the result gathered from the reaction of the subject matter experts it has been an irrepressible fact that fuzzy logic incorporated software application for sugar cane supply cycle is a powerful artifact that most definitely will support the process of developing a profitable sugar factory.

For future there are two main directions that could be exploited:

- to improve establishment phase by proposing a way to get the optimum supply cycle plan, and
- to make the prototype complete and fulfledged with the steady state operation of the supply cycle.

References

[1] Personal Interview & Reply to questioners with Senior Sugar Estate development Consultants at AgriPol International Consultancy & Training Plc.

[2] Roel Bosma, Uzay Kaymak, Jan van den Berg, Henk Udo, and Johan Verreth, "Using Fuzzy Logic Modelling to Simulate Farmers' Decision-Making on Diversification and Integration in the Mekong Delta, Vietnam", Vietnam, 2010.

[3] Harpreet Singh, Madan M. Gupta, Thomas Meitzler, Zeng-Guang Hou, Kum Kum Garg, Ashu M. G. Solo, and Lotfi. A. Zadeh, "Real-Life Applications of Fuzzy Logic, 2013.

[4] R. P. Humbert, "The Growing of Sugarcane", Revised Edition, Elsevier Publishing Company, London, 1968.

[5] R. Julien, "An Evaluation of Methods used for Maturity Testing", International Society of Sugarcane Technologists, Proceeding ISST 15:991-999, 1974.

[6] AgriPol International Consultancy & Training PLC, "Detail Project Report (DPR) for the Establishment of Green-field Sugar Project – Tana-Beles 1, 2 & 3 Sugar Projects", Feasibility Study, Addis Ababa, 2013.

[7] Hussein Fakhry, "A Fuzzy Logic Based Decision Support System for Business Situation Assessment and E-Business Models Selection", University of Dubai, UAE, 2010.

[8] L. A. Zadeh, "Fuzzy Logic", Department of Electrical Engineering and Electronics Research Laboratory, University of California, Berkeley, 1965.

Developing Blockchain Powered Financial Inclusion Model for the Unbanked Population of Ethiopia

Wossen Alemeye
HiLCoE, Computer Science Programme, Ethiopia
wosalem@gmail.com

Tibebe Beshah
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

Financial inclusion refers to the delivery of affordable and usable financial access for unbanked and underbanked people. According to a World Bank report [6], there are approximately 2 billion individuals who lack financial access and an additional 1.5 billion individuals who are underserved by the financial service industry. Out of the total 2 billion, 2.1% (i.e., approximately 42 million adults) are in Ethiopia. This number is too big to ignore. This paper tries to solve some of the barriers to financial inclusion mainly lack of interoperability, high cost and high-risk. Its main objective is developing financial inclusion model for the unbanked population of Ethiopia using distributed ledger technology. Design Science Research Methodology (DSRM) has been used throughout the study. As a proof-of-concept, a prototype was developed using open source blockchain software. The prototype was evaluated by experts with relevant knowhow and experience.

Keywords: Financial Inclusion; Digital Wallet; Digital Currency; Blockchain

1. Introduction

Ethiopia is one of the highly unbanked nations. Since the liberalization of the banking industry in 1996, there has been continuous increase in the number of banked population in Ethiopia. However, we should not be complacent of this achievement as still millions are unbanked. According to a World Bank report in 2014 [6], about 41 million adults in Ethiopia are unbanked. According to [11], one of the primary impediments to providing financial services to the poor through branches and other bank-based delivery channels is the high costs inherent in these traditional banking methods. Research made by Bill & Melinda Gates foundation [11] explicates that using confidential cost and revenue estimations provided by three service providers in Africa, one in Asia, and three in Latin America, it is found out that agent banking does improve the economics for these institutions compared with branches, especially for high-transaction, low-balance accounts that are common among poor users. Though Mobile and Agency Banking in Ethiopia have been contributing a lot to financial inclusion, there are still many barriers [2, 3, 4]. Lack of interoperability, high cost and risk are the main interests of this paper. High mobile phone penetration rate was found out to be the most important potential for the development of mobile banking in Ethiopia [4]. Gallo *et al*. [1] assert that a financial transaction that is highly valued by the unbanked and underbanked is money transfer, whether within a country (i.e., internal payments) or across national borders (i.e., cross border payments). In recent years two important trends are fueling the payment industry: mobile payments and the rise of the blockchain. Thus this paper focuses on developing blockchain-powered financial inclusion model for the unbanked population of Ethiopia.

In line with this, the main objectives of this paper are to:

- explore the potential of blockchain,
- explore how to develop best blockchain-based model for financial inclusion,
- develop Blockchain Powered Mobile and Agency Banking Model, and

- develop and evaluate a prototype using open source blockchain software.

2. Background and Related Work

2.1 Background

Banks and Micro Finance Institutions have been trying to reach the unbanked population of Ethiopia through branches and branchless banking. Henok Arega [4] states that unlike some successful attempts which use Telco-led mobile and agency banking service such as MPesa of Kenya, Ethiopia's mobile and agency banking has not been much successful mainly because of interoperability issue as Ethiopia uses bank-led model. Samman [15] describes that globally 38% of adults remain unbanked. Yet among the survey respondents who do not have an account, only 4% said that the only reason for not having one is that they do not need one. The major reasons include lack of identity, regulations, no banks, cash instead of digital payments and high costs and fees. According to [10], many banks and financial entities envision the blockchain as delivering value to their customers in the form of reduced fees, faster funds transfers and simplified processes. The same research further explains that much less heralded, but potentially far more dramatic, is the transformational impact blockchain could have on the world's massive unbanked population. Among many other uses, blockchain could bolster these efforts by becoming the backbone to open the closed-loop mobile money services. Right now, many of the mobile banking services offered by banks and micro finance institutions are not interoperable. But blockchain could expand interoperability to link these fragmented, closed loop services both domestically and internationally. Swan [9] predicted that blockchain could be the Fifth Disruptive Computing Paradigm following mainframe, PC, the Internet and mobile phone and social networking.

