

**REAL-WORD ERROR DETECTION AND
CORRECTION IN ARABIC TEXT**

BY

MAJED MOHAMMED ABDULQADER AL-JEFRI

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

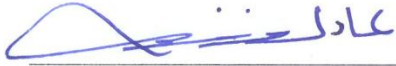
In

INFORMATION AND COMPUTER SCIENCE

May 2013

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN- 31261, SAUDI ARABIA
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **MAJED MOHAMMED ABDULQADER AL-JEFRI** under the direction his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**.



Dr. Adel F. Ahmed
Department Chairman



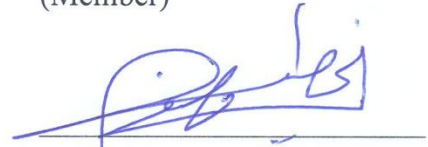
Prof. Sabri A. Mahmoud
(Advisor)

Dr. Salam A. Zummo
Dean of Graduate Studies



Prof. Radwan Abdel-Aal
(Member)

Date



Dr. Wasfi Al-Khatib
(Member)

© Majed Mohammed Abdulqader Al-Jefri

2013

DEDICATED TO

I dedicate this dissertation with all of my love to
Allah first, then to my parents, my wife, my son, my
brothers and sisters.

ACKNOWLEDGMENTS

First and foremost I want to express my deepest gratitude to Allah Almighty who gave me strength, patience and ability to accomplish this work.

I wish to express my appreciation to **Prof. Sabri A. Mahmoud**, for consistent support and guidance through the thesis, his valuable suggestions and encouragement can never be forgotten. Thanks are due to my thesis committee members **Prof. Radwan Abdel-Aal** and **Dr. Wasfi Al-Khatib** for their cooperation, comments and support. Thanks are also due to the Chairman of Information and Computer Science Department **Dr. Adel Ahmed** for providing all the available facilities.

I would like to thank **King Fahd University of Petroleum & Minerals (KFUPM)** for supporting this research and providing the computing facilities.

I also would like to thank **Hadhramout Establishment for Human Development**, which gave me the opportunity for completing my M.Sc. degree at KFUPM. I would like to thank **Dr. Husain Alaidaroos**, the non-native Arabic teacher who provided us with the most common mistakes of non-Arabic speakers and the confusion sets. I would like also to thank **Mr. Irfan Ahmed** for providing us with the OCR data.

I owe thanks to my colleagues and my friends for their help and support. A few of them are Ahmad Aman, Hasan Alkaff, Omer Shaaban, Emran Alaghbari, Adnan Mahdi, and many others; all of whom I will not be able to name here.

Finally I want to thank my parents and my wife whose prayers are always a great source of strength for me.

TABLE OF CONTENTS

| | |
|--|------|
| ACKNOWLEDGMENTS | V |
| TABLE OF CONTENTS | VI |
| LIST OF TABLES | IX |
| LIST OF FIGURES | XI |
| LIST OF ABBREVIATIONS | XII |
| ABSTRACT | XIII |
| CHAPTER 1 INTRODUCTION | 1 |
| 1.1 Introduction | 1 |
| 1.2 Problem Statement | 2 |
| 1.3 Research Objectives | 3 |
| CHAPTER 2 BACKGROUND | 6 |
| 2.1 Introduction | 6 |
| 2.2 Spelling Error Classification..... | 9 |
| 2.2.1 Typing Errors vs. Spelling Errors..... | 9 |
| 2.2.2 Single vs. Multiple Errors..... | 10 |
| 2.2.3 Word Boundary Errors (Run-Ons and Split Words) | 10 |
| 2.2.4 Non-Word Errors vs. Real-Word Errors | 11 |
| 2.2.5 Real-Word Errors Classification..... | 11 |
| 2.3 Arabic Real-Word Spell Checking | 13 |
| CHAPTER 3 LITERATURE REVIEW..... | 15 |

| | |
|---|-----------|
| 3.1 Non-Word Errors | 15 |
| 3.2 Real-Word Errors..... | 17 |
| 3.3 Arabic spell checking and correction | 22 |
| CHAPTER 4 DATA COLLECTION AND PREPARATION | 26 |
| 4.1 Corpus Collection and Preparation..... | 26 |
| 4.2 Dictionary Generation | 29 |
| 4.3 Statistical Language Models..... | 31 |
| 4.4 Confusion Sets | 31 |
| CHAPTER 5 REAL WORD ERROR DETECTION AND CORRECTION USING N-GRAM LANGUAGE MODELS..... | 36 |
| 5.1 Introduction | 36 |
| 5.2 Error Detection Module..... | 37 |
| 5.2.1 Generate candidate words | 40 |
| 5.2.2 Generate candidate sentences using the candidate words | 41 |
| 5.2.3 Ranking candidate corrections..... | 43 |
| 5.2.4 Correct Error Words | 43 |
| 5.3 Experimental Results | 46 |
| 5.3.1 Testing Data..... | 46 |
| 5.3.2 Performance Measures | 47 |
| 5.3.3 Performance Comparison | 52 |
| CHAPTER 6 REAL WORD ERROR DETECTION AND CORRECTION USING SUPERVISED TECHNIQUES | 55 |
| 6.1 Introduction | 55 |
| 6.2 The Baseline Method..... | 56 |
| 6.3 Context Words Co-occurrence Method | 57 |

| | |
|--|-----------|
| 6.4 N-Gram Language Models..... | 60 |
| 6.5 Experimental Results | 61 |
| 6.5.1 Word Co-occurrence..... | 62 |
| 6.5.2 N-gram Language Models..... | 73 |
| 6.5.3 Combining Methods..... | 76 |
| CHAPTER 7 CONCLUSIONS AND FUTURE WORK | 80 |
| 7.1 Conclusions | 80 |
| 7.2 Future Directions..... | 82 |
| REFERENCES | 85 |
| VITAE | 89 |

LIST OF TABLES

| | |
|--|----|
| Table 2.1: Example of several real-words driven from the word ‘أمر’ | 14 |
| Table 4.1: Statistics of our collected corpus | 28 |
| Table 4.2: Sample of number of words' occurrences in the corpus..... | 29 |
| Table 4.3: Statistics of the dictionaries | 30 |
| Table 4.4: The eighteen collected confusion sets misrecognized by the OCR..... | 33 |
| Table 4.5: Common spelling mistakes committed by non-native Arabic speakers | 34 |
| Table 4.6: Nineteen confusion sets from non-native Arabic speakers | 35 |
| Table 5.1: LMs statistics of the corpus | 37 |
| Table 5.2: The word ‘صحبة’ variations | 41 |
| Table 5.3: Statistics of the test sets..... | 46 |
| Table 5.4: Example of TP, FN and FP in a test set with 233 errors..... | 50 |
| Table 5.5: Results on the test sets | 50 |
| Table 5.6: Different cases of false positives | 52 |
| Table 6.1: Baseline method prediction accuracy on the test set for the 28 confusion sets | 58 |
| Table 6.2: Context words method for different values of k using whole words, ignoring stop word and standard deviation used by Bin Othman. | 64 |
| Table 6.3: Comparison between the baseline method and word co-occurrence method with a window size of ± 3 | 69 |
| Table 6.4: Confusion matrix between words in the confusion sets using context words method for $k = 3$ | 70 |
| Table 6.5: Statistics of the language models for the training sentences in the supervised method..... | 73 |

Table 6.6: Comparison between the baseline and the N-gram methods 74

Table 6.7: Comparison between the baseline, the separate LMs and the N-gram methods.
..... 75

Table 6.8: Reducing error rate in the combination method using rejection..... 79

LIST OF FIGURES

| | |
|---|----|
| Figure 4.1: Number of words in each corpus | 28 |
| Figure 4.2: Sizes of the dictionaries constructed | 30 |
| Figure 5.1: Error Detection Algorithm | 39 |
| Figure 5.2: Error Correction Algorithm..... | 44 |
| Figure 5.3: Proposed method for real-word error detection and correction..... | 45 |
| Figure 5.4: Error Generating Algorithm | 47 |
| Figure 5.5: Examples of inducing errors | 47 |
| Figure 5.6: Examples of different corrections | 49 |
| Figure 6.1: Example of context words | 57 |
| Figure 6.2: Training and testing phases..... | 60 |

LIST OF ABBREVIATIONS

OCR : Optical Character Recognition

POS : Part-Of-Speech

LMs : Language Models

TP : True Positive

FP : False Positive

FN : False Negative

ABSTRACT

Full Name : Majed Mohammed Abdulqader Al-Jefri
Thesis Title : Real-Word Error Detection and Correction in Arabic Text
Major Field : Computer Science
Date of Degree : May 2013

Spell checking is the process of finding misspelled words and possibly correcting them. Spell checkers are important tools for document preparation, word processing, searching, and document retrieval. The task of detecting and correcting misspelled words in a text is challenging. Most of the modern commercial spell checkers work on word level with the possibility of detecting and correcting non-word errors. However, few of them use techniques to work on real-word errors. This is one of the challenging problems in text processing. Moreover, most of the proposed techniques so far are on Latin script languages. However, Arabic language has not received much interest, especially for real-word errors.

In this thesis we address the problem of real-word errors using context words and n-gram language models. We implemented an unsupervised model for real-word error detection and correction for Arabic text in which N-gram language models are used. Supervised models are also implemented that use confusion sets to detect and correct real-word errors. In the supervised models, a window based technique is used to estimate the probabilities of the context words of the confusion sets. N-gram language models are also used to detect real-word errors by examining the sequences of n words. The same

language models are also used to choose the best correction for the detected errors. The experimental results of the prototypes showed promising correction accuracy. However, it is not possible to compare our results with other published works as there is no benchmarking dataset for real-word errors correction for Arabic text. In addition, conclusions and future directions are also presented.

ملخص الرسالة

الاسم الكامل: ماجد محمد عبدالقادر الجفري

عنوان الرسالة: التدقيق والتصحيح الإملائي للكلمات الصحيحة الخاطئة في سياق النص العربي

التخصص: علوم الحاسب الآلي

تاريخ الدرجة العلمية: إبريل ٢٠١٣

التدقيق الإملائي هو عملية إيجاد وتصحيح الأخطاء الإملائية، وتعد المدققات الإملائية من الأدوات الهامة لإعداد الوثائق ومعالجة النصوص والبحث واسترجاع الوثائق. وتمثل مهمة كشف وتصحيح الأخطاء الإملائية للكلمات في النص تحدياً كبيراً حيث تعمل معظم المدققات الإملائية على مستوى الكلمة محاولةً كشف وتصحيح الأخطاء التي ليست في القاموس. وقد استخدم عدد قليل منها تقنيات العمل على أخطاء الكلمات الحقيقية (وهي الأخطاء الإملائية والكلمات في النص التي تحدث عندما يعتزم المستخدم على كتابة كلمة ولكن عن طريق الخطأ يقوم بكتابة كلمة صحيحة في القاموس غير الكلمة المرادة وغالبا ما تكون غير مناسبة في السياق). وتعد هذه واحدة من المسائل الصعبة في معالجة النص. علاوة على ذلك، فإن معظم التقنيات المقترحة حتى الآن أجريت على اللغات اللاتينية، بينما لم تلقى اللغة العربية الكثير من الاهتمام، وخاصة بالنسبة لأخطاء الكلمات الحقيقية.

تناولنا في هذه الأطروحة أخطاء الكلمات الحقيقية، فقمنا بتصميم وتطوير نموذج لمدقق لغوي غير معلّم unsupervised لكشف وتصحيح الأخطاء الحقيقية في النص العربي باستخدام نماذج اللغة على مستوى الكلمة word n-grams. كما قمنا بتصميم وتطوير نموذج لمدقق لغوي معلّم supervised لكشف وتصحيح الأخطاء الحقيقية في النص العربي باستخدام تقنية نافذة الكلمات لحساب احتمالات كلمات السياق و نماذج اللغة على مستوى الكلمة word n-grams لمجموعات الالتباس confusion sets (وهي مجموعة من الكلمات التي من المحتمل أن تشكل لبساً مع بعضها البعض للمستخدم) و التي تم جمعها خلال هذا العمل. وقد قمنا بتقييم النموذج و حللنا النتائج.

وقد أظهرت النتائج دقة تصحيح عالية، إلا إنه لم يتسن لنا مقارنة نتائجنا مع غيرها من التقنيات المنشورة لعدم وجود بيانات مرجعية موحدة لتصحيح أخطاء الكلمات الحقيقية في النص العربي.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Conventional spell checking systems detect typing errors by simply comparing each token (word) in a text against a dictionary that contains correctly spelled words. The tokens that match elements of the dictionary are considered as correctly spelled words; other tokens are flagged as errors and corrections are suggested. A correctly spelled token that is not the one that the user intended cannot be detected by such systems. These errors are known as real-word errors or semantic errors. Real-word errors result in morphologically valid words whose use in the context is senseless. These errors may even be caused by the spelling checkers themselves, when correcting non-word spelling errors automatically, as in some word-processors, they change a non-word to the wrong real word (Graeme Hirst and Alexander Budanitsky 2005). Moreover, sometimes the user mistakenly selects a wrong word from the suggested list offered by the word processor. Real-word errors are not a minority, it has been reported that 25%, 30% and 40% of total errors were real-word errors in (Young, Eastman, and Oakman 1991), (Wing and Baddeley 1980) and (Mitton 1987), respectively.

