ENHANCED WORD LENGTH AND MODEL ELIMINATION ALGORITHMS FOR LANGUAGE IDENTIFICATION

NICHOLAS IORNONGU AKOSU

UNIVERSITI TEKNOLOGI MALAYSIA

ENHANCED WORD LENGTH AND MODEL ELIMINATION ALGORITHMS FOR LANGUAGE IDENTIFICATION

NICHOLAS IORNONGU AKOSU

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

OCTOBER, 2014

This thesis is dedicated to my mother, Anna and my lovely wife, Patience.

ACKNOWLEDGEMENT

First and foremost, I would like to record my utmost gratitude to the Federal Polytechnic Nasarawa and Universiti Teknologi Malaysia (UTM) for giving me the opportunity to pursue my studies in my desired area and for sponsorship and publication support. I appreciate my supervisor Prof. Dr. Ali Selamat who has given his tremendous support, guidance and motivation on my thesis. His advice in setting up my research directions was invaluable and his encouragement, critiques and feedback have greatly enhanced and strengthened the quality of my research.

The sacrifice of my family has been very helpful and is hereby acknowledged and appreciated. In particular, I thank my wife for her understanding and love during the past few years. My mother deserves special mention for her untiring support and prayers in spite of her health challenges. I appreciate my mother, Madam Anna Mbasengen Akosu and my late father, Akosu Timeh for having sincerely raised me with their caring and gentle love. I am grateful to my entire family members, especially, Dr. Joe Akosu, for their constant support, encouragement and love. I would like to thank my lovely children, Iwanger, Torkwase, Msendoo, and Zahemen for their sacrifice which was very vital to the pursuit of my PhD degree.

I wish to appreciate all the members of the Software Engineering Research Group, Faculty of Computing, UTM; my friends at UTM and in Nigeria, my colleagues at Federal Polytechnic, Nasarawa for their cooperation and support. Finally, I would like to thank everybody who has helped me not only on my thesis but by also bringing the best memories throughout my study at the UTM.

ABSTRACT

Language identification is the process of determining the natural language of text documents using computational methods. The quality and size of the text available for generating the necessary models has significant impact on the performance of the algorithms used to determine the language of a text. The ability to correctly identify the language of a document is required to ensure the effectiveness of information retrieval systems in a multilingual setting. Unfortunately, existing methods that are used to model natural language have been affected by several limitations. Such limitations include inability to produce reliable models given a small size of training text. Other limitations are: inability to consistently handle multilingual documents, long training times and inability to distinguish closely related languages. The spelling checker technique has been shown to be successful in distinguishing closely related languages but is often hampered by two important constraints: inefficient run time performance and non-availability of spelling checkers for many languages. The aim of this study is to address the problems of language identification by developing improved algorithms that enhance run time performance and accuracy irrespective of the size of corpus available. Therefore, this thesis proposed three algorithms. Firstly, the word length algorithm implements the bag-of-words model using word length information. Secondly, the model elimination algorithm is designed to further improve run time performance by taking advantage of word frequency in training and testing documents. By monitoring the performance of models in the course of processing, this algorithm dynamically selects non-performing models for elimination without compromising accuracy. Thirdly, the linear combination algorithm merges the strengths of the word length and model elimination algorithms by feeding word length features into the model elimination algorithm. Empirical results from the proposed algorithms using test collection from the standard corpora are superior to existing methods in terms of distinguishing closely related languages and multilingual identification. In addition, the word length, model elimination and the linear combination algorithms have better run time performance than the spelling checker method that uses a similar scoring technique, yielding average time gains of 57%, 83% and 98.4% respectively in identification of 140-byte long text.

ABSTRAK

Pengenalpastian hahasa adalah proses menentukan bahasa tabii sesebuah teks dokumen menggunakan kaedah pengkomputeran. Kualiti dan saiz teks sedia ada untuk menjana model, memberi impak yang besar terhadap prestasi algoritma untuk menentukan bahasa sesebuah teks. Kemampuan untuk menentukan bahasa sesebuah dokumen secara tepat adalah perlu untuk memastikan keberkesanan sistem pencarian maklumat dalam pelbagai bahasa. Malangnya, kaedah sedia ada yang diguna pakai mempunyai beberapa kelemahan. Antaranya adalah ketidakupayaan untuk menghasilkan model yang efektif jika menggunakan teks latihan yang kecil, ketidakupayaan untuk mengendali kan dokumen pelbasai Bahasa secara konsisten, jangka masa latihan yang panjang dan ketidakupayaan untuk mengenal pasti bahasa lain yang berkait rapat. Teknik *spelling checker* merupakan kaedah yang berkesan dalam pengenalpastian bahasa yang berkait rapat tetapi di halang oleh dua kekangan utama prestasi *run-time* yang kurang cekap dan ketiadaan *spelling checker* untuk kebanyakan bahasa. Matlamat penyelidikan ini adalah untuk menangani masalah Pengenalpastian hahasa dengan membangunkan algoritma untuk meningkatkan prestasi *run-time* dan ketepatan tanpa mengira saiz korpus yang sedia ada. Tesis ini mengemukakan tiga (3) algoritma. Pertama, algoritma panjang perkataan yang mengimplementasikan model *bag-of-words*. Kedua, algoritma penghapusan model direka untuk meningkatkan lagi prestasi *run-time* dengan mengambil kira frekuensi perkataan dalam dokumen latihan dan ujian. Dengan meneliti prestasi model semasa pemprosesan, algoritma tersebut memilih secara dinamik model-model kurang berprestasi untuk dihapuskan tanpa menjejaskan ketepatan. Ketiga, algoritma kombinasi linear menggabungkan keberkesanan algoritma panjang perkataan dan penghapusan model dengan memasukkan ciri-ciri panjang perkataan ke dalam algoritma penghapusan model. Hasil empirikal daripada algoritma yang dikemukakan menggunakan *test collection* daripada *standard corpora*, adalah lebih unggul daripada kaedah yang sedia ada dari segi pengenalpastian bahasa lain yang berkait rapat dan pengenalpastian pelbagai bahasa. Di samping itu, algoritma panjang perkataan, penghapusan model dan kombinasi linear mempunyai prestasi *run-time* yang lebih tinggi daripada kaedah *spelling checker* yang menggunakan teknik pemarkahan serupa, dengan hasil purata masa 57%, 83% dan 98.4% masingmasing dalam mengenal pasti teks sepanjang 140-byte.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	–	Artificial Neural Network
ART	–	Adaptive Resonance Theory
ARTMAP	–	Supervised Adaptive Resonance Theory model
ASCII	–	American Code for Information Interchange
BM	–	User-specified Benchmark
CLIRS	–	Cross Language Information Retrieval System
CRF	–	Conditional Random Fields
DBCS	–	Double Byte Character Set
DFI	–	Digital Forensic Investigation
DoF	–	Degree of Freedom
DT	–	Decision Trees
ET	–	Elimination Threshold
GA	–	Genetic Algorithms
GMM	–	Gaussian Mixture Model
HCRL	–	Handling Closely Related Languages
HLT	–	Human Language Technology
HMM	–	Hidden Markov Model
HTML	–	Hyper Text Markup Language
ICA	–	Independent Component Analysis
ICT	–	Information and Communication Technology
KB	–	Kilobytes
KNN	–	K-Nearest Neighbor
LID	–	Language Identification
LOP	–	Language Observatory Project
MB	–	Megabytes
ME	–	Model Elimination
MI	–	Multilingual Identification
MLLD	–	Support Minority Languages with Larger Training Data