2.2 Blockchain Technology

At a very high level, blockchain is a decentralized ledger, or list, of all transactions across a peer-to-peer network. This is the technology underlying Bitcoin and other crypto currencies, and it has the potential to disrupt a wide variety of business processes [8]. If the Internet is the foundation for digital innovation of all kinds, blockchain technology is the underpinning of a radical rethinking of how we pay for things—as well as how we verify who owns what and who has the right to buy and sell it [7]. Five components that make up blockchain are: blockchain database, cryptographic tokens, peer-to-peer (P2P) network, consensus formation algorithm and machines/nodes and programming language.

The creator of BitCoin Nakamoto [5] defines blockchain electronic coin as a chain of digital signatures. Each owner transfers the coin to the next by digitally signing a hash of the previous transaction and the public key of the next owner and adding these to the end of the coin. A payee can verify the signatures to verify the chain of ownership.

2.3 Related Work

To the best of our knowledge, no research on blockchain technology has been undertaken in Ethiopia. Voorhies *et al*. [13] conducted a research with the objective of fighting poverty profitably by transforming the economics of payments to build sustainable inclusive financial systems. They stated that blockchain can play a crucial role in fighting poverty profitably. Baruri [7] investigated how financial institutions (FIs) are using blockchain to foster financial inclusion. The paper identified four ways FIs are using blockchain to foster financial inclusion: blockchain-powered economic identity, inclusion through blockchain-powered remittance service, blockchain-powered services for refuges/migrants and blockchain-powered digital identity for citizens in poverty. Gallo *et al.* [1] mentioned two important findings from their research on blockchain First, they suggested that when regulating blockchain and virtual currencies, it may be to the benefit of regulators to engage with financial institutions early and be part of the innovation process, tailor Know Your Customer (KYC) to the risk profile of the unbanked to facilitate

adoption, and apply principled based approaches given blockchain's ever changing nature. Second, FIs using blockchain for internal and cross border payments can lower the costs, shorten settlement time and provide a better user experience.

Probably one of the thorough researches made is by Bill & Melinda Gates Foundation. This research [12] developed what they called the "Level One Project Guide" with the objective to create a more level-playing field to establishing national digital financial services system which benefits everyone – the poor, bankers, mobile operators, payment technology companies, the government and more. The research details important requirements to meet the objective mentioned before. The research suggests that blockchain can be used to move digital assets of all kinds, including fiat currencies. This could be used as an alternative to current transaction switching and settlement models. In theory, it could also be used to lower the costs of providing these functions. The authors found out that blockchain fulfills many of the criteria mentioned in the key findings of the Level One Project Guide. Moreover, Ethiopia can benefit from latecomers' advantage as it hasn't invested a lot on its payment system.

3. Methodology and Approach

The methods employed in this study are literature review and questionnaire followed by interviews. These methods are used to understand the interest of FIs in Ethiopia, in particular commercial banks and MFIs, in regards to blockchain-powered financial inclusion. Also, how the solution is viewed in the eyes of regulators such as National Bank of Ethiopia (NBE) is studied. Moreover, the view of technology service providers was taken into account. Purposeful sampling was used as it is required to reach respondents to address specific purposes related to the research questions. Design science research approach was adopted. Thus important findings from literature review, questionnaires and interviews were taken into account in designing the proposed model.

4. Data Collection, Analysis and Model Design

4.1 Data Collection and Analysis

We distributed questionnaires followed by interviews to 9 commercial banks, two MFIs, one technology service provider and the National Bank of Ethiopia. The respondents gave important ideas based on their knowledge and experience. However, we also noticed that the technology is new to most of them and challenging to some of the respondents.

4.2 Data Analysis and Model Design: Expected Functionalities

Consistent with the recommendations in [12], our findings indicate important features. In general blockchain-based systems can be permissioned or permissionless. Our findings show that the system to be designed be permissioned rather than permissionless. Most respondents chose permissioned blockchains considering the current regulation in the country. However, we found out that there are more compelling reasons as explained in [14] such as limited capacity. For instance BitCoin (a permissionless blockchain) currently supports around 300,000 transactions per day compared to the Visa network which currently handles 150 million transactions per day in the USA. The result of questionnaires/interviews and literature review also point out that the system to be designed be open-loop, be capable of immediate funds transfer, effect push rather than pull payments, enable same-day/real-time settlement, adhere to open international standards, have irrevocable transactions (called immutability in blockchain), have shared fraud service, use tiered KYC and be governed by its direct participants. As to whether to use value-added or not-for-profit/cost recovery model, the cost recovery model was chosen by most respondents. It was also found out from interviews that the system should have regulatory support of tiered KYC and interoperability. Interoperability is currently a requirement by NBE.

According to NBE directive [17] only financial institutions that are licensed by NBE are allowed to

engage in mobile banking services. Hence, FIs shall form consortium for the permissioned blockchain. The same directive stipulates that FIs can carry out mobile banking through their agents. Therefore, we propose involvement of other interested parties (as agents) so that ubiquity can be achieved easily. The use of Ethiopian Postal Services offices as cash-in and cash-out outlets will be highly beneficial.

To achieve ubiquity for the unbanked, any FI in the consortium or its designated agents should be able to provide cash-in and cash-out services. Moreover, the system shall have one common name say "bitBirr System" and each FI can create branded digital wallets to use on the system. Any FI in the consortium should also be able to recruit agents.

Under this system, FIs will deposit cash as collateral with NBE in exchange for NBE-issued digital currency. However, to avoid inflation due to the injection of digital asset, equivalent flat currency shall be removed from circulation by NBE. NBE starts the genesis block and issues the digital asset bitBirr. FIs can then transfer bitBirrs from their accounts to the digital wallets of users and agents by accepting fiat currency. Agents also use digital wallets and give cash-in and cash-out service using the same network. The same wallet can be used to purchase items.

