University of Fort Hare

Together in Excellence

Development of an automatic news summarizer for isiXhosa language.

By

NDYALIVANA ZUKILE

Thesis submitted to the Department of Computer Science in partial fulfilment of the requirements for the Degree Master of Science

Supervised by: Dr Zelalem S Shibeshi

Co-Supervised by: Professor. C Botha

Department of Computer Science

Submitted: April 2018

DECLARATION

I, Zukile Ndyalivana (Student Number: 201007385), the undersigned, thus pronounce that the work contained in this paper is my particular unique work. It has not already been submitted to any institution for a comparative or whatever another degree. I Acknowledged information extracted from different sources.

Signature…… ……..

Date:19-04-2018.....

ACKNOWLEDGEMENTS

In the first place and for most, I might want to send my true because of God almighty for helping me complete this work.

I might want to sincere my genuine appreciation to my supervisor Dr. Zelalem Shibeshi for inspiration, advice and the information he gave me all through this two-year period. Those weeks after week meetings have molded myself to being a dedicated I am today.
I would like to thank my co-supervisor Professor C. Botha for the tremendous support and advice that he gave to make this research success.

I would likewise like to express gratitude towards Telkom Center of Excellence (Coe) and National Research Foundation (NRF) for the financial augment they provided for my study at the University of Fort Hare.

I want to thank the Head of Department (HOD) Mr. S. Scott for allowing me to do my research in the Department of computer science in the Univeristy of Fort Hare.
I send my sincere appreciation to Mr. Zengethwa, who significantly helped in the formation of manual extracts for this study. Without his work this research would have been eccentric.
I, last but not list send my sincere gratitude and appreciation to my family for believing in me ever since I embarked on the research journey.
I want to thank every one of the members who helped during the evaluation period of this study.

Finally, I want to thank everybody who has supported whichever way or the other during this study. I am much indebted to all of you.

DEDICATION

I dedicate this work to my whole family i.e my Grandmother, my Father, my Mother, Brothers, Cousins, and Aunts, and not to forget to mntion my late Uncle, my partner, son, and nephews. They had confidence in me, giving expressions of guidance. All the motivation that all of you you gave throughout the years is greatly apreciated.

ACRONYMS	
NLP	Natural Language Processing
ATS	Automatic Text Summarization
TF	Term Frequency
IDF	Inverse Document Frequency
ICT	Information and Communication Technology
PRM	Pronominal Resolution Module
HMM	Hidden Markov Models
IBM	International Business Machines
NLTK	Natural Language Toolkit
IDE	Integrated Development Environment
SABC	South African Broadcasting Corporation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

ABSTRACT

From practice perspective, given the abundance of digital content nowadays, coming up with a technological solution that sumnmarizes written text without losing its message, coherence and cohesion of ideas is highly essential. The technology saves time for readers as well as gives them a chance to focus on the contents that matter most.

This is one of the reseach areas in natural language proccesing/ information retrieval, which the dessertation tries to contribute to. It tries to contextualize tools and technoologies that are developed for other lamnguages to automatically summarize textual Xhosa news articles. Specifically, the dissetation aims at devloping a text summarizer for textaul Xhosa news articles based on the extraction methods.

In doing so, it examines the litrature and understand the techniques and technologies used to analyze contents of a written text, transform and synthesize it, the phonolgy and morphology of the Xhosa language, and finally, designs, implememnts and test an extraction-based automatic newa article for the Xhosa language. Given comprehension and relevance of the litrature review, the research design , the methods and tools and technologies used to design, implement and test the pilot system.

Two approaches were used to extract relevant sentences, which are, term frequency and sentence position. The Xhosa summarizer is evaluated using a test set. This study has employed both subjective and objective evaluation methods. The results of both methods are satisfactory.
Keywords: Xhosa, Automatic Text Summarization, Term Frequency and Sentence Position.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF LISTINGS

INTRODUCTION

1. INTRODUCTION AND BACKGROUND

1.0 Overview

This dissertation traces the insights made during exploratory research work completed on the subject of making summaries of Xhosa news articles. This chapter highlights the topic by giving a context for the research and by explaining the motivation for initiating the project. Having examined the results of the exploration, the primary objectives of the research are set out, together with its scope and conclusions.

1.1. Automatic Text Summarization(ATS)

The volume of information available for users of the Internet has been increasing on a daily basis. In this, the information age, the growth of electronic information has necessitated intensive research in the area of Natural Language Processing (NLP) and Information Retrieval (IR). The fast growth of information has made it difficult for many users to cope with all the text that potentially is of interest to them. As a result, systems that can automatically summarize one or more documents, have become the focus of interest recently, in the field of automatic summarization [1]. Automatic text summarization has become a suitable tool for assisting people in the task of reading large volumes of textual information.

Examples of summaries that users choose are: news headlines, scientific abstracts, minutes of meetings, and weather forecasts. These are all kinds of summaries people enjoy reading on a daily basis [2].

A summary can help users to get the meaning of a complete text document within a short time. The following are some of the general reasons that support the necessity of text summarization.

- Summarization improves document indexing efficiency
- A Summary or abstract saves reading time
- A Machine generated summary is free from bias
- A Summary or an abstract facilitates document selection and literature searches.

1.2. Motivation

There has been much work done for many languages throughout the world with regard to automatic summarization. The Xhosa language is on the rise in terms of electronic content, especially online. As a result, there is a need for language processing applications for the Xhosa

language, and an automatic text summarizer is one such application that is required by the local community. A large number of people and organizations can benefit from this application by obtaining the most relevant content within the minimum period. As a result of this need, there is a strong motivation to design a Xhosa Text Summarizer.

1.3. The Problem Statement and Justification of the Study

The rate at which information has grown electronically has made it difficult for users to obtain important information in the shortest possible time. In other words, users are highly affected by information overload. Information overload also leads users to read unnecessary details and waste their time. Most of the time they read documents that they are not even interested in, in too much detail. People who speak different languages around the globe are facing this problem, and this is true for people who speak the Xhosa language.

These days many agencies produce a large amount of textual information specifically for Xhosa speakers. These include media bodies such as the South African Broadcasting Corporation (SABC) that broadcasts the news in the Xhosa language, online news publishers and national newspaper publishers. The range of documents goes beyond that of just the publishing of news, it includes reports from government offices, especially from the Provincial Legislatures of the Eastern and Western Cape Provinces. Most of these items are of more than five paragraphs, which in this busy world is not appropriate for users.

As mentioned above, newspapers and other news releases in the Xhosa language reach readers from many sources. It is evident that it is of paramount importance to read news items to keep abreast of what is happening in the world. However, because of the busy lives that people lead and their day-to-day activities, there is frequently insufficient time to read entire documents resulting in important information being missed.

Automatic Text Summarization (ATS) systems are still scarce, especially for the Xhosa language. It would be advantageous if the agencies, some of which host daily shows, that release news in the Xhosa language could have a tool that can provide users with a short version of the original text.

Hence, there is a real need to design a tool that would deal with the problem of information overload, especially in the domain of news in the Xhosa language.

This study presents a possible solution to this particular problem. This work also makes a contribution towards developing Natural Language Processing applications, which can be used by Xhosa native speakers. This work investigates text summarization applications for Xhosa, to increase the scope of the text summarization research for this language.

1.4. Research Questions

How can a computerized system select key sentences?

How can language-based rules be used to extract the salient sentences and reduce content in the text?

1.5. Objectives of the Study

The aim of this study is to investigate, implement, and evaluate a Xhosa text summarizer. The name of the summarizer is IsiXhoSum, which is based on the methods and algorithms put forward by H. P. Luhn[5] and H. P. Edmundson [14]. Changes to language-specific rules to support the Xhosa language were developed and implemented.

The general objectives of this research are: to understand the structure of Xhosa news text, to investigate how a Xhosa text summarizer can be created that can extract relevant sentences from a written text document, and to present these as a readable summary. The aim is to discover an approach that does not require too many linguistic resources but that gives an acceptable result.

The unique methods developed when automatic summarization started did not exploit a significant part of the semantic mechanisms like parsers, annotated corpora and so on. This work aims at concentrating on the primary methods that do not require many semantic resources.

Up to the present time, no work has been done to discover how the elements of the Xhosa language perform in a summarization context. As a significant aspect of the development of an automatic summarizer, this work puts a focus on determining how key information is disseminated throughout a document. Thus, looking at how Xhosa news items are composed using sections or paragraphs, is of particular interest in this research.

1.5.1. Specific Objectives

To understand the field of automatic text summarization and to analyze the related research.

To investigate existing methods and algorithms used to develop extraction-based automatic summarization.

To develop a prototype Xhosa news summarizer that will serve as a model for a full scale/ fully operational Xhosa text summarizer.

To develop a test set to evaluate the system.

To draw conclusions from the results of the research.

To suggest future work on how the system could be used for other documents written in the Xhosa language.

1.6. Significance of the Study

This work will design and develop a complete Xhosa text summarizer. The summarizer will be applicable in different regions in Southern Africa especially Xhosa speaking people and foreign people that want to learn isiXhosa as language. This research has significance in that it could start additional studies in the field of automatic text summarization for Xhosa and other Southern African languages.

1.7. Research Methodology

Before initiating the study, a research methodology has to be chosen. This choice is influenced by reviewing existing literature in the field. The steps to be followed are:

1.8. Literature Review

To achieve the objectives detailed in Section 1.5, an extensive literature review on automatic text summarization was made. This review looked at relevant published documents, materials on the Internet, books, and journal articles, to get a deeper knowledge of the nature of the Xhosa language and of the structure of a document. Chapter two of this thesis contains the literature review.

1.9. Data Source Collection and Preperation

In order to create a corpus, data must be collected. The data used in this study is collected from different sources. This data should be clean and be in readable format. The corpus does not come with clean and readable characters and because of that, the corpus has to be cleaned, and non-characters are excluded. The corpus is cleaned so that only the necessary letters remain in the documents. More information on how the data was collected and prepared to form a Xhosa news corpus is explained in Section 5.5.1 of this work.

1.9.1. Corpus Preparation

This study involves the preparation of a corpus (a collection of written or spoken material stored on a computer and used to find out how language is used) of texts that was used for analysis. These texts were collected from various websites, which publish news using the Xhosa language. These sources are available online and a complete explanation of the how the corpus was created is provided in this study.

1.9.2. Manual Summary Preparation

A portion of the corpus was taken to a linguist to create extracts manually. These texts were randomly select from the corpus. Each text had to be at least two to three paragraphs to be selected. The linguist used his expertise to extract the key sentences in each text. Section 5.5.2 explains in detail how the manual summary was created.

1.10. Summarization Method and Tools used in this Study

The method used in this study is extraction based automatic summarization. Using the extractive approach, salient sentences are extracted from the document and displayed for the user. There is no need for summary regeneration (the rewriting of sentences to form the summary) when the extraction method is used. Sentences are weighted based on the cue phrases the sentences contain, the location of the sentences, and those sentences containing the most frequently used words (term frequency) in the text document. Sentences with the highest weights are kept. Then using the efficient combinations of extraction features, the most important sentences are selected to form a summary.

1.10.1. Development Tools

The Xhosa Text Summarizer is built using the python programming language. There are prerequisites to observe, these includes the readiness of its environments, and settling on a working framework to use. The subsections that follow explain the components that have been put together. The next subsection starts by explaining the Natural Language Toolkit (NTLK) in detail and how it is used in this work.

1.10.2. The Natural Language Toolkit (NLTK)

NLTK is a collection of various language-processing modules and has been developed as an open source library. It is intended to give massive support to a variety of disciplines like researchers, students of empirical linguistics, science, artificial information retrieval, and machine learning in Natural Language Processing [44]. It offers functions and wrappers that are very convenient. The wrappers and functions use building block for common NLP tasks effectively. The toolkit has built-in versions of raw and pre-processed corpora that are mentioned and used in a wide range of NLP literature and courses. This library was used in this study because it gives access to text analysis tools via its toolkit. The toolkit has a variety of built-in tokenizers and statistically based modules for text analysis. It has a wide range of text collections in corpora, for the variety of languages.

1.10.3. The NLTK Installation

The Natural Language Toolkit is a python package and requires the following versions of python: 2.6, 2.7, and 3.7 +. The package installs as a file or as setup when downloaded. For this project, the installation that was chosen is the nltk-3.0.4.win32.exe (md5) installed on a 32-bit windows operating system.

1.10.4. Installing the NLTK data

Installing the NLTK data is a separate package that can be installed without any of the installations outlined in 4.2.1.2. NLKT data comes in the form of a folder where there are subfolders of chunkers, corpora (raw and annotated), stop words, models stemmers, and tokenizer. To make sure that the entire package is downloaded and installed it is first to launch the python interpreter and type a command. After that, the Python interpreter is run as follows:

```
✓ >>> import NLTK
✓ >>> NLTK. Download ()
```

When the command above has been typed and run, a window appears with NLTK downloader. In this downloader, it is possible to select all the files or just a selected number of files. For this study, all the files were selected and installed in the machine directory. In the machine, the complete folder has been located on the Local Disk (C/users/Zukile/AppData/Roaming/nltk_data. In the following subsections, some of the modules that were used in the NLTK are explained.

1.10.5. The NLTK Tokenization

Tokenization as the process of splitting a sentence into its constituent tokens. For segmented languages such as English, the existence of whitespace makes tokenization relatively easy [45]. However, for languages such as Chinese and Arabic, the task is more difficult since there are no explicit boundaries [45]. Xhosa is a Latin based language; it uses the same letters that the English language uses. So tokenizing Xhosa text is as simple as tokenizing in English. The NLTK module was used for the tokenization of the Xhosa text in this study.

1.10.6. The NTLK Corpora

As stated in [45], NLTK comes with several useful text corpora. Already loaded. (NLTK) Despite that fact that there are regular content words, there is also another class of words called stop words, that perform important grammatical functions, but are unlikely to be interesting by themselves, such as prepositions and complementizers. This class has been modified to handle

Xhosa stop words. NLTK comes bundled with the Stop Words Corpus - a list of 2400 stop words across 11 different languages (including English).

1.10.7. The NTLK Corpus Reader Class

NLTK's *corpus reader* classes are used to access the contents of a diverse set of corpora. Each corpus reader class is specialized to handle a specific corpus format [46]. Examples include the *PlaintextCorpusReader*. This class handles corpora that consist of a set of unannotated text files. In addition, the *nltk.corpus* package automatically creates a set of corpus reader instances used to access the corpora in the NLTK data. Each corpus uses a "corpus reader" object from `nltk.corpus`.

Each corpus reader provides a variety of methods to read data from the corpus, depending on the format of the corpus. For example, plaintext corpora support methods to read the corpus of raw text, a list of words, a list of sentences, or a list of paragraphs. The researchers have used the plain text corpus reader to create and read Xhosa text file.

1.10.8. Operating System

The operating system that is utilized is the 32-bit Microsoft windows 8. The NLTK supports different types of operating systems including Windows, Linux, and Macintosh. This work is based on the Natural Language Toolkit, which is written by python. Python has various packages and wants broadened libraries: details one of the libraries will be clarified in section 1.10.10.

1.10.9. The Python Programming Language

Python supports multiple programming paradigms, including object-oriented, imperative, and functional programming or procedural styles. A python 2.7 version was selected and installed. This release supports the NLTK library. The NLTK library also requires the installation of another library called Numpy, which is optional but necessary when instaling python installations.

1.10.10. The Numpy Library

Numpy is an open source library that has a very large collection of modules. Numpy or Numeric Python is a python extension, which supports a wide variety of the major multi-dimensional arrays including matrices [47]. Numpy is equipped with high-level mathematical functions to operate these arrays. It is compulsory to install Numpy in order to use python, as it will give all the functionality of built-in arrays and other modules. This library builds on the original code base, however, it combines feature that have already been created by the Num-array, which also includes added features [48]. When python and Numpy are installed, there follow the NLTK installations.

1.10.11. Charm Integrated Development Environment (IDE)

PyCharm is a smart code editor that provides first-class support for Python, JavaScript, CoffeeScript, TypeScript, CSS, (popular template languages). It takes advantage of language-aware code completion, error detection, and on-the-fly code fixes.

In this study PyCharm IDE has been used to create a user interface in the development of the Xhosa Text Summarizer. It also has some extensive libraries to make the graphical user interface.

1.11. Evaluation Methods

The intrinsic method is an assessment that focusses on the summary itself; it attempts to measure its cohesion, coherence, and in-formativeness. Usually, a comparison is made with another summary of the same text. In this study, the evaluation of the summary is made both subjectively and objectively. This, however, is not the only evaluation method; some authors use subjective evaluation only.

When considering the subjective evaluation, the evaluators (native speakers) look at closely some aspects. They look at the linguistic qualities, such as in-formativeness, and how coherent the summary is. They evaluate the summary in terms of the relevancy of all the summaries. Making use of some pre-defined guidelines, evaluators will allocate a score, using a predefined scale, to each summary that is under evaluation. The evaluators assign quantitative scores to the summaries based on a range of different qualitative features, like content, fluency, etc.

To evaluate the system objectively, a special tool called ROUGE2.0, was installed and configured for the requirements of Xhosa news summaries.

1.12. Scope and Limitations of the Study

This study specifically focuses on the development of an automatic text summarizer for Xhosa news articles. The emphasis is on news articles only. In fact, it is a single document summarization project. Therefore, the scope of this research is limited to apply technologies applicable to languages that do not require the sophistcated langauge based rule, to find out the most suitable factors for achieving precise summaries automatically.

1.13. Outline of the Dissertation

This dissertation comprises several chapters. Chapter one being the introductory chapter, it gives the background, motivation, the problem statement, objectives, significance of the study, methodology and evaluation methods. It also discusses the scope of the research.

CHAPTER TWO presents the literature review undertaken to discover the status of Automatic Text Summarization. It not only gives a profound technical background but also provides a concise overview of different summarization approaches and the resources used over the last decades to automatically summarize the text.

CHAPTER THREE talks about the Xhosa language discusses the consonant and vowel inventory of the language. It talks about the language's orthography, morpheme types, abbreviations, and also the structure of news writing.

CHAPTER FOUR describes the methodology and discusses the tool, the Natural Language Toolkit (NLTK), used to design the summarizer. The modules of the toolkit that were considered are explained in detail in this chapter.

CHAPTER FIVE talks about the Implementation of the Xhosa Text Summarizer. It also shows the interface between the system and the modules used in the Xhosa Text Summarizer. This chapter finishes by explaining how the summarizer was tested.

concludes the study. It starts by giving a research summary. This is followed by a presentation of the conclusions that were drawn from the research. Lastly, it presents ideas for the future work that needs to be carried out, following on from the current work, to produce a fully-fledged tool.

CHAPTER TWO

1.1. LITERATURE REVIEW

2.0 Introduction

The advancements in the Information and Communication Technology (ICT) has resulted in an increase to the production, collection, organization, storage and the dissemination of information which accordingly result in the so-called information overload.

Currently, various technologies used to produce information made it possible for users to access information in multiple formats, multiple sources, and single sources. They also come in single and in multiple languages. This means we need tools to cope with this information explosion. Summarization is an important tool to help us to keep up to date with what is happening in the world. Summarization primarily condenses textual information from one source or more sources and presents it to the reader in short format.

In general, summarization has many uses. The following are how we use summarization in our everyday life Pachantouris G et al [3]:

- Headlines of the news
- Table of contents of a magazine
- Preview of a movie
- Abstract summary of a scientific paper
- Review of a book
- Highlights of a meeting

The last sections of this chapter describe processes, the basic concepts, types, methods and methods of automatic text summarization.

2.1. Automatic Text Summarization

Dinegde, G. D. et al [4] in 2014 defined Automatic Text Summarization (ATS) as the task of taking one or more text document (s) as an input and reduce it in an attempt to produce condensed format (i.e. summary) from the original text document. The most important thing in summarization is to be able to find the most important sentences, and this involves knowing the semantics of written or spoken document(s). Also being able to write a concise and fluent summary needs the capacity to reorganize, transform, and join information expressed in different sentences as input.

A complete interpretation of document(s), followed by the creation of abstracts, is most of the time the most difficult task for people to perform. This has been a difficult task in the area of automatic text summarization. Moreover, the main objective of automatic text summarization is to make a reduction of the compound and lengthy text while the relevancy and content of it are verbatim [5].

H. P. Luhn [5] states that though automatic text summarization is possible and viable, the extractive methods as they are given focus and interest take the attention of researchers to one significant question: How can a system determine the sentences that are most relevant in any text document?

In the past years, the field of Natural Language Processing (NLP) has experienced more advances in the sophistication of machine learning and language processing that ascertains the significance of sentences. The authors in [6] state that the task of determining the information that is imperative to include in the summary has a lot to do with a variety of factors, which are the genre and the nature of the source text.

2.2. Processes of Automatic Text Summarization

R. Boguraev [7] in 2009 described automatic text summation as the process having three phases:

- Analysis of a text
- Transforming the analyzed text
- The synthesis of the output text

Analysis of the text involves the identification of the content to be able to make an internal representation. This may involve the implementation of statistical methods in order to extract the key content and even complex methods that involve deeper natural language processing methods. Statistical methods select salient terms of contextual sentences that make them and connect them to form a summary. Other methods require a complete understanding of the source so that at the end a summary is constructed.

Transformation is changing of text from extraction or abstraction. The transformation step is likely to make some cleaning and conforming to the incoming data to gain accurate data which is correct, complete, consistent, and unambiguous etc. For extraction summaries, the central topics identified in the previous step are forwarded to the next step for further processing. For abstract summaries however, a process of interpretation is performed. This process includes merging or fusing related topics into more general ones, removing redundancies, etc.

The last phase called the synthesis of the output text takes the summary representation and produces a suitable summary, which precisely corresponds to the requirements of the users. This final step in the process deals with the organization of the content. The following sub section reflects on summarization parameter.

2.2.1. Summarization Parameters

The use of text summarizers varies depending upon the user's needs as well as its application. This, therefore, means that there are important things that one needs to take into account when making the design of text summarizer. Various types of summaries are defined based on deferent scenarios. Some of the scenarios include:

- Nature of input text that must be summarized
- Purpose of the summary
- And the output of the summary

Several recent studies [56],[59],[60] have stated that there are two different types of summary called *User-Focused* and *Generic* are defined for this purpose. User-focused is custom-made to the requirements of specific user or group of users .This means that the needs of users are well thought out when developing the summarizer. The user query and background knowledge of the subject is most important factor for user-focused summaries. Generic summaries, alternatively, aim at a wide-ranging readership community.