The following sentences contain real-word errors that are not detected by conventional spell checkers. The words in the brackets are the correct words intended by the user.

- ١- كما سنجد فى المقابل من يعتقد [يبتعد] ثورة يناير بكل نداعياتها ونتائجها
- ٢- التسهيلات المتوفرة ساعدت على الاقبال الكثير [/الكبير] من الدول والافراد

- ٣- تخضع لمرحلة جادة من التقييم لدراسة جواب [جوانب] النجاح العديدة
- ٤- ما خلفته الخسارة من الم تجسد في مواقف الحزن التي لفت نجوم الحريق [الفريق] لحظة تتويجهم
- ٥- لكن التقرير ابدى قلقا بشأن تاخر ارامكو في إزداد [إمداد] الوقود لخطوط الانتاج الجديدة
- ٦- متذرا بان الشركة هي من يمنع هذا الاعلان لان لها استثمارات في جهزت [جهات] اخرى

The first four sentences contain semantic errors while the remaining two contain syntactical errors. Arabic is very inflected language that contains large number of words compared to other languages. The similarity between Arabic words is also high and this raises the importance of the problem of real-word errors in Arabic. Our aim is to detect and correct such errors in Arabic text by considering the context in which those words occur. Since most of the work on real-word errors has been done on English text, our aim is to enrich the field of Arabic spell checking. In the course of this work, prototypes for real-word spelling detection and correction for Arabic text are implemented. This research is a continuation of previous research by a KFUPM colleague (Mahdi 2012) who developed a spell checking model that detects and corrects non-word errors in Arabic text.

In this chapter, we identify the problem statement of the domain of real-word error detection and correction. We also present the thesis objectives and the methodology followed in order to achieve those objectives. The contents of the thesis and its structure are also presented in this chapter.

1.2 Problem Statement

Spell checkers are important tools for document preparation, word processing, searching and document retrieval. The task of detecting and correcting misspelled words in a

written text is challenging. We are not aware of any research that addresses detecting and correcting real-word errors for Arabic text except for (C. Ben Othmane Zribi, Mejri, and M. Ben Ahmed 2010) and their continuation work in (C. Ben Othmane Zribi and M. Ben Ahmed 2012). We are not aware of any stand-alone spell checking for Arabic text, the current implementations are tools in word processors. Consequently, designing a spell checker for Arabic languages is imperative to save time and effort for Arabic language users.

In this thesis, a prototype for spell checking and correction for Arabic text is implemented. This prototype is able to detect and correct real-word errors automatically. N-grams language models and context words method are used to detect spelling errors. Two techniques of addressing real-word errors are discussed. Unsupervised and supervised learning techniques. In the latter, the labels are known in advance in the form of confusion sets that are commonly confused by users.

1.3 Research Objectives

The main objective of this research is to study the problem of spell checking and to investigate the techniques and algorithms used in the literature that address the problem of spell checking and correction. We discuss the spell checking and correction problem in general and focus on detecting and correcting real-word errors in Arabic text. In addition, we aim at designing and implementing a prototype for spell checking and correction for Arabic text, which is capable of detecting and correcting real-word errors. In order to accomplish this objective, the following tasks are conducted:

1. Conducting literature survey on spell checking in general and context-sensitive spell checking techniques in particular.
2. Data Collection and Preparation.
 - a. Building a suitable Arabic text corpus for this work.
 - b. Analyzing the corpus and building a suitable language models for spell checking and correction (N-grams, dictionaries).
 - c. Collecting Arabic confusion sets to be used in the supervised learning techniques.
 - d. Preparing data in which real-word errors are induced to be used in testing the prototype.
3. Prototype Implementation and Evaluation
 - a. Implementing Arabic spell checking prototypes that detect and correct real-word errors.
 - b. Evaluating the performance of the proposed prototypes.
 - c. Identifying factors that can improve the performance of the implemented prototypes.
4. Analyzing the results of the experimental work and presenting conclusions.

The remaining part of this thesis is organized as follows. Chapter 2 presents background information on spell checking and correction in general and Arabic spell checking and correction in particular. We also discuss the terminology used in the literature. In chapter 3 we extensively study the techniques and algorithms used in the literature that address the problem of spell checking and correction. We present the collected and used data in

this thesis in chapter 4. Chapter 5 presents the unsupervised method that addresses real-word errors detection and correction in Arabic text and its prototype. Chapter 6 presents the supervised methods in addressing real-word errors in Arabic text. Finally, the conclusions and future direction are discussed in Chapter 7.

CHAPTER 2

BACKGROUND

2.1 Introduction

Spell checkers identify misspelled words in a document and try to correct them. Detection is the process of parsing the text and finding misspelled words. Correction is the process of correcting errors found in the detection phase. Errors could be either non-word errors (i.e. words that are not in the dictionary) or real-word errors (i.e. words that are in the dictionary but is not what the user intended).

The problem of spell checking is one of the hottest research areas in natural language processing. Research for detecting and correcting spelling errors has started since 1960s (Damerau 1964) and since then many techniques have been proposed to detect and correct spelling errors. Some of the techniques aim only at detecting errors so that the user is aware of the errors and it is his responsibility to correct those errors. While other techniques aim at detecting as well as correcting errors. To this end, automatic spelling error correction systems are classified as:

- Fully automatic spelling error correction systems.
- Interactive spelling error correction systems.

In fully automatic error correction, the system finds candidate words and chooses the most likely one. The interactive system finds candidate words, ranks them, and suggests the most probable ones. The user then chooses the correct word himself.

The difference between the two is that the latter method needs user interaction to choose the correct word. In a fully automatic method, the most likely correction is automatically chosen.

In order to correct a detected error, candidate corrections must be found first, then these corrections are ranked and the most likely correction is chosen (in the case of a fully automatic system) or the first n most probable corrections are suggested (in the case of an interactive system).

The problem of real-word errors is one of the challenging problems in text processing. Most modern commercial spellcheckers work at the word level when trying to detect and correct errors in a text. Hence, they use simple dictionary lookup techniques. When a word is not in the dictionary, it is considered an error. But what if the misspelled word is a valid word in the dictionary; much more effort is needed to handle such errors. In this work the problem of real-word errors is addressed.

Real-word spelling errors are words in a text that occur when a user intends to type a word but mistakenly he types another correctly spelled word. Such errors occur because of spelling or sound similarity between words. They may even be caused by the spelling checkers themselves, when correcting non-word spelling errors automatically, as in some word-processors, they change a non-word to the wrong real word (Graeme Hirst and Alexander Budanitsky 2005). Moreover, sometimes the user mistakenly selects a wrong word from the suggested list offered by the word processor (Wilcox-O'Hearn, G Hirst, and A Budanitsky 2008). In the survey conducted by (Kukich 1992) real-word errors ranged from 15% to 40% of the total spelling errors.

As most spellcheckers deal with words in isolation, they simply accept this type of errors as correct if they are found in the dictionary. They only flag non-words (i.e. sequence of characters that are not a word in the dictionary) as errors as they match none of the dictionary entries. This process is known as dictionary lookup which is, to some extent, sufficient for non-word spelling errors. On the other hand, to detect real-word errors, the spellchecker is required to consider the surrounding context. To that end new research focuses towards making use of context. Thus, techniques that aim at tackling the problem of real-word errors are also referred to as context-sensitive spell checking techniques.

For that, syntactic and semantic knowledge of the language are employed to detect real-word errors. For instance, in the sentence 'مذهب الولد إلى المدرسة', syntactic knowledge could be involved to detect the syntactic error in the sentence. Another example, the sentence 'أكل الرجل الخبر' is semantically incorrect. These types of errors need to be corrected, hence the spellchecker tries to select a closer word as a replacement for the error word as in non-interactive spellcheckers, or the spellchecker suggests a set of candidate words, as in interactive spellcheckers like MS Word, so that the user may choose the intended word by himself.

Research on real-word spell checking for Arabic text is conducted in this work. In the course of this work, prototypes for real-word spelling detection and correction for Arabic text are implemented. This research is a continuation of previous research by a KFUPM colleague (Mahdi 2012) who developed a spell checker that detects and corrects non-word errors in Arabic text. Accordingly, we are considering that the given text is a non-word error free. The prototype will act as a second phase of a spell checking system that addresses real-word errors.

The main idea behind this method is considering the context surrounding the word in error instead of the single word alone. Word N-grams are also used to check misspelling words that result in an unlikely sequence. For example, the word 4-gram ‘عرض عليه مال’ is more frequent than ‘عرض عليه مال كبير’, the hypothesis is that the latter 4-gram is more probable to have a misspelled word(s) and the former is the correct one, because its probability is higher. This probability information is useful to detect unlikely word sequences. They are also useful to suggest corrections for erroneous words in sentences by taking the most probable sequences.

2.2 Spelling Error Classification

Since this research aims at designing and building a prototype for detecting and correcting real-word errors, a proper definition of real-word errors should be agreed on. Different definitions and classifications of errors are found in the literature. The next section shows different kinds of errors and their definition and classification. Most of the following classifications are taken from (Kukich 1992) and (Verberne 2002).

2.2.1 Typing Errors vs. Spelling Errors

Some studies classify errors as typing errors (also called *typos*) and spelling errors. Typing errors are caused by keyboard slips (e.g. ‘عرف’ → ‘عرفغ’). This might happen when a typist misses one key or presses another key mistakenly. Another type of spelling errors results from the writer’s ignorance of the correct spelling. Three possible causes for this type of spelling errors are:

- Phonetic similarity (e.g. ‘ظلام’ → ‘ضلام’)

- Semantic similarity (e.g. 'كثير' → 'كبير').
- Ignorance of grammatical rules (e.g. 'سبعة أحجار' → 'سبع أحجار')

2.2.2 Single vs. Multiple Errors

(Damerau 1964) defined simple errors as words that differ from the intended word by only a single letter. These errors could be a result of four operations:

- *Insertion*: a misspelled word that is a result of inserting an extra character into the intended word. (e.g. 'شجر' → 'شجار')
- *Deletion*: a misspelled word that is a result of omitting a character from the intended word. (e.g. 'فم' → 'فحم')
- *Substitution*: a misspelled word that is a result of replacing a character in the intended word with another character. (e.g. 'قمر' → 'قبر')
- *Transposition*: a misspelled word that is a result of swapping two adjacent characters in the intended word. (e.g. 'محارب' → 'محراب')

Multi-errors refer to misspelling errors that contain more than one character difference (e.g. 'مخبز' → 'مختبر'). The percentage of single error is high. It was found that 80%, 94%, and 69% are single errors in (Damerau 1964), (Zamora 1981) and (Mitton 1987), respectively.

2.2.3 Word Boundary Errors (Run-Ons and Split Words)

A run-on is the result of omitting a space between two words, (e.g. 'من شرح' → 'منشرح').

A split word occurs when a space is inserted in the middle of a word, (e.g. 'بركان' → 'بر كان'). These kinds of errors cause problems for a spellchecker as they consider spaces as

word delimiters. A run-on will be treated as one word while a split word will be treated as two separate words. Consequently, spellcheckers will not flag them as errors if they result in words in the dictionary. (Kukich 1992) found that 15% of all errors were word boundary and (Mitton 1987) found that 13% of errors were word boundary.

2.2.4 Non-Word Errors vs. Real-Word Errors

Another classification is non-word versus real-word errors. A non-word error is an error in a word that yields an undefined word (e.g. 'مقرب' → 'مقزب'). On the other hand, real-word errors are caused by changing a word that results in an existing word in the language (e.g. 'كثير' → 'كبير').

It was found that 25%,30% and 40% of total errors were real-word errors in (Young, Eastman, and Oakman 1991), (Wing and Baddeley 1980) and (Mitton 1987), respectively.

2.2.5 Real-Word Errors Classification

Real word errors are further subclassified in the literature. (Mitton 1987) classifies real-word errors into these subclasses:

- 1- Wrong-word errors
- 2- Wrong-form-of-word errors
- 3- Word-division errors

Wrong-word errors occur when the misspelled words are grammatically and semantically differ from the intended words (e.g. 'مقرب' → 'مقر'). Wrong-form-word errors are errors

that are grammatically different from the intended words (e.g. 'ذهب الولد الى المدرسة' → 'مذهب' (الولد الى المدرسة)). Word-division errors are the word boundary errors, run-on and split words, (e.g. 'من قعر' → 'منقعر'). (Mitton 1987) found that wrong-word errors represent 44% of total errors, while wrong-form-of-word errors represent 24% of total errors and the remaining 32% were word-division errors of which most errors are incorrect splits.

(Kukich 1992) classifies real-word errors by distinguishing between the cause of the error and the result of the error. The following are classes based on the cause of the error:

1. Simple typos (e.g. 'بصر' → 'صبر').
2. Cognitive or phonetic lapses (e.g. 'مرضات' → 'مرضاة').
3. Syntactic or grammatical mistakes (e.g. 'تسعة نساء' → 'تسع نساء').
4. Semantic anomalies (e.g. 'كبير' → 'كثير').
5. Insertions or deletions of whole words
(e.g. 'صرح مدير المدرسة صرح بحاجتها الى المزيد من الموارد').
6. Improper spacing (e.g. 'في هما' → 'فيهما').

