DOCTOR OF PHILOSOPHY

Analysing and Correcting Dyslexic Arabic Texts

Alamri, Maha

Award date:
2019

Awarding institution:
Bangor University

Link to publication

PRIFYSGOL BANGOR UNIVERSITY

School of Computer Science and Electronic Engineering
College of Environmental Sciences and Engineering

Analysing and Correcting Dyslexic Arabic Texts

Maha Marzouq Alamri

Submitted in partial satisfaction of the requirements for the
Degree of Doctor of Philosophy
in Computer Science

Acknowledgements

Always and forever, all praise and thanks be to God (Allah) for everything and for guiding me towards the successful accomplishment of my PhD—a dream that has now become a reality.

I owe many people my sincere gratitude for helping me throughout the process. I should start with expressing my special gratitude to my supervisor, Dr. William J. Teahan, for every single act of support that he has offered for me—thank you for your substantial feedback and dedicated encouragement.

I am especially and profoundly grateful to the provenience of my love, my extended family. Many special thanks and love go to my parents, Marzouq and Faddah, for their infinite love, uncountable prayers and permanent wishes. My brothers and sisters, you know how excited I was throughout the past years in spite of the anticipated obstacles, and you were such huge supporters behind the scene, making me stronger and laugh louder whilst also inspiring me to keep going. I love and thank you all for everything you have done for me.

My special gratitude and appreciation also go directly to my immediate family. My husband, Turki, I know how proud you are feeling right now and

I know it is because my dream has now been achieved. I know there are no words that can describe the uncertainty we had experienced during the long journey of my study, however, you stood by me in everything and always allowed me to nurture my ambition and sustain my enthusiasm towards this moment. My deepest thanks should go to you, and to our daughter, Deema, the light of my life, whose presence with us is the true platform to our happiness.

I would also like to give special thanks to those who participated in this research and contributed to the study in spite of their exhausting academic schedules. Your contributions are key to the successful completion of this thesis, and I am sincerely grateful for your involvement. In addition, I wish to thank everyone who offered me invaluable advice, encouragement and support, including Dukhnah Alamri (lecturer) and Asma Aljathlan, Elham Alsaleh (teachers) and Maha Alammar.

I would like to thank Bangor University for providing an ideal environment for study and research. I am profoundly thankful to all staff and fellow researchers at the School of Computer Science and Electronic Engineering for their help and direction, especially Prof. Ludmila Kuncheva, Dr. Mohammed Mabrook, Dr. Noor Al-Kazaz and Dr. Nadim Ahmed.

Finally, I would like to acknowledge my country, the Kingdom of Saudi Arabia, and Al-Baha University for providing me the invaluable opportunity to study in the United Kingdom.

I dedicate this thesis

To my parents, for their love, prayers, endless support and encouragement;

To my sister Wadha, for standing by me when things turn bleak; and

To people with dyslexia, for inspiring me.

Abstract

Dyslexia is a disorder that involves difficulty with literacy skills and language related skills. It is related to the inability of a person to master the utilisation of written language and affects a significant number of people. This thesis describes the development of the Bangor Dyslexia Arabic Corpus (BDAC) in order to facilitate the analysis and automatic correction of dyslexic Arabic text. This thesis has also developed a new classification of errors made in Arabic by people with dyslexia which was used in the annotation of the BDAC. The dyslexic error classification scheme for Arabic texts (DECA) comprises a list of dyslexia spelling errors classified into 37 types, and grouped into nine categories.

This thesis also investigates a new type of classification – dyslexia text classification – that identifies whether or not a text has been written by a person with dyslexia. The text compression scheme known as prediction by partial matching (PPM) has been applied to the problem of distinguishing dyslexic text from non-dyslexic text. Experimental results show that the F_1 score for PPM-based classification was 0.99 and outperformed other classifiers such as Multinomial Naïve Bayes and Support Vector Machiness.