The following steps summarize the processes discussed above:

1. FIs form Consortium including NBE, agree on blockchain parameters and rules, install/configure blockchain, join the peer-to-peer network. NBE is granted all permissions but can be revoked later based on the agreement of the Consortium.
2. FIs deposit cash as collateral with NBE in exchange for NBE-issued digital currency – bitBirr - after agreed permissions are granted.
3. FIs recruit agents and transfer bitBirrs from their accounts to the digital wallets of agents by accepting fiat currency (Birr) after agents are granted agreed permissions.
4. FIs and Agents transfer bitBirrs from their accounts to the digital wallets of users by accepting fiat currency (Birr) after users are granted connect, send & receive permissions before receiving bitBirrs.
5. Users transfer money among themselves.
6. FIs and Agents transfer bitBirrs from their accounts to the digital wallets of merchants by accepting fiat currency (Birr) after agreed permissions are granted to merchants.
7. Users purchase items using their wallets.
8. Merchants transfer money among themselves.
9. Users Cash-in and Cash-out at FIs or Agents

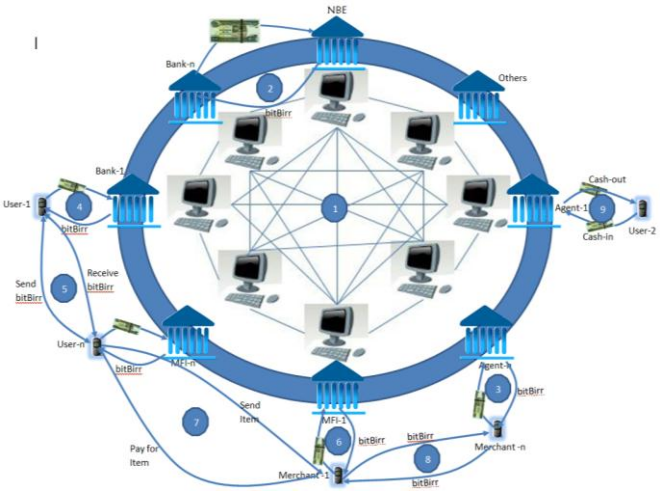

Figure 1: Processes in the bitBirr System

There are basically eight types of permissions in permissioned blockchain. Three of these (connect, send and receive) can be considered as low risk. Users who obtain them can only connect to the blockchain and transact on their own behalf. Three more (issue, create and activate) are medium risk, because users who obtain them can respectively issue assets (e.g., bitBirr), create streams, and manage other user's low risk permissions. Two final permissions (mine and admin) can be considered as high risk – nodes with mine permissions participate in the consensus algorithm, while those with admin permissions can modify all other permissions for users. The purpose of a blockchain is to create decentralized databases, which are not under the control of any individual party. Therefore, it is natural and necessary to extend this philosophy to the

administration of permissions. For all medium risk and high risk permissions, a blockchain can be configured so that individual administrators are unable to change those permissions on their own. Instead, a certain proportion of the administrators have to agree before the permissions change takes place, with the proportions defined by the admin-consensus-blockchain parameters such as 60% of the admins have to agree.

Ttable 1 is a summary of suggested permissions of each participant (responsibility matrix) in the blockchain network based on our findings.

Table 1: Responsibility Matrix in the Blockchain Network

Participant	*Permission*	*Remark*
NBE	Connect, send, receive, issue	Admin-consensus-* and other parameters such as mining diversity and mining turnover have to be agreed by participants.
FIs	Connect, send, receive, create, activate, mine, admin	
MFIs	Connect, send, receive, create, activate, admin	MFIs can also mine provided they have reliable Internet links. Otherwise, it will affect the consensus mechanism.
Agents	Connect, send, receive, activate	
Merchants	Connect, send, receive	
Users	Connect, send, receive	
Others	Connect	Others such as Ministry of Communication and Information Technology

Each user in the system will have private/public key pair as unique identification of the wallet (customer). However, it is also necessary to know who the owner of a certain public key is in the system. In the current situation of Ethiopia, there are no unique identifiers of customers such as national identification cards for KYC purposes. Therefore, mobile phone numbers are suggested to be used as main identifier of customers. But, there must be a way to map these telephone numbers to public key addresses of the wallets and to accounts (if available) on core banking systems. The public key may be used in text or QR code format. Table 2 shows example mapping of addresses:

Table 2: Mapping of Addresses

Public Key	*QR Code*	*Telephone Number*	*Account Number (if any)*
1Chq2vqES4LqiLjvNeQLt33CjafmYqJYQM		+251911225210	

There are many ways to store private keys safely [16]: desktop wallets, mobile wallets, online wallets, hardware wallets and paper wallets. Wallets store the private keys to access a public address and spend funds.

4.3 The Proposed Model

Based on the functional requirements identified using conceptual data from literature review and empirical data from questionnaires and interviews, the proposed model has five layers.

a. The Presentation Layer

The presentation layer is the interface where users interact with the system and have wallets which could be desktop, mobile, online, hardware or paper wallets. Antonopoulos [16] states that at a high level, a wallet is an application that serves as the primary user interface. The wallet controls access to a user's money, managing keys and addresses, tracking the balance, and creating and signing transactions. More narrowly, from a programmer's perspective, the word "wallet" refers to the data structure used to store and manage a user's keys. Based on the type of clients, wallets can be Simplified Payment Verification (SPV) or Full-Node. Participating FIs have to run Full-Node Wallets. Antonopoulos [16] further explains that although nodes in the bitcoin P2P network are equal, they may take on different roles depending on the functionality they are supporting. A bitcoin node is a collection of functions: routing, the

blockchain database, mining, and wallet services. A full node contains all four of these functions. Therefore, a full node contains the functions at the top three layers. All nodes include the routing function to participate in the network and might include other functionality. All nodes validate and propagate transactions and blocks, and discover and maintain connections to peers.

b. The Services Layer

This layer contains services provided by the system such as send, receive, deposit and withdraw bitBirr. According to [1], a financial transaction that is highly valued by the unbanked/underbanked is money transfer, whether within a country or across national borders. Therefore, the model includes these services. As blockchain transactions are immutable, each service is realized by transferring ownership. For instance, if one owns 10 bitBirrs and wants to transfer 8 bitBirrs to another person then this is realized by sending 8 bitBirrs to the receiver and 2 bitBirrs to oneself. To deposit cash, one gives cash to an FI or Agent and the FI or Agent sends the same amount of bitBirr to ones wallet. For withdrawal, the reverse of the previous process is done. So, Send and Receive are the basic operations for all the services.