Classes based on error results are the following:

1. Syntactic errors (e.g. 'صوم الشيخ يوما ويفطر يوما').
2. Semantic errors (e.g. 'ذهب أحمد الى الشوق').
3. Structural errors (e.g. 'مكونات الحاسب الرئيسية خمسة المعالج، الذاكرة، و أجهزة الإدخال والخراج').
4. Pragmatic errors (e.g. 'يقع نهر النيل ثاني أطول نهر في العالم في عمان').

(Verberne 2002) reclassified the classes based on the cause of error by eliminating the last three classes as they are the results of the previous three ones. She also criticizes the classification based on error result; she considers them as two classes, syntactic errors and semantic errors.

2.3 Arabic Real-Word Spell Checking

Arabic is a very inflected natural language that contains huge number of words compared to other languages. Words in Arabic are graphically similar to each other. As a result, the chance of getting semantic errors in texts increases, since a type/spelling error could result in a valid word (C. Ben Othmane Zribi and M. Ben Ahmed 2012).

Table 2.1 shows an example of the inflectional property of Arabic for the word 'أمر'. The word is changed into several different real words by the four operations (i.e. insertion, deletion, substitution of one letter or the transposition of two adjacent letters). This phenomenon was highlighted by a study conducted in (Chiraz Ben Othmane Zribi and Mohamed Ben Ahmed 2003). They took each and every word from the dictionary and applied the four editing operations (insertion of a letter, deletion of a letter, substitution of a letter with another letter and interchanging two adjacent letters).

They calculated the number of correct word forms obtained by applying the four operations. They found that Arabic words are more similar to each other compared to words from other languages such as English and French. It was reported that the average of Arabic word similarity is 10 times greater than English and 14 times greater than French. This gives an indication of the difficulty of treating the problem of real-word errors in Arabic language.

Table 2.1: Example of several real-words driven from the word 'أمر'

| Insertion | Deletion | Substitution | Transposition |
|-----------|----------|--------------|---------------|
| فأمر | مر | تمر | مرأ |
| بأمر | أم | جمر | أرم |
| يأمر | | سمر | |
| نأمر | | أثر | |
| أمرت | | أجر | |
| وأمر | | ضممر | |
| تأمر | | أسر | |
| .. | | .. | |

CHAPTER 3

LITERATURE REVIEW

Most of the researchers of spell checking and correction focused on three difficult problems: (1) non-word error detection; (2) isolated-word error correction; and (3) context-dependent word correction. Many techniques were proposed to address these problems, such as pattern-matching, N-gram analysis techniques, dictionary look up techniques, minimum edit distance, similarity key techniques, probabilistic and rule based techniques (Kukich 1992).

The problem of spelling detection and correction is reviewed in detail in the comprehensive survey of (Kukich 1992) and in the master thesis of (Liang 2008). (Pedler 2007) gave an extensive review of real-word spelling detection and correction in her PhD thesis, and (Graeme Hirst and Alexander Budanitsky 2005) reviewed the problem in detail in their survey .

In this chapter we review non-word errors in general and real-word errors in particular. A separate section discusses Arabic spell checking and correction and the problem of real-word errors detection and correction in Arabic text.

3.1 Non-Word Errors

A non-word may be defined as a sequence of letters that is not a defined word in the language (dictionary). Research on non-word error detection started in the early 1970s.

Most of the research conducted for detecting and correcting non-word errors are based on n-gram analysis and dictionary lookup techniques (Kukich 1992).

(Zamora 1981) presented a study that used tri-gram frequency statistics for detecting spelling errors. He analyzed 50,000 word/misspelling pairs collected from seven abstract service databases. The tri-gram analysis technique was able to determine, in 94% of the time, the error location in a misspelled word accurately within one character. However, the used technique did not distinguish effectively between valid words and misspellings.

(Kernighan, Church, and Gale 1990) and (Church and Gale 1991) devised an algorithm that corrects single-error misspellings by finding a set of candidate corrections that differ from the misspelled word by a single insertion, deletion, substitution or transposition. They implemented their algorithm into a program called CORRECT that uses a reverse minimum edit distance technique to generate a set of candidate corrections. Bayesian formula was used to rank the candidate suggestions. CORRECT was evaluated on a set of 332 misspellings from AP news wire text. Each of these misspellings had exactly two candidate corrections. CORRECT and three human judges were asked to correct the misspellings by choosing the best candidate. CORRECT agreed with at least two of the three judges 87% of the time.

(Brill and Moore 2000) proposed an improved model for spelling correction using the noisy channel model and Bayes' rule. The model used dynamic programming algorithm for finding edit distance between a misspelled word and a dictionary word. A 10,000 word corpus of common English spelling errors, paired with their correct spelling was used. Different context window sizes were used to evaluate the proposed model. The

model achieved 93.6%, 97.4% and 98.5% accuracy in the best one, two and three word candidates respectively. The model gave better results when extended by using a tri-gram language model.

(Lehal 2007) designed and implemented a Punjabi spell checker that detects and corrects non-word errors. He first created a lexicon of correctly spelled words in order to check the spellings as well as to generate suggestions. All the possible forms of words of Punjabi lexicon were sorted then partitioned into sixteen sub-dictionaries based on the word length. Secondly, dictionary lookup technique was used to detect misspelled words. After identifying the misspelled words, reverse minimum edit distance between a misspelled word and a dictionary word was used to generate a list of candidate words. Moreover, words which are phonetically similar to the misspelled words were added to the suggestion list. After that, the suggestion list is sorted based on phonetic similarity between the error word and the suggested word, word frequency of the suggested word, and the smallest minimum edit distance between the misspelled word and the suggested word. The spell checker was evaluated on a test set of 255 most commonly misspelled words. The correct words were on the top of the presented suggestion list 81.14% of the time and 93.4% of the time on the top 10 of the suggested words.

3.2 Real-Word Errors

Real-word errors are typing errors that result in a token that is a correctly spelled word, although not the one that the user intended. Work on real-word detection and correction began in the early 1980s (Kukich 1992). The Unix Writer's Workbench package (L. Cherry and N. Macdonalil 1983) represents one of the first efforts that addressed real

word errors detection and correction in text. It flags common grammatical and stylistic errors and suggests possible corrections (Kukich 1992).

The problem of real-word errors has been discussed in two different perspectives in the literature. In the first one, researchers have considered this problem as the resolution of lexical disambiguation. They used pre-established sets of words that are commonly confused with each other called the confusion sets, like { 'كثير', 'كبير' }. A word is simply suspicious when a member of its confusion set better fits in its context. The correction is made by selecting the most likely member in that set considering the context. The second nature of research is not tied to predefined confusion sets as in the first one. They used other methods that use the context to detect and correct real-word errors by applying unsupervised techniques based on semantic, syntactic or probability information.

(A. Golding 1995) is the originator of lexical disambiguation using predefined confusion sets. He used 18 confusion sets of commonly confused words provided by the Random House Unabridged Dictionary (Flexner 1983). He used a Bayesian hybrid method for real-word spelling correction by identifying the presence of particular words surrounding the ambiguous word. He also used the pattern of words and part-of-speech (POS) tags around the target word. He used these as features to train Bayesian classifiers to select the correct target word. Decision lists are first used to choose the correct word from a confusion set. Golding ran the same experiments with Bayesian classifiers and reported a small improvement over decision lists.

(A. Golding and Schabes 1996) proposed a method called Tribayes. When an occurrence of a word belonging to any of the confusion sets in the test set is examined, Tribayes

substitutes each word from the confusion set into the sentence. For each confusion set member, the method calculates the probability of the resulting sentence based on Part-Of-Speech (POS) trigrams. It selects the word that makes the sentence having the highest probability as a correction. A shortcoming in this method is that the confusion sets only contains sets of words that are commonly confused because of meaning or form similarity. Therefore, typing errors and uncommon errors are not considered. In addition to the limitation of correcting only the limited type of errors described by the confusion sets, this method has another disadvantage, the use of POS tri-grams does not help in case of syntactical errors (i.e. when the words have the same POS tag).

(A. R. Golding and Roth 1996) explored a classification-based approach to the problem of lexical disambiguation. They trained the classifiers to discriminate the intended member of a confusion set by considering the context words around a member of that confusion set. The classifiers were also trained to discriminate the POS tags of the confusion set members. The downside with their approach is that they applied mistake-based classification algorithms to this problem. This requires large amounts of memory for the large features used and can be relatively expensive to train.

(Bergsma, Lin, and Goebel 2008) presented a method on Web-Scale N-gram Models for lexical disambiguation. They used supervised and unsupervised systems that combined information from multiple and overlapping segments of context. The method was used on three tasks viz. preposition selection, context-sensitive spelling correction and non-referential pronoun detection. They reported that the supervised system on the first two reduces disambiguation error by 20-24% over the current state-of-the-art.

Other research is not based on predefined confusion sets and they achieved less effective results since the problem is more difficult. (Mays, Damerau, and Mercer 1991) used dynamically created confusion sets for each word in their 20,000 word vocabulary; the sets are varied in size. They used word tri-gram probabilities from the IBM speech recognition project to capture semantic and syntactic errors. They randomly selected 100 correctly spelled sentences from the AP newswire and transcripts of the Canadian Parliament. They generated 8628 sentences in error using the 100 sentences by successively replacing each word with each member of its associated confusion set. Each sentence contains only one error. By applying their proposed system they reported a detection of 76% of the errors and a correction of 73%. The problem with their approach is that the number of word tri-grams is enormous and it corrects only a single error per sentence. Another limitation is that their errors are simple errors.

(Wilcox-O'Hearn, G Hirst, and A Budanitsky 2008) Analyzed the advantages and limitations of (Mays, Damerau, and Mercer 1991) (MDM) described above, and re-evaluated their method to be comparable with other methods. Then they compared it with the WordNet-based method of (Graeme Hirst and Alexander Budanitsky 2005). Then the vocabulary of the tri-gram model was increased to make it more realistic. In addition, it was applied with a smaller window of the sentence and correcting multiple words within a sentence. The used data is more natural than that of MDM, and the work has good analysis of the implementation factors. The results they reported showed that MDM performs better than their optimized approach, as they got poorer results with multiple corrections. The limitation of their method is addressing only simple errors.

(Fossati and Eugenio 2007) proposed a method of mixed tri-grams model that combines the word-tri-grams model and POS-tri-gram model. They defined confusion sets for all words in the vocabulary using minimum edit distance. The good side of their work is using POS-tri-gram model which solves the data sparseness problem. The limitation of their approach is the lack of using a good smoothing technique for assigning probabilities of unseen tri-grams and the skipping of words with less than three characters.

(Aminul Islam and Diana Inkpen 2009) presented a method for detecting and correcting multiple real-word spelling errors. They presented a normalized and modified version of the string matching algorithm, Longest Common Subsequence (LCS), and a normalized frequency value. Their technique is applied using Google web 1T 3-gram dataset. The proposed method first tries to determine some possible candidates and then sorts them based on string similarity and frequency value in a modified version. Then it selects the best one of these candidates. They stated that Google 3-grams proved to be very useful in detecting and correcting real-word errors. They reported that their proposed method achieved a detection recall of 89% and correction recall of 76%. The used data consists of 500 articles from the 1987–89 Wall Street Journal corpus (approximately 300,000 words). However this data is not enough for such type of analysis. In addition, there is no run-ons nor split errors.

(Verberne 2002) proposed a tri-gram-based method for real-word error detection and correction, using the British National Corpus. The used technique assumes that if a word tri-gram is not in the British National Corpus then it has an error, otherwise it is considered correct without using the probability information of the tri-gram. However,

not every seen tri-gram in the training set is correct; there could be some cases in which the tri-gram is not correct in a given context.

(A. Islam and D. Inkpen 2011) proposed an unsupervised text correction approach that can deal with syntactic and semantic errors in English text using Google Web 1T data set. A limitation of their proposed approach is the dependence on the availability of adequate n-grams.

3.3 Arabic spell checking and correction

Research on spell checking of Arabic language increased dramatically in recent years due to the increased demand for Arabic applications that require spell checking and correction facilities. Few Arabic spell checking research has been reported on non-word error detection and correction and fewer on real-word error detection and correction. In this section, we present some work on Arabic spell checking.

(Haddad and Yaseen 2007) presented a hybrid model for non-word Arabic detection and correction. Their work was based on semi-isolated word recognition and correction techniques considering the morphological characteristics of Arabic in the context of morpho-syntactical, morphographemic and phonetic bi-gram binary rules. Their hybrid approach utilized morphological knowledge in the form of consistent root-pattern relationships and some morpho-syntactical knowledge based on affixation and morphographemic rules recognize the words and correcting non-words.

(A. Hassan, H. Hassan, and Noeman 2008) proposed an approach for correcting spelling mistakes automatically. Their approach used finite state techniques to detect misspelled words. They assumed that the dictionary is represented as deterministic finite state

automata. They build a finite state machine (FSM) that contains a path for each word in the input string. Then the difference between generated FSM and dictionary FSM is calculated. This resulted in an FSM with a path for each misspelled word. They created Levenshtein-transducer to generate a set of candidate corrections with edit distances of 1 and 2 from the misspelled word. Confusion matrix was also used to reduce the number of candidate corrections. They selected the best correction by assigning a score to each candidate correction using a language model. Their prototype was tested on a test set composed of 556 misspelled words of edit distances of 1 and 2 in both Arabic and English text and they reported an accuracy of 89%. However, using the finite-state transducers composition to detect and correct misspelled word is time consuming.