A new system called Sahah is also proposed for the automatic detection and correction of dyslexia errors in Arabic text. The system uses a language model based on the PPM text compression scheme in addition to edit operations (omission, addition, substitution and transposition). The correct alternative for each error word is chosen on the basis of the compression codelength. Two experiments were carried out to evaluate the usefulness of the Sahah system. Firstly, its accuracy was evaluated using the BDAC

containing errors made by people with dyslexia. Secondly, the results of Sahah were compared with the results obtained when using word processing software and the Farasa tool. The results show that the Sahah system significantly outperforms Microsoft Word, Ayaspell and the Farasa tool with an F_1 score of 0.83 for detection and an F_1 score of 0.58 for correction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The word dyslexia originates from the Greek language and signifies difficulty with words (Ghazaleh, 2011), and specifically issues with reading, spelling, and word recognition (Grigorenko, 2001). The earliest consideration of dyslexia was presented by W. Pringle Morgan, in November 1896 (Morgan, 1896). The article described the case of a 14-year-old boy who was apparently at an adequate level of intelligence and logical reasoning for his age, yet struggled considerably in terms of reading and writing skills. This article is one of the first reports concerning congenital word blindness. Therefore, Morgan is often considered as being the pioneer in the field of dyslexia (Guardiola, 2001).

There seems to also be a significant amount of people who have dyslexia, as the International Dyslexia Association (2012) reported that dyslexia affects 15-20% of any given population. It should also be noted that there is no relationship between dyslexia and a person's level of intelligence. Dyslexia is popularly identified with numerous famous figures, such as Richard Branson.

Dyslexia concerns difficulty with acquiring literacy skills, and this difficulty can be present throughout person's life, and might influence their education over the long-term. Furthermore, the disorder is not culture- or language-specific, and is therefore not exclusive to specific cultures, and is observable in all age groups and in all languages (Reid, 2010), such as Arabic.

The Arabic language is one of the most widely used in many parts of the world. A study conducted by Holes (2004) suggested that two fundamental reasons exist for the wide usage of this language, the first that Arabic is the language of the 'Holy Quran', and the second that other languages, such as Urdu and Farsi, employ Arabic letters. Thus, Arabic was selected as the focus of this thesis in addition to it being the researcher's native language. Moreover, the researcher wished to help Arabic people with dyslexia, and to add value for this target group.

Nevertheless, despite the widespread use of the language, academic research concerning dyslexia in Arabic is scarce, because dyslexia is not widely recognised in the Arab region (Aboudan et al., 2011). This is evidenced by the fact that the first dyslexia association in the Arab world was established in Kuwait in 1999, many years after the equivalent association was established in the West, where the oldest such association, the International Dyslexia Association was established in 1949.

There are a number of different approaches that can be employed to help and support people with dyslexia, such as tools to assist in its diagnosis and assessment, and applications such as word prediction software, text classification, and spelling correction. These tools have been developed through different methods such as natural language processing and corpus linguistics.

The use of text corpora has expanded in recent years as it plays a significant role in different aspects such as computational linguistics, and Natural Language Processing (NLP) research. Although the use of text corpora has enjoyed a relatively high level of interest in research, availability of dyslexia corpora is scarce (Pedler, 2007), and only a few studies have considered the potential benefits of using dyslexia corpora.

Furthermore, there is an obvious lack of Arabic dyslexia corpora, which highlights the importance of improving and enlarging the extant resource that was previously developed by the researcher of this thesis (Alamri, 2013). Such a corpus can be used as a starting point for developing a more extensive understanding of dyslexic errors in Arabic, and how they are written and moreover, towards investigating and developing Arabic dyslexia applications. Furthermore, it can serve as a platform for future researchers to develop further studies in the area, or to employ in the creation of applications for dyslexics.

These points formed the main inspiration for conducting this research study.

1.2 Research Questions

The research questions explored for this study are as follows:

1. What is an effective spelling error classification scheme for annotating and analysing Arabic dyslexic corpora?
2. How well dose a compression-based language modelling method, such

as the Prediction by Partial Matching (PPM) text compression method, compare to two well performed algorithms such as Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM) for classifying a text that has been written by a person with dyslexia?

3. Can PPM, in conjunction with other methods, be effectively applied to correcting a text that has been written by a person with dyslexia?

1.3 Aim and Objectives

The aim of this study is to investigate the effectiveness of a new approach to classifying and correcting Arabic dyslexic text, specifically, using the PPM compression method. This study seeks to evaluate how well this approach performs in applications using Arabic dyslexic corpus.