c. The Blockchain Layer

This layer is the core of the model and what differentiates the system from others. It is the core because most of the requirements from literature review as well as empirical data such as interoperability, low cost and low risk features are realized in this layer. It contains a tamper-proof database which essentially is list of all transactions organized into blocks from the beginning (genesis block). This is also the layer where transactions are secured using PKC and are summarized using Merkle Tree (Hashed Pointer Binary Tree). It is also the layer where transactions are verified (mined) using various consensus algorithms such as Proof of Work (PoW) and Proof of Stake (PoS).

d. The Integration Layer

This layer is used to integrate the blockchain layer to traditional database systems. It is the layer where blockchain accounts are mapped to core banking system accounts. Respondents of questionnaires/ interviews were asked whether it is crucial to integrate the DLT network with core-banking system of financial institutions or not and they somehow agree to it. This shows that integration of the blockchain with core banking system is an optional component for serving the unbanked.

e. The Database Layer

This is the layer which contains the core banking system database and traditional bank accounts. This is an optional component as availability of traditional bank accounts is not mandatory. However, since it is important that the unbanked are brought to the formal financial system in the course of time, this layer together with the Integration Layer will be highly beneficial in the future.

4.4 Prototype Development and Evaluation

Multichain, a permissioned/private blockchain platform, has been selected for developing a prototype. As blockchain is a decentralized P2P network, it is required to have more than one node to simulate such network. Therefore, VMware Workstation version 10 was selected to create multiple nodes and implement the model. Windows 7 was installed on each virtual machine. Note that the prototype focuses on the core blockchain functionality and not the presentation layer such as designing a mobile wallet for multichain. However, Multichain Web Demo was used to demonstrate the presentation layer.

The steps followed in the prototype development include: installing and configuring four virtual machines using VMWare and Windows 7, installing Python 3 and PHP 5, downloading and installing the multichain software, creating blockchain, issuing a digital asset called bitBirr, installing and configuring MultiChain Explorer and Multichain Web to manage nodes and wallets.

Four subject matter experts evaluated the prototype developed for the features listed below.

- Transferring assets among nodes.
- Check blockchain parameters such as whether the 'anyone-can-*' parameters are set to "False" as this is a permissioned blockchain.
- Consensus mechanism based on blockchain parameters: admin-consensus-admin, admin-consensus-activate, admin-consensus-mine, admin-consensus-issue or admin-consensus-create.
- Granting/revoking permissions.
- How the system behaves when nodes leave and join the network.
- Test round-robin mining in multichain.
- Test mining diversity and mining turnover.

The evaluators were posed to answer the following questions:

- Can you send and receive the asset –bitBirr? Are these transactions reflected on all available nodes?
- Are the balances reflected on each available node after sending bitBirr?
- What rights does each node has?
- Is the consensus mechanism working as expected?
- Does granting and revoking permissions work? Is the right number of nodes required for consensus?
- Can nodes leave and join the network without affecting the system?
- Do nodes mine/verify transactions in a round-robin fashion?
- Does mining diversity function as expected?
- Is it possible to modify transactions? How does the system respond?

At the beginning, the evaluators were given a brief overview of the system and the basic commands and procedures to test the prototype. After that, a laptop on which the prototype was installed and configured was given to each evaluator to see the features for themselves independently. They started the evaluation with the above questions but were given the freedom to test the prototype for any feature they deemed important. They all tested sending bitBirr and verified that the balances for each wallet are reflected on each node immediately.

In general, all of them verified that the prototype developed behaved as intended. However, they recommended testing it in a real environment using WAN connections instead of virtual machines and LAN environment to have a real feeling of the system.

4.5 Discussion

The purpose of this paper was to show the possibility of creating blockchain-powered interoperable mobile banking system without a trusted third party by a consortium of FIs for reaching the unbanked population of Ethiopia. The research tried to answer how blockchain technology can be used to create interoperable, low cost and low risk mobile and agency banking system in Ethiopia without an intermediary and thereby improve financial inclusion. With blockchain technology interoperability is achieved without a trusted third party which has positive impact in decreasing service costs. Also as the system uses hashes to create shared, tamper-proof ledger, it was shown that it has low risk and is secure. The use of PKC makes the system more secure as only the owner of a private key can transact. The service will be ubiquitous which highly improves service availability to the unbanked. The paper also tried to develop blockchain-powered model for inclusive financing. A prototype was developed to demonstrate how the model can be implemented by the consortium of FIs. The prototype was evaluated by experts and the results show it is a promising system for inclusive financing.

5. Conclusion and Future Work

This paper started with the objective of developing blockchain powered financial inclusion model for reaching the unbanked population of Ethiopia. The model proposed in this paper solves most of the barriers in reaching the unbanked: lack of interoperability, high-cost and high-risk.

Since blockchain-powered system uses shared, tamper-proof ledger, it is automatically interoperable

and reduces risk and fraud. The system is secure as PKC is used. Most existing mobile banking systems in Ethiopia use SMS and send transactions in clear text and hence will suffer security breaches sooner or later. In addition, there is no single point of failure in the blockchain system as the ledger is replicated across multiple nodes. Nodes can also leave and join the network without affecting the system.

The proposed model eliminates trusted third parties. This has a positive effect in reducing costs. However, this doesn't make third parties out of the game as the blockchain system requires continuous maintenance and update. Since settlements are done in real-time, it allows the most efficient use of capital thereby lowering the amount of capital required.

SPV Wallet along with SIM overlay technology were recommended so that people with feature phones can use the system. However, the detailed implementations of overlay SIM for mobile wallets have not been covered in this research. Therefore, it is recommended that these features are handled in the future. The use of overlay SIM requires further research. In addition, how to map public keys to mobile phone numbers and the integration of the blockchain layer with the database layer requires further research.

References

[1] C`harles Gallo, Anna Jumamil, and Pak Aranyawat, "Blockchain and Financial Inclusion: The role blockchain technology can play in accelerating financial inclusion", http://finpolicy.georgetown.edu/sites/finpolicy.georgetown.edu/files/Blockchain_and_Financial_Inclusion.pdf, Last accessed on March 29, 2017.