MNB	–	Multinomial Naive Bayes
NLP	–	Natural Language Processing
NLTK	–	Natural Language Tool Kit
OC LID	–	Open Class Language Identification
OCR	–	Optical Character Recognition
OOVs	–	Out-Of-Vocabulary words
PCA	–	Principal Component Analysis
POS	–	Part of Speech Tagging
PPM	–	Prediction by Partial Matching
SALAMA	–	Swahili Language Manager
SEC	–	Standard Evaluation Corpora
SILC	–	Identification System for Language and Encoding (in French)
SOM	–	Self Organizing Maps
SPLID	–	Spelling Checker Language Identification
SVM	–	Support Vector Machines
TTP	–	Text to Phoneme
TPE	–	Time Performance Evaluation
TPL	–	Targets Popular Languages
UCS	–	Universal Character Set
UDHR	–	Universal Declaration of Human Rights Act
ULMD	–	Targets Under-resourced Languages with Minimal Training Data
UNESCO	–	United Nations Educational Scientific and Cultural Organization
UNICODE	–	Unique, Universal, and Uniform Character enCoding
UTF	–	Unicode Transformation Format
VE	–	Vocabulary Extension
VQ	–	Vector Quantization
WWW	–	World Wide Web
XML	–	eXtensible Markup Language

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

In spite of major advances in technologies that facilitate information sharing across the globe, linguistic differences often constitute significant barriers to information access. These barriers may be addressed by using translation systems. However, to achieve any meaningful translation the language of the target document must first be ascertained (Newman, 1987). To overcome these challenges, several computational methods have been developed to solve the problem of language identification. Such methods include N-gram models, naive Bayes event models, artificial neural networks (ANN), self-organizing map (SOM), fuzzy ARTMAP, adaptive resonance theory (ART), support vector machines (SVM), independent component analysis (ICA), decision tree (DT), hidden markov models (HMM)(Susperregi, 2010). Selamat and Ng (2011) define language identification as a process of determining the natural language of text documents. Automatic language identification is usually done using computational methods and available corpus or linguistic data. However, the quality and size of the corpus data available for generating the necessary language models has significant impact on the performance of the algorithms used to determine the language of a text (Brown, 2012; Botha and Barnard, 2012; Hughes *et al.*, 2006).

The ability to correctly identify the language of a document is required to ensure the effectiveness of information retrieval systems in a multilingual setting such as the Internet (Selamat, 2011). Unfortunately, existing methods including N-gram and its classifier variants that are used to model natural languages for identification have been affected by several limitations. Such limitations include: (a) Inability to produce reliable models given a small sample size of training text and consistently handle multilingual document. (b) Data sparseness problem. (c) Long training time.

(d) Inability to distinguish closely related languages (Hammarstr-om, 2007; Ljubesic *et al.*, 2007; da Silva and Lopes, 2006; Ranaivo-Malancon, 2006; Zampieri, 2013) .

It has been noted from previous studies that accuracy of language identification is almost 100% for different language identification techniques (Takci and Gungor, 2012). However, most studies did not report results in terms of time performance because research in the area of language identification has focused on two main directions: exploring new techniques suitable for the task and advancing on the level of accuracy achievable in using these techniques. This is very important for information retrieval applications (Yang and Wu, 2012; Sun and Liu, 2011). As stated earlier much success has been achieved in the direction of accuracy. Some level of progress has also been made in the direction of increasing the language coverage of the available techniques. However, investigating the computational speed performance of the various methods has been apparently neglected. It is noted that even in comparative studies only *accuracy* tends to be the yard stick for comparison. Consequently, this study considers it timely for research to change direction to the investigation of speed performance. Issues of language identification are as follows.

a) Run Time Performance of Language Identification

This research investigates the run time performance of the lexicon based approach for language identification. The research is focused on language identification using limited training text. This is often a problem with natural languages with little or no digital resources, which are also called under-resourced languages. These are mainly minority languages i.e., spoken by a few, but which are gaining in importance due to increasing and widespread use of the Internet and the possibility of such languages being used for communication over the Internet. So far, not much research has been done on these languages because they were previously perceived as being less important than the popular languages. However, the research by Pienaar and Snyman (2010) was a good beginning and also pointed the direction for further research on resource-poor languages. The special nature of this class of languages has also influenced the choice of technique and data set for this research.

In their research, Winkelmolen and Mascardi (2011) proposed investigation of spelling checker based language identification by running text through spellcheckers of different target languages and using the number of errors in each language (i.e. the

Hamming distance from the corrected text). They noted that such approach could give very accurate results, but would be very inefficient. However, this thesis takes the position that, only a critical study of the performance of this approach can confirm or disprove such opinions.

b) Identification of Minority Languages

Supporting minority languages and resource-poor languages is one of the outstanding problems in language identification research. Minority languages are generally understood to be languages that are spoken by a small number of people (Pienaar and Snyman, 2010; Prinsloo, 2000; Barbaresi, 2013). There are languages on earth today that have as few as 1000 speakers or even less (Lewis, 2009). Language identification research has not been extended to cover even languages that have native speakers running into millions; however, languages with few speakers should still be considered in language identification research for historical reasons and for reasons of cultural preservation (Pavan *et al.*, 2010; Jothilakshmi and Palanivel, 2012). Moreover wide spread use of the Internet has continued to grow very rapidly and as more people of various linguistic origin become players in cyber space they are bound to bring content in their native languages (Varma, 2010). There is the need to research on how such documents can be identified for effective information sharing. Attempts to solve this problem using N-gram modeling and SVM for the South African languages (Botha and Barnard, 2006) have failed to produce acceptable results, by giving accuracy of less than 70% in language identification.

c) Sparse or Impoverished Training Data and Multilingual Documents

This problem is closely related to the issues of resource-scarce languages(also known as under-resourced languages). Hughes *et al.* (2006) wonder if it is possible to develop algorithms that can handle as few as 50, 100, 250 words OR 50, 100, 250 characters for language identification. This would be particularly useful in identification of resource-poor languages. There are indeed situations where NLP tasks have to deal with such a small amount of input text for identification (Roux, 2008). For example when dealing with multilingual advertisements, menus and short messages (SMS texts). Research by (Carter *et al.*, 2011; Tromp and Pechenizkiy, 2011; Gottron and Lipka, 2010) addressed the question of identification of short text but did not tackle the case of small training samples. However, Vatanen *et al.* (2010) have shown that small training samples yield poor accuracy of less than 80%. Also research by Brown (2012) has revealed that error rates increase sharply when training text is below 500k, while Botha and Barnard (2012) has shown that accuracy of language identification stabilizes only after training text is more than 200k.