(C. Ben Othmane Zribi, Mejri, and M. Ben Ahmed 2010) proposed a method for detecting and correcting semantic hidden errors in Arabic text based on their previous work of Multi-Agent-System (MAS) (Ben Othmane Z C Ben Fraj F 2005). Their technique is based on checking the semantic validity of each word in a text. They combined four statistical and linguistic methods to represent the distance of each word to its surrounding context. These methods are co-occurrence-collocation, context-vector method, vocabulary-vector method and Latent Semantic Analysis method. They compared this representation with the ones obtained from a textual corpus made of 30 economic texts (29,332 words). They assumed that there is only one error in each sentence and based on that they used a voting method to select one from the suspected errors found by each method. Once an error is detected, all candidate suggestions of one minimum edit distance are generated in order to correct the error. A list of all candidates is maintained and substituted with the erroneous word forming a set of candidate sentences. Sentences

with semantic anomalies are eliminated from the list using the detection module of the system. The remaining sentences are then sorted using combined criteria of classification namely, typographical distance, proximity value and position of error. The system was tested on a test set of 1,564 words and 50 hidden errors in 100 sentences and a result of 97.05% accuracy was reported. The limitation of their work is assuming that a sentence can have a maximum of one error. In addition, the corpus used in training phase is small and the number of errors in testing is limited.

(Shaalán, Aref, and Fahmy 2010) proposed an approach for detecting and correcting non-word spelling errors made by non-native Arabic learners. They utilized Buckwalter's Arabic morphological analyzer to detect the spelling errors. To correct the misspelled word, they used the edit distance techniques in conjunction with rule-based transformation approach. They applied edit distance algorithm to generate all possible corrections and transformation rules to convert the misspelled word into a possible word correction. Their rules were based on common spelling mistakes made by Arabic learners. After that, they applied a multiple filtering mechanism to reduce the proposed correction word lists. They evaluated their approach using a test data that is composed of 190 misspelled words. The test set was designed to cover only common errors made by non-native Arabic learners, such as Tanween errors, Phonetic errors and Shadda errors. They evaluated their system based on precision and recall measures for both spelling error detection and correction to measure the performance of the system. They achieved 80+% recall and a 90+% precision as reported.

(Alkanhal et al. 2012) presented a stochastic-based technique for correcting misspelled words in Arabic texts, targeting non word-errors. They also considered the problem of

space insertion and deletion in Arabic text. Their system consists of two components, one for generating candidates and the other for correcting the spelling error. In the first component, the Damerau–Levenshtein edit distance was used to rank possible candidates for misspelled words. This component also addresses merged and split word errors by utilizing the A* lattice search and 15-gram language model at letter level to split merged words. For the split words the component finds all possible merging choices to produce the correct word. In the second component they used the A* lattice search and 3-gram language model at the word level to find the most probable candidate. They reported that their system achieved 97.9% F₁ score for detection and 92.3% F₁ score for correction.

(C. Ben Othmane Zribi and M. Ben Ahmed 2012) proposed an approach for detecting and correcting real-word errors by combining four contextual methods. They used statistics and linguistic information to check whether the word is semantically valid in a sentence. They implemented their approach on a distributed architecture with reported precision and recall rates of 90% and 83%, respectively. They focused only on errors that cause total semantic inconsistencies; this can be considered as a limitation as they ignored partial semantic inconsistencies and semantic incompleteness errors. In addition they assumed that a sentence can have one error at most. Moreover, the used corpus is relatively small (1,134,632 words long) containing only economics articles (i.e. no variations in topics).

CHAPTER 4

DATA COLLECTION AND PREPARATION

A dataset is an essential resource that is used in spell checking research. In this chapter we will describe the used data in this thesis. The used data set passed through two main phases, the first phase is data collection and preparation in which we collected the corpus and made the preprocessing needed; the second phase is building the language models, dictionary generation, and collecting the confusion sets.

4.1 Corpus Collection and Preparation

We are not aware of any benchmarking Arabic dataset for spell checking and correction. Hence, a collection of Arabic text is very important for building well-trained n-gram language models to get best results and good performance in the detection and correction phases for spell checking and correction.

Manual collection of data is time consuming and error prone, hence we developed Crawler. Crawler is a program that is able to collect a huge amount of data from web sites. In our project we choose Al-Riyadh newspaper because it has many topics in different fields. One can get those topics by easily going directly to the archived library of the web site. The topics of the collected dataset are sport, health and economics. Moreover, our Crawler is able to fetch data from other sources if needed.

A large corpus was collected from Al-Riyadh newspaper on three topics, namely health, economic and sport of (4,136,833), (24,440,419) and (12,593,426) words each, taken

from (7,462), (49,108), (50,075) articles respectively. We will assume that this data is error free and address this issue by taking words with above a minimum number of occurrences. The Crawler was used to extract the body texts automatically (i.e. only the articles body texts were extracted without the articles titles). The total sizes for the Health, Economic and Sport corpora are 42 MB, 261 MB and 131 MB, respectively. Table 4.1 shows the statistics of our Al-Riyadh newspaper corpus for the three topics.

We added to our corpus another smaller corpus that was collected in (Mahdi 2012). This corpus consists of Arabic texts collected from different subjects such as news, short stories, and books. In addition, Arabic Gigaword Third Edition, a rich corpora compiled from different sources of Arabic newswire, Corpus of Contemporary Arabic (CCA), a corpus collected by Latifa AlSulaiti in her master thesis (Al-Sulaiti 2004), and Watan-2004 corpus which contains about 20000 different articles from different topics were also used. In addition, the text of the Holy Quraan was added to the corpus in estimating the n-gram models to correct errors in writing Ayahs of Quraan. These corpora were combined and added to form one complete corpus of 10,820,312 words of total size of 124 MB. For more details about this added corpus reference may be made to (Mahdi 2012).

All these corpora are combined into a single comprehensive corpus of size of 508 MB; it is the largest corpus for Arabic text to our knowledge. The corpus is normalized by removing diacritics, numbers, symbols and punctuation marks, English letters were also removed from the corpus. Detailed information for each corpus is shown in Table 4.1. Figure 4.1 shows the number of words in each topic in the corpus. Table 4.2 shows a sample of the number of words' occurrences in the corpus sorted in a descending order.

Table 4.1: Statistics of our collected corpus

| Topic | Number of words | Number of articles | Size on disk | Source |
|-------------------|-----------------|--------------------|--------------|--------------|
| Health | 4,136,833 | 7,462 | 42 MB | Al-Riyadh |
| Sport | 12,593,426 | 50,075 | 131 MB | Al-Riyadh |
| Economic | 24,440,419 | 49,108 | 261 MB | Al-Riyadh |
| News | 9,025,403 | NA | 69.7 MB | (Mahdi 2012) |
| Stories | 106,185 | NA | 5.1 MB | (Mahdi 2012) |
| Medicine | 612,824 | NA | 5.4 MB | (Mahdi 2012) |
| History | 236,370 | NA | 2.76 MB | (Mahdi 2012) |
| Variety of topics | 750,131 | NA | 4.8 MB | (Mahdi 2012) |
| General | 51,990,990 | NA | 508 MB | All previous |

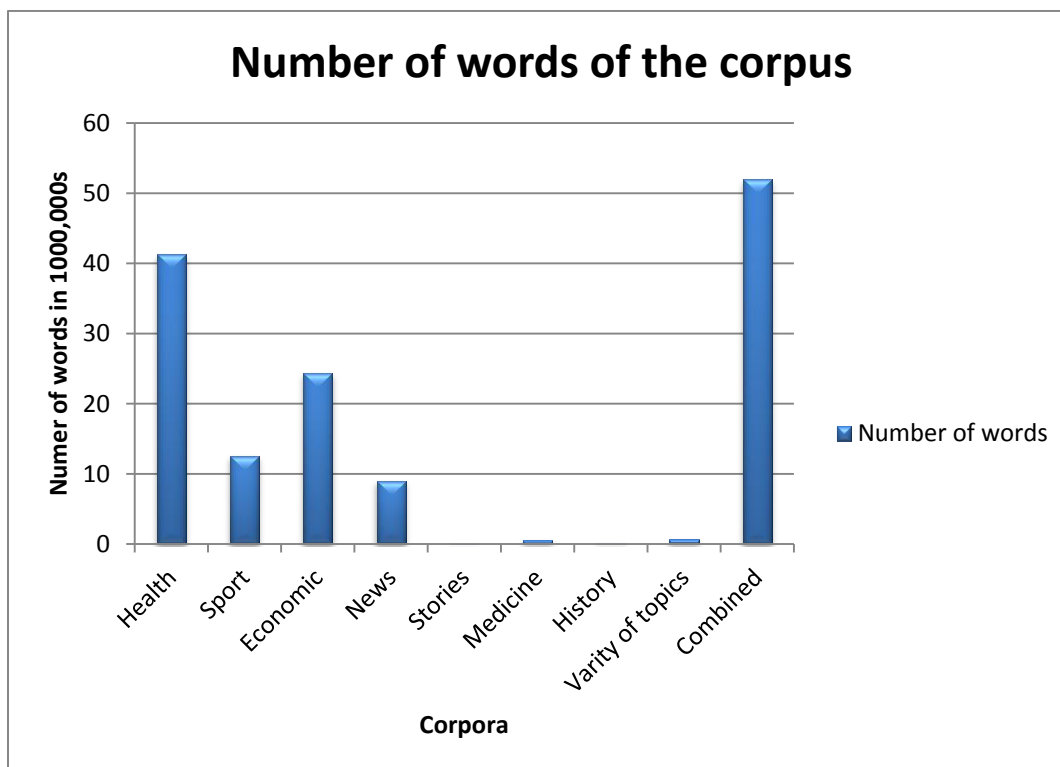


Figure 4.1: Number of words in each corpus

Table 4.2: Sample of number of words' occurrences in the corpus

| Word | Count | Word | Count | Word | Count |
|------|--------|----------|-------|----------|-------|
| في | 352127 | الانسان | 3726 | الشامل | 689 |
| من | 282292 | الوطنية | 3708 | الحقيقي | 688 |
| على | 155428 | لقد | 3662 | اندية | 687 |
| ان | 153057 | المختلفة | 3654 | الامام | 687 |
| الى | 119263 | الخارجية | 3631 | الاعداد | 687 |
| التي | 69378 | اهمية | 3614 | الخابورة | 687 |
| عن | 55429 | المالية | 3603 | التنسيق | 686 |
| الذي | 45094 | عدة | 3598 | اليد | 685 |
| مع | 43445 | الشركات | 3598 | يعتمد | 685 |
| .. | .. | .. | .. | .. | .. |
| .. | .. | .. | .. | .. | .. |

4.2 Dictionary Generation

In this phase we generate dictionaries of words from our collected corpus. We extracted all the words from the corpus and counted their occurrences. Then the words were sorted in a descending order based on their number of occurrences. Different dictionaries with different sizes were generated from these distinct words with respect to the number of occurrences as shown in Table 4.3. For instance, the total number of words in dictionary 5 is 88,645, each word is repeated at least 20 times. Naturally the dictionary size decreases as the minimum number of occurrences increases. For example, dictionary 1 is larger than dictionary 2 and so on. However, the correctness of words in the smaller dictionaries is higher than that in the larger ones. For instance, the words 'سسيجارة' in dictionary 1 and the word 'ردولار' in dictionary2 are mistyped although we assumed that the corpus is error free. Using dictionaries with higher word occurrences results in

reduced spelling errors. Figure 4.2 shows a graph representation of the dictionaries and their sizes.