[2] Elfagid Aregahegne, "The Challenges and Prospects of Mobile and Agent Banking in Ethiopia", 2015.

[3] Afework Gugsa, "Assessment of Adoption of Agency Banking Innovation in Ethiopia: Barriers and Drivers", Addis Ababa University, 2015.

[4] Henok Arega, "Financial Inclusion through Mobile Banking: Challenges and Prospects", IJSRST, 2015.

[5] Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", 2008.

[6] World Bank Group, "The Global Findex Database 2014", 2015.

[7] Pani Baruri, "Blockchain Powered Financial Inclusion", Cognizant, 2016.

[8] PwC FinTech, "What is blockchain?", https://www.pwc.lu/en/fintech/docs/pwc-fintech-qa-what-is-blockchain.pdf, Last Accessed on April 3, 2017.

[9] Melanie Swan, "Blockchain –Blueprint for a New Economy", O'Reilly, USA, 2015.

[10] American Banker, "Blockchain can Bring the Unbanked into the Global Economy", https://www.americanbanker.com/opinion/blockchain-can-bring-the-unbanked-into-the-global-economy, Last Accessed, May 15, 2017.

[11] Global Savings Forum, "How Agent Banking Changes the Economics of Small Accounts", https://docs.gatesfoundation.org/Documents/agent-banking.pdf, Last Accessed on May 22, 2017.

[12] Bill & Melinda Gates Foundation, "The Level One Project Guide: Designing a New System for Financial Inclusion", https://leveloneproject.org/wp-content/uploads/2015/04/The-Level-One-Project-Guide-Designing-a-New-System-for-Financial-Inclusion1.pdf, Last Accessed on June 21, 2017.

[13] Rodger Voorhies, Jason Lamb, and Megan Oxman, "Fighting poverty, profitably: Transforming the economics of payments to build sustainable, inclusive financial systems", https://docs.gatesfoundation.org/Documents/Fighting%20Poverty%20Profitably%20Full%20Report.pdf, Last Accessed on July 4, 2017.

[14] Gideon Greenspan, "MultiChain Private Blockchain -White Paper", https://www.multichain.com/download/MultiChain-White-Paper.pdf , Last Accessed on Aug 31, 2017.

[15] George Samuel Samman, "Blockchain and the Unbanked: The Road to Financial Inclusion MultiChain Private Blockchain", https://www.slideshare.net/gsamman/blockchain-and-the-unbanked-the-road-to-financial-inclusion, Last Accessed on Aug 31, 2017.

[16] Andreas M. Antonopoulos, "Mastering Bitcoin - Programming the Open Blockchain", O'Reilly, USA, 2017.

[17] EthNews, "Proof-of-Work vs. Proof-of-Stake Explained", https://www.ethnews.com/proof-of-work-vs-proof-of-stake-explained, Last Accessed on Nov 30, 2017.

Improving Android's Device Security using Behavioral Biometrics

Yonatan Mekonnen
HiLCoE, Software Engineering Programme, Ethiopia
yoniecco@gmail.com

Gezahegne Dugassa
HiLCoE, Ethiopia
gdugassa@gmail.com

Abstract

Mobile phone technology dominantly controls our daily life whether for solemn business matter or for our social life through which we access the whole universe just in a tap away. Thus, the power of connectivity using the Internet through GSM (Global System for Mobile)/LTE (Latest Technology Extended) enhanced its usage in a way not having one can be considered as handicapped. Losing our device, leaving the device unattended and exposure of our password/pattern can make us not only lose our reputation in our social network presence but also might be a risk allowing others to have control over our activities by the features we use through applications in our phones.

Hence, authentication becomes one of the core issues in security as one is privileged to do any activities and tasks so long as s/he provides the credentials and is authenticated regardless of the real identity. Relying on the default screen-lock option that Android devices have is common. The problem here is that this screen-lock authentication feature is breakable and the device can be unlocked easily.

Thus, the proposed solution in this paper is designed to act as a layer for the device's security improving confidentiality of its content when the screen lock feature is compromised due to various reasons. Our application will manage to work on behavioral authentication by studying and analyzing the characteristics of the owner of the apparatus with a combination of signature and character usage scheme. This paper adopted the concept of machine learning from natural language processing so that an intelligent application can secure the device's content from unauthorized access.

Keywords: Authentication; Security; Mobile Apparatus; Screen Lock

1. Introduction

1.1 Background

Mobile devices usually contain information like the contact list which were intended to be kept by ourselves up to confirmation and verification messages that will enable us change passwords for our official/personal emails, mobile banking applications and social media accounts. Though we use the applied "Screen-Lock" option by Android, bypassing it is actually as easy as plugging the locked device into a computer that runs applications which can remove the Screen-Lock in just a click [1].

The irony here is that this Screen-Lock removing software are available in different options and feature containers with options like simple designs to a very complex one for free. Therefore, it is no more a wonder to see one's device being "Unlocked". Android actually has shown improvement in their versions for security but it is not as strong as the potential that exist working against it. Android is open-source and the Software Development Kit (SDK) it uses allows third-party applications and developers to gain access deep into the Unix-kernel through rooting the device or even by just asking for permission upon installation [2].

As stated in [1], because of Android's powerful development framework, users as well as software developers are able to create their own applications for wide range of devices. Some of the key features of Android operating system are: Application Framework, Dalvik virtual machine, Integrated browser, Optimized Graphics, SQLite (Server Query Language), Media Support, GSM Technology, Bluetooth, Edge (Enhanced Data for GSM

Evolution), 3G (third Generation), WiFi, Camera, and GPS (Global Positioning System). To help the developers for better software development, Android provides SDK.

The intention we have here is as to how can we improve the device security when the authentication that Android implemented gets breached by designing an application to act as another layer of security so that we at least can maintain Confidentiality.

The following research questions are formulated:

1. What is Android's Screen-Lock authenticator?
2. How does Android's Screen-Lock authentication work?
3. How strong is Android's Screen-Lock authenticator and what are its limitations?
4. How can we improve the limitations?

This paper puts a focus mainly on understanding and analyzing the gap and limitation that the default Android Screen-Lock authentication has. The solution that is proposed will be applied only for devices that run Android operating system which allows integration of third-party applications.