The problem of language identification of Multilingual documents still needs further research. There have been a number of research efforts aimed at solving this problem (Yamaguchi and Tanaka-Ishii, 2012). The use of independent component analysis for reduction of document features proves to be time consuming Selamat and Zhi-Sam (2008). On the other hand using a fine-grained model for such a simple task as language identification produced a heavy system that is rather slow, Hammarstr-om (2007). Both researchers report varying levels of accuracy, but considerable computational costs. The method used by Yang and Liang (2010) was particularly expensive in processing time during training.

This thesis focuses on solving the problem of using a small data set to build robust language models that can be used to identify documents by language. This is necessary because is it often difficult to find large amounts of training text in many natural languages and this could make it impossible to identify documents written in such languages. Such a state of affairs would make these languages vulnerable for information hiding and other uses. This needs to be addressed quickly especially since under-resourced languages are becoming widely used on the internet. Thus, improved computational methods are needed to capture all the necessary details and organize the models in such a way that permits accurate and efficient identification of resourced-poor languages, as well as taking cognizance of the need to distinguish closely related languages and handle multilingual identification (Susperregi, 2010). However, it is necessary to test the improved method on heavily resourced languages, such as English, to ensure that the developed method remains valid for a long time and for many languages.

1.2 Motivation

The growing number of electronic documents on the Internet is one of the major reasons that motivated research into automatic language identification. For example: a student, business man or any other person, conducting research in some area could issue a query (or request) through a browser to find information of interest on any subject. Such a search need not be limited to the language in which the query was issued, but should in general, be able to search the entire World Wide Web (WWW) and return pages with the relevant information. Upon assembly of all the relevant documents the researcher should be able to arrange translation of the documents to any language of his/her choice. This is only possible and viable with availability of good language identification tools. Indeed, Lewandowski (2008) insists that if the language of a document is incorrectly identified, subsequent translation will also be meaningless, leading to gabbled translation.

This research is motivated by the fact that only about 12% of all the languages in the world have been studied in language identification research (Brown, 2012) because in most cases only the popular languages are investigated. This means most natural languages cannot be distinguished by digital methods due to none availability of identification tools. However, consideration of automatic language identification of many natural languages also needs to contend with the fact that these languages are resource-poor as there are no corpora available in these languages. This raises the question of which technique to use for the language identification research because most techniques for language identification are statistical approaches which require a large corpus in the language to be studied.

Pienaar and Snyman (2010) applied second generation spellcheckers to perform language identification on the 11 official languages of South Africa. A second generation spelling checker (Prinsloo and Schryver, 2003b) not only uses a lexicon to check words, but also uses a morphological analyzer to check compounds and other complex morphological forms such as inflections and extensions of words. Their choice of technique was predicated on the fact that African languages are resource-poor because digital resources are not available. Their experimental results were impressive with respect to identification of closely related languages and multilingual documents. This encouraged Pienaar and Snyman (2010) to suggest extension of this method to other resource-poor languages like Wolof, Yoruba, Igbo, Hausa and Kinyarwanda. Pienaar and Snyman (2010) are of the view that research into language identification of under-resourced languages would make a search for resources in these languages a precise task. This indeed stresses the fact that there must be some minimum level of digital resources available in any language to enable language identification research on such a language. Tables 1.1 and 1.3 exhibit regional statistics on languages and some selected African countries.

Table 1.1: Distribution of world languages by region of origin, source: Ethnologue (Lewis, 2009)

Region	No. of Languages	%	No. of Speakers	%
Africa	2.146	30.2	789,138,977	12.7
Americas	1,060	14.9	51,109,910	0.8
Asia	2,304	32.4	3,742,996,641	60.0
Europe	284	4.0	1,646,624,761	26.4
Pacific	1,311	18.5	6,551,278	0.1
Totals	7,105	100.0	6,236,421,567	100.0

Table 1.2: Language distribution for selected African countries , source: Ethnologue (Lewis, 2009)

Country	No. of languages	Number of Speakers
Algeria	21	33,001,300
Angola	38	16,070,730
Cameroon	281	8,931,726
Chad	132	6,594,079
Congo	215	39,906,030
Egypt	28	81,716,600
Morocco	14	26,653,930
Nigeria	529	104,138,885
South Africa	44	44,637,399
Zimbabwe	23	15,712,470

In line with the discovered gaps, this study aims to address the problems and limitations of developing digital resources for natural languages, improving on run time performance of models and multilingual identification. The motivation of this study is based on the need to properly integrate the newly emerging languages in the digital community by developing modeling strategies that are effective and efficient in solving existing problems of language identification. Such methods must be so developed as to work effectively also for the heavily resourced languages.

1.3 Problem Statement

With the ever-increasing number of users of the Internet across the globe, the possibilities of information sharing among the Internet population are often hindered due to linguistic diversity. The main cause of this setback is the fact that language creates barriers to information access across the various linguistic boundaries. Considering that there are over 7000 languages in the world (Lewis, 2009), it is easy to see that within a very short time there could be a sizable volume of content scattered across many languages on the Internet. Due to the lack of technologies to break the linguistic boundaries there is bound to be much content that cannot be shared because of the language digital divide.

The development of language identification tools is meant to solve this problem. However, language identification tools currently exist for a small number of languages (Brown, 2012; Lui and Baldwin, 2012). For the past 3 decades several research works have attempted to provide solutions to the language identification problem. However, it is now clear that the problem of language identification is multifaceted. There are morphological issues, what script is used to write a particular language, the fact that some languages can be written using several scripts. Furthermore, there are encoding issues: is a document stored in latin1 or utf8? The challenge of distinguishing between closely related languages has not been overcome. Other outstanding issues in language identification research include identifying multilingual documents, coping with minority languages (Hughes *et al.*, 2006; Scannell, 2007) and the closely related problem of under-resourced languages (Pienaar and Snyman, 2010).