Table 4.3: Statistics of the dictionaries

| Dictionary | Minimum # of occurrences | Dictionary size |
|-------------------|---------------------------------|------------------------|
| Dictionary 1 | 1 | 576,793 |
| Dictionary 2 | 2 | 324,456 |
| Dictionary 3 | 5 | 187,312 |
| Dictionary 4 | 10 | 128,684 |
| Dictionary 5 | 20 | 88,645 |
| Dictionary 6 | 50 | 52,754 |
| Dictionary 7 | 100 | 34,461 |

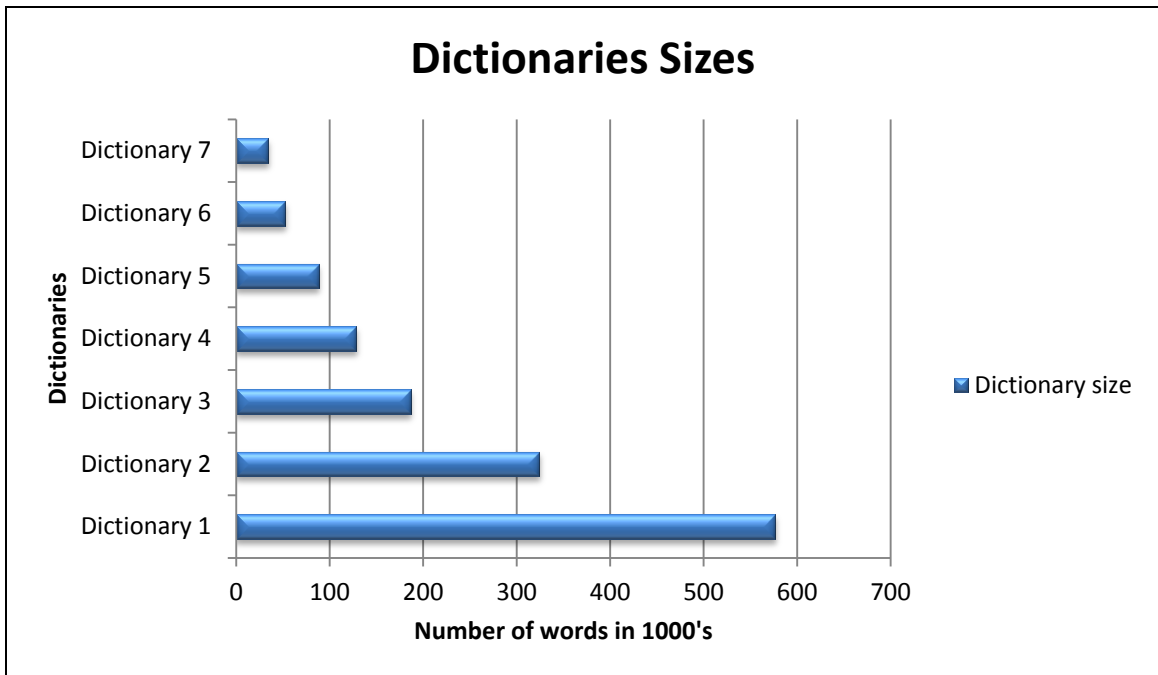


Figure 4.2: Sizes of the dictionaries constructed

4.3 Statistical Language Models

Statistical language models are used in many natural language processing applications such as speech recognition, machine translation, part-of-speech tagging, parsing and information retrieval. Such models try to capture the properties of the language, and may be used to predict the next word in a sentence.

Many language model (LM) toolkits are used to build statistical language models. Among these is the SRILM Toolkit¹, which is a toolkit for building and applying statistical language models (LMs). SRILM was primarily used for speech recognition, statistical tagging and segmentation, and machine translation (Stolcke 2002). It has been under development in the SRI Speech Technology and Research Laboratory² since 1995.

Using the SRILM Toolkit, we generated the n-gram language models of our corpus. The language models consist of uni-grams, bi-grams and tri-grams. SRILM was also helpful in generating the dictionaries for the corpus as it counts the number of words' occurrences.

4.4 Confusion Sets

A collection of confusion sets is normally used in addressing real-word errors in case of supervised learning techniques. There are several ways to obtain the confusion sets. One way is to find words from dictionary that have one letter variation from others (Mays, Damerau, and Mercer 1991) . { 'علم', 'قلم' } is an example of this type of confusion sets.

¹ <http://www.speech.sri.com/projects/srilm/>

² <http://www.speech.sri.com/>

Another way to find such sets is to gather words that have the same pronunciation (A. Golding 1995) (A. R. Golding and Roth 1996). For example, {’ناظرة‘, ’ناضرة‘}.

In our work we collected sets from both types, in two ways. In the first type we used words recognized wrongly by an Arabic OCR system (i.e. words that are recognized as correctly spelled words by the OCR system while they were not the same words in the original text). Table 4.4 shows a sample of the confusion sets wrongly recognized by the OCR system. We excluded the corpus taken from (Mahdi 2012) as it contains the text of the Holy Quran and we don’t want to manipulate the words of Quran. Our Al-Riyadh newspaper corpus was used to extract the confusion sets. Note that not all confusion sets in the Table are used in our experiments, some are ignored because there are no sufficient occurrences for one of the confusion set words as in {’الرملة‘, ’الرحلة‘}, where the word ’الرملة‘ has not occurred in the collected corpus. The same case for {’تدخلت‘, ’تدخنت‘} in which the word ’تدخنت‘ occurred only once in the corpus.

In the second type, we obtained the confusion sets by gathering words that have the same sound; we collected a set of the most common confused words made by non-native Arabic speakers. Table 4.5 shows some common spelling mistakes committed by non-native Arabic speakers. At the beginning of this list, we include misspelling generalities that should be stated. These words were collected and classified into groups according to their sounds (place of articulation), resulting from the nearness of some letters sounds which cause confusion with other words of different meanings. Another group is changing one character that results in changing the word meaning with no similarity between letters in sound, this is known in the science of rhetoric as anagram, for instance ’ينفذ‘, ’ينفذ‘ and ’يحتقي‘, ’يحتقي‘.

Table 4.4: The eighteen collected confusion sets misrecognized by the OCR

| Confusion Set | No. of occurrences of each word | No. of occurrences in the corpus |
|-----------------------|---------------------------------|----------------------------------|
| مصر – مضر | 667 - 6544 | 7211 |
| القرعة – القرحة | 188 - 797 | 985 |
| الرسم – الرحم | 2749 - 1091 | 3840 |
| مال – حال | 11641 - 3267 | 14908 |
| يفرق – يغرق | 52 - 177 | 229 |
| عسل – غسل | 803 - 327 | 1130 |
| الشرق – الحرق – الأرق | 470 - 109 - 9959 | 10538 |
| الرملة – الرحلة | 854 - 0 | 854 |
| العتيقة – العريقة | 480 - 23 | 503 |
| تدخلت – تدخلت | 129 - 1 | 130 |
| الإسهال – الإهمال | 211 - 519 | 730 |
| مسحوق – مسبوق | 647 - 511 | 1158 |
| يعصر – يصر | 206 - 20 | 226 |
| عروق – حروق | 60 - 66 | 126 |
| موضع – مرضع | 5 - 768 | 773 |
| حيوان – حيران | 108 - 5 | 113 |
| العارضة – الارضة | 4 - 1116 | 1120 |
| يبلغ – يبلغ | 5605 - 12 | 5617 |

A sample of nineteen confusion sets from non-native Arabic speakers is chosen in our experiments. Table 4.6 shows the confusion set of this type. More words may be added to this list in the future.

Moreover we obtained additional sets generated by non-word Arabic spell checker that corrects non-words to real-word errors not intended by the user.

Table 4.5: Common spelling mistakes committed by non-native Arabic speakers

| Original letter | Replaced Letter | Examples | |
|-----------------|-----------------|---|--|
| ع | أ | عمارة عَلَّمَ العَرْض عَنْ (ظهر) | إمارة أَلَّمَ الأَرْض أَنَّ أو أَنْ |
| ح | هـ | محنة حَمَام حاوية حمزة | مهنة هَمَام هاوية همزة |
| ث | س | أثاث إثم ثلاثة | أساس إسم سلسلة |
| ش | س | الشَّعْر شرب الشُّكْر | السَّعْر سَراب السُّكْر |
| ظ, ض | ز, د | مضمار رضي الظهور | مزمار ردي الزهور |
| ص | س | مصير إصرار يصب | مسير إسرار يسب |
| ط | ت | الطلاق مسطور | التلاق مستور |
| خ | ح | يختفي | يحتفي |
| ذ | د | ينفذ | ينفد |

Table 4.6: Nineteen confusion sets from non-native Arabic speakers

| Confusion Set | No. of occurrences of each word | Total No. of occurrences in the corpus |
|-----------------|---------------------------------|--|
| كبير – كثير | 7203 - 25366 | 32569 |
| صغير – قصير | 803 - 989 | 1792 |
| أساس – أثاث | 139 - 3713 | 3852 |
| عمارة – إمارة | 1526 - 310 | 1836 |
| ثمن – سمن | 27 - 1225 | 1252 |
| أشعار – أسعار | 27930 - 465 | 28395 |
| تفسير – تصوير | 48 - 1648 | 1696 |
| إصرار – إصرار | 570 - 278 | 848 |
| الزهور – الظهور | 825 - 359 | 1184 |
| هاوية – حاوية | 349 - 25 | 374 |
| سراب – شراب | 204 - 36 | 240 |
| مسير – مصير | 660 - 23 | 683 |
| محنة – مهنة | 1147 - 29 | 1176 |
| يسب – يصب | 540 - 5 | 545 |
| ألم – علم | 2291 - 1392 | 3683 |
| السعر – الشعر | 2819 - 4250 | 7069 |
| الأرض – العرض | 7140 - 4950 | 12090 |
| السكر – الشكر | 2581 - 3387 | 5968 |
| ثلاثة – سلاسة | 89 - 2204 | 2293 |

CHAPTER 5

REAL WORD ERROR DETECTION AND CORRECTION

USING N-GRAM LANGUAGE MODELS

Language modeling is used in many natural language processing applications such as speech recognition, machine translation, part-of-speech tagging, parsing and information retrieval to name a few. To the best of our knowledge, there is no completely unsupervised approach or method that corrects text containing real-word errors in Arabic text, either using syntactic or semantic information of the language.

In this chapter we address the problem of real-word errors in Arabic text using an unsupervised technique in which n-gram language models are used to detect and correct real-word errors.

5.1 Introduction

N-gram statistical language models try to capture the properties of a language and to predict the next word in a sentence. They assign probability to a sequence of n words $P(w_1, w_2, \dots, w_n)$ by means of probability distribution.

In a tri-gram model, the probability of a sequence of n words $\{w_1, w_2, \dots, w_n\}$ is given by:

$$P(S) = P(w_1) P(w_2|w_1) P(w_3|w_1w_2) \dots P(w_n|w_{n-2}w_{n-1}) = \prod_{i=1}^n P(w_i|w_{i-2}w_{i-1}) \quad (5.1)$$

Where $P(s)$ is the probability of the sentence, $P(w_1)$ is the probability of w_1 , $P(w_2/w_1)$ is the probability of w_2 given w_1 , and so on.

We are going to use the tri-gram model to calculate the likelihood that a word sequence would appear. Using the tri-gram model, we will be able to detect as well as to correct words in error in a given context. For instance, if the word ‘الخبز’ is replaced with the word ‘الخبير’ we will be able to detect this error as the tri-gram ‘أكل الولد الخبز’ is more likely to appear than ‘أكل الولد الخبير’.

Building well trained language models requires a huge corpus to capture the language properties and to estimate the probabilities of the n-grams. The general corpus that combines all the collected texts is used in building the language models (LMs). Details on the corpus are discussed in detail in chapter 4. LMs statistics of the corpus are shown in Table 5.1.

Table 5.1: LMs statistics of the corpus

| Number of words | Uni-grams | Bi-grams | Tri-grams |
|-----------------|-----------|------------|------------|
| 51,990,990 | 576,796 | 13,196,695 | 30,587,630 |

We proposed and implemented a new method for real-word error detection and correction for Arabic text in which N-gram language models (from uni-grams to tri-grams) are used. Our method consists of two main modules, the first module detects real-word errors in a context; the second module corrects the errors detected by the error detection module.

5.2 Error Detection Module

We are proposing an algorithm for detecting real-word errors using the n-grams language models. The algorithm finds suspicious words in a given text by checking the availability of the tri-grams in a sentence.

To detect suspicious words in a sentence, the algorithm checks for the presence of the tri-grams $\{w_{i-2}, w_{i-1}, w_i\}$ to validate w_i in the target sentence by looking it up in the tri-gram language model. We assume that the first word in the sentence is correct and there should be at least three words in each sentence to be treated. If the tri-gram is not found, the algorithm further checks for the availability of the two bi-grams $\{w_{i-1}, w_i\}$ and $\{w_i, w_{i+1}\}$ in the bi-gram language model, if both of them are not found, provided that w_{i-1} and w_{i+1} are not frequent words (i.e. frequent words have high possibility to come with many words) in this case, then w_i is considered suspicious.

Once all suspicious words are flagged, the second step is to verify whether they are true errors or not. For each suspicious word s , we find all its spelling variations $\{v_1, v_2, \dots, v_n\}$. We define the spelling variations of a word w to be the words in the dictionary that are derived from w by insertion, deletion, or replacement of one character, or the transposition of two adjacent characters. Dictionary 7 is used to find the spelling variations for a suspicious word³. Each time the suspicious word s is replaced by one of its spelling variations v_i and its probability is calculated. Five words are actually considered in the sentence, two words to the left of the word (w_{i-2}, w_{i-1}) , the word itself and two words to the right (w_{i+1}, w_{i+2}) . For example, if v_i is one of the suspect word variations, the three tri-grams that make the difference in calculating the probability are considered which are $\{w_{i-2}, w_{i-1}, v_i\}$, $\{w_{i-1}, v_i, w_{i+1}\}$, and $\{v_i, w_{i+1}, w_{i+2}\}$. We add the log probabilities of these three tri-grams to calculate the probability of the five words sequence. The same is done for all w_i variations. In case that the tri-gram is not found, bi-grams back off is used. For instance if the tri-gram $\{w_{i-1}, v_i, w_{i+1}\}$ is not found we back

³ Details of Dictionary 7 can be found in chapter 4.

off to the two bi-grams $\{w_{i-1}, v_i\}$ and $\{v_i, w_{i+1}\}$, and if a bi-gram is not found we further back off to the uni-grams of each word in that bi-gram. For example, if the latter bi-gram is not found, the two uni-grams w_i and w_{i+1} are considered in the calculation. The highest probability obtained by the most probable spelling variation in the context is compared with the probability of the original word (i.e. the suspicious word). If the probability of the former is higher than the later, we take this as an indication that the variation is more likely to be the intended word and the suspicious word is raised as a detected real-word error.