The proposed solution in this paper will equip the end user with the following.

- A cost-effective and user-friendly security improvement mechanism.
- Secure the device so that it will be safer to save data and use applications.
- Avoid the requirement of additional hardware for authentication.
- Save the hassle of fighting to secure the content of the device.

1.2 Methodology

Experimental research method is recommended by most scholars as it is highly effective on problems that require complex software solutions such as the creation of software development environments, the organization of data that is not tabular, or the construction of tools to solve constrained optimization problems.

While applying it to our specific problem area, the method was used to extract flaws of Android's screen-lock authenticator and assess the possibility of improvement. Also, the method was used to test the sample taken with the algorithms proposed and choose one from the three. As this approach is largely used to identify concepts that facilitate solutions to a problem and then evaluate the solutions through the construction of prototype systems, it was found convenient and applicable with better throughput than other research methods.

Thus, we used qualitative research method enabling us to fulfill our objectives, use philosophical assumptions for understanding how people make sense of their worlds and the experiences people have along with knowing or understanding from the participants' perspectives on what is needed to be addressed. It also has enabled us to address what is needed to be given a key focus for understanding, rather than predicting or controlling, social settings or social phenomena while intending to address the problem specified.

Opportunity sampling was used to select the test subjects as this sampling method uses people from the target population available at the time and willing to take part. However, for the evaluation of the experiment, we used stratified sampling as we identified the different types of people that make up the target population and works out the proportions needed for the sample to be representative. Since gathering such a sample would be extremely time consuming and difficult to do, we recreated the samples by generating conditions using a simple randomization algorithm.

We added random sampling to strengthen the evaluation using Weka and MS Excel along with simple randomization algorithm so that we can measure cost, efficiency, accuracy and more detailed performance related aspects. The advantage is that our sample will represent the target population and eliminate sampling bias caused by opportunity sampling even though it requires more time and effort.

As natural language processing is applied for our application, we gave emphasis for the following 6 key attributes in which our decision will be determined by.

1. TP = true positives: number of examples predicted positive that are actually positive
2. FP = false positives: number of examples predicted positive that are actually negative
3. TN = true negatives: number of examples predicted negative that are actually negative
4. FN = false negatives: number of examples predicted negative that are actually positive.
5. Accuracy: the degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard.
6. Precision: refinement in a measurement, calculation, or specification, especially as represented by the number of digits given.

We then examined and got a data of 31 records for each of the 4 conditions in our equation specified as Signature (S), Typing Behavior (T), Favorite A/V (A), and Frequently Accessed apps (F). We then populated a data of 6000 for each condition with total evaluation data of 6031. Out of the 6031, we only fed 40 true (valid) conditions using a logical operator AND to our CSV file where the A cells contain Signature, the B is for Typing Behavior, C contains Favorite Audio/Video files and D is for Frequently accessed applications as follows: AND(A2>=94, A2<=100, B2>=85, B2<=100, C2>=80, C2<=100, D2>=90, D2<=100).

2. Related Work

Application markets such as Apple's App Store and Google's Android Market provide point and click access to hundreds of thousands of paid and free applications. Markets streamline software marketing, installation, and update therein creating low barriers to bring applications to market, and even lower barriers for users to obtain and use them. The fluidity of the markets also presents enormous security challenges. Rapidly developed and deployed applications [3], coarse permission systems [4], privacy invading behaviors [5, 6, 7], malware [8, 9, 10], and limited security models have led to exploitable phones and applications. Although users seemingly desire it, markets are not in a position to provide security in more than a superficial way [11]. The lack of a common definition for security and the volume of applications ensure that some malicious, questionable, and vulnerable applications will find their way to the market [9].

Biometric authentication technologies are used for machine identification of individuals. The human-generated patterns used may be primarily physiological or behavioral, but usually contain elements of both components. Examples include voice, handwriting, face, eye and fingerprint identification. The search is to find a uniquely identifying characteristic not of what we are, but of what we do. An example is gait analyze someone's walking style and one can easily determine their identity. However, that is not going to work on a Smartphone. But if we consider a person's signature that is the only security layer in banking. It can be analyzed since exact handwriting style is unique to everyone. It is possible that devices could analyze the speed, style and exact position on the screen of how one signs a name, probably using a stylus [12].

There are two types of signature features that are used for signature verification: Functions and Parameters.

1. Function features are used in terms of time functions whose values constitute the feature set. Parameter features are used in terms of a vector of elements, each one representative of the value of a feature. Function features allow better performance than parameters but require higher system cost for matching [13].
2. Parameter features are classified into two categories: Global and Local. Global parameters concern features that are related to whole signature, i.e., total writing time, number of pen strokes in the signature and total written distance. Local features are like

global features but retrieved only from the specific part.

Depending on the level of details, local parameters can be divided into two categories: component-oriented and point (pixel) oriented.

1. Component oriented parameters are extracted at the level of component, i.e., stroke height, width ratio and relative position of signature parts.
2. Point oriented are extracted at levels of point (pixel) resolution, i.e., pixel density and the total count of the stylus points. This feature classification is not strict. Some features can be used globally and locally.

The following are some of the most common function and parameter features found in the literature [14].

- Position: X and Y position of pen tip
- Speed: Overall speed of writing or in X and Y direction
- Acceleration: Overall acceleration or in X and Y direction
- Pressure: pressure on which pen leaves to the signing area
- Direction of pen movement: direction of pen speed vector

Our application will use offline verification though the attributes position, speed and acceleration are widely used for online signature verification. But for our application, velocity and acceleration function can be obtained from a specific device or numerically derived from position. Pen inclination is considered as a most consistent feature. However, it can be captured only by some specific devices. Finally, we will take a percentage for the match after comparison is made and will use it for our later equation as a variable S.

Text Prediction method of authentication model will help us in design level as our proposed solution will learn the behavior of one's way of word usage in such a way that a relation can be made between words. After a sensible set of data is developed in the application database, the application will generate or re-use a phrase as a question to be asked during authentication as it will determine the most likely response. Behavioral biometrics will then be applied by calculating how close the similarity is with some tolerable deviation in a percentage scale.