Another important way to look at summaries is in terms of the difference between *Indicative* and *Informative* summaries. Borko, H., & Bernier, C. L. , H [8] in 1975 stated that based on the content of text document to be summarized, the content can be either an informative or an indicative summary. An *informative summary* is meant to represent (and often replace) the original document. In view of that, it must contain entirely the appropriate information needed to deliver the structural information. The focus of an *indicative summary* is to suggest the contents of the article without taking away details on the substance of the article. It helps to attract the user into reading the full document. User-focused summaries had gained wide popularity, because of their ability to capture the user's requirements and their interests.

Summarization systems can be viewed in two ways: single text document and multiple text documents. A single text document only takes one documents as an input. In the case of multiple text documents, more than one text documents are taken as input [9], [10],[11].

There are two ways of viewing text summaries, Eduard Hovy et al [12] in 1997 stated the difference between Extracts and Abstracts. An extract involves selecting important sentences as they are in the original text and put them in summary whereas an abstract includes breaking the text down making a number of various key ideas, merging of certain ideas in order to obtain more ones than that are general, and creation of new sentences different from the original text(s). However, in abstraction, the focus should be more on semantic meaning and cohesion. This method produces new sentences that are completely not from the source text. The sentences are put together from the existing content.

A typical example would be the phrase "She ate an *orange*, *peach*, and *apple*." In trying to bring about a more concise form of the sentence, we would get the following summarized phrase as "She ate fruit." This abstract wants to produce a more general concept 'fruit', two or more topics, orange; peach and apple joined. Implementation of abstract methods necessitates symbolic world knowledge which is by far the most difficult to obtain on a large scale to provide a summarization.

An extract is created by picking up certain sentences or phrases accurately from the original text to form a summary [2]. An extract is a collection of meaningful sentences in a document, reproduced verbatim [13].

However, extraction methods have been the point of focus in the area of automatic text summarization, but then again there is an issue of cohesion and balance [7]. While the use of extraction method has been the point of focus for many researchers, many challenges arise when creating such an extractive summary. These challenges are:

- How to select an important sentence from a long text.
- Creating a summary that is coherent.
- Redundancy of terms in the summary.

Extraction method has its usefulness and up until now, it is a workable method. Extracting sentences from a text with the statistical keyword approach often brings the problem of cohesion (explained later in the linguistic concepts section). To be able to improve the quality of the result i.e. the summary, some methods are usually combined with other methods. The following section explains such methods.

2.2.2. Methods of Summarization

One of the most important concepts in automatic text summarization is the decision on the use of an appropriate method to create a summary. Many researchers have used numerous extraction

features and weighing methods to find ways of creating a fluent summary. Summarization systems use a number of methods that are independent components. These methods include positional methods, cue and phrase methods, query, Word, and Phrase Frequency Methods.

- **Sentence Positional Methods**

Mentioned before, the way a title, sentence, paragraph are positioned in the document have some great deal of significance. This happened most of the times in the way newspapers are written for instance the first sentence in the first paragraph of the paper portrays a significant meaning. It has the first priority when making a summary[19].

- **Cue Word or Phrase Method**

In some certain genres, words such as significant, conclusion have some level of importance. It is in the sense that these words are prioritized, and they should be extracted. H.P Edmundson [14] in his work has used three types of cue words that he put for experimentation: 783 words which were called bonus words (The author said that these words positively affect the relevance of the sentence) e.g. “Greatest,” “Incidentally” and “Significant ,” 73 of the phrase were called stigma words because they negatively affect the relevance of the sentence for example “hardly,” “impossible” and “inadequate”.193 were just null words which were also termed as irrelevant. H. P. Edmundson [14] then computed a cue weight for each sentence. The cue weight is the summation of the individual cue word in the phrase. Same method was also adopted by The scholars in [15] which they applied to their study, and reported that this approach was their best method. This is based on the 64 % of joint precision and recall that they got. What they did was a collection of manually build scientific text of cue phrases from a specific domain. Subsequently, they rated each cue phrase for relevance to the text unit by allocating a so-called ‘goodness score’ which ranges from one to three.

- **Query Method**

R. M. Alguliev et al [16] in 2007 stated that a query method is used to query text based summarization systems. Given the text document, all (the sentences are scored based on the frequency in the text document). In addition, those sentences that carry the query phrases get the higher scores while sentences with single query phrase get lower scores. Those sentences with the highest score make it into the summary with their structured context included. Most importantly, these portions of text are taken out from various section and subsections. This says what is in the summary is the collection extracts. The number of sentences extracted highly relies on the summary.

- **Word and Phrase Frequency Method**

The scholar H. P. Luhn [5] uses the Law provided by Zipf, which is the law of word distribution. The Zipf''s states that:

Few words that occur very often

Fewer words that occur somewhat often,

Many words that occur infrequently

In order to develop the following extraction benchmark: if there are unusual words in the text, then there is the likelihood of those sentences in the text being necessary. The systems created by the scholars in [5],[14],[15] put into use various frequency measures, and make a report based on the performance that is between 15 percent and 35 percent recall and precision.

- **Title Method**

This method is close to a query method except the fact that here the interest is only in the words that are on the titles and headings. In the work provided by H. P. Edmundson [14], there was a combination of word and phrase method. In his method, each title word is assigned the same score and then a sum of text unit is made. According to the author in [15], the score is the mean frequency of title word occurrences in the sentences.

- **Machine Learning Methods**

With the robust rise of machine learning methods in the field of Natural Language Processing in 1990, researchers have written and published many papers that incorporate an infinite number of statistical methods to create documents extracts.

Despite the fact, many systems have based their reliance on the feature independence; some have based their reliance on the Naïve Bays methods. Some systems based their reliance on the choice of relevant features, and some focused more on the learning algorithms to reduce the assumption of independence.

Other researchers have also considered using models like Hidden Markov model and Log-Linear Models, neural networks and to enhance extractive summarization significantly.

- **Naive Bayes Approach**

Julian Kupiec et al [17] described a method that was trained and was able to learn from data. This approach was derived from work provided H. P. Edmundson [14]. The authors in [17] used a

classification function to classify the worthiness of the sentence and this is done using the naïve-Bayes classifier.

- S1 be a particular sentence,
- S be set of sentences that make up the summary,
- Moreover, F1, F2 …, Fk are the features.

The following formula assumes independence of features:

$$P(s \in sF2, \ldots \ldots Fk) = \frac{\prod_i^k 1P(FiS), P(s \in S)}{\prod_i^K 1P(Fi)}$$

➢ Sentence Position Method

The scholars in [18] deliberate the significance of single features called the sentence position. The method is to weigh a sentence by its position in the text. The authors named it the "position method." This rose from the idea that text generally follows a predictable discourse structure. Additionally, the sentences of greater topic centrality tend to occur in certain specifiable locations (e.g. title, abstracts, etc.). The authors argued that since the discourse structure significantly varies over domains, the positioning method could not be as naively as in [19].

➢ Hidden Markov Models (HMM)

An approach, which describes that a given set of feature computes a posterior probability that treats each sentence as summary sentence. The HMM has fewer assumptions of independence, in particular, it does not assume that the probability that a sentence *i* in the summary is independent of whether sentence i -1 is included in the summary.

There are five features used for the development of the HMM. The features used in the HMM (built into the state structure of the HMM) are position of the sentence in the document, position of the sentence in the paragraph, number of terms in the sentence and the probability of terms given the document.

➢ Neural Networks and Third Party features

The author in [52] used the neural networks to train the system to learn the types of sentences that should be included in the summary. The network is with sentences in several test paragraphs. In this method, the neural network is trained with sentences that are located in several paragraphs where each sentence is identified whether it should be included in the summary or not. A human reader does this.

- **Combination of different techniques**

Some of the methods researchers combine are the ones outlined above. Researchers have also discovered that no single method can outdo well in terms of scoring the text more than the way human extracts are created.

However, combining various methods requires different evidence of various sources. Additionally, the incorporation of the different methods put into use seems to do well. This implies that there is finest approach that can do well alone. Julian Kupiec et al [17] developed a Bayesian classifier based on a principle that any sentence will be contained within in the final output i.e. a summary .This was possible provided that there is some certain feature such as paragraph position , cue phrase indicators , word frequency , upper-case words as well as a sentence length. Short sentences were generally excluded in this regard. The experimental results they obtained were that 33 % when paragraph position and 29 were from cue phrase indicator and when methods were combined 42 %.

On the other hand, the authors in [18] made a comparison of eighteen combination of the features an optimal combination was achieved. This was possible with an incorporation of machine learning algorithm. The combined features are the same as the ones presented above and few others that are indicating the prominence of names, dates, quantities pronouns, as well as quotes in the sentence. The method explained above all is the learned function. The term query method became the second best score achieved. The third best score (up to 20 percent length) attained correspondingly by using the word frequency, the lead method, as well as the naïve combination function. The scholars in [18] further says that summaries should not exceed 35 percent and they should not be shorter than 15 percent.

2.3. Linguistic Concepts to Consider

When people write text, they write to bring about a specific idea, concept, and event. In a sensible text, the text document does not just contain the bag of sentences that bear no meaning. The manner in which text is created is even, it has grammatical structure and meaning, or relevance is not the same. A text has a necessary component referred to as a semantic structure. It is natural that every specific text document revolves around a specific idea.

In an excellent presentation, the major idea can be presented and divided into subsidiary concept and ideas. Ideas connect collectively to bring about a broader picture. The topic should be flow. They should move in a proper manner in order to drive the reader to the general idea in any easy

and comprehensive way. This means that a text needs to reveal some higher level of coherency that will attract the reader to understand the general concept.

2.3.1. Coherence

In linguistics, coherence is the semantic integrity of a particular text and is an essential component in a well-composed text. An element provides a feeling that a text document is written in a logical manner. Coherence is a semantic structure of a text. Modeling coherence requires an interpretation of the text.

Coherence relationship, which is also called semantic relation, can be used to create a model of a text. The current relationships are an elaboration, cause, support, exemplification, contrast, and result. Classifying the relationship for sentences is also a very complex process. Typically, efforts regarding coherence analysis lead to trees where the nodes are regarded as text segments (paragraphs, sentences, and phrases) linked to these relationships.

Coherence structure of the text can be represented excellently using discourse structure and rhetorical parsing. The author Daniel Marcu [21] presents an efficient summarization system that makes use of models of coherence. Daniel Marcu [21] takes the help of cue phrases and name them as discourse markers. A tree-like model is formed by the local discourse structures they are the ones that form the global discourse structure of the text.

Coherence structure is a complex feature to deal with because it is necessary that there should be more knowledge than the information that could be obtained from the text. Therefore using both coherence and cohesion is crucial in understanding the dynamics of text.

2.3.2. Cohesion

Cohesion is better simpler than coherence; it helps to establish the discourse structure in a particular text. It is considered as a surface level feature. Coherence deals specifically with the entire semantic structure of the text, whereas cohesion, only deals with relationships among the peer units of the text. If there is a text document, cohesion ascertains whether a unit of text has a connection with the other units in the same text or not. Ruqaiya Hasan et al [22] state five types of cohesion relationship that are found in a text.

- **Conjunction**

Usage of conjunctive structures like ’and’ to present two pieces of evidence in a cohesive manner. An example, the sentence ’I have a cat, and his name is Felix’, two facts are connected with the conjunctive ’and’.

- **Reference**

The use of pronouns for entities for example ’Dr. Kenny lives in London. He is a doctor.’ the pronoun ’he’ in the second sentence refers to ’Dr. Kenny’ in the original sentence.

  - **Lexical Cohesion**

    The use of related words. For instance sentence ’Prince is the succeeding leader of the kingdom.’ ’Leader’ is a more common word for ’prince.'

  - **Substitution**

    Making use of the indefinite article for a noun. In the example ’As soon as John was given a cup of tea, Mary wanted one too.’ the word ’one’ denotes to the phrase ’cup of tea’.

  - **Ellipsis**

    Point towards a noun without reiterating. For example ’Do you have a car? No, I don’t ’, the word ’car’ is indirect without starting in the second sentence.

    From the above cohesion structures, lexical cohesion is the best definite and easiest to catch.

2.3.3. Lexical Cohesion

A lexicon is defined as a structured knowledge base keeping semantic data about a series of words Cohesion is based on the type of relation between units of a text document. These units are referred to as words and phrases. Lexical cohesion is known as a phrase or word in a text document that shows the semantic relationship. Creating lexical cohesion relies on ascertaining the semantic relationship between words or phrases.

2.4. News Writing Structure

Usually, when writing news, you write an account of what has been happening from one place to another. The news may also give information about many issues; these issues may include new projects, ongoing projects, initiatives, and discoveries. The writing of news aims at being able to respond to any basic questions about any certain event: which are who, where, which, why and what. “How” is always put at the beginning of an article .The way news are structured portray a

tone and its relative importance to its intended user. The author in [23] articulates that the major concern is with the structure of vocabulary and sentences.

The authors also state that News stories also contain at least one of the following essential characteristics relative to the intended audience: proximity, prominence, timeliness, human interest, oddity, or consequence. This form of a structure is an inverted pyramid. It refers to the decreasing information in the subsequent paragraphs in the news [23]. Discussing these characteristics further is not in the scope of this work.

Newspapers are generally adhered to an expository writing style. Expository writing is a type of writing where the purpose is to explain, inform, or even describe. As mentioned in [23], the purpose of expository writing is to explain and analyze information by presenting an idea, relevant evidence, and appropriate discussion.

Whatever structures a news item follows, the first sentence is the one that carries the most significant structural element of the story. The lead sentence is usually the first sentence, in some instances, the first two sentences become the lead sentences, and ideally, these sentences are 20 – 25 words in length [23].

2.5. Evaluation Methods used in Automatic Summarization

Although researchers are attempting to create real human replaceable summaries using computers over the last several years, the subject of evaluation is one area with an unresolved problem [61]. It is hard to define a better summary even based on the perception but at times; it is easier to state if a summary is poor or good.

There are two types of summary evaluation: extrinsic and intrinsic. An extrinsic method of evaluation is where the quality of the summary is judged on how well it helps a person performing other task such as information retrieval. An intrinsic evaluation is where humans judge the quality of summarization directly on an analysis of the auto-generated summary.

Comparing system output to some ideal summary was performed in works of [14],[21],[17]. To simplify evaluating extracts, Daniel Marcu [21] independently developed and automated method to create extracts corresponding to abstracts (ideal summary).

The other way to use intrinsic method is to have evaluators rate systems' summaries responsiveness and /or linguistic quality using some scale (readability, grammar, informativeness, fluency, coverage, redundancy) [24].

The method of comparing system output to some ideal summary was carried out in work provided by [14],[21], [17]. Daniel Marcu [21] independently developed an automated method to create extracts same as abstracts (ideal summary).

2.6. Discusion on Related works

The main objective of text summarization is to be able to identify the most importtant sentence in a docuemnent and eventually generate a a concise form of it. The work presented in Dinegde, G. D. et al [4] is a complete explanation of the abtratiction summarization that was caried out to create automatic summaries. Abstraction summarization requires deep semantic and understanding of the at a greater depth. This means deep linguistic feature was be thouroughly studied. On the other side, extraction summarization summaries are created by picking up certain sentences or phrases accurately from the original text to form a summary [2]. In this method, sentences are extracted verbatim [13] from a text document. Abtraction is still a difficult method to implement and a lot of work has be done using extyraction methods.

There are many summarization methods and systems available for languages such as English. Although some of them claim to be language-independent, they need at least language resources to work with. So far there is work done for isiXhosa langauge and there are langauge based based resource such stemmers [34] etc. So with such resources Xhosa can be expanded for other Natural Language Proccessing technologies.

The sentence position method was also adopted in this study, this is because most of the times the way newspapers are written for instance the first sentence in the first paragraph of the paper portrays a significant meaning. It has the first priority when making a summary[19].

2.7. Conclusion

In this chapter, we have looked at different approaches carried out by various researchers in the area of automatic text summarization. We have looked at major approaches on summarization such as statistical approaches as they are of high importance in the research. We also gave some insight on linguistic concepts that are imperative to understanding the whole text and its

summary. We then close the chapter by talking about evaluation methods, which are the methods used in this the research.

CHAPTER THREE

THE XHOSA LANGUAGE

3.0 Introduction

Xhosa (called isiXhosa in the language) is a language spoken, for the most part, in the Eastern and Western Cape of the Republic of South Africa. It is one of South Africa's eleven official languages. The Census Department of South Africa [43] reports that around 18% of the nation's populace speaks the Xhosa language. The language has a rich morphology. As indicated in a study [25], Xhosa is a southeastern Bantu language and part of the Nguni language family that incorporates isiZulu, isiNdebele, and siSwati.

According to work provided in [62] Same as Zulu language, Xhosa is one of the Nguni languages which is a group that share a significant degree of commonality. However, these languages have some etymological contrasts, for example, phonology, morphology, vocabulary, and sentence structure [55]. Due to this, they are viewed as independent languages with individual characteristics and their own particular word references and linguistic uses. Xhosa is one of the Bantu dialects that is significantly endowed with the refinement of click sounds e.g. X, c, and q.

The authors in [26] state that the language has a tonal component, as one of its notable elements. The authors explain the significance of consonants and vowels which depends on whether they are being said utilizing a rising or falling voice. The scholar in [25] that the orthography of the language is Latin based and has a composition framework created by the Christian teachers in the nineteenth century.

The first paper to be published in Xhosa and distributed is believed to have been in 1834. The Xhosa langauge has several dialects which are Ngqika, Gcaleka, Mfengu, Thembu, Bomvana, and Mpondomise. However, the author in [27] points out that other authors say that the Xhosa language is based on the Gcaleka, Ndlambe, and Gaika dialects.

Xhosa, as well as other Bantu languages, has borrowed words generously from Khoisan (languages of the Southern African, aboriginal hunter-gatherer populations) and in modern times from English and Afrikaans [28]. Some scholars believe that the existence of clicks in the Xhosa language is because there was close interaction and socializing of Xhosa and Khoisan people.

Being one South Africa’s eleven official languages, the Xhosa language is a medium of instruction in various areas of the country starting from grade one up to senior levels.

According to Europa Publications provided in [29], literary work in Xhosa including prose and poetry has been developed. The South African Broadcasting Corporation offers a domestic service in Xhosa on both radio (129 hours per week) and television (15 hours a week in TV2). There is also a Radio programme"Umhlobo Wenene FM.", which is broadcast at the national level The broadcasts in Xhosa alone are concentrated in 27 community FM radios stations in the Eastern Cape. A number of publications and newspapers are published in Xhosa and English or other African languages.

The online presence of the language is increasing significantly; this includes newspapers, online dictionaries, and online courses. Religious documents (online Xhosa bibles), research articles, and journals are published and are available online in the language. Xhosa and other African vernacular languages are used in numerous organizations for legislative, judicial, and administrative purposes. This growth, therefore, suggests the need for a means to filter the most important content for interested users to read, hence the necessity of tools such as automatic text summarizers.

The next sections of this chapter discuss the Xhosa writing system and its alphabets, this is followed by the punctuation marks and their usage, the morphology of Xhosa, and Xhosa word boundaries.

3.1. Xhosa Consonants and Vowels

The scholars in [30] state, "Every language of the world contains the two basic classes of speech sounds often referred to as consonants and vowels." In the writing of Xhosa, it is not easy to distinguish vowels from consonants, as the common word syllables in this language contain common consonants and vowels. The work put forward in [28] states clearly that the phonology of Xhosa has a simple vowel inventory as well as a highly marked consonantal system, which contains ejectives, implosives, and clicks. In the following subsections, there is information about the vowel inventory and consonantal system.

3.1.1. The Vowel System

Yule [31, p.40] states that "while the consonant sounds are mostly articulated via closure or obstruction in the vocal tract, vowel sounds are produced with the relatively free flow of air." The author in Finegan [52, p.89] agrees by stating that, "Vowel sounds are produced by passing air through different shapes of the mouth, with the various positions of the tongue and of the lips, and with the air stream relatively unobstructed by narrow passages except at the glottis."

The author in Fromkin [30, p.88] further state, “Vowels usually constitute the main core or are the nucleus of syllables. Vowels, like glides, are [- consonantal] and [+sonorant]. They differ from glides because they constitute syllable peaks; so vowels are [+syllabic], whereas glides are [-syllabic].”

A vowel is a letter of the alphabet that represents the sound of a spoken vowel. Xhosa speech possesses seven distinct vowels; however, in the writing system, they are represented by only five symbols [32].

The length of the vowel is predictable and contrastive in most cases. All vowels appear as short vowels in Xhosa [33, p.73]. Zerbian [33], however, says that depending on its position in the sentence an underlying short vowel can become long in the penultimate syllable of a word. The Xhosa language features second to the last lengthening in which the second to last syllable of one word features a longer vowel than the rest of the others in the word. The length is contrastive, when it is within a set of noun class markers. The single marker becomes [i], and its plural is [I:].

Table 1:Vowels in Xhosa Vanderstouwe [34, p.3]

Vowels	**Front**	**Central**	**Back**
High	I		u
Mid	E		c
Low		A	

3.1.2. Consonants

The scholars in [35, p.45] states, “A consonant has been described as a sound in which the air passage is either stopped entirely at some point or narrowed so as to give rise to audible friction. Consonants are classified according to the manner in which they are formed, i.e. according to the state of the air passage, and according to the organs which articulate them.”

Xhosa has a rich collection of consonants sounds: the pulmonic ingressive sounds (like those found in English), velaric ingressive sounds referred to as clicks and lastly one glottic ingressive sound called implosive [b]. Consonants are sounds produced by partly or entirely blocking air in its passage from the lungs through the vocal tract. Xhosa is a tone language with two characteristic tones referred to as low and high [53].

Some consonants sound like velar nasal [ŋ], and other nasal click sounds (in all the places of articulation) may be allophonic when with other phonemes. Error! Reference source not found. indicates all the consonant sounds found in the Xhosa language.

Error! Reference source not found. shows the click sounds found in the Xhosa language, the clicks are represented using the phonetic symbols.