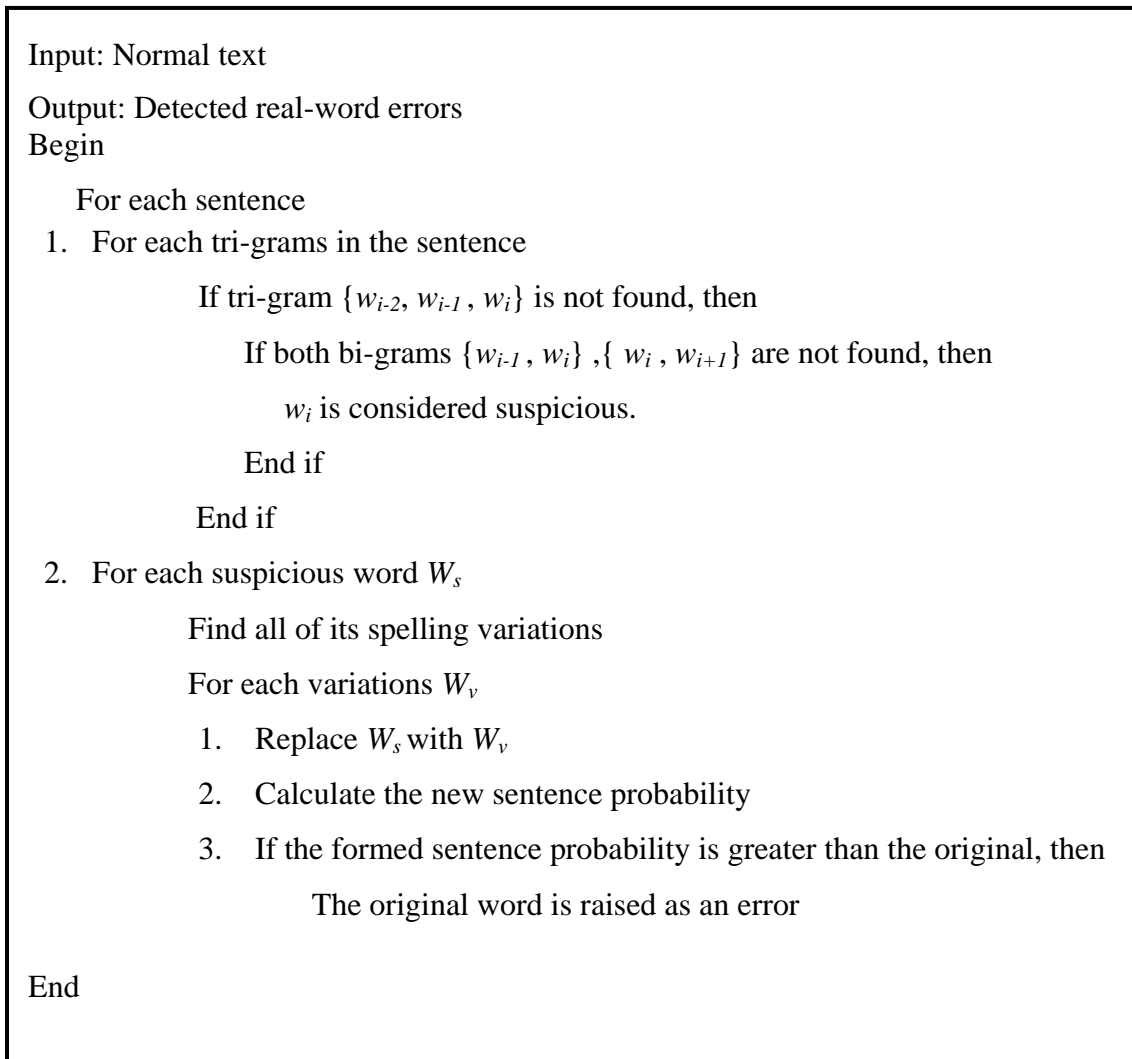


Figure 5.1: Error Detection Algorithm

This module is used to correct the errors detected by the error detection module. In order to correct a misspelled word, the correction module: (1) generates a list of candidate words; (2) generates candidate sentences using the candidate words; (3) ranks the candidate sentences; and (4) possibly replaces the sentence with a candidate with the highest probability.

5.2.1 Generate candidate words

Once a word has been detected as an error, candidate correction words are generated in order to correct it. Quite a few algorithms have been used for finding candidate corrections in the literature. The minimum edit distance is by far the most popular one. The minimum edit distance is the minimum number of editing operations (i.e. insertions, deletions, substitutions and transposition) required to transform one string into another. (Damerau 1964) implemented the first minimum edit distance based spelling correction algorithm based on the first three types of character transformation, (Levenshtein 1966) developed a similar algorithm for correcting deletions, insertions and transpositions. (Wagner and Fischer 1974) generalized the technique to cover also multi-error misspellings.

To correct a detected error, we look for the spelling variation of that error which would fit better into the context than the original word. All word variations for each detected error (i.e. words that have a minimum edit distance of one from the erroneous word) are fetched from the dictionary. These variations are considered as candidate corrections for the erroneous word. For example, in the sentence:

’وقالت الشركات الثلاث في بيانات صحية أمس أن التغطية التأمينية‘

The word 'صحبة' is erroneous and its variations in the dictionary are shown in Table 5.2. (From now on we refer to the detected errors by the detection module as erroneous word).

Table 5.2: The word 'صحبة' variations

| | | |
|-------|-------|-------|
| وصحية | صحيفة | صحبة |
| صحيح | ضحية | صحياً |
| تحية | حية | صحي |
| صحوة | صحة | صحفية |

5.2.2 Generate candidate sentences using the candidate words

After generating candidate words, new sentences are formed by replacing each erroneous word by all of its variations. The probabilities of the new sentences in addition to the original sentence are calculated. The sentence that gives the highest probability is considered as the correct one. We take this as an indication that the word variation in that sentence is a better fit to the context and hence more likely to be the intended word; this is the case of fully automated system. Five words are actually considered in the sentence, as in the detection phase, for example, if w_i is the erroneous word, the three tri-grams that make the difference in calculating the probability are considered (viz. $\{w_{i-2}, w_{i-1}, w_i\}$, $\{w_{i-1}, w_i, w_{i+1}\}$, and $\{w_i, w_{i+1}, w_{i+2}\}$). We add the log probabilities of these three tri-grams to calculate the probability of the five word sequence. The log of probability is used to avoid underflow and to save computation time by using addition instead of multiplication. The same is done for all w_i variations. If the tri-gram is not found, bi-grams back off is used. For instance if the tri-gram $\{w_{i-1}, w_i, w_{i+1}\}$ is not found we back off to the two bi-grams $\{w_{i-1}, w_i\}$ and $\{w_i, w_{i+1}\}$, and if a bi-gram is not found we further

back off to the uni-grams of the words of that bi-gram. For example, if the latter bi-gram is not found, the two uni-grams w_i and w_{i+1} are considered in the calculation.

For the previous example 'وقالت الشركات الثلاث في بيانات صحبة أمس أن التغطية التأمينية' with a suspicious word 'صحبة', the three tri-grams which make the difference in the probability calculation are:

في بيانات صحبة
بيانات صحبة أمس
صحبة أمس أن

The probability for the original sentence is calculated. Then the erroneous word 'صحبة' is replaced each time with one of its variations, in this case twelve variations for the word 'صحبة' resulting in twelve different sentences. For instance, the suspect word 'صحبة' is replaced with word 'ضحبة' forming a new sentence 'وقالت الشركات الثلاث في بيانات ضحبة أمس أن التغطية التأمينية'.

Because the remaining words are the same, their probabilities will be the same. Hence, the calculation is done only for the following three tri-grams:

في بيانات ضحبة
بيانات ضحبة أمس
ضحبة أمس أن

The same is done for all variations in Table 5.2. The variation that gave the highest probability is 'صحفية' which was the correct replacement for the erroneous word 'صحبة'; therefore the correct sentence is:

'وقالت الشركات الثلاث في بيانات صحفية أمس أن التغطية التأمينية'.

5.2.3 Ranking candidate corrections

In the case of interactive systems, the list of candidate words is ranked such that the most probable one comes first (considering the context). The user is provided with the top n suggestions for choosing the most suitable one.

If the top two choices have equal probabilities; ranking could be based on a similarity measure, like minimum edit distance between the suggestions and the erroneous word. In other words, the candidate correction that has the smallest minimum edit distance with the erroneous word will have the highest rank and will be put at the top of the suggestion list. Ranking could also be based on suggested word n -gram frequencies. For example, the frequency of 'من' is higher than the frequency of 'منن', so 'من' is ranked higher than 'منن'.

Minimum edit distance and word n -gram frequency could be combined together. In case of equal minimum edit distance, the most frequent will be considered highest or they could be interpolated to rank the candidates.

5.2.4 Correct Error Words

In the case of fully automatic system, the detected error words are replaced with the words given in the sentence with the highest probability. However, in the case of interactive system the top n candidate words for each suspicious word are suggested to the user to choose the best correction from.

In a fully automatic system, which we follow in this thesis, the variation that gives the highest probability in the context is compared with the original suspect word. If the

variation probability in the context is higher than the original probability (with respect to a threshold value) as in Equation 5.2, then the variation is more likely to be the intended word and the original word is replaced with that variation. We tried different threshold values of 0, 0.1, 0.2 and 0.3. Figure 5.3 shows our proposed method.

$$\frac{\mathbf{Probability}_{\mathbf{variation}} - \mathbf{Probability}_{\mathbf{original}}}{\mathbf{Probability}_{\mathbf{variation}}} > \mathbf{Threshold} \quad (5.2)$$