Several previous works have developed techniques for many of the above challenges using N-gram models, naive Bayes event models, artificial neural networks (ANN), fuzzy ARTMAP, support vector machines (SVM), independent component analysis (ICA), decision tree (DT), hidden markov models (HMM), etc. However, most of these studies end up with one research gap or the other based on the weaknesses of the applied methods (Baldwin and Lui, 2010; Hughes *et al.*, 2006). In particular the problems of time consuming models, unsatisfactory levels of accuracy in multilingual identification and the unbending constraints of under-resourced languages have remained intractable. Therefore the following questions need to be answered in order to achieve the purpose of the study.

> *"Can the naive Bayes multinomial model be used to improve language identification performance in the case of limited training text?"*

In order to formulate research objectives it is necessary to define the following detailed research questions from the primary research question:

RQ1: What language identification approaches currently exist?

RQ2: What are the relative strengths / weaknesses of existing language identification methods?

RQ3: Why is language identification in the case of limited training text important?

RQ4: Can a viable lexicon model be implemented from a small corpus?

RQ5: How to apply the 'bag-of-words' model for language identification in cases of extreme scarcity of linguistic resources?

RQ6: What metrics can be applied to improve the run time performance of the 'bag-of-words' model for language identification?

RQ7: Can the 80/20 rule and Zipf's law be used to improve performance of 'bag-of-words' model for language identification?

RQ8: Is the model elimination technique suitable for Language identification of multilingual documents?

RQ9: How to implement a linear combination of word length and model elimination algorithms to improve performance of language identification?

RQ10: Can a linear combination of word length / model elimination approach handle language identification of multilingual documents?

1.4 Research Aim

The aim of this study is to develop improved lexicon models and word length and model elimination algorithms to enhance the efficiency and effectiveness of language identification irrespective of the size of corpus available. By addressing the specific and peculiar constraints of natural languages, the research strives to develop algorithms that improve the running time performance of language identification while maintaining accuracy and consistency of results with the ultimate goal of improving multilingual identification and distinguishing closely related languages to enable effective information sharing in multilingual settings such as the Internet.

Table 1.3: Research questions and where they are addressed

Research Questions	Treated in
RQ1: What language identification approaches currently exist?	Chapter 2
RQ2: What are the relative strengths / weaknesses of existing language identification methods?	Chapter 2
RQ3: Why is language identification in the case of limited training text important?	Chapter 2
RQ4: Can a viable lexicon model be implemented from a small corpus?	Chapter 4
RQ5: How to apply the 'bag-of-words' model for language identification in cases of extreme scarcity of linguistic resources?	Chapter 4
RQ6: What metrics can be applied to improve the run time performance of the 'bag-of-words' model for language identification?	Chapter 4
RQ7: Can the 80/20 rule and Zipf's law be used to improve performance of 'bag-of-words' model for language identification?	Chapter 5
RQ8: Is the model elimination technique suitable for Language identification of multilingual documents?	Chapter 5
RQ9: How to implement a linear combination of word length and model elimination algorithms to improve performance of language identification?	Chapter 6
RQ10: Can a linear combination of word length / model elimination approach handle language identification of multilingual documents?	Chapter 6
Summary	
Research Questions 1 - 3	Chapter 2
Research Questions 4 - 6	Chapter 4
Research Questions 7 - 8	Chapter 5
Research Questions 9 - 10	Chapter 6

1.5 Research Objectives

In pursuance of the stated research aim, the following objectives have been set:

i. To propose an enhanced word length algorithm for language identification.

ii. To propose an enhanced model elimination algorithm for language identification.

iii. To propose a linear combination of word length and model elimination algorithms that will complement the strengths of both algorithms to achieve optimum run time performance for language identification.

1.6 Contributions of the Research

The need for improved modeling and classification methods for language identification is imperative in view of the increasing widespread use of the Internet in the various regions of the world. Effective language identification tools have significant impact on the overall information sharing potentials to the benefit of the global Internet community. Securing improved methods of modeling the various languages under the constraints of reduced training data in order to achieve high performance accuracy in multilingual identification, distinguishing closely related languages, while maintaining low running time, have been major concerns to practitioners in this field of research because these are the most critical challenges of the earlier used approaches.

This research focused on developing enhanced computational models using small training data to facilitate language identification in order to achieve the desired high performance accuracy in multilingual identification, distinguishing closely related languages, and maintaining low running time with the ultimate goal of facilitating effective information sharing in a multilingual setting. In addition, implementing the improved bag-of-words model for language identification using minimal training text has potential for increasing digital resources for many natural languages. Moreover, as a further advantage of the proposed solutions, the improved models contribute to expanding the language identification coverage among world languages thereby increasing the capacity of crime prevention systems and digital forensic investigation strategies.

1.7 Research Scopes

The scope of this research is limited to the following:

i. The proposed study will focus on reviewing previous work related to language modeling using statistical and computational intelligence approaches for purposes of language identification.

ii. In this study an improved lexicon based model is proposed along with computational algorithms for language identification.

iii. The proposed method was tested and analyzed using the publicly available universal declaration of human rights (UDHR) corpus obtained from UNESCO website (UNESCO, 2011). This data set is considered suitable for this study since it consists of standard legal genre text in over 300 languages. The UDHR has been tagged the most translated document in the world (UNESCO, 2011). This delivers two advantages. Firstly, the document consists of only 30 articles of the law, which means that it is not very large. This fits perfectly into the requirement of this research by allowing the testing of the *small data set* constraint. Secondly, the fact that it is a translation means that it is a good data set for testing the ability of the algorithms to distinguish closely related languages, because being a translation implies that the documents are semantically identical. Therefore the only thing that would distinguish any two documents in this corpus is the words and style of writing the respective languages.

iv. In addition, this research only experiments on identification of 15 languages, comprising nine African languages (Hausa, Igbo, Tiv, Yoruba Asante, Akuapem, Ndebele, Zulu, and Swahili), two Asian languages (Malay and Indonesian) and, four European languages (Serbian, Slovak, Croatian, and English). This selection was deliberate in including two Asian languages which are strictly not resource-poor but are closely related languages. The same can be said of Serbian and Croatian. The English language is possibly the most resourced language but is included here to test the viability of the proposed approaches to the richly resourced languages of the world.

v. Developed models were validated using many data sets, namely the (a) Universal declaration of human rights act (UNESCO, 2011). (b) Documents downloaded in the 11 official languages of South Africa (from the South African government services website) to obtain text of other genre e.g. history, science, medicine, and politics for testing language identification using spellchecker technique. (c) A Large data set on English language downloaded from 'Project Gutenberg'

(Bird, 2006). This was used to test the performance of the algorithms on large data sets. (d) Text in 9 European languages from the Europarl corpus Tiedemann (2012). (e) Text in 9 European languages from the Leipzig corpus Quasthoff *et al.* (2006).

vi. Performance of the algorithms was evaluated based on accuracy, processing speed, precision, recall, and F_1 measures.