The Xhosa languages have a huge number of borrowed words, and the [r] consonant is only found in those borrowed words e.g. [iɔrendʒi] 'orange,'[ifestire] 'window venster.' Explaining the consonant and vowel aspect of Xhosa is not enough. An understanding of the Xhosa orthography and its origin also needed .

Table 2: Xhosa Pulmonic Sounds and Implosive Sounds Vanderstouwe [34, p.8]

Consonants	Bilabial		Labio-dent.		Alveolar		Postalv.		Palatal		Velar		Glottal	
Plosive	p pʰ	b			t tʰ	d			c cʰ	ɟ	k kʰ	g		
Implosive		ɓ												
Nasal		m				n				ɲ		ŋ		
Trill						r*								
Fricative			f	v	s	z	ʃ				x	ɣ		ɦ
Lateral Fricative					ɬ	ɮ								
Affricate					tɬ		tʃ	dʒ						
Approx.	(w)									j	(w)			
Lateral Approx.						l								

Table 2:Velaric Sounds (Clicks) Vanderstouwe [34, p.8]

Dental	**ǀ**	**ǀh**	**nǀ**	**ŋgǀ**	**ŋkǀ**	**gꝑǀ**
Alveolar	**ǃ**	**ǃ**	**ŋǃ**	**ŋǃ**	**ŋkǃ**	**gꝑǃ**
Lateral	**ǁ**	**ǁǁ**	**ǁǁh**	**ŋǁǁ**	**ngkǁǁ**	**gꝑǁǁ**

The following section gives an overview of Xhosa orthography.

3.2. Overview of Xhosa Orthography

Saul [36, p.48] describes Orthography as the specialty of spelling individual words and the effective arrangement of groups of words . Saul [36] added that the term "Orthography" does not refer to individual words but rather to the way in which groups of words are arranged whether as a phrase or a sentence. This implies that word division needs to be considered as well as hyphenation and the utilization of punctuation. In the accompanying sections, a brief history of the orthography of the Xhosa language will be given.

John Bennie put the Xhosa language into writing interestingly around 1823. John Bennie's work is disjunctive [50]. It has a perplexing framework. The framework represents every syllable of every individual word [36]. Here is an example of the disjunctive structure.

In ko mo on ke ze zi ka- Ti xo: un gum ni ni zo ye na.

"All the cattle belong to God: He is their owner."

This type of writing demonstrates that the nasal compound was left out even when it was supposed to be included. The author in [36] gives a few examples that reveal the absence of the nasal consonant in many words like **go ko** rather than **ngoko** " therefore", **ga yo** rather than **ngayo** "with it" and **go ku ba** rather than **ngokuba** "in light of the fact that". The scholar in [36] also says that the reason for the absence of the nasal compound is that a foreigner is not capable of hearing the nasal-voiced velar nasal compound since it is not in his aural vocabulary.

Saul [36] suggests that this might be the reason why Bennie wrote that way, he wrote the way he perceived the sounds. Bennie introduced hyphens in the way he wrote sentences. Thirty years later, in 1953 John Bennie [50] devised a new form of writing. Bennie [50] abandoned the syllabic writing style and produced a conjunctive way of writing.

Ngennxa "as a result of."

Ummoya "air."

Inncwadi "a book."

In 1915, Kropf introduced a new form of writing which was called diacritics. He used this form of writing to represent the voicing of sounds in some words. The following examples show Kropf's way of writing:

Bala "write

Kulu "great."

This form of writing introduced by Kropf ignored the aspiration of sounds, which is present in the modern way of writing the Xhosa language. Saul gives examples of visible diacritic found in the language as it is written these days: the usage of the symbol /h/ that indicates aspirated sounds in the writing of words such as:

Iphepha (paper)

Phaphama (be watchful), etc.

Later, in 1925 the Xhosa author Mqhayi [49] introduced his new form of writing where he presented sounds without diacritics. Examples below show Mqhay's new form of writing:

Ingqeqesho "discipline."

e-Ncera "at Ncera."

In the following paragraphs, the Xhosa morpheme types are explained in detail.

3.3. Xhosa Morpheme Types

3.3.1. Xhosa Nouns

In Xhosa, nouns play a major role in a sentence. This is because other words in the sentence must agree with the noun. They do that by concatenating affixes known as concords. The noun is comprised of two parts: the prefix and the stem. The stem of the subject noun never changes. Xhosa is an agglutinative language, which means that the grammatical information transference is done by attaching prefixes and suffixes to roots and stems [53]. The authors in **Error! Reference source not found.** state that "Xhosa has SVO word order but allows many variations of this order for stylistic and literary purposes as well as emphasis." Bellow is an example of SVO word order.

U-m-fundi u-funa i-moto

Class 1-N SAgr 1-V Class 9-N

"The student wants the car"

In the next subsections, the prefixes, stems, and affixes in the Xhosa language will be discussed. Later, there will be a discussion about pronouns, personal pronouns, possessive pronouns, and verbs.

3.3.2. Xhosa Prefixes

The author in [37, p.3] defines a prefix as a morpheme that is placed in front of the stem of a word. The scholars in [38, p.121] state that the prefix of the noun indicates the group to which the noun belongs. Moreover, the prefix also regulates what subject agreement is used with the verb. The prefix changes according to the usage of the noun in a sentence. Nouns in the Xhosa

language commence with a prefix. The prefix starts with a vowel (e.g. “Umama” Mother, Utata ‘Father’, Inja ‘Dog’, Igusha’ Sheep’). Xhosa has about fifteen noun class prefixes.

Satyo [37] refers to these noun class prefixes as “Amahlelo.” The first noun class prefix has the prefix **Um-.** Below are the examples of the first noun class prefix:

Um*hlobo “a friend.”*

Um*lungu “English person.”*

Um*thunzi “shade.”*

There are two more noun class prefixes found in the Xhosa language. These noun class prefixes are known as 1(a) and 2(a). Noun class 2(a) is the plural of 1(a). There is no prefix in these two categories. In Xhosa, there is an augment, which is known as “Iceba.” These two categories use nouns that refer to human beings and folktale characters that are animals. Below are examples of these two noun classes:

1(a)	2(a)
Usisi “sister”	*Oosisi “sisters”*
Umalume “uncle”	*” uncles”*

The Xhosa language has 15 nouns classes. The scholars in [38, p.114] explain that these classes dictate the agreement marking that attaches to the verb stem. The researchers in [38, p.114] continue to explain that Xhosa noun classes are not referentially transparent which means that the semantic classification of the noun classes is neither systematic nor consistent. Study of nouns in this research is imperative, as nouns are the words that contain the major meaning in every sentence.

3.3.3. The Xhosa Noun Stems

A stem in the Xhosa language is the one that carries the meaning. Adding a prefix and a suffix gives a clear and a whole meaning to a word. A stem also remains in place when changing a word to another part of speech [39]. The example below illustrates the formation of the word “Ixhegokazi” old female:

I: Prefix

-Xheg- : Root

-o- : final vowel

-kazi: Feminine Suffix

As mentioned above the stem carries the actual meaning of the word. The meaning can only be modified by adding a prefix or a suffix to the stem of the word. The new meaning depends on the prefix or the type of prefix used, and designates the suffix that can be added.

For example, the stem or the root (-Xhel-) can have a number of meanings and parts of speech according to the sort of prefix that is attached. The examples below show this variety.

Ukuxhela*"to slaughter" (infinitive verb)*

Umxheli *"the one who slaughters'' (noun)*

Oxhelayo *"the one who slaughters" (relative)*

3.3.4. Xhosa Suffixes

Xhosa suffixes are located at the end of a word. In Xhosa, suffixes are known as "Izimamva." The authors in [39] say that this term is used because a suffix is located at the end of a word. In the Xhosa language, verbs and nouns always conclude with a vowel. **Table 4** shows some nouns that end with a vowel.

Table 3: Examples of Xhosa Nouns that end with a Vowel. Mtuze et al. [39]

***Isimaphambili*/ prefix**	**Root/*ingcambu***	**Suffix/*isimamva***	**Final Vowel/ Isigqibelo**	**Stems**
I-	-sel-	-	-e	-sele "frog"
Um-	-thung-	-	-i	-thungi "tailor"
Im-	-bonis-	-el-	-o	-bonisel "to see"
Um-	-bon-	-is-	-o	-boniso "show"
U-	-cang-	-	-o	-cango "door"

3.3.5. Pronouns

Pronouns indicate the case, nominative or accusative in the Xhosa language. This depends on their location in the phrase. They can be independent, prefixes, or affixes. There are different pronouns: singular and plural, which are for the fifteen different groups or classes of nouns. The Xhosa language does not have articles; instead, the language has noun class prefixes. The prefixes and their respective nouns are conjunctive. Nouns in Xhosa are used with prefixes. The use of possessive pronouns also poses no challenges.

In Xhosa the possessive pronouns take the same form. Possessive pronouns can appear by themselves in response to a question such as, “Nguban lo”/“Who is this?”, or they attach to a possessive copular as a predicate indicating possession. For example:

Aba nga-bantwana- bam/“They are my children”

Personal pronouns a refered to in P. Kese [65] as (izimnini) in the language designate nominative, accusative and dative. This depends on the form and the position they are in the phrase. Pronouns can be affixes or prefixes. Personal pronouns have their agreement based on the number, singular/plural and also the person, male/female. Examples are : (+) mna,(+) thina, (+) wena, (+) nina, (+) yena, (+) bona;(meaning I, we; you (singular), you (plural); and he/she,them.

3.3.6. Verbs

According to the explanation provided in [40], verbs are words that describe a particular action(s). In the context of a sentence, verbs typically express a relation involving the referents of one or more noun phrases. P. H Swart [64] state that “ The verb in Xhosa also has a complex derivational structure in that it takes verbal derivational morphemes (such as the applicative, causative, reciprocal etc.) that influence the argument structure of the sentence.” In Xhosa, as in other languages verbs change according to certain aspects of time. Verbs change to show tense. The present tense is indicated by an affix “a” at the end of the word. Here are examples:

U-beth-a thina

“She/he hits us.”

Zi-se_l-a ama_-nzi

“They drink water.”

1.4.7 Tenses

Tense specifies the time of the predication relative to some particular moment [41]. Secondly, the author states that an aspect indicates whether an event, state, process or action that is denoted by a verb is completed or in progress, and it overlaps the inflection / derivation dichotomy. The author [41] says mood describes an event in terms of whether it is necessary, possible, permissible, or desirable.

Xhosa has two past tenses called the recent past and the distant past. There is no set time limit determining what is recent and what is distant, only the speaker’s perception of the event. The recent past tense in the Xhosa language is indicated by using the “é” which is called the final tense marking affix. The example below shows the recent past tense:

Ndi-sel-é ama_-nzi

“I drank water.”

The other past tense that exists in the Xhosa language is called the distant past tense. The tense is indicated by using a first vowel quality ‘an’, and this is usually in the nominative pronoun prefix and by a word-final “an” affix. The last tense, called the future tense, is marked by the presence of two prefixes, ***za*** and the other one is known as the future marker ***k’u-*** which is a preposition. The two prefixes markers ‘za’ and ‘ku’ usually appear together and are never separated after the nominative marker, the pronoun marker nor before the accusative pronoun marker. Here is an example of the future tense of the verb eat:

Si-za-k’u-ty__’-an in-yama

“We will eat meat.” In the passive form

3.3.7. Adjectives

Adjectives usually follow the nouns they modify. For example, the boy is chasing the white goat / Inkwenkwe ileqa ibhokhwe ***emhlophe***. However, an adjective can be placed at the beginning of the sentence followed by the object, e.g. “Emhlophe ibhokhwe ileqwe yinja.”/ “The white goat was chased by a dog.” An adjective can also appear before the object, e.g., “Inja ileqe emhlophe ibhokhwe/the dog is chasing the white goat.”[37].

Possessive adjectives have a possessive stem and a prefix .The possessive stem and prefix must agree with the noun the adjective is modifying. The possessive adjective follows the noun it modifies. For example**:** *Akho’your’/Akhe’His/Her’ etc.* The frequent use of possessives is a feature of the English language. However, in Xhosa, possessive adjectives are handled in a different manner. For example, we love our beautiful country”Thina, sithanda ilizwe lethu elihle. Siyalithanda ilizwe lethu elihle thina.” However, it becomes a problem in English when one tries to translate this sentence word by word, e.g. “we love the country of ours that is beautiful us.”

The Xhosa language, like any other language, has its borders. These include white spaces (blank character) at the end of each word. Furthermore, the language makes use of brackets, quotes, and parenthesis, which it uses as word boundaries. Just like English, sentences always end with a full stop (.). Question marks and exclamation marks are also sentence boundaries.

3.3.8. Apostrophe

According to defination made in [66], the common use of the apostrophe is “to show the omission of letters in a contraction”. In general Saul [36] states that, an apostrophe is put every time a vowel is omitted in the written Xhosa language. For example:

Ndifun’ukuhamba (I want to go),

ufun'inyama (he/she wants meat).

3.4. Abbreviation

Saul [36] states that "In writing, an abbreviation is any shortened form of a word or phrase." As stated in [63], an abbreviation of a word consists only of the first part of the word. It is often used where certain words must be written (and read) consecutively. Examples: Tues. = Tuesday; Dec. = December; Minn. = Minnesota; Eur = Europe, European. Saul [36] further states that" IsiXhosa too employs a number of abbreviations. In the consulted isiXhosa texts, various abbreviations were identified, but it was observed that uniformity is lacking as far as the use of a full stop in the writing of these abbreviations is concerned"

These examples illustrate the point:

Finy isifinyezo (PanSALB, 2008:60) (abbreviation)

Vum isivumelanisi (PanSALB, 2008:60) (concord)

3.5 Summary

In this chapter, some background was given of the written Xhosa language. The standard consonants and vowels of the language were described. Nouns were explained, together with verbs, which are the significance transporters of the language. The following chapter discusses the methodology and system design that are used in this study.

CHAPTER FOUR

METHODOLGY AND SYSTEM DESIGN

4.0 Introduction

This chapter sets out the vital instruments used to the Xhosa text summarizer. These include the framework, the methodology and the system design.

4.1. Methodology

In this study a number of investigation and development of the system was followed. These include literature review, data source collection, algorithm development, testing, and evaluation. The development of the Xhosa Text Summarizer follows an iterative process until it is completed. The Xhosa text Summarizer is extractive based; an extractive based automatic text summarizer has three stages: preprocessing, sentence ranking and summary generation. The preprocessing stage also involves three stages tokenization, stop word removal, and stemming. The following subsections will give a brief discussion about all the methods used in this work.

4.2. Proposed Algorithm

The algorithm developed has to follow a sequence of stages in order to create a summary. These stages are in a predefined order. The input document should be in .txt format. The algorithm has to put the document through the preprocessing stage where first it splits the text into tokens, secondly it removes words that do not contain meaning using a pre-defined list, and thirdly it stems the individual words. These stages are listed below:

4.4.1. How the Algorithm Works

- Preprocessing
  - A plain text is entered into the system. The system then:
  - Splits the text into sentences and splits sentences into individual words or tokens
  - Removes stop words using a stop word filter
  - Stems each word and keeps count of how many times each stem occurs in the text
- Sentence Ranking
  - Scores (sentence) = sum ([free (word) for words in sentence])
  - Rates each sentence by the words it contains
  - Takes the X% highest rated sentences.
- Summary Generation
  - Collects the X% top sentences to make a summary of the text.

4.3. Preprocessing

In this section, the preprocessing stages of Tokenization, Stop words, and Stemming are explained. **Section 4.6 and 4.7** explain the two stages in automatic text summarization called sentences ranking and summary generation.

4.3.1. Tokenization

Usually, it is compulsory to break the text into words (tokens) in order to classify these boundaries between clauses, phrases, or sentences. This step ensures the breaking down of text into individual tokens. When the text is divided into individual tokens, it becomes easy to apply further other linguistic steps, like stop word removal.

4.3.2. Stop Words

Stop words are a collection of prepositions, conjunctions, articles, and particles. A stop word list was made for this study. Xhosa stop words are explained in detail in **Section 5.2** of this research.

4.3.3. Stemming

Taking out the suffixes and affixes, by automatic means, is an operation, which is especially useful in extractive summarization and information retrieval. This study used a lightweight stemmer for Xhosa that strips the suffixes, on a “longest match” basis. The stemmer used was previously designed by the authors in [42] which, at the time, focused on stemming nouns only. If this stemmer were to be used it would have an effect on the accuracy of the summary, since it only

focused on nouns, therefore some adjustments were made, so that the stemmer was able to stem verbs as well as nouns. The explanation of the rules of the stemmer is in **Section 5.3** .

4.6 Sentence Ranking

Once the text has been broken down into individual terms (i.e. tokens) and it has been formatted correctly, the stop words are excluded with the help of a prepared stop word list, and stemming is done using the Xhosa stemmer. The tokens are then scored and ranked accordingly. When the preprocessing stage is done, the terms get their weight value. This weight value is allocated to individual tokens. The following formula is used to calculate the weight of each term:

$$wt = \frac{\sum frequency of term}{\sum terms \in the\ document}$$

This method is based on the view that sentences that portray meaning to the entire text contain the most common words in the document. When the weight is assigned to the individual terms, sentences are ranked. This implies that the sentences will be ranked according to how significant the weight of the sentence is to the other sentences. The weight of the sentence is calculated by adding the weight of all the terms in the sentence and dividing it by the total number of terms in that sentence. The sentence with the highest weight is ranked first. . The following formula illustrates that:

$$wt_s = \sum_{i=1}^{n} (wt_i)/n$$

The abbreviations used in the formula above are as follows: wt_s= is the weight of each sentence and is equal to the summation of $wt_1, wt_2, wt_3, wt_4 \ldots\ldots\ldots\ldots\ldots\ldots wt_n$

These are the weights of individual terms in each sentence. The value of (n) is the total number of term in a sentence (it is an average). It is important to note here that a term has different forms. However, each term is treated individually in order to be able to assign a weight value to it.

The final step in the process is for the summarizer to extract the highly ranked sentences. The first sentence will be included in the summary. This is because it is known to have the most relevant information; as a result, it requires priority. The system is based on a user-centered principle, where the user can specify the number of sentences that he/she wishes to see in the summary. This is done through user input. The percentage of the summary is calculated by dividing the percentage specified by the user by the total number of ranked sentences, and then taking the maximum number of that result.

4.7 Summary Generation

For readability of the summary, the sentences in the summary are rearranged based on their forms in the original text, e.g. the sentence, which occurs first in the original text is given high priority and is also assigned a higher score that the rest of the sentennces. This is with the notion that first sentence in new articles ahs th most imofrmation and this is also the case in the Xhosa news articles.

To weigh the sentences the authors used the sentence postion method explained in section 2.2.2 of this study. This feature scores the sentences according to their position in the text. In this work, we assume that the first sentences of the text are the most important ones. So, the first sentence of a document gets a score value of 1, the second sentence gets 0.9, the tenth sentence gets 0.1 and the rest of the sentences.

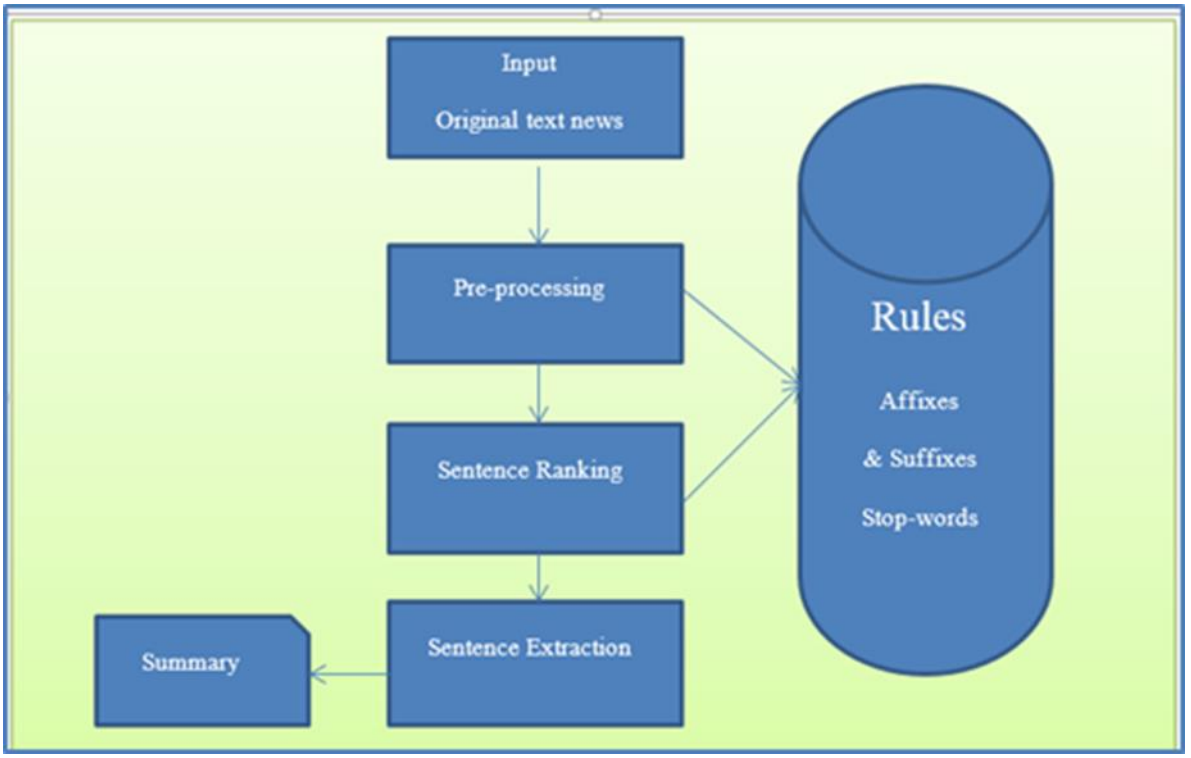
4.8 System Design

The back-end supports the front-end services and allows users to interact with the front-end system. This also helps in the response request that proceeds between the front-end and on the back-end system. The back-end design of the Xhosa Text Summarizer controls the interaction of users with the front end of the system; this includes taking the whole text from the text area to the back end for processing. Error! Reference source not found. illustrates the system design of the Xhosa Text

4.10 Summary

The aim of this chapter was to outline the methodology and to present the system design of the scheme. This chapter also gave insight on the modules and classes that were selected for text analysis. In the next chapter, the implementation of the system is discussed, together withthe techniques used in this work.