Therefore, this study's objectives in investigating the research questions are as follows:

- Review the extant literature regarding dyslexia, Arabic language, dyslexia spelling errors, corpus linguistics, text classification, spelling correction, and text compression (see Chapter 2);
- Improve the existing Arabic corpus of texts written by people with dyslexia (the Bangor Dyslexia Arabic Corpus (BDAC)) (see Chapter 3);
- Create a new dyslexic error classification scheme for Arabic dyslexic texts (DECA) (see Chapter 4);
- Develop and evaluate a method to classify whether or not a text has been written by a person with dyslexia, using the PPM compression

scheme, and compare the performance of the PPM with other classification methods, such as the Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM), when they are employed for the purpose of classifying dyslexic text (see Chapter 5);

- Design and evaluate an automatic spelling correction system for correcting spelling errors in Arabic texts, produced by people with dyslexia, by comparing them with other spelling correction tools (see Chapter 6).

1.4 Contributions

Since there is currently a lack of an Arabic dyslexia corpus, this study has the potential to make valuable contributions to the field. The specific contributions are as follows:

- The first and foremost contribution of this research study is the enlargement of the Arabic dyslexia corpus (the BDAC) to comprise 28,203 words written by both male and female with dyslexia aged between 8 to 13 year olds. Based on the literature review, the BDAC is the first dyslexia corpus for Arabic.

- The second contribution is the development of dyslexic error classification scheme for Arabic texts (DECA) that can provide a framework to help analysing and annotating specific errors committed by writers with dyslexia. Also, this has been used to provide an annotated dyslexic corpus (the BDAC) and then analysis of Arabic dyslexic errors, based on the corpus.

- The third contribution is the creation of Bangor Non-Dyslexia Arabic

Corpus (BNDAC), consisting of 9,099 words written by non-dyslexic male and female between the ages of 8 and 13.

- The fourth contribution is the investigation of an effective, new method for classifying dyslexic text, based on the PPM compression method.
- The final contribution is the development and testing of a new system called Sahah to automatically correct Arabic dyslexic text by using PPM text compression scheme and an edit operation approach using compression codelength.

1.5 Publications

The researcher has already published one conference paper based on this study. In addition, a further two journal papers have been submitted for publication. Table 1.1 shows specific papers which relate to this study.

The first paper, entitled "A New Error Annotation for Dyslexic Texts in Arabic", is included in Chapter 4. The paper describes a new classification scheme of errors made in Arabic by people with dyslexia to be used in the annotation of the Arabic dyslexia corpus (BDAC). The dyslexic error classification scheme for Arabic texts (DECA) comprises a list of spelling errors extracted from previous studies and a collection of texts written by people with dyslexia that can provide a framework to help analyse specific errors committed by writers with dyslexia. The classification comprises 37 types of errors, grouped into nine categories. The paper also discusses building a corpus of dyslexic Arabic texts that uses the error annotation scheme and provides an analysis of the errors that were found in the corpus. The paper was presented at the Third Arabic Natural Language Processing Workshop

Table 1.1: Publications that relate to this study.

1	**Title**	A New Error Annotation for Dyslexic Texts in Arabic
	Authors	Maha M. Alamri, and William J. Teahan
	In	The Third Arabic Natural Language Processing Workshop (WANLP)
	Publisher	Association for Computational Linguistics (ACL)
	Year	2017
	Status	Published
2	**Title**	Distinguishing Dyslexic Text from Non-dyslexic Text
	Authors	Maha M. Alamri, and William J. Teahan
	In	Transactions on Computers (TC) Journal
	Publisher	IEEE
	Year	2019
	Status	Submitted
3	**Title**	Automatic Correction of Arabic Dyslexic Text
	Authors	Maha M. Alamri, and William J. Teahan
	In	Computers Journal
	Publisher	Multidisciplinary Digital Publishing Institute (MDPI)
	Year	2019
	Status	Published

(WANLP) co-located with EACL 2017, held in Valencia, Spain.

The second paper, entitled "Distinguishing Dyslexic Text from Non-dyslexic Text", which Chapter 5 is based upon, investigates a classification problem, specifically dyslexia text classification, which involves identifying whether or not a text has been written by a person with dyslexia. For this purpose, we apply the PPM text compression scheme for the binary classification problem of distinguishing dyslexic text from non-dyslexic text. Various experiments were conducted to evaluate the method using three corpora. Experimental results show that the accuracy for PPM-based classification significantly outperformed standard feature-based classifiers such as Multi-