1.8 Structure of Thesis

This thesis is made up of eight chapters as follows:

i. Chapter 1 presents a general introduction, the problem statement, motivation, aims, objectives, scope, significance and expected contributions. The chapter also narrates the organization of the entire thesis.

ii. Chapter 2 discusses a review of the literature related to this study. Outstanding problems in language identification research are highlighted along with the weaknesses and strengths of earlier used approaches. Also discussed is the spelling checker method as a new entrant into language identification research with special appeal to language identification of under-resourced languages. The present direction of language identification research is discussed along with coverage of language identification among the languages of the world.

iii. Chapter 3 describes the methodology employed in this research and presents the operational framework for the entire study. The methodological steps are discussed including the data gathering, pre-processing and the processing steps for each method beginning with vocabulary extension and going on to word length and model elimination. The universal declaration of human rights acts (UDHR) and details of other corpora are presented as the data set for this research.

iv. In Chapter 4 presents a detailed description of the word list based model and highlights the steps for exploitation of a document's inherent structure for language identification. A detailed experimental validation of this model is discussed along with results for the 15 languages studied. Also presented in this chapter, is the algorithm for vocabulary extension. The strength of a spellchecker lies in the size of its lexicon. This chapter includes detailed results and discussion of experimental analysis of vocabulary extension.

v. Chapter 5 describes the word length algorithm as an enhancement of the bag-of-words model. The chapter presents the theoretical background for the algorithm as well as experimental results and discussions for the word length algorithm. A detailed comparison of the performance of this approach with other standard approaches is presented.

vi. In Chapter 6, a discussion of the dynamic model selection (model elimination) algorithm is presented along with the motivating principles for this approach, namely: the ”Pareto principle”, ”Zipfs law” and power laws. This chapter discusses how to leverage on the content of the text to be identified in order to reduce processing time and memory utilization. Detailed comparison of results from the proposed approach with results from other standard approaches is presented.

vii. Chapter 7 explains a linear combination of the word length and model elimination algorithms for language identification. The idea is to leverage on the strengths of the word length algorithm and the model elimination algorithm by developing a linear combination that feeds the output of the word length algorithm into the model elimination algorithm. Chapter 6 presents experimental setup and results for the proposed approach with detailed comparison of the proposed approach and other standard approaches for language identification.

viii. Chapter 8 discusses research findings, overall thesis contributions in conjunction with the earlier set objectives, conclusions, and recommendations for future work.

1.9 Summary

This chapter discusses the introduction of the research presented in this thesis. The introduction covered the problem statement, motivation, aims, objectives, scope, significance and expected contributions of the research. The chapter highlighted the methods that will be investigated with a view to accomplishing language identification. Peculiar considerations for resource-poor languages along with methodological constraints involved in automatic identification of this class of languages have been enumerated. Improved approaches have been proposed and presented in the brief summaries of each of the chapters of this thesis covering the methods that will be investigated and enhanced in order to carry out language identification in line with the objectives earlier set for this research.

REFERENCES

Adegbola, T. (2009). Building capacities in human language technology for African languages. In *Proceedings of the First Workshop on Language Technologies for African Languages*. Athens, Greece: Association for Computational Linguistics, 53–58.

Al-Salman, A. M., Alkanhal, M. and AlOhali, Y. (2007). Mubser: a bilingual Braille to text translation with an Arabic interface. *International Journal of Web Information Systems*. 3(3), 257–271.

Amine, A., Elberrichi, Z. and Simonet, M. (2010). Automatic Language identification: An alternative unsupervised approach using a new hybrid algorithm. *International Journal of Computer Science and Applications. Technomathematics Research Foundations*. 7(1), 94–107.

Baldwin, T. and Lui, M. (2010). Language identification: The long and the short of the matter. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics*.

Barbaresi, A. (2013). Challenges in web corpus construction for low-resource languages in a post-BootCaT world. In *6th Language & Technology Conference, Less Resourced Languages special track, Poznan : Poland (2013)*.

Barroso, N., Delpi, K. L., Ezeiza, A., Barroso, O. and Susperregi, U. (2010). Hybrid approach for language identification oriented to multilingual speech recognition in the basque contextc. In *Proceedings of the 5th international conference on Hybrid Artificial Intelligence Systems - Volume Part I, San Sebasti225, Spain.*

Beesley, K. (1988). Language identifier: A computer program for automatic natural language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*. 47–54. doi:doi=10.1.1.22.9868.

Bergsma, S., Mcname, P., Bagdouri, M., Fink, C. and Wilson, T. (2012). Language identification for creating language-specific twitter collections. In *Proc. Second Workshop on Language in Social Media*. 65–74.

Bilcu, E. B. and Astola, J. (2006). A hybrid neural network for language identification from text. In *Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*. IEEE Comput. Soc. Press.

Bird, S. (2006.). NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics*.

Botha, G. R. and Barnard, E. (2006). Text-based language identification for the South African languages. In *Proceedings of the 17th Annual Symposium of the Pattern Recognition Association of South Africa. Parys, South Africa*. 7–13.

Botha, G. R. and Barnard, E. (2012). Factors that affect the accuracy of text-based Language Identification. *In Computer Speech and Language (In press uncorrected proof available on 16/1/2012.*

Boudin, F., Huet, S. and Torres-Moreno, J.-M. (2011). Graph-based Approach to Cross-language Multi-document Summarization. *Polibits*. 43, 113–118.

Brown, R. D. (2012). Finding and identifying text in 900+ languages. *Digital Investigation*. 9, 34–43. doi:10.1016/j.diin.2012.05.004.

Carnegie Mellon University, C. M. U. (2008). CMU Pronouncing Dictionary. Retrievable at `http://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/logios/`.

Carter, S., Tsagkias, M. and Weerkamp., W. (2011). Semi-supervised priors for microblog language identification. In *Dutch-Belgian information retrieval workshop(DIR-2011)*. Amsterdam. Retrievable at `http://www.wouter.weerkamp.com/downloads/poster-dir2011-lid.pdf`.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, Nevada, USA*. 161–175.

Ceylan, H. and Kim, Y. (2009). Language identification of search engine queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2.*, vol. 2. Association for Computational Linguistics.

Choong, C. Y., Mikami, Y., Marasinghe, C. A. and Nandasara, S. T. (2009). Optimizing n-gram Order of an n-gram Based Language Identification Algorithm for 68 Written Languages. *The Intl. Journal on Advances for ICT for Emerging Regions.* 02, 21–28.

Choong, C. Y., Mikami, Y. and Nagano, R. L. (2011). Language Identification of web Pages Based on Improved N-gram Algorithm. *Intl. Journal of Computer Science*

Issues, 8(3(1)), . 8, 1694–0814.