Figure 1: Architecture Design of Xhosa Text Summarizer



CHAPTER FIVE

IMPLEMENTATION

5.0 Introduction

This Chapter explains all the processes that were necessary to develop and implement the Xhosa text summarizer. This chapter also explains the preprocessing steps in detail than in Chapter 4.

5.1.Tokenization

Before examining the in-depth linguistic behaviour of a text, the units of that text must be defined and put into groups. In an extraction-based summarization, it is necessary to decide on what the granularity of the extraction fragments will be. Granularity refers to the "size" of the textual elements that will be copied from the original text and put into the final document, which is a summary. These fragments could be in the form of a paragraph, a sentence, a phrase or even a clause, although the most commonly extracted textual element is probably the sentence.

5.2.Stop Word Removal

During the pre-processing stage, as already explained, the stop words are removed from a text document. There are 158 Xhosa stop words in the NLTK folder, which includes a combination of prepossessions, articles, adjectives, etc. The Xhosa stop word list is imported from the NLTK data folder. A complete list of the most common Xhosa words can found in Appendix C:**. Table 4** shows the most common Xhosa stop words identified by researcher, together with their English meaning and their part of speech.

Table 4: Sample of Common Xhosa Stop Words used in this Research

Word	**English Meaning**	**Part of Speech**
Ngaphezulu	above	Adjective
Ngaphantsi	under	Adverb
Ngaphambili	before	Adverb
Emva	Back	Adverb
Phambi kwe	Before the	Locative
Emva kwe	Behind the	Adverb
Ukuze	So that	Conjunctive
Kufuphi	Nearby	Verb
Phakathi	inside	Adverb
Nam	me	Locative

5.3.Stemming

As mentioned in **section 4.3.3**, this study uses a Xhosa stemmer. The Xhosa stemmer takes a word as input and removes its suffixes according to a rule-based algorithm. The algorithm follows the known Porter algorithm for the English language and it is developed according to the

grammatical rules of the Xhosa language. The rules include a pre-defined list of suffixes and affixes used to change words into their root form. These rules are presented as follows: The rules are defined using python programming language.

```
If word.startswith ('asingo') or word.startswith ('ayingo'):
        return 6
    if word.startswith('nga') or word.startswith("asi"):
        return 3
    if word.startswith("ku")  or word.startswith("en"):
        return 2
    if word.startswith ("aba") or word.startswith ("abe"):
        return 3
    if word.startswith("um") or word.startswith ("em"):
        return 2
    if word.startswith ("oo") or word.startswith ("im"):
        return 2
    if word.startswith ("imi"):
        return 3
    if word.startswith("ama")or word.startswith("ili"):
        return 3
    if word.startswith("ame")or word.startswith("isi"):
        return 3
    if word.startswith("is")or word.startswith("iz"):
        return 2
    if word.startswith("izin") or word.startswith("izim"):
        return 4
    if word.startswith("in")or word.startswith("im"):
        return 2
    if word.startswith("iin")or word.startswith("iim"):
        return 3
    if word.startswith("ii")or word.startswith("uk"):
        return 2
    if word.startswith("ulu") or word.startswith("ulw"):
        return 3
    if word.startswith("ul")or word.startswith("ub"):
```

```
        return 2
    if word.startswith("ubu"):
        return 3
    if word.startswith("utyw"):
        return 4
    if word.startswith("uty"):
        return 3
    if word.startswith ("uku" or word.startswith ("ukw")):
        return 3
    if word.startswith ("utyw"):
        return 4
```

The Xhosa stemmer is tested on 127 verbs and 327 nouns. **Table 5 and Table 7** shows a portion of the list of verbs tested on the Xhosa stemmer. The complete list of Xhosa nouns and verbs is in **Appendix B:.**

Table 5: Sample of Xhosa Stemmed Verbs

Verb	**Stem**	**Meaning**
uyahamba	hamba	Go, going
uyabaleka	baleka	Run, running
ulele	lele	asleep
liyaduduma	duduma	lightning
luyakhala	khala	cry
uyapheka	pheka	cooking
ziyadilika	dilika	falling

Table 6:Sample of Xhosa Stemmed Nouns

Noun	**Stem**	**Meaning**
abantu	ntu	people
abathwa	thwa	bushman
abazala	zala	cousins
abazali	zali	parents
abembi	mbi	diggers
aboni	oni	sinners
afrika	afrika	africa
amanzana	nz	water
amaqaqa	qaqa	hills
amashwa	shwa	bad luck

amathumba	thumba	opportunities
amatye	tye	stones
amazwi	zwi	voices

5.4.Implementation

A simple interface is created to allow an easy interaction for a user.

5.4.1. The IsiXhoSum Interface

The interface allows a user to input Xhosa news item of his or her choice. The interface has three buttons (Summarize, Clear, and Quit), and two text boxes (one for the input and another for the output). The first one is for the user to paste a news item into and the second is for the output - the summary. The interface of the summarizer is shown in Figure 2**.**

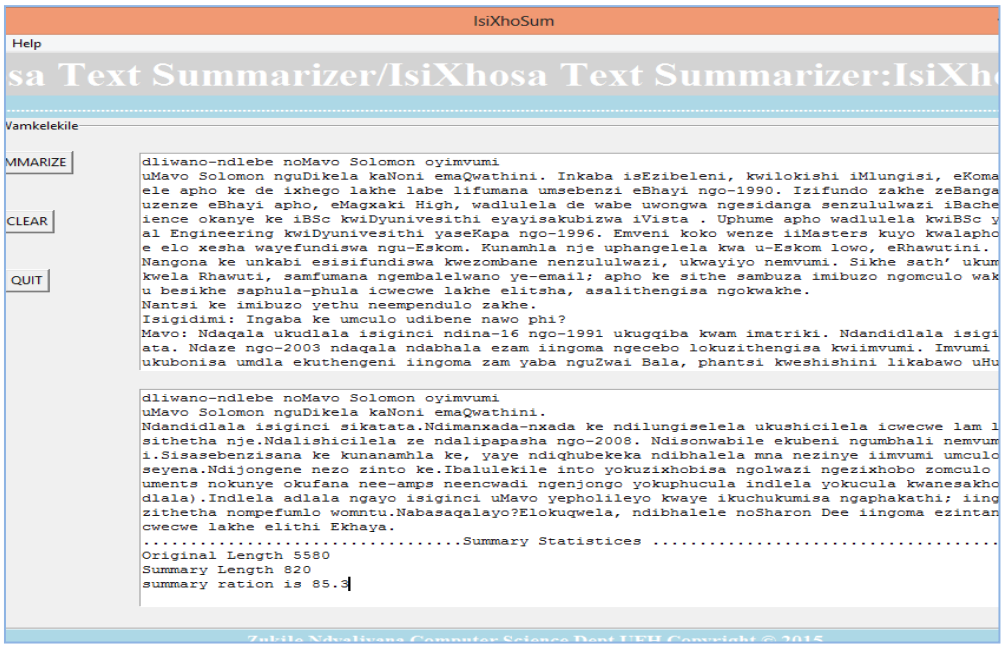


Figure 2: Xhosa Text Summarizer interface

5.4.2. Modules of the Xhosa Text Summarizer

The summarization process starts when a user inserts the news item on the interface and presses the SUMMARIZE button. Before the summary is generated, the text will undergo three phases: Preprocessing, sentence ranking and summary generation as discussed in the Design section of this thesis.

As mentioned before, this study used the NLTK toolkit to develop the summarizers. For the preprocessing part, the toolkit's tokenizer module was used, which has all the Xhosa sentence boundaries, like comma, semicolon, and question marks. This tokenizer is called the

wordpunct_tokenizer. The tokenizer first splits the text into words or tokens, using the code presented in **Listing 1:**

```
def tokenize_words(text):
    stemmed=[]
    xhosa_stop_set = set(stopwords.words("isiXhosa"))
    text1 = "".join([ch for ch in text if ch not in string.punctuation])
    tokens = wordpunct_tokenize(text1)
```

Listing 1: Sentence Tokenisation Code

- **Stop Word Removal**

  The Xhosa stop words are stored in a folder of the NLTK toolkit and are imported using the nltk.corpus import stop words.

  The code in **Listing** Error! No sequence specified. checks if a particular word is in the file.

  If it is in the file, it will be removed.

  If it is not in the file, the word will be preserved.

```
def tokenize_words(text):
    stemmed=[]
    xhosa_stop_set = set(stopwords.words("isiXhosa"))
    text1 = "".join([ch for ch in text if ch not in string.punctuation])
    tokens = wordpunct_tokenize(text1)
```

Listing 2: Removing Noisy Words from the Text

- **Stemming**

According to the definition of stemming, a stemming algorithm does not need to identify the linguistically correct stem, but it is sufficient to map all the forms of a word to a single form. The intention of this work is to look for a stemmer that is able to bring all words with the same stem into a single form. Nouns and verbs have been stemmed separately.

When the stop words have been removed, and the remaining words have been tokenized, the stemmer then checks the length of each word and passes the word to the next rule on the list for matching the prefixes and suffixes against the ones provided.

The code in **Listing 3** shows how the stemming method is called when tokenization and stemming is done.

```
for item in cleanup:
        stemmed.append(stem(item))
        return item
```

Listing 3: Stemming of words Code Method

After the text is preprocessed using the above two techniques, the next step is sentence ranking, which is discussed in the next section.

- **Sentence Ranking**

A sentence, which has the highest number of high frequency words in it, will be ranked higher than any other sentence. The other sentences will be ranked according their frequency. The code below shows how sentences are ranked in Xhosa Text Summarizer.

```
def sentence_score(sentence, frequencies):
    return sum((frequencies[token] for token in tokenize_words(sentence)))

def create_summary(sentences, max_length):
    summary = []
    size = 0
    for sentence in sentences:
        summary.append(sentence)
        size += len(sentence)
        if size > max_length:
            break
```

Listing 1: Sentences Ranking Code

- **Summary Creation**

The last part in automatic text summarization is the creation of a summary that will contain only the most highly ranked sentences. The summary also provides and statistical data about the summary: (information about) the number of text lines, the most frequent words, etc. In the creation of the summary, the user can specify if he/she wants 10, 20, 30, 50…100 % of the text as a summary. The summarizer returns the summary based on the maximum size specified by the user.

5.5. Experimentation

This section explains the preparation of the Xhosa corpus as well as the preparation of manual extracts.

5.5.1. Corpus Preparation

As mentioned in **Section** Error! Reference source not found., this study involves the collection and preparation of Xhosa news items in order to form a corpus. The aim of the corpus is to evaluate the summarizer after it has been developed. The corpus is comprised of two hundred news items from online Xhosa news websites called ISIGIDIMI (www .isiGidimi.co.za), NALIBALI (www.Nalibali.org), and IOL (www.iol.co.za/isolezwe). The electronic versions of the articles were transformed to plain text format. **Table 7** shows statistics on the organized corpus.

Table 7: Basic Statistics of the Xhosa Test Set

Number of articles	200
Number of words	50901
Total number of sentences in the file	3289
Total number of characters	310302

5.5.2. Creation of Manual Summaries

Creation of Manual Summaries the automatic text summarization for Xhosa news articles is evaluated against summaries produces manually from selected news items manually selected extracts. The language expert marks the manual extracts. These news texts are of various contents and lengths. As these items are selected, they are given to a Xhosa linguist to make manual summaries. The linguist reads the whole text and picks out sentences that contain the overall meaning.

The text files for manual summarization were taken straight from the prepared Xhosa corpus of 200 texts. Fifteen text files were randomly selected from the corpus, but there were conditions, described by an algorithm as follows:

The paragraphs should contain between 4 and 10 sentences Sentences should contain between three and six. The algorithm is illustrated in **Figure 3**. If the above statements are met, the article should be selected for manual summarization.

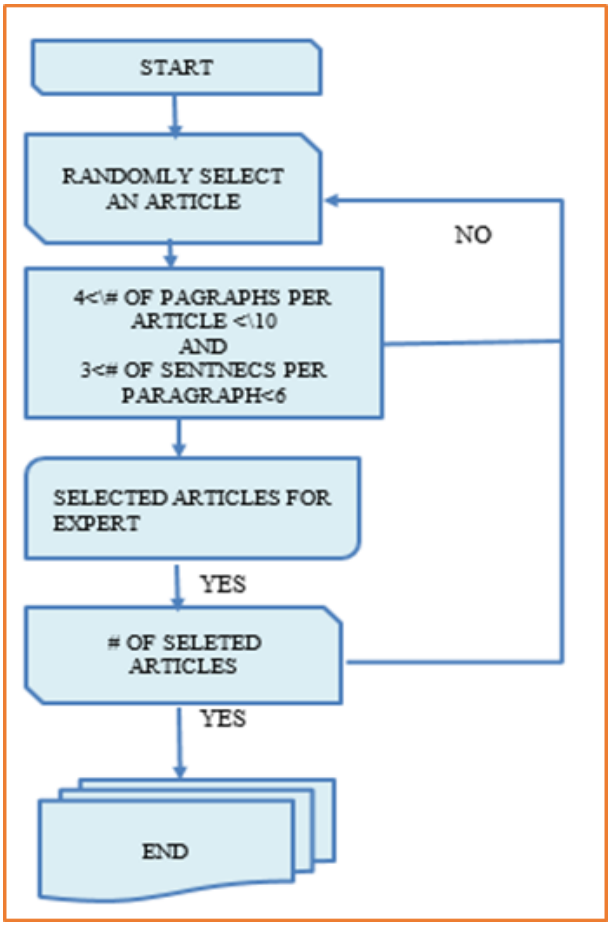


Figure 3: Algorithm for Selecting Article in the Corpus

Table 8 shows the number of sentences in the original text and the number of sentences in those summaries that were prepared manually from the 15 selected texts.

Table 8: Manual Summaries Used for Evaluation.

News Item ID	**Original Text (# of sentences)**	**Summary (# of sentences)**
News item one	25	9
News item two	42	15
News item three	12	4
News item four	36	15
News item five	30	10
News item six	14	10
News item seven	14	12
News item eight	43	24
News item nine	20	7
News item ten	50	12
News item eleven	66	23
News item twelve	69	20
News item thirteen	20	7
News item fourteen	20	10
News item fifteen	11	5

5.6.Summary

In this chapter, the objective was to give a detailed explanation of the implementation of the Xhosa text summarizer. This study followed up-to-date methods to implement the Xhosa text summarizer. This chapter also explained the preparation of both the Xhosa corpus and the manual summaries.

CHAPTER SIX
TESTING, RESULTS, AND DISCUSSION

6.0 Introduction

In the preceding chapter, the details of the implementation of Xhosa Text Summarizer were explained. The objective of this Chapter is to analyze and evaluate the results and give meaning to them. This chapter has three subsections: Testing, Results, and Discussion.

6.1. Testing

The Xhosa Text Summarizer was evaluated using Xhosa news text. The texts taken from the corpus were sent to a Xhosa linguist who prepared manual summaries. The linguist used his knowledge to make the extractive summaries. Since this study is following the extraction-based method, the linguist did not need to rewrite the sentences, as this would be an abstractive method, which is not the focus of this research. Two types of summaries were provided to the evaluators: automatic and manual.

In order to rate, the two summaries were distributed to five Xhosa native speakers (Xhosa lecturers, Xhosa radio news presenters, and three Xhosa students). An intra-group evaluation was codncuted to achieve a reasonable evaluation proccess. The results of the students was evaluated seperatly and the results of the language experts was also done seperately.

The two summaries were compared subjectively. The Xhosa native speakers first read the original text, and then the two summaries. The compression rate on Xhosa Text Summarizer was at the three ranges of 10%, 20%, and 30%. There were guidelines to help the readers with the process of evaluating the two summaries, as follows:

The two summaries are to be assessed based on these qualities:

- Informativeness (In which one of the summaries is the most important information being kept?). Informativeness means that only the best sentences, that contain the most valuable information of the topic sentence, are selected for the summary. Read the topic statement and all associated summaries generated using different methods. Then, chose the method / methods that create a more informative summary.
- Linguistic quality (what score would you assign to each summary. This is assessed on a five-point scale from “1” to “5 where “5” indicates that the summary is right, “1” indicates that the summary is bad, and “2” to “4” show the grades in between.
- Coherence structure (Which summary is more coherent?). In a coherent summary, there is a smooth transition between sentences. While reading the sentences in their rank order, there

should not just be a collection of related information, but the sentences should build a coherent body of information about a topic. You are required to read the original news item and the two summaries and mark which of the method/methods produce the most coherent summaries.

The following questions were asked after the participants had read all the news items with their respective summaries

- Which summary is better?
- In which one is the most important information being kept, better?
- Which summary is more coherent?
- On a scale of 1-5, where five is the best, what score would you assign to the linguistic quality of each summary?

Table 9 shows the numerical details of the subjective assessment. The two methods that were used were called **M1**, which is the manual method and **M2**, which is the automatic method.

Table 9: Text files tested on Xhosa Text Summarizer.

News Item ID	**Original Text(length(Number of words)**	**Summary Text(length(Number of words)**
News item one	2059	530
News item two	3552	392
News item three	1022	438
News item four	3025	412
News item five	2625	444
News item six	1178	413
News item seven	1289	529
News item eight	3724	389
News item nine	1838	381
News item ten	4318	490
News item eleven	4332	490
News item twelve	5649	330
News item thirteen	6267	323
News item fourteen	1495	429
News item fifteen	1927	360

6.2. Results

The evaluation method used in this study is both subjective and objective.

6.2.1. Results of Subjective Evaluation

This section gives details of the responses provided by the evaluators during the assessment phase of the Xhosa text summarizer (IsiXhoSum). It also gives statistical data on what the evaluators thought of the two different kinds of summarization methods.

- Which summary is better?

  The calculation of the scores is as follows:

  If three evaluators out of the ten have chosen the method M1 as a better method for making a news summary, this means that 3/10 = 0, 3, or 30% of the evaluators chose M1.

Table 10: Shows Results of a Better Method

TID	Total(M1 and M2)	(Manual Summary)M1	Automatic SummaryM2	M1	M2
News one	M2(3),M1(2)	0,40	0,60	40%	60%
News two	M2(4),M1(1)	0,20	0,80	20%	80%
News three	M2(3),M1(2)	0,40	0,60	40%	60%
News four	M2(3),M1(2)	0,40	0,60	40%	60%
News five	M2(4),M1(1)	0,80	0,20	80%	20%
News six	M2(4),M1(1)	0,80	0,20	80%	20%
News seven	M2(1),M1(4)	0,80	0,20	80%	20%
News eight	M2(3),M1(2)	0,40	0,60	40%	60%
News nine	M2(4),M1(1)	0,20	0,80	20%	80%
News ten	M2(4),M1(1)	0,80	0,20	80%	20%

News eleven	M2(1),M1(4)	0,20	0,80	20%	80%
News twelve	M2(1),M1(4)	0,20	0,80	20%	80%
News thirteen	M2(3),M1(2)	0,40	0,60	40%	60%
News fourteen	M2(4),M1(1)	0,20	0,80	20%	80%
News fifteen	M2(4),M1(1)	0,80	0,20	80%	20%
Average		**0,47**	**0,53**	**47%**	**53%**

- In the two summaries provided, which one contains the most important information?

  The Informativeness of the summary created by each of the two methods for each test item is scaled to 100 (out of the expected total five votes by the evaluators). For example, if M1 is selected by three of the five evaluators, the percentage of the Informativeness of the summary is measured as 3/10= 0.3 i.e. 30%. Error! Reference source not found. displays such results**.**

Table 11: Important Information Preserved

	Total(M1 and M2)	**M1**	**M2**
News one	M2(1),M1(4)	80%	20%
News two	M2(2),M1(3)	60%	40%
News three	M2(2),M1(3)	60%	40%
News four	M2(3),M1(2)	40%	60%
News five	M2(3),M1(2)	40%	60%
News six	M2(1),M1(4)	80%	20%
News seven	M2(3),M1(2)	40%	60%
News eight	M2(2),M1(3)	40%	60%
News nine	M2(2),M1(3)	60%	40%
News ten	M2(2),M1(3)	60%	40%
News eleven	M2(3),M1(2)	40%	60%
News twelve	M2(5),M1(0)	-	100%
News thirteen	M2(3),M1(2)	40%	60%
News fourteen	M2(4),M1(1)	20%	80%

News one	M2(4),M1(2)	20%	80%
Average		**45%**	**55%**

Which summary is more coherent?

Again, in the case of the coherent summary, the calculation is as follows: if five evaluators out of ten chose method M1, that means that 50% of the evaluators believe that M1 is the more coherent.

Table 11: Results of a coherent summary

TID	**Total(M1 and M2)**	**M1**	**M2**
News one	M2(3),M1(2)	40%	60%
News two	M2(3),M1(2)	40%	60%
News three	M2(4),M1(1)	40%	60%
News four	M2(3),M1(2)	40%	60%
News five	M2(1),M1(4)	80%	20%
News six	M2(2),M1(3)	80%	20%
News seven	M2(4),M1(1)	80%	20%
News eight	M2(3),M1(2)	40%	60%
News nine	M2(4),M1(1)	20%	80%
News ten	M2(1),M1(4)	80%	20%
News eleven	M2(1),M1(4)	20%	80%
News twelve	M2(4),M1(1)	20%	80%
News thirteen	M2(2),M1(3)	60%	40%
News fourteen	M2(4),M1(1)	20%	80%
News one	M2(3),M1(2)	40%	60%
Average	**47%**		**53%**

- On a scale of 1-5, where five is the best, what score would you assign to the linguistic quality of each summary?

This method ensures that all the linguistic qualities of the summary are considered including grammar and non-redundancy (see **0**). For example, calculating the final score of the first news item on row one. The results from the users are turned into a percentage based on the sum of the scores of the evaluators. For instance, if the summary of Test1 using M1 is scored 3 by **Evaluator One**, 4 by **Evaluator Two**, 5 by **Evaluator Three**, 3 by **Evaluator Four**, and 3 by **Evaluator Five** the percentage of the overall linguistic quality of the M1 summary for News One is the average of the sum of the scores divided by the maximum score possible, in this instance, 20] i.e. 3 + 4 + 5 + 3+3 = 11/20*100 = 55%.