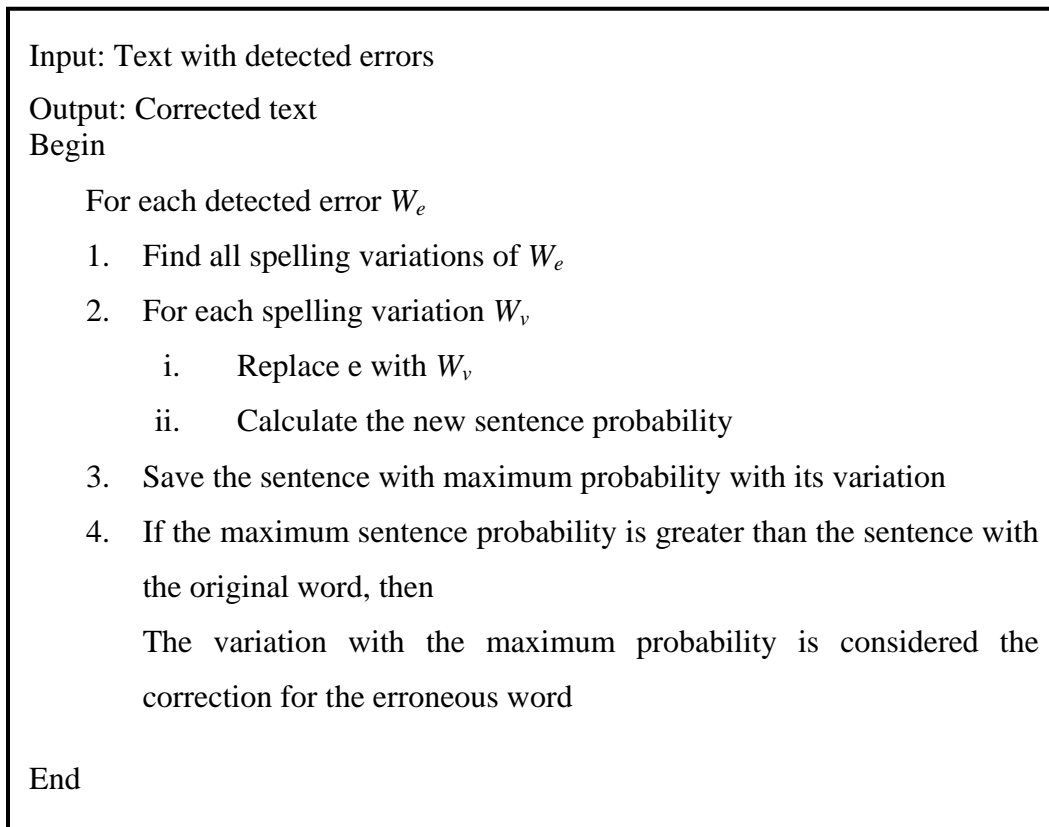


Figure 5.2: Error Correction Algorithm

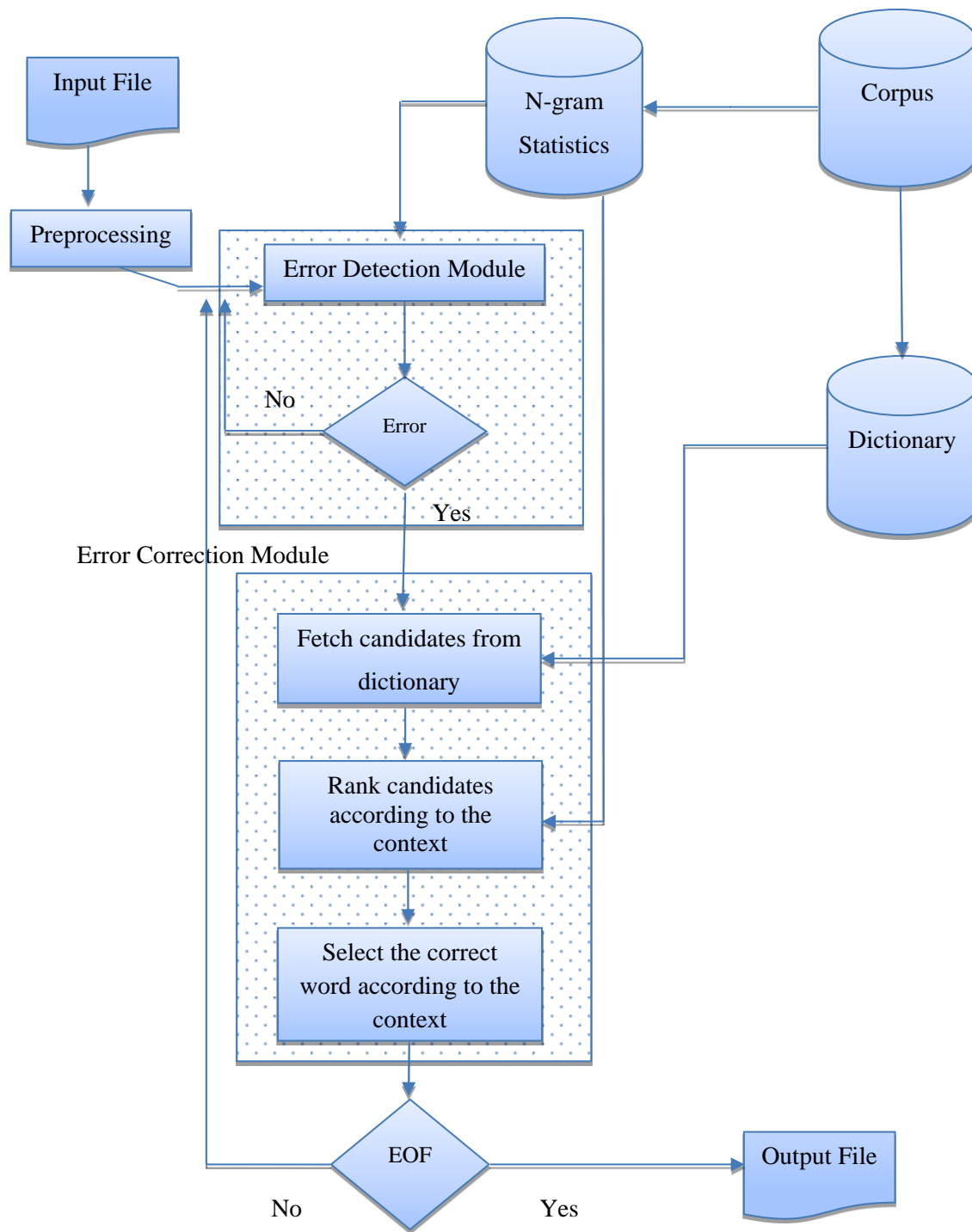


Figure 5.3: Proposed method for real-word error detection and correction

5.3 Experimental Results

In this section our experimental results using the unsupervised N-gram technique are presented.

5.3.1 Testing Data

To create realistic test sets that have real-word errors, we chose 75 articles from the Al-Arabiya website, 25 for each subject (health, sport and economic), then we automatically induced real-word errors in these articles by randomly replacing one word by one of its spelling variations (i.e. one of its confusion set members) in approximately 1 and 2 percent of total words in the text file. Figure 5.5 shows some examples of the induced errors in the test sets. Dictionary 7 is used to find words' variations in the error inducing process⁴. The process of inducing errors is done six times, three times with 1% error rate and three with 2% error rate, resulting in six different test sets, three for each error rate. We defined a spelling variation to be a single character insertion, deletion, replacement, or the transposition of two adjacent characters that results in another real word. Table 5.3 shows the statistics of the test sets.

Table 5.3: Statistics of the test sets

| Total number of words | Average words in each article | Number of errors | |
|-----------------------|-------------------------------|------------------|-----------|
| | | Sets (1%) | Sets (2%) |
| 27,115 | 362 | 233 | 509 |

⁴ Details on Dictionary 7 are presented in chapter 4.

Input: Normal text

Output: Text with errors

Begin

1. Select a word randomly from the input text;
2. Find all variations of the selected word from the dictionary;
3. Select one variation randomly from the fetched variations;
4. Replace the original word with the selected variation.
5. Repeat 1 to 4 until the number of errors to be induced is reached.

End

Figure 5.4: Error Generating Algorithm

- يمكن مقارنته مع الارقام المعلنة للتأكد من مدى دقتها ومدى تحسن مستوى شفافية المعلومات ← وقتها
- المخصص لها مليار ريال لبناء وحدة سكنية ← ركنية
- كان قادرا على العودة للمباراة بعد التعادل ← التعامل
- كنتسجيل مستوى غير طبيعي للسكر في الدم قد يؤدي الى امراض في القلب وجلطات ← القطب
- وان الارقام الصادرة عن الاتحاد العالمي للسكري تشير الى ان المرض تسبب بحوالي ← العرض

Figure 5.5: Examples of inducing errors

5.3.2 Performance Measures

The prototype is evaluated on the test sets, and we measured the performance by means of detection and correction recall and precision.

Detection recall is defined as the fraction of induced errors correctly detected (i.e. the number of correctly detected errors divided by the number of errors that should have been detected). Detection precision is defined as the fraction of detections that are correct (i.e. the number of correctly detected errors divided by the number of all detected errors).

Correction recall is defined as the fraction of errors correctly amended (i.e. the number of correctly amended errors divided by the number of errors that should have been corrected). Correction precision is defined as the fraction of amendments that are correct (i.e. the number of correctly amended errors divided by the number of all corrected errors). F1-measure is also used in measuring performance and can be interpreted as a weighted average of both precision and recall. Equations 5.3 and 5.4 show the recall measure while Equations 5.5 and 5.6 show the precision measure. F₁-measure is shown in Equation 5.7.

$$\text{Recall} = \frac{\text{Number of actual detected misspelled words}}{\text{Number of all misspelled words in the data}} \quad (5.3)$$

$$\text{Recall} = \frac{\text{true positives}}{(\text{true positives} + \text{false negatives})} \times 100 \quad (5.4)$$

$$\text{Precision} = \frac{\text{Number of actual detected misspelled words}}{\text{Total number of detected words}} \quad (5.5)$$

$$\text{Precision} = \frac{\text{true positives}}{(\text{true positives} + \text{false positives})} \times 100 \quad (5.6)$$

$$\text{F}_1 - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \quad (5.7)$$

Figure 5.6 shows some examples of successful corrections, false positive, false negative and true positive detection, false positive correction.

| | |
|--|---|
| SUCCESSFUL CORRECTION | <ul style="list-style-type: none"> • مع ابقاء سقف الانتاج المعقول به والبالغ ← المعمول [المعمول] • خصوصا لدول اوبك التي تستطيع ان تتعايش مع هذا <u>المعدة</u> لتغطي اجمالي متطلباتها المالية ← المعدل [المعدل] • تدل دلالة واضحة على مدى الحرص والاهتمام الذي توليه الحكومة الرشيدة في <u>توفر</u> العيش الرغيد للمواطنين ← توفير [توفير] |
| FALSE POSITIVE | <ul style="list-style-type: none"> • وتتعامل مع المتطلبات <u>النفطية</u> على صعيد العرض والطلب بالارقام والاحصاءات ← النفسية [النفطية] • لكن لم يكن هناك اي اثر لخلايا سرطانات الكلى <u>الخالصة</u> ← الخاصة [الخالصة] • النتائج التي توصلنا اليها <u>تدعم</u> التوصيات الغذائية للوقاية من السرطان ← وتدعم [تدعم] • هذا التوجه قائم لدى المسؤولين في <u>القناة</u> الذين يرغبون طبعا في تطوير برامجهم ← القضاة [القناة] |
| FALSE NEGATIVE | <ul style="list-style-type: none"> • وخلص باحثون نشروا دراستهم في الدورية الامريكية للتغذية السريرية الى ان البالغين في منتصف <u>العشر</u> الاكثر تناولا للحوم الحمراء يكونون اكثر عرضة للاصابة [العمر] • باستخدام افران المايكروويف <u>لطي</u> اللحوم جزئيا قبل تعريضها لدرجات حرارة عالية [لطي] • والمنطق <u>الفكري</u> السليم يرفض ذلك بل ومنطقة المفهوم الشامل للعلاج [الفطري] |
| TRUE POSITIVE DETECTION, FALSE POSITIVE CORRECTION | <ul style="list-style-type: none"> • وربطت الدراسة وجود محتوى اكبر من المواد الكيميائية في اللحوم <u>المعوية</u> بزيادة خطر الاصابة بالمرض ← المعنية [المشوية] • تتسم ميزانية الدولة دائما بمميزات عدة يتمثل اهمها في الشمولية وتركيز الصرف <u>الكبير</u> على جوانب عدة حساسة تمس حياة المواطنين اليومية ومتطلباتهم ← اكبر [الاكبر] • وكانت شركات غربية منها توتال الفرنسية وشل الهولندية وسكور الكندية <u>المختصر</u> بانتاج الغاز ← المختص [المختصة] |

Figure 5.6: Examples of different corrections

Table 5.4 shows an example of true positives (TP), false negatives (FN) and false positives (FP) in a test set with 233 total errors. Table 5.5 shows the average results of all test sets for different threshold values. Detection and correction recall, precision and F_1

measure are presented in the Table. 60.5% to 76.2% of all real-word errors were detected using our prototype, 56 % to 70.7% of the detected real-word errors have been correctly amended. Unfortunately, precision is very low, only 7.8% to 12.3 of total detected errors are rightly detected. This point is addressed below.

The table shows low precision rates that are caused by the large number of false positive detections. Table 5.6 shows some examples of false positives with their causes.

Table 5.4: Example of TP, FN and FP in a test set with 233 errors

| | Detection | Correction |
|-----------|-----------|------------|
| TP | 168 | 134 |
| FN | 65 | 99 |
| FP | 2436 | 2470 |

Table 5.5: Results on the test sets

| Threshold | Detection | | | Correction | | |
|-----------|-----------|-------|-------|------------|-------|-------|
| | R | P | F | R | P | F |
| 0 | 76.2% | 7.8% | 13.9% | 70.7% | 6.0% | 10.9% |
| 0.1 | 66.5% | 10.8% | 18.1% | 61.4% | 8.9% | 15.0% |
| 0.2 | 63.2% | 11.7% | 19.4% | 58.4% | 9.7% | 16.4% |
| 0.3 | 60.5% | 12.3% | 20.2% | 56.0% | 10.4% | 17.3% |

We noticed that some of the false positives are non-word errors that were detected and corrected by our prototype. Although we assumed that the test sets are non-word error free, but there were some of these errors. For example, 'الخضراوات' was replaced with 'الخضروات' and 'وأبدا الرشيد أمله' was corrected by 'وأبدي الرشيد أمله'. These were correctly

detected and corrected by our prototype, so that is a proof that our method can work with non-word errors as well.

Other false positives are proper nouns that are replaced based on their high probability; replacing 'محمود' with 'محمد' is an example of this proper noun false positive, another example is replacing 'أوميغا' with 'أوميغا'. Grammatical false positive errors are also detected as in 'الاونة' which was replaced with the word 'الاونة'.

There are also some real-word errors in the original text that were detected by our prototype as in 'خطاء' which is actually a real word error; the intended word was 'خطا'. A run on error was also detected 'تعتبرهي' in which the typist missed a space in between the two word 'تعتبر' and 'هي'. The prototype treated it as an error and replaced it with the word 'تعتبره'.

Limitations of our method:

We believe that our method lacks the following:

- 1- Although the used corpus is large, it is still not large-enough for such type of applications. There are always many correct tri-grams in a text that are not in the tri-gram database, this is known as data sparseness problem. The larger the corpus the better the expected results. We believe if Google lunched the 1T n-grams for Arabic as it did for English languages (Thorsten and Alex 2006) and other European languages like Czech, Dutch and Spanish (Thorsten and Alex 2009), Arabic Google 1T n-grams will help as Google has huge resources for Arabic texts.
- 2- We found that our method does not work well with proper nouns as shown.
- 3- Large space is needed to store the language models. An efficient way

Table 5.6: Different cases of false positives

| False Positive | Amending | Interpretation |
|----------------|-----------|-------------------|
| أوميجا | أوميغا | Proper noun |
| السيلكون | السيليكون | Proper noun |
| البايا | اسبانيا | Proper noun |
| الخضراوات | الخضروات | Non-word error |
| مليارا | مليار | Grammatical error |
| خطاء | خطا | Real-word error |
| بدر | بندر | Proper noun |
| تعتبرهي | تعتبره | Run on error |
| ستعيد | سيعيد | Different pronoun |
| يقتولون | يقتلون | Non-word error |
| الشراثن | الشريان | Proper noun |
| الاونه | الاونة | Non-word error |
| واشطن | واشنطن | Proper noun |
| الحميضي | الحميدي | Proper noun |
| لصحيفة | لصحيفة | Kashida |
| محمود | محمد | Proper noun |

5.3.3 Performance Comparison

The experimental results of the prototype showed promising correction recall. However, it is not possible to compare our results with other published works as there is no benchmarking dataset for real-word errors correction for Arabic text. Most of the research has been done to English language, we will view some of the results obtained in English context-sensitive spell checking in this section, although it is not a fair comparison.