Clement, R. and Sharp, D. (2003). Ngram and Bayesian Classification of Documents for Topic and Authorship. *Literary and Linguistic Computing*. 18, 423–447.

CompuFocus, S. D. (2011). *Ispell vir Afrikaans*. Retrievable at `http://compufocus.co.za/afrik/ispell/`.

da Silva, J. F. and Lopes, G. P. (2006). Identification of Document Language is Not yet a Completely Solved Problem. In *Proceedings of the International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce. IEEE.* IEEE Computer Society, 212. doi:10.1109/cimca.2006.117.

Dahl, H. (1979). *Word Frequencies of Spoken American English.* Verbatim, Essex, CT.

Dehzangi, O., Bin, M. and Haizhou, L. (2010). Error corrective classifier fusion for spoken Language Recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. ISBN 1520-6149, 1994–1997. doi:10.1109/icassp.2010.5495235.

Ehsan, N. and Faili., H. (2010). Towards grammar checker development for Persian language. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on. IEEE, 2010*. 1–8. doi:10.1109/nlpke.2010.5587839.

Estrella, P. S., Hamon, O. and Popescu-Belis, A. (2007). How much data is needed for reliable MT evaluation? Using bootstrapping to study human and automatic metrics., 167–174.

Fiol-Roig, G., Mir-Juli, M. and Herraiz, E. (2011). Data Mining Techniques for Web Page Classification. In *Highlights in Practical Applications of Agents and Multiagent Systems.* Springer Berlin Heidelberg, 61–68.

Gaultney, J. V. (2011). *Gentium Typeface*. Retrievable at `http://www.sil.org/~gaultney/gentium/download.html`.

Gebre, B. G., Zampieri, M., Wittenburg, P. and Heskes, T. (2013). Improving native language identification with tf-idf weighting. *NAACL/HLT 2013*, 216.

Gen, M. and Cheng, R. (2002). *Genetic Algorithms and Engineering Optimization*. Willey, New York.

Gordon, R. G. (2005). *Ethnologue: Languages of the world.* SIL International. Dallas, TX.

Gottron, T. and Lipka, N. (2010). A Comparison of Language Identification Approaches on Short, Query-Style Texts. In Gurrin Y.; Kazai, G.; Kruschwitz, U.; Little, S.; Roelleke, T.; Ruger, S.; and van Rijsbergen, K.,, C. H. (Ed.) *Advances in Information Retrieval, Lecture Notes in Computer Science*, vol. 5993. Springer Berlin / Heidelberg, 611–614.

Grefenstette, E. and Pulman, S. (2010). *Analysing Document Similarity Measures*. Master's Thesis. University of Oxford.

Grefenstette, G. (1995). Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data.*

Grothe, L., Luca, E. and Nrnberger, A. (2008). A Comparative Study on Language Identification Methods. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).*

Hammarstr-om, H. (2007). A Fine-Grained Model for Language Identification. In *Workshop of Improving Non English Web Searching. Amsterdam, The Netherlands*. 14–20.

Huang, C. and Lee, L. (2008). Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of PACLIC 2008*. 404410.

Hughes, B., Baldwin, T., Bird, S. and Nicholson, S., J.and MacKinlay (2006). Reconsidering language identification for written language resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy. , 485-488.*

Hurskainen, A. (1999). SALAMA: Swahili Language Manager. *Nordic Journal of African Studies*. 8(2), 139–157.

Ingle, N. C. (1996). A language identification table. *The Incorporated Linguist.* 15(4), 98–101.

Jiang, J. J. and Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of ROCLING-X, the ROCLING 1997 conference on Research on Computational Linguistics*.

Jothilakshmi, S. and Palanivel, S. (2012). A hierarchical Language Identification system for Indian Languages. *In Digital Signal Processing, In press. Corrected proof available on 27/1/2012*. doi:10.10.1016/j.dsp.2011.11.008.

Klein, E. and Loper, E. (2009). *Natural Language Processing with Python.* O'reilly Media.

Kockmann, M., Burget, L. and Ernock, J. H. (2011). Application of speaker- and language identification state-of-the-art techniques for emotion recognition. *Speech Commun.* 53, 1172–1185.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence, vol. 14. 1137-1145*, vol. 14. 1137–1145.

Koller, D. and Sahami, M. (1997). Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine learning*. Stanford InfoLab.

Komisin, M. C. (2012.). *Identifying Personality Types Using Document Classification Methods.* Master's Thesis. University of North Carolina.

Kralisch, A. and Mandl, T. (2006). Barriers to Information Access across Languages on the Internet: Network and Language Effects. In *System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on*, vol. 3. ISBN 1530-1605, 54b–54b. doi:10.1109/hicss.2006.71.

Kruengkrai, C., Srichaivattana, P., Sornlertlamvanich, V. and Isahara, H. (2005). Language identification based on string kernels. In *Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on*, vol. 2. 926–929. doi:10.1109/iscit.2005.1567018.

Lee, Y. S., Shamsuddin, S. M. and Hamed, H. N. (2008). Bounded PSO vmax function in neural network learning. In *Proceedings of the 2008 Eighth International Conference on Intelligent Systems Design and Applications. Washington, DC, USA: IEEE Computer Society. ISBN 978-0-7695-3382-7, 474-479.*

Levow, G.-A., Oard, D. W. and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*. 41(3), 523–547. doi:10.1016/j.ipm.2004.06.012. Retrievable at `http://www.sciencedirect.com/science/article/pii/S0306457304000810`.

Lewandowski, D. (2008). Problems with the use of web search engines to find results in foreign languages. *Online Information Review*. 32(5), 668–672. doi:10.1108/14684520810914034(PermanentURL).

Lewis, D. (1992). An Evaluation of phrasal and clustered representations on a text categorization task. In *SIGIR-92*.

Lewis, M. P. (2009). *Ethnologue: Languages of the world sixteenth edition.* Dallas, TX.: SIL International. Retrievable at `Onlineversion:http://www.ethnologue.com(2009).`

LexTek-International (2012). *LexTek language identifier SDK*. Retrievable at `http://www.lextek.com/langid/li/languages.htm`.

Li, W. (1992). Random texts exhibit Zipfs-law-like word frequency distribution. *IEEE Transactions on Information Theory*. (38), 1842–1845.

Ljubesic, N., Mikelic, N. and Boras, D. (2007). Language Indentification: How to Distinguish Similar Languages? In *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*. ISBN 1330-1012, 541–546. doi:10.1109/iti.2007.4283829.

Lui, M. and Baldwin., T. (2011). Cross-domain Feature Selection for Language Identification. *IJCNLP.*, 553–561.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *ACL (System Demonstrations)*. 25–30.