All these results are shown in Table 14.

Table 12: Linguistic Quality Results

TID	EvaOne	EvaTwo	EvaThree	EvaFour	EvaFive	Result	%
News one (M1)	3	1	5	1	3	13	66%
News one (M2)	5	3	4	4	1	17	87%
News two (M1)	3	4	5	4	3	19	97%
News two (M2)	4	5	3	1	4	17	87%
News three(M1)	2	4	4	3	4	17	87%
News three(M2)	5	5	3	2	3	18	92%
News four(M1)	4	4	4	3	3	18	92%
News four(M2)	5	2	5	1	5	18	92%
News five(M1)	3	1	5	1	4	14	71%
News five(M2)	4	5	3	3	4	19	97%
News six(M1)	2	4	4	4	4	18	92%
News six(M2)	4	3	4	4	4	19	97%
News seven(M1)	3	4	4	4	3	18	92%
News seven(M2)	5	2	5	2	4	18	92%
News eight(M1)	3	5	3	4	4	19	97%
News eight(M2)	5	4	2	3	3	17	87%
News nine(M1)	1	1	3	3	4	12	61%
News nine(M2)	5	5	2	2	4	18	92%
News ten(M1)	2	3	4	3	4	16	82%
News ten(M2)	5	5	1	3	5	19	97%
News eleven(M1)	4	1	2	4	3	14	71%
News eleven(M2)	5	4	3	4	3	19	97%
News twelve(M1)	3	3	4	3	4	17	87%
News twelve(M2)	5	4	5	1	2	17	87%
News thirteen(M1)	3	5	4	3	4	19	97%
News thirteen(M2)	1	4	2	4	3	14	71%

News fourteen(M1)	3	4	3	3	3	16	82%
News fourteen(M2)	4	3	4	3	4	18	92%
News fifteen(M1)	2	4	5	3	3	17	87%
News fifteen (M2)	2	1	2	3	2	10	51%

6.2.2. Results of Objective Evaluation

This section, gives the results of the objective assessment. The objective assessment method adopts one best summary and compares the automatic summary against the reference/manual summary. It is intended to measure the system's summary approximation to the reference summary on the basis of recall (R), precision (P) and F-measure (F).

The standard recall and precision measures are calculated as follows and the f-measure is calculated based on the values of precision and recall:

- Recall(R) = correct/(correct +missed)
- Precision(P) = correct/(correct + wrong)
- F-measure(F) = 2*RP(R+P)

Where:

Correct = the number of sentences in both the summarizer's summary and the reference summary,

Wrong = the number of sentences in the summarizer's summary but not in the reference summary,

Missed = the number of sentences in the reference summary but not in the summarizer's summary.

Summaries are required to be generated at four compression rates 10%, 30%, 40% and 50%. From the 15 news articles prepared for experimentation, one pair of news items were randomly selected to be input to the system for a given extraction rate.

Rouge2.0 [51], developed in java and, is packaged as jar file. Major changes were made to the file called rouge Properties. This is where you specify the ROUGE-N type that you want to evaluate, stop words to use, output file, synonyms. The rouge package also outputs the results in a CSV file for the purpose of statistical analysis.

Table 13 shows the statistical results of the fifteen files that are analyzed using ROUGE2.0 tool. It also shows the specific comparisons between the recall, precision, and f measure averages.

Table 13: Output of the ROUGE2.0 tool

ID	(Method 1=M1)			(Method 2=M2)		
	Avg_Recall	**Avg_Precison**	**Avg_F-Score**	**Avg_Recall**	**Avg_Precison**	**Avg_F-Score**
News one	4%	3%	2%	66%	22%	12%
News two	3%	2%	2%	71%	21%	11%
News three	25%	35%	29%	34%	45%	29%
News four	8%	2%	3%	32%	71%	8%
News five	35%	35%	25%	35%	35%	39%
News six	8%	2%	3%	2%	71%	70%
News seven	35%	35%	25%	35%	49%	34%
News eight	70%	2%	3%	82%	8%	8%
News nine	34%	5%	25%	35%	29%	82%
News ten	8%	2%	3%	70%	8%	35%
News eleven	70%	35%	25%	35%	29%	70%
News twelve	34%	8%	3%	2%	8%	35%
News thirteen	34%	35%	25%	35%	29%	72%
News fourteen	7%	2%	3%	22%	70%	34%
News fifteen	34%	35%	25%	35%	34%	62%
Averages	**27%**	**16%**	**13%**	**39%**	**35%**	**40%**

6.3. Discussion of the Results

This sub-section discusses the results obtained from the evaluators who made their subjective judgements.

A better summary means that most of the key points in the original text have been well extracted, and proves that the summarizer works for Xhosa news texts. **Tables** 10 to 14 list the results of the subjective evaluation. These results show that the M2 summary is more informative than the

M1, as judged by the evaluators, 45% to 55%. M2 is preferred copmared to M1 with an average of 63 %. These results show that, based on the evaluators that M2 does better than M1.

When coherency is evaluated, M2 scores 53% compared with 47% for M1. This shows that the content in M2 is clear; there is flow and fluency in the sentences. The flow of the content is influenced by the inclusion of language-based tools such as the stop word list and the stemmer for the Xhosa language.

Based on these observations it can be seen that both methods, automatic and manual, display good linguistic qualities with M2 slightly outperforming M1 **Table 12**. In both methods, the use of punctuation marks, full stops, etc. are being maintained. Despite the good performance brought forward by both methods, the observation is that there is an issue of pronoun resolution. That has influenced the performance of the summarizer.

the objective results in **Table 13** show that the average F-measures between the two methods are M1 13%, and M2 40%, so that M2 outperforms M1 Furthermore, the gap between the performances of M1 and M2 shows the improvements made by the summarizer with the inclusion of an improved term frequency method using the language specific rules. It was good to see that in the study the summarizer worked well for the objective as well as the subjective evaluation. This result is emphasized in the statistical representations of the recall and precision of the summarizer. The scores achieved by M2 are much higher than those achieved by M1.

Other observations that are made on the results are the following:
On the manual method, which is the method performed by a linguist, it transpired that in most instances, 60% of the time the first sentence would appear first and 40% of the time, it appeared second or somewhere else in the paragraph. The manual summaries are attached in **Appendix G**:, and show the position of the first sentence. The first sentence appeared 73 % as the first in the first paragraph and 33% in another paragraph. This shows that the method of sentence position works in the news domain and worked well for the Xhosa text summarizer.

It was observed that the Xhosa stemmer, used some verbs that could not be stemmed. This is because the stemmer does not have a large number of verbs yet as the original stemmer only stemmed nouns, adding new verbs to the stemmer requires some deeper investigations so that suitable rules are made for the Xhosa stemmer.

The author also notices that, some results depend on the compression rates to which the source text is subjected for example: At a 60 % compression rate provided by the user, the summarizer can only extract four sentences. The summarizer ranks those sentences highest. At 40% compression rate, again provided by the user, the summarizer only extracted two sentences with first sentence being at the top. Lastly, at 20% compression rate, the summarizer extracted only one sentence that is the first sentence.

The use of nouns and maintaining the capital letter when referring to the name of a person, or a place are some of the things that have been maintained in the summary. Examples are UFikile Mbalula 'Name of person', eKapa 'Cape Town: name of place'. The author collected a maximum number of stop words which means that most stop words in the text were taken out thus leaving a readable summary. The automatically generated summaries are informative just like the human-generated summaries because after reading such an informative summary, the native speakers have a good knowledge of the content, and are able to describe parts of the original text. This is proved by the averages of method 2 (M2) and method 1 (M1).

6.4. Discsion on Coherence and Cohesion

The presented evaluation successfully shows the improvements or integrating cohesion and coherence, but it has two weak points. First, the size of the corpus and the fact that it represents a single type, which does not allow rare generalisations. Second, the fact that evaluation metrics fall short in assessing the improvements yielded by the combination of these two discursive informations, since they cannot account for quantitative improvements at granulatity levels different from the unit used in the golden standard, and therefore a full evaluation of summaries involving sentence compression is excluded. Furthermore, qualitative improvements on general text coherence cannot be captured, nor their influence on summary readability.

The authors have tried to address this problem by identifying text segments which carry non-useful information, but the presented metrics do not capture this improvement.

6.5. Summary

This chapter explains how the testing was conducted, it lists all the results and these are discussed. The aim of the discussion is to give meaning to and deliberate on the results obtained.

CHAPTER SEVEN

5 CONCLUSION AND FUTURE WORK

7.0 Introduction

In the preceeding chapter the testing of the summarizer was outlined, and the results were given and discussed. In this chapter, the Conclusion and suggestions for Future work are presented.

7.1. Research Summary

The principal aim at the start of this exploratory work was to conduct a comprehensive literature review to assemble information and acquire a comprehension of different text summarization methods and algorithms. This study makes utilization of the extraction approach, rather than the abstractions approach, for automatic summarization, to develop the Xhosa Text Summarization tool.

The position of the first sentence of the original text was observed in the summaries generated by the summarizer. The findings revealed that 73% of the time the first sentence would appear first, and 27% of the time the summarizer would place the first sentence in the second ,third or fourth paragraphs it elsewhere. Whereas is 40% and 60% respectively on the manual method.

Two evaluation techniques, known as subjective and objective were used in this study. The overall subjective evaluation results show the effectiveness of the features that were added to supplement the Xhosa Text Summarizer, the Xhosa stemming rule and the stop-word list brought some improvement to the summarizer. From the experiments, a conclusion is made that the addition of more advanced methods would help improve the summarizer, so that it would be able to create a better and more coherent summary. The results of the objective evaluation have shown relatively coherent results about the effectiveness of the summarizer.

The major objectives of this study were to investigate, implement and evaluate the Xhosa Text Summarizer. In this study, there were sub-objectives that needed to be fulfilled.

- **To do research about existing and well-known methods and algorithms to extraction-based automatic summarization.**

An intensive investigation was carried out on work that has already been done in this field, and the relevant parts of it are presented in Chapter Two of this dissertation. Various methods and algorithms that are used to make effective text summarizers were discovered and investigated. In the observations, it was noted that most scholars combine statistical methods, which in this case is term frequency and sentence position, with language based rules like the inclusion of a stemmer and a stop word list. The methods, when combined, try to produce a summary that is more refined

than one confined to statistical methods alone. This study necessitated the study of the Xhosa language: this is presented in **CHAPTER THREE** of this dissertation. The nouns and verbs of the language, together with its orthography and sentence construction, were studied in detail. Chapter Three has a full explanation.

- **Develop a prototype summarizer for the Xhosa language that will serve as a model for Xhosa news text summarizer.**

After careful consideration of the language based rules and also by choosing the appropriate methods to use, this was done. **CHAPTER FOUR** of this dissertation explains the Methodology carried out in this work. The implementation took place with the integration of all the components to produce a fully functional prototype. **CHAPTER FIVE** of this dissertation has a full explanation of the implementation phase of the Xhosa text summarizer.

- **Develop a test set of texts to evaluate the system.**

A test set of texts for this study was collected and prepared. The test data that was collected was taken from online Xhosa news websites called isiGidimi.co.za and Nalibali.org. The websites publish news purely in the Xhosa language. About two hundred news items from various news categories were prepared. The news items were taken to Xhosa linguists for manual creation of extraction based summaries and were used to compare with the summaries produced by the automatic news summarizer. The results, which were obtained, are explained in detail in **CHAPTER SIX** of this dissertation.

7.2. Conclusion and Future Work

In the present day's information era, text summaries have become a necessity. This is because they save time, and they are a useful tool for managing the increasing amount of content that confronts people on daily basis.

This study focuses particularly on the investigation and development of a first automatic text summarizer for Xhosa news articles and the first ever for Nguni languages. In this study, both subjective and objective assessments are utilized and it has been found that using a blend of approaches, enhanced term frequency and positional methods, produces great results. The results of the evaluation show that the Xhosa Text Summarizer performed well. However there are still problems and it is yet not a perfect text summarizer, although it has contributed to the Xhosa community and to the country as a whole, by filling an existing gap.

limitations of the Xhosa summarizer such as the inability of controlling the length of the summary and inability of selecting constituents smaller than a sentence,however, evaluation of the Xhosa

summarizer against an ideal human constructed summaries has proved that out summarizer gives better performance.

The investigation and implementation of the prototype Xhosa Text Summarizer was successful. As the system is still in a prototype state, and like any other summarizers developed for other languages it has drawbacks that need careful attention, especially in future.

In the following paragraphs, some future improvements are addressed that need to be included in the implementation of the next version of the Xhosa Text Summarizer:
Incoherencies happen specifically when the summaries are below 30% of the original texts and, for example, when a pronoun reference hangs free with no reference in the text. The pronoun resolution methods are of crucial consideration to make the summary more coherent. Pronoun resolution will resolve the pronouns in the text and replace them with the original noun when necessary.

The stemmer used in the prototype is a lightweight stemmer. Some additional strict rules are needed. These will come from a thorough study of deeper linguistic rules. These rules are needed to make the summarizer more useful for a broad NLP community. In addition, the continued development of Xhosa morphology means that the stemmer will need continuous revision in the future to accommodate such new words.

This study serves as a benchmark for undertaking research in automatic text summarization for South African languages. In future, it would be good if the summarizer could be web based, to make it available to a wide public audience. In this regard, advanced methods like abstract based summarization methods are recommended. Abstract methods give a summary a more coherent structure, and the semantics of the language are kept.

.

REFERENCES

[1] DINEGDE, GIRMA DEBELE. 'Afan Oromo News Text Summarizer.' MASTER OF SCIENCE. ADDIS ABABA UNIVERSITY, 2012. Print.

[2] Mani, I., & Maybury, M. T. (1999). Advances in Automatic Text Summarization. (I. Mani & M. T. Maybury, Eds.)Computational Linguistics (Vol. 26). MIT Press. http://doi.org/10.1162/coli.2000.26.2.280

[3] Pachantouris G. and Dalianis H. (2005), "Greek Sum A Greek Text Summarizer," Word Journal of the International Linguistic Association, pp. 1-45.

[4] Dinegde, G. D., & Tachbelie, M. Y. (2014). Afan Oromo News Text Summarizer, 103(4), 1–6.

[5] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159{165. [2, 3, 6, 8].

[6] Mani I., Bloedorn E.and Gates B. (1998), "Using cohesion and coherence models for text

[7] R. Boguraev, "Evolutionary Algorithm for Extractive Text Summarization," Intell. Inf. Manag., vol. 1, no. 2, pp. 128–138, 2009

[8] Borko, H., & Bernier, C. L. (1975). Abstracting Concepts and Methods. San Diego, California: Academic Press.

[9] Ganapathiraju M. (2002), "Relevance of Cluster size in MMR based Summarizer", a report in proceedings of the Second NTCIR Workshop.

[10] Schlesinger J. and Baker D. (2001), "Using Document features and Statistical Modelling toShell, M. (2014).Summarization", In AAAI 98 Spring Symposium on Intelligent text summarization Swedish. In Proceedings of NODALIDA 03 - 14th Nordic Conference on Computational Linguistics, May 30-31 2003, Uppsala, Sweden.

[11] Manabu O. and Hajime M. (2000), "Query-Biased summarization based on lexical chaining", Computational Linguistics, 16, 578-585.

[12] E. Hovy and C-Y Lin. 1997. Automated Text Summarization in SUMMARIST. In Proceedings of the Workshop of Intelligent Scalable Text Summarization, July.

[13] E. Hovy, E. Hovy, C.-Y. Lin, and C.-Y. Lin, “Automated Text Summarization in \textscsummarist,” pp. 81–94, 1999.

[14] H. P. Edmundson, “New methods in automatic extracting,” *J. Assoc. Comput. Mach.*, vol. 16, no. 2, pp. 264–285, 1969.

[15] Teufel S. and Moens M. (1997), “Sentence extraction as a classification task”, Proceedings of the ACL Workshop on Intelligent Text Summarization, Madrid.

[16] Z. Ahmed, R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, C. A. Mehdiyev, R. Al-hashemi, E. Branny, J. M. Conroy, D. P. O. Leary, , G. Pachantouris, H. Dalianis, “The acquisition of subject agreement in Xhosa,” Proc. 2nd Conf. Gener. Approaches to Lang. Acquis. North Am., vol. 2, no. 2, pp. 114–123, 2007.

[17] Kupiec J., Pedersen J. and Chen F. (1995), “A trainable document summarizer”, Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR).

[18] E. Hovy, E. Hovy, C.-Y. Lin, and C.-Y. Lin, “Automated Text Summarization in \textscsummarist,” pp. 81–94, 1999.

[19] Baxendale, P. (1958). Machine-made index for technical literature - an experiment. IBM Journal of Research Development, 2(4):354{361. [2, 3, 5].

[20] J. M. Conroy, D. P. O. Leary, “Text Summarization via Hidden Markov Models ,”

Eckroth, J. et al., 2012. NewsFinder : Automating an AI News Service. , pp.43–54.

[21] Daniel Marcu. Discourse-based summarization in duc-2001. In DUC01, New Orleans, LA, 2001.M. Halliday and R. Hasan. Cohesion in English. Longman, London, 1976.

[22] Halliday, M. A. K & Hasan, R. 1976. Cohesion in English. English Language Series, Longman, London.

[23] Parks, William (2009). 'BASIC NEWS WRITING.' http://www.ohlone.edu/people/bparks/. N.p, Web. 29 Oct. 2015. Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization.

[24] Brandow R., Mitze K.and Rau L. (1995), “Automatic condensation of electronic publications By sentence selection”, Information processing and Management.

[25] Saul, Zandisile Wilberforce, (2013). The significance of accuracy in the orthographical development of isiXhosa in a post-democratic South Africa. 1st ed. Alice, South Africa: University of Fort Hare.

[26] Swart, P.H., 2000. PROSODIC FEATURES OF IMPERATIVES IN XHOSA : IMPLICA TIONS FOR A TEXT -TO-SPEECH SYSTEM by. , (March).

[27] Vuyokazi Sylvia Nomlomo, 1993. ve rs ity e w ve rs ity e To w. UNIVERSITY OF CAPE TOWN LANGUAGE.

[28] Lmp.ucla.edu, (2015). UCLA Language Materials Project: Language Profile. [Online] Available at: http://www.lmp.ucla.edu/Profile.aspx?LangID=21&menu=004 [Accessed 9 Nov. 2015].

[29] Europa Publications. 1993. "South Africa" in the Europa World Year Book 1993, Vol. 2, 2567 2591. London: Europa Publications Limited.

[30] FROMKIN, V. and RODMAN, R. 1988. An introduction to the language, fourth edition. London: Holt, Rinehart, and Winston.

[31] YULE, G. 1996. The study of language: an introduction. Cambridge: Cambridge University Press.

[32] TheFreeDictionary.com, 'Vowel'. N.p., 2015. Web. 29 Oct. 2015.

[33] Zerbian, S. (2004). Phonological Phrases in Xhosa (Southern Bantu) 1, 71–99.

[34] Vanderstouwe, C. (2009). A phonetic and phonological report on the Xhosa language.

[35] WESTERMANN, D. & WARD, I.C. 1964. *Practical phonetics for students of African languages*. Oxford: Oxford University Press.

[36] Z. W. SAUL, “THE SIGNIFICANCE OF ACCURACY IN THE ORTHOGRAPHICAL,” Univeristy of Fort Hare, 2013.

[37] Satyo, S. (1988). Igrama noncwadi lwesixhosa ibanga 10. Via Afrika Limited.

[38] S. Gxilish, P. De Villier, and J. De Villiers, “The acquisition of subject agreement in Xhosa,” Proc. 2nd Conf. Gener. Approaches to Lang. Acquis. North Am., pp. 114–123, 2007.

[39] Mtuze, P. T., Tshabe, S.L., Putu, B.D., Mini, B.M. & Mkonto, N.V. (1987). IsiXhosa Sezikhuthali. De Jager-Haum.

[40] Nltk.org, (2015). 5. Categorizing and Tagging Words. [Online] Available at: http://www.nltk.org/book/ch05.html [Accessed 13 Nov. 2015].

[41] Mletshe, L. K. (2010). DEVERBAL NOMINALS IN XHOSA. Stellenbosch University.

[42] M. Nogwina and Z. Shibeshi, “Towards Developing a Stemmer for the IsiXhosa Language Towards Developing a Stemmer for the IsiXhosa Language,” no. June, 2015.

[43] Statistics South Africa (2012) Available at: http://www.statssa.gov.za/census2011/default.asp.

[44] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O’Reilly Media Inc.

[45] Madnani, N. (2015). Getting Started on Natural Language Processing with Python. 1st ed. [eBook] pp.3-4. Available at: http://desilinguist.org/pdf/crossroads.pdf [Accessed 11 Nov. 2015].

[46] Nltk.org, (2015). Corpus Readers. [Online] Available at: http://www.nltk.org/howto/corpus.html#corpus-reader-objects [Accessed 12 Nov. 2015].

[47] About.com Education, 'Correct and Effective Punctuation & Mechanics.' N.p., 2015. Web. 29 Oct. 2015.

[48] FINEGAN, E. 2008. Language: its structure and use, fifth edition. United States

of America: Thomson Wadsworth.

[49] FINEGAN, E., and BESNIER, N. 1989. Language: its structure and use. New York: Harcourt Brace Jovanovich Publishers.

[50] Bennie, W.G. 1953. A grammar of Xhosa for the Xhosa-speaking. Lovedale: The Lovedale Press.

[51] Kavita-ganesan.com, (2015). *ROUGE 2.0 - Java Package for Evaluation of Summarization Tasks with Updated ROUGE Measures | Kavita Ganesan*. [Online] Available at: http://kavita-ganesan.com/content/rouge-2.0 [Accessed 25 Nov. 2015].

[52] Stanford NLP Group, "Stanford log-linear part of speech tagger", Available at: Http://nlp.stanford.edu/software/tagger.shtml [June 21, 2016].

[53] M. Pascoe and M. Smouse, "Masithethe : Speech and language development and difficulties in isiXhosa," vol. 102, no. 6, pp. 469–471, 2012.