In its most basic form, keyboard prediction uses text that one enters over time to build a custom, local "dictionary" of words and phrases that are typed repeatedly. It then "scores" those words by the probability one uses or needs it again. The first or second time one ignores the word, it will assume it is not a misspelling, but not a word one uses often enough to be presented within similar usage patterns. If one ignores it a third or fourth time (how many times depends on the specific keyboard), the keyboard will mark it as a future probable choice, and start presenting with it when one types similar words or sentences [14].

3. The Proposed Solution

For a developer of Android applications, the SDK will work in a standard and convenient way so long as application permission wizard is designed and configured properly for the SDK to comply with it. As a developer, we are able to design an application which replaces any existing system applications including the pre-installed application like SMS client application, default-dialler, the web browser and contact manager.

The favour we have benefited from Android is that there is no such thing as a compulsory application. Also there is equality among applications and all applications can run in parallel as per their pre-defined priority. Since Android operating system is open source and allows various permissions, an Android application developer can:

1. Integrate a new application along with an existing one.
2. Extend an application by adding some new features in the new one.
3. Replace an existing application with a new one but with similar functionalities.

The good thing about Android's operating system is that a third-party application can be designed in such a way that it can take a full control over the device. This is to mean that if a given application is designed with some specific permission settings, i.e., manifestation for the SDK, then that application will be considered as a trusted application and can run in the same place that the operating system executes.

Therefore, we have designed an application that can act as a layer between the applications and the user. This layer contains the following functions.

1. Hides confirmation messages sent to the device by scanning keywords and phrases that are contained as main structures of confirmation messages.
2. Hides applications that are sensitive and used for authentication like Google's one-time password generator.
3. Initiates device activities remotely including encryption, shred-data, factory-reset and wipe data through USSD (Unstructured Supplementary Service Data) and SMS commands.
4. Sends SMS containing new SIM (Subscriber Identity Module) status to the actual owner of the device once another SIM is inserted into the device by capturing the new SIM-Card's ICCD (Intensified Charge-Coupled Device) and service number.
5. Forces log out any accounts that are logged in when the device detects security breach.
6. Clears cache memories for applications by having control over the application manager module.
7. Re-authenticates the user after a breach is detected using behavioral authentication.

4. Discussion

So far, we have investigated and analyzed the core and basics of Android operating system. By doing so, we are able to assure that the problem does actually exist and many users are being affected by it. For instance, if we take one-day statistical data from ethio telecom and Facebook joint monitoring dashboard portal, approximately 30% of all Facebook users use Android phones on average per day. This is huge number as the "all" refers to those who use the desktop web browser and other mobile operating systems like iOS (iPhone Operating System). This also implies that at the minimum 1000 people per day that ethio telecom gives a SIM card replacement, the same percentage are Android based devices. As our motive mainly relies on a scenario in which a user's device falls in the wrong hand regardless of the cause, the Android based devices will be vulnerable if intended for easy unlock of the Screen-Lock Authenticator.

The proposed application is implemented having the following components for authentication as our application will have 4 authentication attributes:

1. The hand-written signature noted as S
2. Typing behavior noted as T
3. Favorite audio/video file noted as A
4. Frequently accessed application noted as F

While using our application, the user will be asked to select authentication preferences from the above four in a check-list. The user will be forced to select a minimum of two authenticators from the four provided that signature will be considered as a prerequisite since it is the core attribute for our application. Then from the chosen ones, the application will create an arithmetic percentage and then will base its decision on the deviation from the acceptable result. Thus, our authentication equation noted as A will be: A = (R ≤ (%M+(%O)))/O+1 and is mathematically represented as.

$$A = R \leq \frac{(\%M + (\%O))}{O + 1}$$

where:

- A is our application authentication as whole,
- R is Threshold in which is validated to be authenticated or not,

- M is a constant mandatory attribute chosen for authentication, and
- O is an optional variable attribute or attribute chosen for authentication.

Then, deviation will be calculated by defining a range from which a minimum and a maximum can be registered. For instance, in the case of signature verification, the 5 pre-signed signatures will be taken as a benchmark and then any signature signed after that will be compared with those five and the difference and variance it has towards the others will be calculated in percent. Then if the difference is greater than 10%, then the newly signed signature will be rejected and the authentication will be considered as a failure. We have defined a range after we apply experiments using sample subjects to study their behavior and signature. The range can then be used for decision in which a result can deviate from the 100% that we have derived for the 4 conditions earlier as per Table 1.

Table 1: Range Definition

Attribute	*Min*	*Max*	*Deviation*
Signature (S)	94	100	6
Typing Behavior (T)	85	100	15
Favorite A/V (A)	80	100	20
Frequently Accessed apps (F)	90	100	10

By doing so, we can verify if the person being authenticated is actually the real user or not. For this to be applied, we can implement it using the code fragment shown below for a scenario in which the user has chosen all the 3 optional authentication parameters where all the "X" variables represent the minimum threshold deviation and all "Y" variables represent the maximum threshold deviation.

Finally, the rule X = ((S=TRUE) AND (T=TRUE) AND (A=TRUE) AND (F=TRUE)) will be applied for our proposed application as a validation control for decision.

After we have the structure above, we implement it using the following algorithm that can be used in any programming language.

```
function Authenticate($Signature, $Typing_Behaviour, $Favourite_AV, $Frequently_Accessed)
   {
       if($Signature && $Typing_Behaviour && $Favourite_AV && $Frequently_Accessed)
       {
           if(($Signature >= X) && (($Signature <= Y))
           {
               if(($Typing_Behaviour >= X) && (($Typing_Behaviour <= Y))
               {
                   if($Favourite_AV >= X) && (($Favourite_AV <= Y))
                           {
    if($Frequently_Accessed >= X) && (($Frequently_Accessed <= Y))
    {
                           return TRUE;
                       }
                   }
               }
           }
       }
   }
```

Out of 8 available algorithms, three were chosen as they fit most for this experimental research. We first tested the data for evaluation using the classifier rules-part. Rule-based machine learning (RBML) is a term in computer science intended to encompass any machine learning method that identifies, learns, or evolves 'rules' to store, manipulate or apply [6, 7, 8]. As our proposed application derived a rule for decision, it implies that the rules typically take the form of an {IF: THEN} expression, (e.g., {IF 'condition' THEN 'result'}, or as a more specific example, {IF 'Signature matched' AND 'favorite application chosen right' THEN 'grant access'}). We then summarize the important elements in Table 2 so

that we can make a decision by which algorithm we should be working with.