Mahapatra, L., Mohan, M., Khapra, M. M. and Bhattacharyya, P. (2010). OWNS: Crosslingual word sense disambiguation using weighted overlap counts and Wordnet based similarity measures. In *Proceedings of the 5th International Workshopon Semantic Evaluation (SemEval 2010), Uppsala,Sweden.*

Manning, C. D. and Schutze, H. (2002). *Foundations of statistical natural language processing*. Cambridge, Massachusetts: The MIT Press.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *In AAAI-98 workshop on learning for text categorization, vol. 752,*. 41–48.

McNamee, P. (2005). Language identification: a solved problem suitable for undergraduate instruction. *J. Comput. Small Coll.* 20(3), 94–101.

Mishra, G., Nitharwal, S. L. and Kaur, S. (2010). Language identification using Fuzzy-SVM technique. In *Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on*. 1–5. doi:10.1109/icccnt.2010.5592553.

Murthy, K. and Kumar, G. (2006). Language identification from small text samples. *Journal of Quantitative Linguistics.* 13(1), 57–80.

Nagano, R. L. (2011). Language Identification of web Pages Based on Improved N-gram Algorithm. *Intl. Journal of Computer Science Issues*. 8(3(1)), 814–1694.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp. Phy.* 46, 323–351.

Newman, P. (1987). Foreign language identification - A first step in translation. In *Proceedings of the 28th Annual Conference of the American Translators Association.* 509–516.

Ng, C.-C. and Selamat, A. (2011). Improving Language Identification of Web Page

Using Optimum Profile. In Zain Wan Mohd, Wan Maseri bt, El-Qawasmeh, Eyas, J. M. (Ed.) *Communications in Computer and Information Science Vol. 180*, vol. 180. Springer Berlin Heidelberg. ISBN 978-3-642-22191-0, 157–166. doi: 10.1007/978-3-642-22191-0_14. Retrievable at http://dx.doi.org/10.1007/978-3-642-22191-0_14.

Nguyen, D. and Dogruoz, S. (2013). Word level language identification in online multilingual communication. In *Proceedings of the 2013 EMNLP2013 Seatle, USA*.

Nicholson, J. and S.MacKinlay (2006). Reconsidering language identification for written language resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy.* 485–488.

Padro, M. and Padro, L. (2004). Comparing methods for language identification. In *proceedings of the XX Congreso de la Sociedad Espanol apara el Procesamieno del language Natural, (SEPLN 04), Barcelona, Spain).*

Pavan, K., Tandon, N. and Varma, V. (2010). Addressing challenges in automatic Language Identification of Romanized Text. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing.* Macmillan Publishers, India.

Pienaar, W. and Snyman, D. P. (2010). Spelling Checker-based Language Identification for the Eleven Official South African Languages. In *Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa. 22-23 November 2010. Stellenbosch, South Africa*. 213–216. Retrievable at http://www.prasa.org/proceedings/2010/prasa2010-36.pdf.

Prager, J. M. (1999). Linguini: language identification for multilingual documents. In *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*, vol. Track2. 11 pp. doi:10.1109/hicss.1999.772689.

Prinsloo, D. (2000). The Compilation of Electronic Corpora with special reference to African Languages. *South African Linguistic and Applied Language Studies*. 18(1-4), 89–106.

Prinsloo, D. J. and Schryver, G. M. (2003a). Non-word error detection in current South African Spellcheckers. *Southern African Linguistic and Applied Language Studies*. 21(4), 307–326.

Prinsloo, D. J. and Schryver, G. M. (2003b). Towards Second-Generation Spellcheckers for the South African Languages. In *De Schryver, G-M (ed.). 2003. TAMA 2003 South Africa: Conference Proceedings: Pretoria: (SF)2 Press.* 135–141.

Quasthoff, U., Richter, M. and Biemann, C. (2006). Corpus Portal for Search in

Monolingual Corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006, Genoa*. 1799–1802.

Ralf, S., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006*.

Ramisch, C. (2008). *N-gram models for language detection.* Retrievable at `https://docs.google.com/...mohammadzadeh.info/media/blogs/snf/Resources/LangID/N-gram`.

Ranaivo-Malancon, B. (2006). Automatic Identification of Close Languages - Case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*. 2(2), 126–134.

Read, J., Velldal, E. and vrelid., L. (2012). Topic Classification for Suicidology. *JCSE*. 6(2), 143–150.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J.* (B 4), 131–134.

Reese, S., Boleda, G., Cuadros, M., Padr, L. and Rigau., G. (2010). Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10), La Valleta, Malta.*

Rehurek, R. and Kolkus, M. (2009). Language Identification on the Web: Extending the Dictionary Method. In Gelbukh, A. (Ed.) *Computational Linguistics and Intelligent Text Processing, 10th International Conference, CICLing 2009, Proceedings. Vyd. první. Mexico City, Mexico: Springer-Verlag, 2009. ISBN 978-3-642-00381-3, pp. 357-368*, vol. 5449. Springer Berlin / Heidelberg. ISBN 978-3-642-00381-3, 357–368. doi:10.1007/978-3-642-00382-0_29. Retrievable at `http://dx.doi.org/10.1007/978-3-642-00382-0\_29`.

Roux, J. C. (2008). *HLT Development in Sub-saharan Africa.* Technical report. Marrakesh. Retrievable at `http://www.ilc.cnr.it/flarenet/documents/lrec2008\_cocosda-write\_worshop\_roux.pdf`.

Sahni, S. (2005). *Analysis of Algorithms*. Chapman & Hall/CRC Computer and information Sc series.

Scannell, K. P. (2007). The Crubadan Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, vol. 4. 515.

Selamat, A. (2011). Improved N-grams Approach for Web Page Language

Identification. *Transactions on Computational Collective Intelligence*. 6910, 1–26. doi:10.1007/978-3-642-24016-4_1. Retrievable at http://dx.doi.org/10.1007/978-3-642-24016-4_1.

Selamat, A. and Lee, Z. S. (2008). Language identifications of Arabic script web documents using independent component analysis. In *Proceedings of the Second Asia International Conference on Modeling & Simulation*. 427–432.

Selamat, A., Lee, Z. S., Maarof, M. A. and Shamsuddin, S. (2011). Improved Web page identification method using Neural Networks. *Int. J. Comp. Intel. Appl.* 10.

Selamat, A. and Ng, C.-C. (2009a). Improved Letter Weighting Feature Selection on Arabic Script Language Identification. In *Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference on*. 150–154. doi:10.1109/aciids.2009.33.

Selamat, A. and Ng, C.-C. (2009b). Improved Letter Weighting Feature Selection on Arabic Script Language Identification. In *Intelligent Information and Database Systems, 2009. ACIIDS 2009. First Asian Conference on*.