[54] "Appendix:Xhosa parts of speech - Wiktionary", En.wiktionary.org, 2017. [Online]. Available: https://en.wiktionary.org/wiki/Appendix:Xhosa_parts_of_speech. [Accessed: 13- Jun- 2016].

[55] S. Spiegler and P. A. Flach, "Ukwabelana - An open-source morphological Zulu corpus," no. August, pp. 1020–1028, 2010.

[56] I. Mani, T. Mitre, S. Hills, and R. Reston, "Summarization Evaluation : An Overview," 1960.

[57] I. Mani, T. Mitre, S. Hills, and R. Reston, " Summarization Evaluation: An Overview,", International Journal of Computer Applications, vol. 171, no. 10, pp. 1-17, 1960.

[58] B., L. and Venkata, P. (2017). An Overview of Text Summarization. International Journal of Computer Applications, 171(10), pp.1-17.

[59] R. Alguliev and R. Aliguliyev, "Evolutionary Algorithm for Extractive Text Summarization," vol. 2009, no. November, pp. 128–138, 2009.

[60] J. Li, L. Sun, C. Kit, and J. Webster, "A query-focused multi-document

summarizer based on lexical chains," Proceedings of the Document Understanding Conference (DUC'07), New York, USA, 4p, 26–27 April 2007.

[61] I. Mani, Automatic Summarization. John Benjamins, 2001.

[62] L. Pretorius and S. Bosch, "Exploiting Cross-linguistic Similarities in Zulu and Xhosa Computational Morphology," no. March, pp. 96–103, 2009.

[63] (http://www.lyberty.com/encyc/articles/abbr.html (Accessed 19/10/2012).

[64] P. H. Swart, "PROSODIC FEATURES OF IMPERATIVES IN XHOSA : IMPLICA TIONS FOR A TEXT -TO-SPEECH SYSTEM by," no. March, 2000.

[65] P. Kese, "TEACHING NON-INDIGENOUS SPEAKERS OF ISIXHOSA : A CRITICAL EVALUATION OF OWN PRACTICE," vol. 28, no. 1, pp. 92–110, 2012.

[66] (http://grammar.about.com/od/punctuationandmechanics).(Accessed 08/08/2015).

LIST OF APPENDIXES

Appendix A: List of Publications

List of publications related to this research.

Ndyalivana, Z, & Shibeshi Z. (2014). *Xhosa Text Summarizer: An Extractive Based Automatic Text Summarizer for Xhosa News Articles*. *WIP*. SATNAC Conference 2014, 31 August -03 September 2014, Boardwalk, Port Elizabeth, Eastern Cape, South Africa.

Ndyalivana, Z; Shibeshi, Z & Christofel Botha. (2016). *IsiXhoSum: An Extractive Based Automatic Text Summarizer for Xhosa News Items. Full Paper*. SATNAC Conference 2016, 4-7 September 2016 Fan court, George, Western Cape, South Africa.

Appendix B: Xhosa Stemmed Nouns and Verbs

Nouns	**Corresponding Stems**
1. abafo	bafo
2. abakhwetha	khwetha
3. abantu	ntu
4. abathwa	thwa
5. abazala	zala
6. abazali	zali
7. abembi	mbi
8. aboni	oni
9. afrika	afrika
10. amacholi	choli
11. amaciko	ciko
12. amacirha	cirha
13. amafokofoko	fokofoko
14. amahayihayi	hayihayi
15. amanzana	nz
16. amaqaqa	qaqa
17. amashwa	shwa
18. amathumba	thumba
19. amatye	tye
20. amawa	wa
21. amaxaxaviti	xaxavithi

22. amaXhosa	Xhosa
23. amayeye	yeye
24. amazwi	zwi
25. asibantu	bantu
26. asibantu	bantu
27. asingobantu	bantu
28. asingobantu	bantu
29. asingomakhwenkwe	makhwenkwe
30. ayingobantu	bantu
31. bantu	bantu
32. ebantwini	ntwini
33. emlanjeni	lanjeni
34. emlanjeni	lanjeni
35. emthonjeni	thonjeni
36. emthonjeni	thonjeni
37. entabeni	tabeni
38. entabeni	tabeni
39. fa	fa
40. gugulethu	gugulethu
41. ibakwana	bakw
42. ibandla	bandla
43. ibanga	banga
44. ibanjwa	banjwa
45. ibatha	batha
46. ibekelo	bekelo
47. ibele	bele
48. ibhinqa	bhinqa
49. ibhobhile	bhobhile
50. icabanga	cabanga
51. icala	cala
52. icamagu	camagu
53. icandelo	candelo
54. icawa	cawa
55. iceba	ceba
56. icebiso	cebiso
57. icekwa	cekwa
58. icephe	cephe
59. iceya	ceya
60. ichele	chele

61. ichitywa	chitywa
62. ichule	chule
63. icongwane	congwane
64. idangatye	dangatye
65. idelakufa	delakufa
66. idlavu	dlavu
67. idlelo	dlelo
68. idliso	dliso
69. idlolo	dlolo
70. idobo	dobo
71. idolophana	Doloph
72. idwalana	dwala
73. idyudyu	dyudyu
74. ifanankosi	fanankosi
75. ifokofoko	fokofoko
76. igaba	gaba
77. igalelo	galelo
78. igazi	gazi
79. igcisa	gcisa
80. igeza	geza
81. igqirhakazi	qqirha
82. ihempe	hempe
83. ihlaba	hlaba
84. ihlaba	hlaba
85. ihoko	hoko
86. iilwimi	lwimi
87. iimbabala	mbabala
88. iimboya	mboya
89. iindima	dima
90. iindywala	dywala
91. iinkawu	nkawu
92. iitafile	tafile
93. ijaja	jaja
94. ijaji	jaji
95. ijengxeba	jengxeba
96. ijikazi	ji
97. ikamva	kamva
98. ikhaba	khaba
99. ikhakhasholo	khakhasholo

100.ikhangala	khangala
101.ikhatshu	khatshu
102.ilahleko	lahleko
103.ilalela	lalela
104.ilangatya	langatya
105.ilaphu	laphu
106.ilifana	fa
107.ilishwa	shwa
108.ilitye	tye
109.iliwa	wa
110.iliwa	wa
111.ilizwi	zwi
112.imbabala	babala
113.imfeketho	feketho
114.imibhalo	bhalo
115.imililo	lilo
116.imithi	thi
117.imithombo	thombo
118.imizi	zi
119.imvaba	vaba
120.imvana	v
121.imvelo	velo
122.incebakazi	ceba
123.inceku	ceku
124.incindi	cindi
125.incwadana	cwad
126.indawo	dawo
127.indebe	debe
128.indlamanzi	dlamanzi
129.indlazi	dlazi
130.indlezana	dlez
131.indlovukazi	dlovu
132.indodana	dod
133.indulana	ndul
134.indwadunge	dwadunge
135.indwalutho	dwalutho
136.inembana	emb
137.ingcathu	gcathu
138.ingxangxa	gxangxa

139.ingxangxasi	gxangxasi
140.ingxangxosi	gxangxosi
141.injakazi	nja
142.inkawu	kawu
143.inkobonkobana	kobonkob
144.inkobonkobo	kobonkobo
145.inkondekazi	konde
146.inkosikazi	kosi
147.inqilo	qilo
148.intabakazi	taba
149.intando	tando
150.into	to
151.intso	tso
152.ipasika	pasika
153.ipasile	pasile
154.ipatyutyu	patyutyu
155.iphanyazo	phanyazo
156.iqabakazi	qaba
157.iqabaza	qabaza
158.iqadi	qadi
159.iqaqa	qaqa
160.iqaqoba	qaqoba
161.iqegu	qegu
162.iqela	qela
163.iqobokazana	iqobo
164.irhabasa	rhabasa
165.irhafu	rhafu
166.irhamncwa	rhamncwa
167.irharha	rharha
168.irhengqe	rhengqe
169.irhewu	rhewu
170.isabatha	abatha
171.isabelo	abelo
172.isabelo	abelo
173.isabhalala	abhalala
174.isabhokhwe	abhokhwe
175.isabhongo	abhongo
176.isabhunge	abhunge
177.isabhungqu	abhungqu

178.isabile	abile
179.isacholo	acholo
180.isaci	aci
181.isadunge	adunge
182.isadyenge	adyenge
183.isadyudyu	adyudyu
184.isamani	amani
185.isanda	anda
186.isandanda	andanda
187.isandle	andle
188.isandulela	andulela
189.isazi	azi
190.iselwa	elwa
191.ishwangusha	hwangusha
192.isibaluli	baluli
193.isibeleko	beleko
194.isibetho	betho
195.isibhakabhaka	bhakabhaka
196.isibhanxa	bhanxa
197.isibhaxu	bhaxu
198.isibotshwa	botshwa
199.isicaka	caka
200.isicengo	cengo
201.isichenene	chenene
202.isichenge	chenge
203.isichopho	chopho
204.isichotho	chotho
205.isichumiso	chumiso
206.isicithi	cithi
207.isicithi	cithi
208.isidabane	dabane
209.isidala	dala
210.isidalwa	dalwa
211.isidamba	damba
212.isidanga	danga
213.isidenge	denge
214.isidikimfa	dikimfa
215.isidima	dima
216.isidlangalala	dlangalala

217.isidlo	dlo
218.isidlokolo	dlokolo
219.isidlwengu	dlwengu
220.isidodo	dodo
221.isidudla	dudla
222.isifede	fede
223.isifingo	fingo
224.isifinyezo	finyezo
225.isihlanganisi	hlanganisi
226.isihlangu	hlangu
227.isihlava	hlava
228.isihlomelo	hlomelo
229.isikhaka	khaka
230.isikhakhamela	khakhamela
231.isikhala	khala
232.isikhalazo	khalazo
233.isikhali	khali
234.isikhutali	khutali
235.isimahla	mahla
236.isimamva	mamva
237.isimelabizo	melabizo
238.isinayimesi	nayimesi
239.isinciphiso	nciphiso
240.isipeliti	peliti
241.isiphako	phako
242.isiphango	phango
243.isiqabu	qabu
244.isiqhamo	qhamo
245.isiqhazolo	qhazolo
246.isishiqi	shiqi
247.isishuba	shuba
248.isisila	sila
249.isitha	tha
250.isithako	thako
251.isithili	thili
252.isithilikazi	thili
253.isithombo	thombo
254.isithonga	thonga
255.isithunywa	thunywa

256.isitiya	tiya
257.isiwiliwili	wiliwili
258.isixando	xando
259.isiza	za
260.isizamva	zamva
261.isizananina	zananina
262.isizekabani	zekabani
263.isiziba	ziba
264.isizungu	zungu
265.isizwe	zwe
266.isizwekazi	zwe
267.isomelezo	omelezo
268.isonka	onka
269.isono	ono
270.itafile	tafile
271.ithambeka	thambeka
272.ithambo	thambo
273.ithemba	themba
274.ithimla	thimla
275.ithokazi	tho
276.ithongo	thongo
277.ithumba	thumba
278.itshivela	tshivela
279.ivamna	vamna
280.iwaka	waka
281.ixandeka	xandeka
282.ixaxavithi	xaxavithi
283.ixethuka	xethuka
284.ixhegokazi	xhego
285.iyelenqe	yelenqe
286.izabelo	abelo
287.izabhungqu	abhungqu
288.izazi	azi
289.izibaluli	baluli
290.izibazana	baz
291.izibuko	buko
292.izicheko	cheko
293.izicithi	cithi
294.izikhuthali	khuthali

295.izimvo	mvo
296.izinayimesi	ayimesi
297.izinto	nto
298.izintso	tso
299.iziqhamo	qhamo
300.izitiya	tiya
301.izizwe	zwe
302.izola	zola
303.kubantu	bantu
304.ndebele	ndebele
305.ngabantu	bantu
306.ngabantu	bantu
307.ngamakhwenke	makhwenke
308.nja	nja
309.nomasele	nomasele
310.oogugulethu	gugulethu
311.oonyoko	nyoko
312.somandla	somandla
313.somanyala	somanyala
314.sonininanini	sonininanini
315.sosuthu	sosuthu
316.sotheko	sotheko
317.ubabalo	babalo
318.ubafazini	afazini
319.ubandlululo	andlululo
320.ubende	ende
321.ubengezelo	engezelo
322.ubengu	bengu
323.ubhengxeshe	hengxeshe
324.ubhobhoyi	hobhoyi
325.ubhontsi	hontsi
326.uboya	oya
327.ububanzi	banzi
328.ububele	bele
329.ubuchopo	chopo
330.ubuchule	chule
331.ubuchwepheshe	chwepheshe
332.ubuciko	ciko
333.ubude	de

334.ubudoda	doda
335.ubufutshane	futshane
336.ubuhilihili	hilihili
337.ubuhlobo	hlobo
338.ubukhulu	khulu
339.ubukhwenkwe	khwenkwe
340.ubulembu	lembu
341.ubumanzi	manzi
342.ubungqoqo	ngqoqo
343.ubusi	si
344.ubusika	sika
345.ucaca	caca
346.ucanzibe	canzibe
347.uchuku	chuku
348.uchulumanco	chulumanco
349.ucikicane	cikicane
350.ucingo	cingo
351.udaba	daba
352.udakada	dakada
353.udlalani	dlalani
354.udlezinye	dlezinye
355.udlomdlayo	dlomdlayo
356.udlwabevu	dlwabevu
357.ufefe	fefe
358.ugcado	gcado
359.ugovo	govo
360.ugugulethu	gugulethu
361.uhlaselo	hlaselo
362.ujingi	jingi
363.ujobela	jobela
364.ukona	ona
365.ukudliwa	dliwa

Verbs	**Stem**
1. abuhlungu	hlungu
2. benzani	nzani
3. iyabhabha	bhabha
4. ibolile	ibolile
5. bukekela	bukekela

6. butha	butha
7. dada	dada
8. dubula	dubula
9. funda	funda
10. fundisa	fundisa
11. galela	galela
12. gugutha	gugutha
13. uyagungxula	uyagungxula
14. uyaguqula	uyaguqula
15. ndimhlebele	ndimhlebele
16. hlumela	hlumela
17. ibuzwa	ibuzwa
18. iyaluma	iyaluma
19. jika	jika
20. kekela	kekela
21. khaba	khaba
22. Khawundixelele	Khawundixelele
23. khula	khula
24. khulula	khulula
25. khwankqisa	khwankqisa
26. kubuya	kubuya
27. kuyaduduma	kuyaduduma
28. kwanqisseka	kwanqisseka
29. uyaloba	loba
30. uyalamba	lamba
31. landa	landa
32. landa	landa
33. lifikile	fikile
34. ziyalima	lima
35. uzalinda	linda
36. luphumela	phumela
37. ndimamele	mamele
38. iyamangalisa	mangalisa
39. ayampompoza	mpompoza
40. Ndibone	Ndibon
41. Ndiyacela	cela
42. Ndiyayifuna	yifuna
43. Ndizohlala	hlala
44. ndikhetha	ndikhetha
45. ndilinde	ndilinde
46. ndimamele	ndimamele
47. Ndimbonile	Ndimbonile

48. ndinike	ndinike
49. Ndisebenza	Ndisebenza
50. Ndivela	Ndivela
51. ndivile	ndivile
52. ndixolele	ndixolele
53. ndiyambona	ndiyambona
54. Ndiyayazi	yazi
55. ndiyosela	sela
56. ndiyozela	zela
57. Ndizakubona	bona
58. uyangena	ngena
59. uyothusa	othusa
60. uyaphaka	phaka
61. uyaphefumla	phefumla
62. uyaphehlelela	phehlelela
63. uyaPheka	Pheka
64. phuphumala	phuphuma
65. qukuqela	qukuqela
66. sabela	sabela
67. sasaza	sasaza
68. sebenza	sebenza
69. uyasela	sela
70. Siyabulisa	bulisa
71. Sobonana	Sobona
72. iyasuka	suka
73. swelekile	kile
74. uyateleka	teleka
75. uyandithanda	thanda
76. uyathetha	thetha
77. yandithuka	thuka
78. uthule	thule
79. thumela	thumela
80. thutha	thutha
81. tsalela	tsalela
82. tyhila	tyhila
83. ukwazi	ukwazi
84. ulele	ulele
85. undixelele	undixelele
86. Uyaqala	qala
87. uqalile	qalile
88. usebenza	sebenza
89. utyelele	tyelele

90. uvela	uvela
91. uyagoduka	goduka
92. uyabala	bala
93. uyabaleka	baleka
94. uyabelka	belka
95. uyabona	bona
96. uyabulala	bulala
97. uyabulela	bulela
98. uyabulisa	bulisa
99. uyabuya	buya
100. uyacela	cela
101. uyadada	dada
102. uyagula	gula
103. uyahambisa	hambisa
104. uyajikela	jikela
105. uyakhazimla	khazimla
106. uyalila	lila
107. uyalima	lima
108. uyalingana	ngana
109. uyandibetha	betha
110. uyandibethelela	bethelela
111. uyandifuna	ndifuna
112. uyandijonga	jonga
113. uyapheka	pheka
114. uyaphupha	phupha
115. uyathengaisa	thengaisa
116. uyathengisa	uyathengisa
117. uyandithethisa	thethisa
118. uyeza	uyeza
119. uyavala	vala
120. yavela	vela
121. uyavula	vula
122. uyavuma	vuma
123. iyavuza	vuza
124. uyavuselela	vuselela
125. yima	yima
126. uzimela	zimele
127. uyasimelela	simelela

Appendix C: The Xhosa Stop Word List

Words

1. ne
2. nhe
3. ukuze
4. ngaphezulu'
5. ngaphantsi
6. ngaphambi
7. emva
8. kwaye
9. ethe
10. phambi kwe
11. emva kwe
12. ude ne
13. kufuphi
14. phakathi
15. ngaphakathi
16. ngaphandle
17. kanye
18. ngaphandle
19. ge Audio
20. hakathi
21. aye
22. pha
23. kodwa
24. yathi
25. sabo
26. awe
27. futhi
28. yenye
29. ye
30. ke
31. suka
32. phambi
33. kuya
34. kodwa
35. njani
36. intoni
37. ubani
38. ngoba
39. Phi?

40. bona
41. ngoku
42. ukuba
43. mbo
44. amhlanje
45. ebusuku
46. ngomso
47. kungekudala
48. noba
49. inoba
50. abanye
51. ngokukhawuleza
52. kancinane
53. no
54. nabanye
55. nani
56. inokuba
57. isenokwenzek
58. nam
59. naye
60. wakhe
61. wake
62. kunye
63. sonke
64. injalo
65. obabini
66. kakhulu
67. phantse
68. yase
69. oloko
70. uqhele
71. ngamanye
72. amaxesha
73. hayi
74. rhoqo
75. soze
76. zatywala
77. mna
78. lwazo

79. iye
80. kwabe
81. be
82. komnye
83. kukhona
84. umo
85. osuka
86. mba
87. lwaze
88. lakhe
89. kuwonke
90. xaki
91. waka
92. yase
93. sithi
94. ngelo
95. esise
96. ziz
97. nanku
98. kwezi
99. bobunye
100.nokunye
101.ka
102.ngale
103.ngantoni
104.joost
105.wena
106.lithe
107.kutsha
108.anee
109.okokuba
110.ena
111.zabo
112.ngoku
113.izinto
114.bona
115.nabo
116.into
117.nina

118.yiza
119.apha
120.yena
121.yena
122.thina
123.wethu
124.bonke
125.bona
126.okanye
127.eyam
128.yakho
129.yakhe
130.yena
131.yethu
132.yabo
133.kwi
134.inye
135.ndini
136.zona
137.zabo
138.zonke
139.zimbini
140.zintathu
141.zine
142.zintlanu
143.zintandathu
144.isixhenxe
145.sisibhozo
146.lithoba
147.lishumi
148.okokuqala
149.okwesibi

Subjective Evaluation Results

News item one up to news item fifteen are items that are taken to the evaluators of the systems. Ten evaluators evaluated the system summaries. We call them **Evaluator One(EvaOne)**, **Evaluator Two(EvaTwo)**, **Evaluator Three(EvaThree)**, **Evaluator Four(EvaFour)**, **Evaluator Five(EvaFive)**. M1 and M2 are the two methods that are used in this study Manual

(M1) and automatic (M2) methods. The evaluators were asked question and the results are presented in the following tables etc.

Which summary do you think is better?

TID	**Total(M1 and M2)**	**M1**	**M2**	**M1**	**M2**
News one	M2(3),M1(2)	0,40	0,60	40%	60%
News two	M2(4),M1(1)	0,20	0,80	20%	80%
News three	M2(3),M1(2)	0,40	0,60	40%	60%
News four	M2(3),M1(2)	0,40	0,60	40%	60%
News five	M2(4),M1(1)	0,80	0,20	80%	20%
News six	M2(4),M1(1)	0,80	0,20	80%	20%
News seven	M2(1),M1(4)	0,80	0,20	80%	20%
News eight	M2(3),M1(2)	0,40	0,60	40%	60%
News nine	M2(4),M1(1)	0,20	0,80	20%	80%
News ten	M2(4),M1(1)	0,80	0,20	80%	20%
News eleven	M2(1),M1(4)	0,20	0,80	20%	80%
News twelve	M2(1),M1(4)	0,20	0,80	20%	80%
News thirteen	M2(3),M1(2)	0,40	0,60	40%	60%
News fourteen	M2(4),M1(1)	0,20	0,80	20%	80%

News one	M2(4),M1(1)	0,80	0,20	80%	20%
Average		**0,47**	**0,53**	**47%**	**53%**

In which one the most important information is being kept?