Table 2: Decision Table for Algorithm Selection

Algorithm (Classifier)	*True Positive Rate*	*False Positive Rate*	*Precision*	*Recall*	*Confusion Matrix*	*Response Time (in Seconds)*
Rules PART	0.999	0.05	0.999	0.999	Apr-31	0.11
J48 graft	1	0	1	1	0/6031	0.08
Naïve Bayes	0.997	0.422	0.997	0.997	18/6031	0.03

As a result, we have chosen J48 graft classifier as it granted us with 100% precision, true positive, and recall with 0% false positive assuring us effectiveness. It has slower response time compared to Naïve Bayes, but a slower yet accurate response is better than a faster but compromised one.

5. Conclusions and Future Work

This paper designed a solution as an additional layer for device security for Android running devices to improve the mobile device's security to maintain privacy and content confidentiality. By applying behavioral authentication options, including signature, one's typing behavior and the individual's interest for accessing applications and media, the content residing in the device can get additional security. The proposed application for device authentication will not only secure the privacy of the device content but also will allow the end user to feel more secured. Though this paper cannot grant option to re-gain access to a lost device, it can minimize the hassle the user has to go through to re-gain access to an account if a third party has control over the lost device, i.e., a stolen/lost mobile apparatus.

The proposed solution is a security layer by which addition authentication will be applied. But still, the effectiveness of this application will highly depend on the user's willingness to use the application, the knowledge of the user on basic application usage and the combination of authentication options that the user chooses from the application. If a case study focusing on awareness creation is done, end users can manage the level of vulnerability to which they are exposed.

Future work will mainly focus on integrating the proposed solution with artificial intelligence and neural-biology along with face recognition algorithms so that a device independent yet effective application can be developed.

This solution can borrow and adhere logics from psychology and neural biology so that the level of intelligence can be enhanced by advancing its decision-making schema. In addition, the efficiency of the chosen algorithm might vary based on the data input, thus a better experiment involving different test subjects will be more accurate.

References

[1] "Android Application Security with OWASP Mobile Top 10 2014", Luleå University of Technology, 2014.

[2] William Enck, "A Study of Android Application Security and Internet Infrastructure Security", Pennsylvania State University.

[3] University of Minnesota, "Classification Methods", http://www.d.umn.edu/~padhy005/Chapter5.html, 2015.

[4] E. George, "On-line Handwritten Signature Verification Using Wavelets nd Back Propagation Neural Networks," Sixth International Conference on Document Analysis and Recognition, 2001.

[5] R. Abbas, "A Prototype System for Offline Signature Verification Using Multilayered Feed Forward Neural Networks", Michigan, 1996.

[6] https://source.android.com/security/, Last. Accessed on 13/06/2017.

[7] C. Zhong-Hua Quan, "Hybrid HMM/ANN Based Approach for Online Signature Verification," International Joint Conference on Neural Networks, IJCNN 2007.

[8] L. Z. Xia, "Efficient Object Recognition Using Boundary Representation and Wavelet Neural Network," 4th International Conference On ASIC, 2001.

[9] M. G. Kresimir Delac, "A Survey of Biometric Recognition Methods," 46th International Symposium Electronics in Marine, ELMAR-2004, 16-18 June 2004.

[10] Z. S. Zhang, "Association Rule Mining: Models and Algorithms," 2002.

[11] R. N. Totty, "Study of the Variation in the Dimensions of Genuine Signatures," Journal of Forensic Science Society, Vol. 25, pp. 207-215, 1985.

[12] https://www.android.com/, Last Accessed on 12/06/2017.

[13] G. W. Bassel, E. Glaab, J. Marquez, M. J. Holdsworth, and J. Bacardit, "Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets," The Plant Cell, 23 (9): 3101–3116.

[14] Hovemeyer, D. and Pugh, W., "Finding Bugs is Easy", In Proceedings of the ACM Conference on Object-Oriented Programming Systems, Languages, and Applications, 2004.

HiLCoE

School of Computer Science and Technology

HiLCoE, one of the first higher education institutions in Ethiopia, was established in July 1997. It is a specialized Computer Science and Technology College that adheres to computing and quality of education. It is now widely recognized that HiLCoE prepares students to become industry and academic leaders who can apply technology and computer science principles across a wide variety of fields.

HiLCoE, since its inception has a vision of being Center of Excellence in ICT Education, Research and Development (ER&D). To realize this vision, in addition to its role in academia, it offers cutting-edge consultancy in the areas of Information Systems Security, Enterprise Architecture, Strategic and Solution Architecture, Service Engineering, Software Development, Project Management, Distributed Systems and Mobile Applications and many more.

Opportunities are tremendous with attractive reward for every effort you put to know and implement the evergreen field of Computer Science and Technology. Your dreams become realized by choice to join this pioneer and **Center of Excellence** in ICT ER&D in parallel with cutting-edge consultancy.

A commitment to excellence, coupled with continuous improvement, will result in HiLCoE being recognized as innovative, dynamic, and exciting community to learn, teach, and work with.

HiLCoE offers:

Three undergraduate programmes:

- **BSc Degree in Computer Science** and
- **BSc Degree in Information Systems** (in progress).
- **BSc Degree in Software Engineering** (in progress).

Three postgraduate programmes:

- **MSc Degree in Computer Science**,
- **MSc Degree in Software Engineering**, and
- **MSc Degree in Information System Security** (in progress).

HiLCoE Computer Systems Engineering, from its industry link, offers **specialized tailored training and consultancy** in:

- Enterprise Architecture,
- Strategic IT Architecture,
- Information Systems Security, Policy and Architecture,
- Project Management,
- Systems Development,
- Cloud Computing, and
- Mobile Application and Computer Network Design and Implementation.

HiLCoE paves a fabulous highway that makes you capable to be partner and contributor to the development arena.