Selamat, A. and Ng, C. C. (2011). Arabic script web page language identifications using decision tree neural networks. *Pattern Recognition*. 44, 133–144.

Selamat, A., Ng, C.-C., Selamat, M. D. and Bujang, S. (2009a). Agent Architecture for Criminal Mobile Devices Identification Systems. In Nguyen, N., Katarzyniak, R. and Janiak, A. (Eds.) *New Challenges in Computational Collective Intelligence, Springer BerlinHeidelberg*. (pp. 281–290). vol. 244. Springer Berlin / Heidelberg. ISBN 978-3-642-03957-7. doi:10.1007/978-3-642-03958-4_24. Retrievable at http://dx.doi.org/10.1007/978-3-642-03958-4_24.

Selamat, A., Subroto, I. M. I. and Ng, C.-C. (2009b). Arabic Script Web Page Language Identification Using Hybrid-KNN Method. *International Journal of Computational Intelligence and Applications*. 18(3), 315–343.

Selamat, A. and Zhi-Sam, L. (2008). Language Identifications of Arabic Script Web Documents Using Independent Component Analysis. In *Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on*. 427–432. doi:10.1109/ams.2008.46.

Smucker, J. A., Mark D. and Carterette., B. (2007). ”A comparison of statistical significance tests for information retrieval evaluation.”. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM*.

Souter, C., Churcher, G., Hayes, J., Hughes, J. and Johnson, S. (1994). Natural language identification using corpus-based models. *Hermes Journal of Linguistics.*

13, 183– 203.

Sun, A. and Liu, Y. (2011). Web Classification of Conceptual Entities using Co-training. *Expert Systems with Applications*. 38, 14367–14375.

Susperregi, U. (2010). Hybrid approach for language identification oriented to multilingual speech recognition in the basque context. In *Proceedings of the 5th international conference on Hybrid Artificial Intelligence Systems - Volume Part I*. San Sebastián, Spain: Springer-Verlag, 196–204. doi:10.1007/978-3-642-13769-3_24.

Takci, H. and Gungor, T. (2012). A high performance centroid-based classification approach for language identification. *Pattern Recognition Letters*. 33(16), 2077–2084. doi:http://dx.doi.org/10.1016/j.patrec.2012.06.012. Retrievable at `http://www.sciencedirect.com/science/article/pii/S0167865512002000`.

Takci, H. and Soukpnar, A. (2005). Letter based text scoring method for language identification. *Advances in Information Systems*. doi:10.1007/978-3-540-30198-1_29. Retrievable at `http://link.springer.com/chapter/10.1007/978-3-540-30198-1\_29`.

Tiedemann, J. (2009). *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces.*, John Benjamins, Amsterdam/Philadelphia, chap. Advances in Natural Language Processing (vol V). 237–248.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*.

Tiedemann, J. and Ljubesic, N. (2012). Efficient Discrimination Between Closely Related Languages. In *Proceedings of COLING 2012: Technical Papers Mumbai, December 2012*. 26192634.

Tromp, E. and Pechenizkiy, M. (2011). Graph-based N-gram language identification on short texts. In *The 20th annual Belgian-dutch conference on machine learning (BENELEARN-2011)*. The Netherlands, 27–34.

UNESCO (2011). *Office of the High commissioner for Human Rights 1948-, Universal Declaration of Human Rights*. Retrievable at `http://193.194.134.190/udhr/index.htm[Accessedon19thAugust,2011]`.

Varma, V. (2010). Addressing challenges in automatic Language Identification of Romanized Text. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India.* Retrievable at `http://ltrc.iiit.ac.in/proceedings/ICON-2010`.

Vatanen, T., Väyrynen, J. J. and Virpioja, S. (2010). Language identification of short text segments with n-gram models. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. 3423–3430.

Veken, A. V. d. and Schryver, G. M. (2003). Les langues africaines sur la Toile. Etudie des cas Haoussa, Somali, Lingala et isi-xhosa. In *Cahiers du Rifal 23 (Theme: Le traitement informatique des langues Africaines): pp. 33-45, 2003*.

Wang, S.-J. (2007). Measures of retaining digital evidence to prosecute computer-based cyber-crimes. *Comput. Stand. Interfaces*. 29(2), 216–223. doi:10.1016/j.csi.2006.03.008.

Windisch, G. and Csink, L. (2005). Language identification using global statistics of natural languages. In *SACI. 2005*.

Winkelmolen, F. and Mascardi, V. (2011). Statistical Language Identification of Short Texts. In *ICAART*, vol. 1. 498–503.

Xi, Y. and Wenxin, L. (2010). An N-Gram-and-Wikipedia joint approach to Natural Language Identification. In *Proceedings of the 2010 Universal Communication Symposium (IUCS), 2010 4th International. 2010.* 332–339.

Xia, F., Lewis, W. D. and Poon, H. (2009). Language ID in the context of harvesting data off the web. In *proceedings of The 12th Conference of the European Chapter of the Association of Computational Linguistics.(EACL 2009). Athens, Greece*.

Yamaguchi, H. and Tanaka-Ishii, K. (2012). Text segmentation by language using minimum description length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Jeju Island, Korea: Association for Computational Linguistics, 969–978.

Yang, C.-Y. and Wu, S.-J. (2012). Semantic Web Information Retrieval Based on the Wordnet. *JDCTA: International journal of Digital Content Technology and its Applications*. 6(6), 294–302.

Yang, J., Zhang, X., Suo, H., Lu, L., Zhang, J. and Yan, Y. (2012). Maximum A Posteriori Linear Regression for language recognition. *Expert Systems with Applications*. 39(4), 4287–4291. ISSN 09574174. doi:10.1016/j.eswa.2011.09.104. Retrievable at `http://www.sciencedirect.com/science/article/pii/S0957417411014278http://linkinghub.elsevier.com/retrieve/pii/S0957417411014278`.

Yang, X. and Liang, W. (2010). An N-Gram-and-Wikipedia joint approach to Natural Language Identification. In *Universal Communication Symposium (IUCS), 2010 4th International*. 332–339. doi:10.1109/iucs.2010.5666010.

Zampieri, M. (2013). Using bag-of-words to distinguish similar languages: How efficient are they? In *Computational Intelligence and Informatics (CINTI), 2013 IEEE 14th International Symposium on*. doi:10.1109/CINTI.2013.6705230.

Zampieri, M., Gebre, B. G. and Nijmegen, H. (2012). Automatic identification of language varieties: The case of Portuguese. In *Jancsary, J., editor, Proceedings of KONVENS, ÖGAI*. 233–237.

Zhai, L. F., Siu, M. H., Yang, X. and Gish, H. (2006). Discriminatively trained language models using support vector machines for language identification. In *Proceedings of the IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*. 1–6.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Reading, MA.