TID	**EvaOne**	**EvaTwo**	**EvaThree**	**EvaFour**	**EvaFive**	**Result**	**M1**	**M2**
News one	M2	M2	M1	M2	M2	M2(1),M1(4)	**0,80**	**0,20**
News two	M1	M2	M2	M1	M1	M2(2),M1(3)	**0,60**	**0,40**
News three	M2	M1	M1	M2	M2	M2(2),M1(3)	**0,60**	**0,40**
News four	M2	M2	M2	M2	M1	M2(3),M1(2)	**0,40**	**0,60**
News five	M1	M1	M2	M2	M2	M2(3),M1(2)	**0,40**	**0,60**
News six	M2	M1	M2	M2	M2	M2(1),M1(4)	**0,80**	**0,20**
News seven	M1	M1	M2	M2	M2	M2(3),M1(2)	**0,40**	**0,60**
News eight	M2	M2	M2	M2	M1	M2(2),M1(3)	**0,40**	**0,60**
News nine	M1	M2	M2	M2	M1	M2(2),M1(3)	**0,60**	**0,40**
News ten	M1	M1	M1	M2	M2	M2(2),M1(3)	**0,60**	**0,40**
News eleven	M1	M2	M1	M2	M2	M2(3),M1(2)	**0,40**	**0,60**
News twelve	M1	M1	M1	M2	M2	M2(5),M1(0)	**-**	**1,00**
News thirteen	M1	M2	M2	M2	M1	M2(3),M1(2)	**0,40**	**0,60**
News fourteen	M2	M2	M2	M1	M2	M2(4),M1(1)	**0,20**	**0,80**
News fifteen	M2	M2	M2	M1	M1	M2(4),M1(2)	**0,20**	**0,80**

							0,45	0,55

Which summary is more coherent?

TID	**Total(M1 and M2)**	**M1**	**M2**	**M1**	**M2**
News one	M2(3),M1(2)	0,40	0,60	40%	60%
News two	M2(4),M1(1)	0,20	0,80	20%	80%
News three	M2(3),M1(2)	0,40	0,60	40%	60%
News four	M2(3),M1(2)	0,40	0,60	40%	60%
News five	M2(4),M1(1)	0,80	0,20	80%	20%
News six	M2(4),M1(1)	0,80	0,20	80%	20%
News seven	M2(1),M1(4)	0,80	0,20	80%	20%
News eight	M2(3),M1(2)	0,40	0,60	40%	60%
News nine	M2(4),M1(1)	0,20	0,80	20%	80%
News ten	M2(4),M1(1)	0,80	0,20	80%	20%
News eleven	M2(1),M1(4)	0,20	0,80	20%	80%
News twelve	M2(1),M1(4)	0,20	0,80	20%	80%
News thirteen	M2(3),M1(2)	0,40	0,60	40%	60%
News fourteen	M2(4),M1(1)	0,20	0,80	20%	80%
News one	M2(4),M1(1)	0,80	0,20	80%	20%
Average		**0,47**	**0,53**	**47%**	**53%**

Out of a scale from 1-5, where five is the best, what score would you assign to each summary?

TID	**EvaOne**	**EvaTwo**	**EvaThree**	**EvaFour**	**EvaFive**	**Result**	Decmals	Percentage
News one (M1)	3	1	5	1	3	13	0,663265306	66%
News one (M2)	5	3	4	4	1	17	0,867346939	87%
News two (M1)	3	4	5	4	3	19	0,969387755	97%
News two (M2)	4	5	3	1	4	17	0,867346939	87%
News three(M1)	2	4	4	3	4	17	0,867346939	87%
News three(M2)	5	5	3	2	3	18	1,224489796	92%
News four(M1)	4	4	4	3	3	18	0,969387755	92%
News four(M2)	5	2	5	1	5	18	1,173469388	92%
News five(M1)	3	1	5	1	4	14	0,714285714	71%
News five(M2)	4	5	3	3	4	19	0,969387755	97%
News six(M1)	2	4	4	4	4	18	0,918367347	92%
News six(M2)	4	3	4	4	4	19	1,173469388	97%
News seven(M1)	3	4	4	4	3	18	0,918367347	92%
News seven(M2)	5	2	5	2	4	18	0,918367347	92%
News eight(M1)	3	5	3	4	4	19	0,969387755	97%
News eight(M2)	5	4	2	3	3	17	0,867346939	87%

News nine(M1)	1	1	3	3	4	12	0,612244898	61%
News nine(M2)	5	5	2	2	4	18	1,020408163	92%
News ten(M1)	2	3	4	3	4	16	0,816326531	82%
News ten(M2)	5	5	1	3	5	19	1,020408163	97%
News eleven(M1)	4	1	2	4	3	14	0,714285714	71%
News eleven(M2)	5	4	3	4	3	19	0,969387755	97%
News twelve(M1)	3	3	4	3	4	17	0,867346939	87%
News twelve(M2)	5	4	5	1	2	17	0,867346939	87%
News thirteen(M1)	3	5	4	3	4	19	0,969387755	97%
News thirteen(M2)	1	4	2	4	3	14	0,714285714	71%
News fourteen(M1)	3	4	3	3	3	16	0,816326531	82%
News fourteen(M2)	4	3	4	3	4	18	0,918367347	92%
News fifteen(M1)	2	4	5	3	3	17	0,867346939	87%
News fifteen (M2)	2	1	2	3	2	10	0,510204082	51%

Appendix D: Comparison of the Methods in keeping the first sentence

ID	**The first sentence kept. M2**
News one	Yes
News two	Yes
News three	No
News four	No
News five	Yes
News six	Yes
News seven	No
News eight	Yes
News nine	Yes
News ten	Yes
News eleven	Yes
News twelve	Yes

News thirteen	No	
News fourteen	Yes	
News fifteen	Yes	
Total(No)	5	33,0%
Total(yes)	11	73,0%

ID	**First sentence kept.M1**	
News one	**No**	
News two	**No**	
News three	**Yes**	
News four	**No**	
News five	**No**	
News six	**Yes**	
News seven	**Yes**	
News eight	**Yes**	
News nine	**Yes**	
News ten	**yes**	
News eleven	**yes**	
News twelve	**No**	
News thirteen	**No**	
News fourteen	**Yes**	
News fifteen	**Yes**	
Total(No)	**6**	**40%**
Total(yes)	**9**	**60%**

Appendix E: Objective Evaluation results

ID	**(Method 1=M1)**			**(Method 2=M2)**		
	Avg_Recall	**Avg_Precison**	**Avg_F-Score**	**Avg_Recall**	**Avg_Precison**	**Avg_F-Score**
News one	4%	3%	2%	66%	22%	12%
News two	3%	2%	2%	71%	21%	11%

News three	25%	35%	29%	34%	45%	29%
News four	8%	2%	3%	32%	71%	8%
News five	35%	35%	25%	35%	35%	39%
News six	8%	2%	3%	2%	71%	70%
News seven	35%	35%	25%	35%	49%	34%
News eight	70%	2%	3%	82%	8%	8%
News nine	34%	5%	25%	35%	29%	82%
News ten	8%	2%	3%	70%	8%	35%
News eleven	70%	35%	25%	35%	29%	70%
News twelve	34%	8%	3%	2%	8%	35%
News thirteen	34%	35%	25%	35%	29%	72%
News fourteen	7%	2%	3%	22%	70%	34%
News fifteen	34%	35%	25%	35%	34%	62%
Averages	**27%**	**16%**	**13%**	**39%**	**35%**	**40%**

Appendix F: Example Summary

Original Text

Source: www.IsiGidimi.co.za

News Item 8

Emven' kweminyaka bezama ukuba ngoosomashishini eKapa - sebevela nokuthengisa iimbadada nokucoca imizi - uLonwabo, xesh' eyokukhangela umsebenzi kwishishini leekhompyutha (eMica) wathetha into evanayo nomnikazi ngoba kakade wayefundele ezobuxhaka-xhaka eCPUT. "Umnikazi wandibuza ukuba ndingakwazi na ukuthengisa iikhomyutha. Ndathi 'ewe'. Wabuza 'phi?' Ndathi 'koomaKhayelitsha, Gugulethu, ezilokishini'." Wavuya ke lo mnikazi wa banika iikhompyutha ezimbini esithi ukuba bangazithengisa ingekapheli iveki, bangasebenzisana.Naye uLuvuyo wazithakazelela ezi ndaba, kwaye yaba nguye oyokuzithengisela ootitshala ezikolweni. Yaba kukuqala kukaSILULO. Kukwalapho baqala ukwanda, ngoo2008, bethengisela uwonke-wonke iikhompyutha; bayokuvula neevenkile zomnathawonxibelelwano (ze-intanethi). Namhlanje baneevenkilike ezili-19.Zange baphelele ekuthengiseni qha, bavule nezikolo ezili-12 zokufundisa abantu basezilokishini iindlela ngeendlela zokusebenzisa iikhompyutha. Emven' kweminyaka emithathu befundisa abahlali bade bavunywa na yi SETA. Namhlanje bafundisa abafundi abanga-2400 ngonyaka. Xeshikweni besenza uphando ngezinye izikolo zeekhompyutha bafumanisa ukuba zixabisa kakhulu, kwaye nezizinto zifundiswa kwezi zikolo zee khhompyutha yayizizinto abanokukwazi nabo ukuzifundisa. Kulapho uSigqibo, owayeyingcali kwezobuxhaka-xhaka,

wabaluncedo olukhulu khona. Into entle ngalo mbono waba bafo yile nto yokuba bazisa ezi zikolo kwiilokishi nasemaphandleni; apho abantu bakuthi bahlala khona.

Iivenkile zabo uninzi lwazo zibaziivenkile ze-intanethi nokufundisa ngeekhompyutha. Sebenwenwela nakwiMpuma Koloni. KweyoMdumba, kulonyaka, baqale ukushishina eKomani. KweyeKhala bavule enye yeevenkile zabo eMonti, pha ngakwamasipala edolophini. EGcuwa naseMthatha sebegalelekile, baqala ukushishina kulenyanga, kwaye bafika namaxabiso afikelelekayo. Umzekelo: i-intanethi yabo ibetha PHA kooma-R6 ngeyure, ukushicilela ikhasi ibaziisenti ezingama-60.uLuvuyo ucacisa ngelithi: “EKapa kukho amathala eencwadi anee khompyutha njalo-njalo kodwa eMpuma Koloni akukho nto ikhoyo. EMonti iivenkile ze-intanthi zimbalwa kakhulu. Andithethi ke ngoMthatha, inoba zimbini qha. Akukho bantu bazimiseleyo ngezi zinto pha. Thina sizimisele ukuphathela abantu bakuthi iinkonzo ezinokubasebenzela” Kulo nyaka bavule iivenkile ezintandathu. Kulo nyaka uzayo bajonge ukuvula ezingama-20 eMpuma naseNtshona Koloni.uLonwabo wongeze ngelithi: “Kwaye na le nkampani yethu sifuna ukuyi-Franchise-a. Xa kuphela lo nyaka uzayo sakube sinee venkile ezingama-50; kuzakufika ixesha lokuba sizinikezele kwabanye abaphathi. Baziqhubele nabo ooSilulo abo.”Ngelabo, eyona nto ibalulekileyo kwivenkile ye-intanethi ngumashini wokushicilela amakhasi. “Zama uku fumana owona mashini ukumgangatho ophezulu. Ungawuthengi, wuqeshe kumnikazi – ngoba uyaxabisa okumgangatho ophezulu; uhamba pha koomaR80 000. Thetha nabenzi baba mashini bokushicilela; nenze isivumelwano sokuba bazakuqeshisa umshini, bakulungisele wona xa wokonakala, njalo-njalo. Yonke into ephuma kula mashini yeyona nto izakuphathela ushishino - ngoko ke kufuneka ufumane owona mashini ungazokunika zingxaki.” Siyazi ke nathi kwesisigidimi ukuba uninzi lweevenkile zithenga oomashini abafikelelekayo boomaR1000, abo mashini babuya babanike iingxaki ezininzi kwaye bacothe nokucotha. Kucacile ukuba kufuneka umntu kufuneka azi iindidi ngeendidi zabamashini kwaye akhengele okumgangatho ophezulu, qha. Nabo ababafo bakwaSilulo baqala ngezaa ngezaasheleni zisibhozi (80c) zokushicilela amakhasi. Sibazi kakuhle kwaye uhambo lwabo luqala bengakh.

The following figures show a Xhosa summary created by Xhosa Text Summarizer at different compression rates (i.e. 10%, 40% and 50%).

Summary with 50% compression rate

Emven’ kweminyaka bezama ukuba ngoosomashishini eKapa - sebevela nokuthengisa iimbadada nokucoca imizi - uLonwabo, xesh’ eyokukhangela umsebenzi kwishishini leekhompyutha (eMica) wathetha into evanayo nomnikazi ngoba kakade wayefundele ezobuxhaka-xhaka eCPUT.uLuvuyo ucacisa ngelithi: “EKapa kukho amathala eencwadi anee khompyutha njalo-njalo kodwa eMpuma Koloni akukho nto ikhoyo.’ Ndathi ‘koomaKhayelitsha, Gugulethu, ezilokishini’.” Siyazi ke nathi kwesisigidimi ukuba uninzi lweevenkile zithenga oomashini abafikelelekayo boomaR1000, abo mashini babuya babanike iingxaki ezininzi kwaye

Summary with 40% compression rate

Emven' kweminyaka bezama ukuba ngoosomashishini eKapa - sebevela nokuthengisa iimbadada nokucoca imizi - uLonwabo, xesh' eyokukhangela umsebenzi kwishishini leekhompyutha (eMica) wathetha into evanayo nomnikazi ngoba kakade wayefundele ezobuxhaka-xhaka eCPUT.uLuvuyo ucacisa ngelithi: "EKapa kukho amathala eencwadi anee khompyutha njalo-njalo kodwa eMpuma Koloni akukho nto ikhoyo.' Ndathi 'koomaKhayelitsha, Gugulethu, ezilokishini'

Summary with 10% compression rate

Emven' kweminyaka bezama ukuba ngoosomashishini eKapa - sebevela nokuthengisa iimbadada nokucoca imizi - uLonwabo, xesh' eyokukhangela umsebenzi kwishishini leekhompyutha (eMica) wathetha into evanayo nomnikazi ngoba kakade wayefundele ezobuxhaka-xhaka eCPUT.

Appendix G: Manual Summaries

Manual summary for news item one

Umphathiswa wezothutho eMpuma Koloni unkosikazi uWeziwe Tikana uthi isebe lakhe lizawukwenza konke ukuqiniseka ukuba abantu bayafundiswa ukuba ukulwa ingozi ezindleleni luxanduva lomntu wonke. Iqela lentambula iBobby Rush laseRhawutini livule ngokusesikweni iqonga kulomsitho apho licule ingoma eyodwa ebilungiselelwe lomcimbi wokumiselwa kweliphulo, ngeenjongo zokusebenzisa umculo nje ngesixhobo sokufikelela kuye wonke ubani ochaphazeleka kwingozi zendlela zelizwe loMzantsi Afrika. Othethe egameni lesebe umnumzana uNcedo Kumbaca uthi benze inzame zokuqamba ingoma ngenjongo zokwenza banzi amalinge okulwa ingozi zendlela ngelixesha lepasika. Ngeengoma zevangeli becela amazulu ukuba angenelele ukuphungula iingozi zendlela.

Manual summary for news item two

Vuka Mz'Ontsundu: Awuboni ukuba udyojwa ngentshong'emehlweni? Ubutyebi baseNtshona bunobungozi, ingakumbi xa bukufikela ngequbuliso okanye ungaxhobanga, kuba bukukhuthaza ukuba usoloko uthe gcobho ebhotolweni. Uninzi lwethu luyayiqonda indima edlalwa ngamajelo osasazo ingakumbi iintengiso enjongo yazo ikukubumba ubudoda bomntu omnyama. Xa ubeka ingqalelo kwezi ntengiso zeBlack Label uyakufika zibonisa ityendyana eliMnyama nelisaze ngobuso elizweni .Nangona iBlack Label inesiphako (okanye ithathelwa phantsi nje ezilokishini), kwezi ntengiso iboniswa njengophawu lobutyebi kwanolwempumelelo. Nangona uninzi lwezi ntengiso lusenza ingathi lulandela ifashoni ethile, siyayazi ukuba ludala udidi oluthile lokuphila kwintsha eMnyama. Eyona nto ibuhlungu kukuba 'ubuhle' bobu 'bomi' okanye bobu 'butyebi butsha' abucaciswa ngaphandle kokugadlwa phambi kwabantu. Abantu abaye bayicinge into yokuba obu bubomi 'buyinkohliso' buyakwazi ukukukhotha buphinde bukuxathule ungalindelanga. Esingakuqondiyo thina siyiNtsha eNtsundu kukuba sibethwa ngemf'iphindiwe zezi ntengiso kuba buyasetyenzelwa obu bomi sibuboniswa kwezi ntengiso.

Manual summary for news three

Ngaphambili iPAC ityhole iANC ngokuthabathela kuyo esikhumbuzo; isithi lo mbutho ukhokheleyo ufuna ukucima igalelo le-PAC kwimbali yeli. Kaloku yi-PAC eyaququzelela imatshi eyayiqhankqalazela umthetho wokuphathwa kwamapasi eSharpeville nakwa-Langa, Olibambela mongameli kwi-PAC, uMike Muendane.

Manual summary for news four

Ndandithanda kakhulu ukudlala ibhola kodwa ndandingathandi ukusoloko ndibizwa inkomo. Lonto leyo yandenza ndingafuni ukudlala ibhola ekhatywayo Tu. Ngelakha yam ndaphela ndisiya kwesinye isikolo, isikolo saseQonce. Apho ke kwesosikolo kwakudlalwa umbhoxo ingekho ibhola ekhatywayo. Apha kulombhoxo kwakungahlekiswa ngobukhulu okanye ngobuncinci kwesiqu, kwakudlalwa qha. Enye into eyandenza ndawuthanda umbhoxo yindebe yehlabathi yango 1995 apha eMzantsi Afrika. UChester Mornay Williams's wazalwa ePaarl eNtshona Koloni ngo 1970. Emveni kokukhululwa entolongweni kuka tata uNelson Mandela kwa-nyanzeleka abantu abamhlophe bavumele abantu abamnyama bonyulwe kwiqela lama Bhokobhoko. Kwabe ke kuvuleke amasango okuba kungene iinkwenkwezi ezi njengoo, Chester Williams. Siphinde sibuyele kulento yalendebe yehlabathi ngo 1995. Wonzakala uChester sekushiyeke iinyanga itumente iqale. Kwanyenzeleka kufakwe uPieter Hendricks endaweni yakhe. Ngelishwa lika Hendricks walwa edlala neCanada waphela egxothwa kwitumente. AmaBhokobhoko abetha iFrance agagana neNew Zealand emdlalweni wamanqam. Apho ke abafana baka Francois Pienaar baphela bezintshatsheli zehlabathi, n uChester Williams's wavulela umzi ontsundu amasango wokudlalela amaBhokobhoko go1995.

Manual summary for news five

Ibhola yomboxo Inyanga yeDwarha ibiyinyanga yeCurrie Cup. Lena ke yitumente yombhoxo wamaphondo walapha eMzantsi Afrika eyaqala ngonyaka ka1892. Kulo nyaka iCurrie Cup ibigiba iminyaka engama120 ikhona. Iintshatsheli zokuqala ngonyaka ka1892 yayiyiNtshona Koloni. Umdlalo wamanqam ubeseThekwini eKings Park ngomhla wama27 kweyeDwarha. Lo mdlalo ubuphakathi kooKrebe neNtshona Koloni. OoKrebe bebesele beqalisa ukuba sisithunzela kumqeqeshi weNtshona Koloni, uAllister Coetzee. Singalibali ke ukuba ooKrebe basuka kukhiqa iiStormers (iNtshona Koloni) kwitumente yeSuper Rugby kulo nyaka kumdlalo obusandulela lona wamanqam. Into ebesingayo iphele xa, uJuan de Jongh ekore eyona try intle endakhe ndayibona kulo nyaka. Umdlalo uphele iNtshona Koloni ibethe ooKrebe ngamanqaku athi 25-18. Iqela lase Mpuma Koloni, iiKings, nalo liphumelele laphela liziintshatsheli kwisebe lesibini kwiCurrie Cup. Ekupheleni kwempelaveki yomdlalo wamanqam uHeyneke Meyer (umqeqeshi wamaBhokobhoko) ukhethe iqela lamaBhokobhoko elizakusimela ngaphesheya eYurophu.

Manual summary for news six

Ezimnyama ngenkani 'zijabulile' emva kokubetha AmaKhosi. IArsenal izibethele iFulham ngo 3 – 1. Iqela laseLiverpool lona libethe elaseAshton Villa ngo 1 – 0. Amabhokobhoko asindiswe ziipenalti ezimbini kwimizuzu yokugqibela yomdlalo wawo neqela laseArgentina phezolo (siphumelele ngo 22

– 17)… Kweminye imidlalo iiLions zidwaxushe iiBlue Bulls ngo 62 – 23, Western Province ibhungce ngo 20 – 19 kumdlalo wayo neGriquas. URik de Voest wakweli ukhiqiwe kwi US Open emven' kokuthiwa qhwi qhwi ngo 6-4, 6-2 NGU Peter Gojowczyk wase Germany, izolo. Bayaphi ooUSSASA ababenceda abantwana besikolo kwezemidlalo? Owayesakuba yintshatsheli kwezamanqinidi, uMike Tyson, uthi akakwazi ukuzinqanda kwibhelu lomsele.

Manual summary for news seven

UNgconde Balfour uxhome ibhatyi kwisikhundla sakhe sikasihlalo weBoxing South Africa. Zikhiqiwe iiSouthern Kings ZaseMpuma Koloni kwi Super 15 yombhoxo nangona zibethe iLions izolo. Owona mdlalo ubujongwe ngamehlo abomvu izolo ibingulo weChiefs ne Brumbies. Chief's ziintshatsheli zeSuper Rugby. Maritzburg Utd ibethe iPolokwane City ngo 1 - 0. Golden Arrows ne Ajax Cape Town zibambene ku 1 - 1. Undowns ibethe iCeltic ngo 3 - 1; Omnye umdlalo obujongwe ngabomvu NGU lo we Pirates ne Al Ahly. Amabhaka-khaka abuye nawo onke amanqaku kumdlalo wabo nentshatsheli zase Egypt. Amanqaku e Pirates afakwe nguNtshumayelo Ali no Myeni kwesesibini.

Manual summary for news eight

Emven' kweminyaka bezama ukuba ngoosomashishini eKapa. ULonwabo, xesh' eyokukhangela umsebenzi kwishishini leekhompyutha (eMica). Umnikazi wandibuza ukuba ndingakwazi Na ukuthengisa iikhomyutha. Wabuza 'phi?' Ndathi 'koomaKhayelitsha, Gugulethu, ezilokishini'. Wavuya ke lo mnikazi. ULuvuyo wazithakazelela ezi ndaba, kwaye yaba nguye oyokuzithengisela ootitshala ezikolweni. Yaba kukuqala kukaSILULO. Baqala ukwanda, ngoo2008. Bayokuvula neevenkile zomnatha wonxibelelwano (ze-intanethi).

Namhlanje baneevenkilike ezili-19. Zange baphelele ekuthengiseni qha, bavule nezikolo ezili-12 zokufundisa abantu. Namhlanje bafundisa abafundi abanga-2400 ngonyaka. Into entle ngalo mbono waba bafo yile nto yokuba bazisa ezi zikolo kwiilokishi nasemaphandleni; apho abantu bakuthi bahlala khona. Sebenwenwela nakwiMpuma Koloni. KweyoMdumba, kulonyaka, baqale ukushishina eKomani. I-intanethi yabo ibetha PHA kooma-R6 ngeyure, ukushicilela ikhasi ibaziisenti ezingama-60. EMonti iivenkile ze-intanthi zimbalwa kakhulu. Kulo nyaka uzayo bajonge ukuvula ezingama-20 eMpuma naseNtshona Koloni. Eyona nto ibalulekileyo kwivenkile ye-intanethi ngumashini wokushicilela amakhasi. "Zama uku fumana owona mashini ukumgangatho ophezulu. Ungawuthengi, wuqeshe kumnikazi. Kufuneka ufumane owona mashini ungazokunika zingxaki.

Manual summary for news nine

Qho xa kuphela inyanga ndilungiselela ukubhatala iindleko zokushicilela nokusasaza eliphepha. Umnt'omnyama ukhohlakele, kwaye ubhuqwa yindlala, zondo nochuku olungazenzisiyo. La mazwi andenze ndakhe ndayithandabuza kancinci yonke lento ndiyenzayo. Ingaba thina sisonke mzontsundu

singenza ntoni Na ukuvuselela lamathemba afayo – nezidima zabantu bakuthi? Eliphepha sisiqalo. Lisebenziseni. Masibhaleni. Masakhane. Masimanyane. Kulenyanga liyakwaphuphuphuma zizibhalo ngezibhalo zabemmi beli jikelele eliphepha.

Manual summary for news item ten

Andizimiselanga ukubonga okanye ukuncoma nokuthethelela amabhulu kwesisibhalo. Inkohlakalo yawo siyazi kak'hle. Ntokunayo nje kukho izinto esinokhe siziboleke PHA kwimigaqo nemibutho yabo yamandulo. Ithemba lam kubantu bakuthi limfiliba. Asikwazi Kwa ukuncedisana siqale imibutho yotshintsho kutshanje. Sonke sineenkanuko zokwenza umahluko kwizwe lethu kodwa nkqi ngezenzo. Wofika abantu bakuthi abanesidima, nesibajongela phezulu, bekhwinikhwiniza okweentsana phamb' kweentsana ezimhlophe, ez'dolophini. Luloyika nokungazithandi okwabethelelwa ezintlokweni zabo kudaloo. Wofika betweza imilomo. bencuma-ncumezela izinto ezingekhoyo xa bephamb' kwabelungu. Ngoba kunanamhlanje sisephants' kolawulo lwamaNgesi namaBhulu. Ngonyaka ka1902 emven' kwemfazwe yes'bini. Kwatshatyalaliswa amawaka amakhaya amabhulu kwabulalwa abantwana nabafazi babo abayi-26000. Ezorhwebo, qoqosho okanye ushishino ngeloxesha lwaluqhutywa ngesiNgesi sodwa. Yayiwatya kanobom ke lento amabhulu. Ngo1914 kwabunjwa umbutho weNational Party (NP). Umsindo oxutywe noloyiko wazala iBroederbond yamabhulu ngo1918. Babekwaqhutywa nayinkolo enzulu kubuKrestu – yokuba bona balelona hlanga lukaThixo. Zizinto apha ezithi zivuselele ithemba ebantwini. Izigqibo yayizezokuba ibhulu kufuneka lingamnkeli kwanto engesosibhulu. Injongo yayikukwakhana. Lamadoda, ekhokhelwa nguKlopper, awezimisele ukudibanisa kwanto elibhulu eMzantsi. Baseka imvisiswano nezigqibo ezifanayo. Andithi kaloku xa kuthethwa ngazwi linye, izenzo zibalula? Ekuqaleni basokola ooKlopper, bade baqonda ukuba mabasebenzise ubuKrestu.

Manual summary for news item 11

Ityala lobuqhetseba yintloko ye-Arhente yophuhliso yolutsha kazwelonke (NYDA), uAndile Lungisa, limiselwe umhla wama-30 kweyoMqungu. ULungisa inkosikazi yakhe u-Ursula Sali, uThabo Shogolo noXolisile Guquza-bajongene netyala lobuqhetseba. ULungisa nabanye babanjwa emva kokuzinikezela kookhetshe kwinyanga ephelileyo. Ukusukela ngoko wacelwa ukuba akhululwe kumsebenzi wakhe. Ukanti uLungisa usemi kwelokuba akanatyala.

Manual summary for news item 12

UMavo Solomon nguDikela kaNoni emaQwathini. Izifundo zakhe zeBanga Leshumi uzenze eBhayi. EMagxaki High. Uwongwa ngesidanga senzululwazi iBachelor of Science kwiDyunivesithi iVista kwiBSc yeMechanical Engineering. Wenze iiMasters kuyo kwalapho. Wayefundiswa ngu-Eskom. Ukwayiyo nemvumi. Ikwangubawo uMasekela ke nowandikhuthaza ukuba ndizicule ngokwam iingoma endizibhalayo. Licwecwe lam lokuqala Eli, libizwa iSiGiDiMi. Ndalishicilela ze

ndalipapasha ngo-2008. Iingoma ezikweli cwecwe zilishumi. Ingoma ethi IMBALI YETHU yeyona ibanga inkxalabo kum ngekamva leentsana esizikhulisayo. Le ngoba ke isikhumbuza ukuba singabalibali abo basakhulayo ukuba sifuna basikhumbule nabo bakuba semandleni. KwiNcwadi Yobuso kule dilesi: www.facebook.com/mavosolomon. Ndaziseke Elam ishishini nje. Imvumi nganye iqiniseke ngeenjongo zayo zokuba yimvumi. UZwai Bala ushicilele enye ingoma yam ethi "Ndize?" kwicwecwe lakhe elitsha (The Indigo Child). Imvumi entsha engu Nthabiseng. Nayo eshicilele iingoma zam ezisibhozo endiyibhalele zona. Ndibhala ngeelwimi zonke zoMzantsi Afrika ngaphandle nje kwesiQhakancu nesiBhulu. Yonke imvumi mayizixhobise. Eyona ngoma isichukumisileyo iseyile ithi “Uthando”. Nethi “Ngonaphakade”. Nenye ethi “Kaloku”. Iingoma zakhe zithetha nompefumlo womntu. Eyona ngoma sinokuthi yimamele wena mfundi yile ebekhe wathi gqaba-gqaba ngayo, ethi “Imbali Yethu.”

Manual summary for news item 13

SIYAZONYANYA singumzi kaNtu. Thina kuqala sizidelisa ngokuba ziz’khwenene ezilinganisa. Kumalinge omcinezeli. Izinto zakwaNtu asinamsebenzi nazo. Apha ndithetha ngobugqirha. Kwantokazi esand’ ukuziqond’ ukuba ingumntu izibon’ ubugqirha. Utyhudisana “negqirha” eziteksini lithengisa amayeza. Evathe ezo ntsimbi kula Facebook. “Gqirha” linye lithe qhiwu ibhotile yotywala lihlek’ isqhazolo nootshomi. Ubugqirha bungcoliswa kwasithi sizukulwana. Sithi abehlisa isidima solu bizo lundiliseke kangaka kwaNtu. Ubugqirha abuyomfashoni, abufani nokunxiba iskinny jeans ezo. “Akufanelang’ba abantu badlale ngobugqirha. Akuamelanga amagqirha asele ubusuku bonke. “Igqirha elinyanisekileyo lizithembile izinyanya zalo.” Igqirha alihambi lizithengisa. Amagqirha ale mihla axabise imal. Amagqirha angaphandisisiyo ngezimo zempilo yabantu abaze kuwo .Amagqirha angawaziyo amayeza awasebenzisayo. Abantu kuvavanyelwa izifo zonke phamb' kokuba banikwe igunya lokunxiba iintsimbi. ISebe Lezempilo lilugniselela ukuba necandelo ukuphucula intsebenziswano phakathi kwamagqirha norhulumente. Asibukokosanga ubugqirha. Sihleke kwakuhlekiswa ngabo. Masizazini izinto zokudlala Bantu bakuthi, zidla ngokuba namavili.

Manual summary for news item 14

Abaphangi kuthiwa baqhekeze kwikhaya eliseMilnerton, eKapa, lomxolelanisi welilizwe, uArchbishop Emeritus #Desmond. Asikwazi ukuxela okubiweyo, uArchbishop nowakwakhe bebengekho sekhaya. Ngokumemelela uluntu loMzantsi Afrika lulandele umzekelo kaMadiba. Ixesha ekwenziwe ngalo lomkhwa kuzintloni eMzantsi Afrika. NgeyeThupha, abaqhekezi bangena kulendlu ngelixa uArchbishop Tutu nowakwakhe uLeah bekobentlombe. Kuvele nezokuba itoliki ibisebenzisa intetho yezandla. Lendoda kuthiwa yayikhona kwinkomfa yaseMangaung yombutho weANC.

Manual summary for news item 15

U-Itumeleng Khune, uchongwe njenge-2013 SA Sport Star of the Year bomsitho ububanjelwe eSun City phezolo. Uvotelwe kwezokuqubha, uChad Le Clos. UHashim Amla; umdlali wentenetya, uLucas Sithole nembaleki kwi-Athletic, uMapaseka Makhanya. UKune uzihambele nomgrugra we-BMW1 series nesigidi seerandi. UFikile Mbalula ewonge abadlali abatshatsheleyo kwiinyanga ezilishumi eninesibini ezigqithileyo. UMapaseka Makhanya uphumelele njengeSports woman of the Year. KuMakhanya njengentshatsheli kwi 1500m ne5000m kwiSA Senior Championships eStellenbosch nakwiSPAR Grand Pix series. UCameron van der Burgh yena uphumelele njengeSportsman of the Year. Ebutsheni balonyaka uphumelele igolide kwi-Men's 50m. Ukanti kwicala labaneziphako zomzimba iimbasa ziye kuKgothatso Montjane no Lucas Sithole. IiProteas laphumelela njenge-Team of the Year.

Appendix H: System generated Summaries

.....Summaryis/IsiCatshulwa.....
....

System generated summary for news item one (highlighted sentences are first sentences kept in system generated summaries**)**

Iqela lentambula iBobby Rush laseRhawutini livule ngokusesikweni iqonga kulomsitho apho licule ingoma eyodwa ebilungiselelwe lomcimbi wokumiselwa kweliphulo, ngeenjongo zokusebenzisa umculo nje ngesixhobo sokufikelela kuye wonke ubani ochaphazeleka kwingozi zendlela zelizwe loMzantsi Afrika.Ukanti amaqela omculo wevangeli afana noLusanda Mcinga, Dumi Mkostad,Veliswa Skeyi, Butho Vuthela,Bulela M kwakunye noBethusile Mcinga onwabise lenginginya ngeengoma zevangeli becela amazulu ukuba angenelele ukuphungula iingozi zendlela.

.....
....

System generated summary for news two

Eyona nto ibuhlungu kukuba ‘ubuhle’ bobu ‘bomi’ okanye bobu ‘butyebi butsha’ abucaciswa ngaphandle kokugadlwa phambi kwabantu.Uyakufika ooKenny bephelekwa yimpi yabo nangamagqiyazana xa betyelela iindawo ezinjengooMzoli; phofu ke kubenzima ukubabona bebhencisa le mpilo bayiphilayo emaCamps Bay okanye emaSea Point kuba kaloku akukho mntu ukhathalele le ntlalo bayiphilayo phaya.

.....
.....

System generated summary for news three

Umbutho we-African National Congress (ANC) iwakhabe ngawo omane amabango eqela le-Pan Africanist Congress (PAC) wokuba ikhethwe okwephel'emasini ngexesha kukhunjulwa isiganeko semini yamalungelo oluntu, eSharpeville ne-Langa Massacre.Ngo-Lwesihlanu umbutho we-PAC ubambisene nombutho we-Economic Freedom Fighters (EFF) bebambe umcimbi eSharpeville; kufutshane nalapho uMongameli uGedleyihlekisa Zuma ebebambe khona isikhumbuzo sale mini.

.Summary

is/IsiCatshulwa.....

System generated summary for news four

Ndaqala udlala umbhoxo ngo1992, wathi efika u1995 ndabe sendisentweni qha ke babengekho abantu abamnyama kwiqela lamaBhokobhoko ngoko.UChester Williams wanika abantu abaninzi ithemba lokuba noba usuka phi noba ungubani, xa uzimisele uyakwazi ukuphumelela.Emveni kokukhululwa entolongweni kuka tata uNelson Mandela kwa-nyanzeleka abantu abamhlophe bavumele abantu abamnyama bonyulwe kwiqela lama Bhokobhoko.

.....

.....

System generated summary for news five

Phambi kwalo mdlali, ibali belisithi iNtshona Koloni yagqibela ukubuya nayo le ndebe ngo2001 (apho babetha ooKrebe) kwaye bagqibela ukudlala kumdlalo wamanqam ngonyaka ka2010, apho babethwa ngooKrebe kwakula Kings Park bekudlalelwa kuyo kulo nyaka.Lo mfana wakwa De Jongh usuke wathi gqi ephaphatheka ngathi nguloliwe wombane – loliwe lo ubungenobonwa nanguZahara, ngoba sithe sisothuka wabe uDe Jongh sele eyibeka phantsi kweempondo ibhola.

.....

...

System generated summary for news six

Ezimnyama ngenkani ‘zijabulile’ emva kokubetha AmaKhosi ngenqaku elinye (fakwe ngu Daine Klate).Mbhoxo: Amabhokobhoko asindiswe ziipenalti ezimbini kwimizuzu yokugqibela

yomdlalo wawo neqela laseArgentina phezolo (siphumelele ngo 22 – 17)… Kweminye imidlalo iiLions zidwaxushe iiBlue Bulls ngo 62 – 23, Western Province ibhungce ngo 20 – 19 kumdlalo wayo neGriquas.

Ntenetya: uRik de Voest wakweli ukhiqiwe kwi US Open emven' kokuthiwa qhwi qhwi ngo 6-4, 6-2 ngu Peter Gojowczyk wase Germany, izolo.

Bekukho nemidlalo yeBarclays Premier League; apho iArsenal izibethele iFulham ngo 3 – 1.

Kunini kwathiwa Rik de Voest – ingaba ayikho na nyani italente entsha kwintenetya kweli lethu?

Iqela laseLiverpool lona libethe elaseAshton Villa ngo 1 – 0.

.Summary is/IsiCatshulwa.....

System generated summary for news seven

Sundowns ibethe iCeltic ngo 3 - 1; Omnye umdlalo obujongwe ngabomvu ngu lo we Pirates ne Al Ahly – iPirates iphumelele ngo 3 – 0.Mbhoxo: Zikhiqiwe iiSouthern Kings zaseMpuma Koloni kwi Super 15 yombhoxo nangona zibethe iLions izolo – azikwazanga ukuqokelela amanqaku oneleyo.Al Ahly ifumene isohlwayo(ukhutshelwe ngaphandle umdlali wabo) emva kokudlala rhabaxa umdlali we Pirates phambi kokuphela kwesiqingatha sokuqala.

.....

System generated summary for news eight

Emven' kweminyaka bezama ukuba ngoosomashishini eKapa - sebevela nokuthengisa iimbadada nokucoca imizi - uLonwabo, xesh' eyokukhangela umsebenzi kwishishini leekhompyutha (eMica) wathetha into evanayo nomnikazi ngoba kakade wayefundele ezobuxhaka-xhaka eCPUT.' Ndathi 'koomaKhayelitsha, Gugulethu, ezilokishini'.uLuvuyo ucacisa ngelithi: "EKapa kukho amathala eencwadi anee khompyutha njalo-njalo kodwa eMpuma Koloni akukho nto ikhoyo." Siyazi ke nathi kwesisigidimi ukuba uninzi lweevenkile zithenga oomashini abafikelelekayo boomaR1000, abo mashini babuya babanike iingxaki ezininzi kwaye bacothe nokucotha.

……

……

System generated summary for news nine

.Summary is/IsiCatshulwa.....

Qho xa kuphela inyanga ndilungiselela ukubhatala iindleko zokushicilela nokusasaza eliphepha; ngemalana endiziqokelela kwiintengiso nemisebenzi yam emihlanu kweliKapa.Ndazibuza ukuba ingaba inene na ndiyamazi na umntu omnyama? Ingaba thina sisonke mzontsundu singenza ntoni na ukuvuselela lamathemba afayo – nezidima zabantu bakuthi? Ndithe ndisazibuza njalo suke

ndabethwa ngalempendulo: “Udlala ngemali yakho, soze ukwazi umnceda umntu omnyama Unathi.”“Umnt’omnyama ukhohlakele, kwaye ubhuqwa yindlala, zondo nochuku olungazenzisiyo.Umbuzo ke ngowokuba ingaba njeng’ba sohlulwa-hlulwa kangaka lucalucalulo noonyawontle bamandulo; yintoni esinokuyenza ukuze sikwazi ukwakha intsebenziswano nokuthandana phakathi kwabantu abamnyama?”La mazwi andenze ndakhe ndayithandabuza kancinci yonke lento ndiyenzayo, yokupapasha eliphepha.Impendulo yalombuzo ayikho kurhulumente – impendulo yalombuzo ikuthi.....……
……
……
…

System generated summary for news item ten

Kodwa, okumangalisayo kukuba basoloko bezihleka ezi nzingo bantsonteleke kuzo bekwabek’ ithemba phofu, kubuhle bothando olulambathayo kubayeni babo; bade bathembe ngamanye amaxesha ukuba loo mlilwana wobutshivela uvutha ezintliziyweni zabayeni (okanye zabalingane) babo uya kucima ngenye imini.Ukanti, nangona bethwele intaba yemisebenzi emagxeni abo, kusafuneka bazinike ithuba lokufundela okanye lokubalisela abantwana babo amabali, ukukhaliphisa nokuvuselela ezo ngqondwana zincinane.Ngaphezu koko, kufuneka babe ngabaphulaphuli abanomonde baphinde bamelane nento yokuba abantwana babo baza kufikelela ebudaleni.Ukuphuma kwabo emisebenzini, bangqala ngqo emakhaya ngokukhawuleza, baphekele iintsapho zabo, bahlambe abantwana bekwaqinisekisa ukuba abantwana bayawenza umsebenzi wesikolo.Kule minyaka idlulileyo kuye kwabakho ulwando loomama abakhulisa abantwana babo bodwa (single parents, ngolwasemzini).
.....
.....

.....Summary is/IsiCatshulwa.....

System generated summary for news item sixteen

Ityala lobuqhetseba nelokuhlisa imali ngomlenze kwalowo wayesakuba yintloko ye-Arhente yophuhliso yolutsha kazwelonke (NYDA),uAndile Lungisa,limiselwe umhla wama-30 kweyoMqungu,kunyaka ozayo kwinkundla echophela amatyala anje,iCommercial Crimes Court,eRhawutini.uLungisa nabanye abathathu atyholwa nabo,inkosikazi yakhe u-Ursula

Sali,uThabo Shogolo noXolisile Guquza-bajongene netyala lobuqhetseba emva kokuthembisa ukuzisa imvumi yaseMelika,uR.

.....

Summary is/IsiCatshulwa.....

System generated summary for news item twelve

Imvumi yokuqala ukubonisa umdla ekuthengeni iingoma zam yaba nguZwai Bala, phantsi kweshishini likabawo uHugh Masekela, iChissa Records, eselabhangayo kungoku nje.
Ndithe ekumameleni iingoma eziculwa zezinye iimvumi zalapha eMzantsi, ndabona ukuba nam ndinganegalelo ngokuhlomla kumba wepolitiki, wezokukhula phantsi kwengcinezelo nocalu-calulo nangekamva lethu njengabemi boMzantsi Afrika.
Mavo: Inzima into yeRecord Company ngoba indlela abasebenzisana neemvumi ngayo yindlela enye, akukhathalisekanga nokuba ucula ikwaito, ijazz, igospel okanye ntoni na.

……

……

.....Summary is/IsiCatshulwa.....

System generated summary for news item thirteen

Uyakufik'umntu esindwa ziintsimbi, edakas' idolophu le yonke ekhangel' int' angayaziyo kuba kaloku kufunek' ejongiwe kukhwazwe kuthiwe "Yhu! Uyakumbona eligqirha elithinzileyo; ez'the ntonga kuloo ndawo, efake iintsimbi zakhe kwedini, esis'mumu, de abe neendaw' ezingathi uyoyik' ukuphakamis' intloko."Umam' uNobulawu yena, oligqirha laseLusikisiki, naye owaathwasa isiNdawu (eDzaneen, kwaMaake) nesiXhosa (eNgqamakhwe kwaPhuphuma), ungqinelene nomam' uMahlasela, esithi "igqirha elinyanisekileyo lizithembile izinyanya zalo.

.....

.....Summary is/IsiCatshulwa.....

System generated summary for news item fourteen

Abaphangi kuthiwa baqhekeze kwikhaya eliseMilnerton,eKapa,lomxolelanisi welilizwe,uArchbishop Emeritus

.....Summary is/IsiCatshulwa.....

System generated summary for news item fifteen

IKapteni yeBafana-Bafana neKaizer Chiefs,u-Itumeleng Khune,uchongwe njenge-2013 SA Sport Star Of the Year,kubukhazi-khazi bomsitho ububanjelwe eSun City phezolo.uMapaseka Makhanya,nongoyena mdlali wasetyhini bekhuphisana namadoda,nobechongelwe iSports Star of the Year,uphumelele njengeSports woman of the Year,ehlalisa phantsi uAnne Peace noMandisa Williams. UKhune uthi ukuphumelela iSA Sports Star of the Year, kuthetha ukuba umdlalo webhola ekhatywayo usengowona uhamba phambili kwimidlalo yeli.