Low performance problem, especially low precision rate problem, in the unsupervised approaches is not only a problem we faced, this problem is reported by many researchers.

In (G. Hirst and Budanitski 2001), the performance of the system that detects and corrects malapropisms in English text was measured using detection recall, correction recall and precision. They reported a detection recall varying from 23.1% to 50%, a correction recall varying from 2.6% to 8% and a precision varying from 18.4% to 24.7%.

(Verberne 2002) developed a word-tri-gram method that, she considered a tri-gram to be probably wrong if and only if it does not occur in the British National Corpus. Her evaluation of the method showed a recall of 51% of detection recall of 33% for correction at the price of a precision of only 5%.

(St-Onge 1995) developed a method for detecting and correcting malapropisms in English text. He also measured the performance of his method using detection recall, correction recall and precision. A detection recall of 28.5% was reported, the correction recall obtained was 24.8% and the precision was 12.5%.

(Mays, Damerau, and Mercer 1991) used a tri-gram model to detect and correct real-word errors. However, the test data they used to measure the performance of their spell checking prototype was not realistic as they induce errors in the sentences they chose for testing. They knew in advance that the sentence has only one error. For that reason, we cannot really compare their results to ours. The detection recall they reported was 76% and the correction recall was 74%.

(Wilcox-O'Hearn, G Hirst, and A Budanitsky 2008) re-evaluated the method of (Mays, Damerau, and Mercer 1991) to make it comparable with other methods. Then they compared it with the WordNet-based method of (Graeme Hirst and Alexander

Budanitsky 2005). They obtained a detection recall of 30.6%, a correction recall of 28.1% and a precision of 20.7%.

All these studies show that our unsupervised method performs relatively well if we bear in mind the inflectional property and the problem of words similarity in Arabic language as discussed in section 2.3.

CHAPTER 6

REAL WORD ERROR DETECTION AND CORRECTION USING SUPERVISED TECHNIQUES

Ambiguity between words can be detected by the set of words surrounding them. For instance, if the target word is ambiguous between 'سمن' and 'ثمن', and we observe words like 'البيع', 'مشتريات', 'يدفع', and 'البضاعة' nearby, this indicates that the target word should be 'ثمن'. On the other hand, words such as 'تعجن', 'النباتية', and 'ملعقة' in the context more probably imply 'سمن'. This observation is the idea behind the method of context words, which is also known as word co-occurrence. 'سمن' and 'ثمن' are called a confusion set. Each word w_i in the confusion set has a characteristic distribution of words that occur in its context. In order to judge an ambiguous word, we look at the set of words surrounding it, then we see which w_i 's distribution the context most closely follow (A. Golding 1995).

In this chapter we address the problem of detecting and correcting real-word errors in a context using two supervised methods, namely the word co-occurrence method and the n-gram language models. Word co-occurrence method uses the context words surrounding the target words from predefined confusion sets. The n-gram language models method is explained in detail in the previous chapter.

6.1 Introduction

In order to address real-word errors, information from the surrounding context is used for detecting the erroneous words as well as to correct them. We need to identify words that

are semantically unrelated to their context. Afterwards, we need to find out which of the word variations is more related to that context and could be the best replacement for the erroneous (suspicious) word. Relatedness of a word to its context is determined by a measure of semantic distance initially proposed by (Jiang and Conrath 1997).

A collection of confusion sets is used in addressing real-word errors using the surrounding context. We chose twenty eight confusion sets to be used in our experiments. These confusion sets are chosen from the different types of confusion sets mentioned early in chapter 4 based on their availability in our corpus.

6.2 The Baseline Method

The baseline method disambiguates words in the confusion set using the Maximum Likelihood Estimate (MLE). It selects the word most encountered in the training corpus and simply ignores the context information (i.e. words are predicted by their prior probabilities). For instance, in the confusion set {'غسل', 'عسل'}, 'غسل' occurred more often than 'عسل' in the training corpus. Using MLE, the method predicts every occurrence of 'غسل' or 'عسل' in the test corpus as 'غسل' as it is more probable in the training corpus.

Table 6.1 shows the performance of the baseline method for 28 confusion sets. This collection of confusion sets will be used for evaluating the remaining methods with the same training and testing sets. Each row of the table gives the results for one confusion set: the words in the confusion set; the number of occurrences of all words in the confusion set in the training and in the test sets; the word in the confusion set that occurred most often in the training corpus along with the number of instances; and the prediction accuracy of the baseline method for the test set. Prediction accuracy is the

number of times that the method predicted the correct word, divided by the total number of test cases. For example, the members of the confusion set { 'كثير', 'كبير' } occurred 2770 times in the test corpus; out of which 'كبير' occurred 2183 times and 'كثير' occurred 587 times. The baseline method predicts 'كبير' each time, and thus is right 2183 times, $2183 / 2770 = 0.788$, therefore the accuracy is 78.8%. The baseline method will be used for comparison with other methods.

6.3 Context Words Co-occurrence Method

Given the context words c_j where j is from 1 to n using a k -word window of the target word. We need to find out the proper word w_i from the confusion set that is most probable to that context. Figure 6.1 shows an example of predicting a word from the confusion set { 'سمن', 'ثمن' }, given a context window size of ± 3 words. The probability of each word w_i in the confusion set is calculated using Bayes' rule:

$$p(w_i | c_{-k}, \dots, c_{-1}, c_1, \dots, c_k) = \frac{p(c_{-k}, \dots, c_{-1}, c_1, \dots, c_k | w_i) p(w_i)}{p(c_{-k}, \dots, c_{-1}, c_1, \dots, c_k)} \quad (6.1)$$

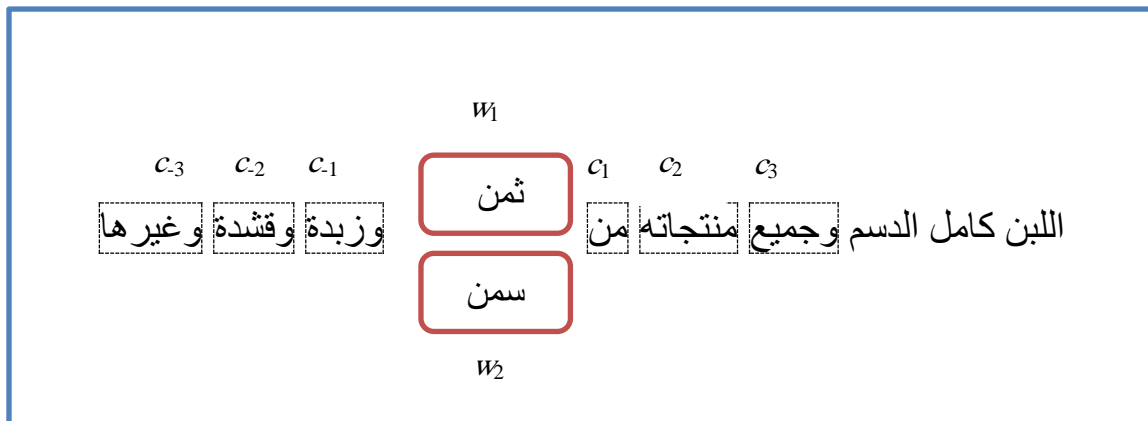


Figure 6.1: Example of context words

Table 6.1: Baseline method prediction accuracy on the test set for the 28 confusion sets

| Confusion Set | No. of training sentences | Most frequent word (No. of sentences) | No. of test sentences | Baseline Accuracy % |
|-------------------|---------------------------|---------------------------------------|-----------------------|---------------------|
| كبير – كثير | 6463 | كبير (5080) | 2770 | 78.8 |
| صغير – قصير | 1241 | صغير (671) | 532 | 54.1 |
| أساس – أثاث | 2567 | أساس (2469) | 1100 | 96.9 |
| عمارة – إمارة | 1179 | إمارة (988) | 505 | 81.2 |
| ثمن – سمن | 861 | ثمن (839) | 369 | 98.7 |
| العمارة – الإمارة | 283 | الإمارة (472) | 284 | 74.3 |
| ثروة – ثورة | 537 | ثروة (326) | 230 | 57.4 |
| الأرق – الحرق | 348 | الأرق (281) | 149 | 79.2 |
| مضر – مصر | 1310 | مصر (1017) | 562 | 74.9 |
| أشعار – أسعار | 6805 | أسعار (6580) | 2916 | 96.1 |
| القرعة – القرحة | 610 | القرعة (505) | 262 | 77.9 |
| الرحم – الرسم | 1919 | الرحم (1231) | 823 | 62.2 |
| حال – مال | 9785 | مال (7647) | 4193 | 78.8 |
| يغرق – يفرق | 160 | يغرق (123) | 69 | 87.3 |
| غسل – غسل | 654 | غسل (435) | 281 | 70.5 |
| تسير – تصير | 867 | تسير (837) | 351 | 95.4 |
| إصرار – إصرار | 587 | إصرار (397) | 250 | 66.8 |
| الزهور – الظهور | 766 | الظهور (563) | 327 | 75.2 |
| هاوية – حاوية | 220 | حاوية (204) | 94 | 90.4 |
| سراب – شراب | 152 | شراب (128) | 65 | 83.1 |
| مسير – مصير | 471 | مصير (458) | 202 | 95.6 |
| محنة – مهنة | 734 | مهنة (711) | 315 | 98.1 |
| يسب – يصب | 377 | يصب (373) | 161 | 99.4 |
| ألم – علم | 2388 | علم (1487) | 1023 | 63.7 |
| الشعر – السعر | 4023 | السعر (2627) | 1724 | 65.8 |
| الأرض – العرض | 5206 | العرض (3056) | 2231 | 59.6 |
| الشكر – الشكر | 3472 | الشكر (1766) | 1488 | 50.5 |
| ثلاثة – سلاسة | 1539 | ثلاثة (1471) | 660 | 96.2 |

It is difficult to estimate the probability $p(c_{-k}, \dots, c_{-1}, c_1, \dots, c_k | w_i)$ due to data sparseness. Instead, we assume that the presence of a word in the context is independent from the presence of the others. By this assumption the estimated probability is calculated as:

$$\mathbf{p}(c_{-k}, \dots, c_{-1}, c_1, \dots, c_k | w_i) = \prod_{j \in \{-k, \dots, -1, 1, \dots, k\}} \mathbf{p}(c_j | w_i) \quad (6.2)$$

We use the MLE to estimate the probabilities of the context words surrounding w_i . We count the number of occurrences of a word c_j within the context of each w_i that occur within $\pm k$ words window in the training corpus. For each context word c_j , we calculate the probability $p(c_j | w_i)$ by dividing its number of occurrences by the total number of w_i occurrences.

Once we observe a word in the confusion set within a sentence in the correction phase, we look for the words within the same window. Based on the probability, we classify the word to be any of the confusion set members. The probability is calculated as follows: if a word is observed within the context of the word w_i , in the training phase, we sum the log probability of that word given w_i . The probabilities of the sentences for every word w_i in the confusion set are calculated. The log of probability is used to avoid underflow and to save computation time by using addition instead of multiplication. The word in the confusion set with the highest probability is chosen. Figure 6.2 shows the proposed algorithm for the training and the testing phases.

Training phase

For each word in the confusion set:

- (1) Extract all words in a $\pm k$ window.
- (2) Count the number of occurrences for each context word.
- (3) Store context words along with their statistics.

Testing Phase

For any encountered word in the confusion set,

- (1) Extract the words c_j in a $\pm k$ window.
- (2) If the word c_j is encountered during the training phase, then
 - a) Calculate its probability.
 - b) Calculate the cumulative probability summing up log probabilities of all encountered context words.
- (3) Repeat (1) and (2) for all confusion set members
- (4) The word of the sentence with the highest probability is considered as the best fit in that context.

Figure 6.2: Training and testing phases.

6.4 N-Gram Language Models

N-gram language models is a supervised technique used for real word error detection and correction. The same mechanism in the unsupervised method is followed when detecting and correcting real-word errors. The target words here are the words belonging to any of the confusion sets.

To detect or to predict the proper word of the confusion set in the tested sentence, we use the same procedure followed in the unsupervised case. For each target word w_i in the confusion set, the four words surrounding it are utilized to predict the proper word in that

sentence. For each word in the confusion set, a new sentence is generated by placing the confusion set word in the target word. The probabilities of all the sentences with respect to the confusion set words are calculated. Three tri-grams are considered viz. $\{w_{i-2}, w_{i-1}, w_i\}$, $\{w_{i-1}, w_i, w_{i+1}\}$, and $\{w_i, w_{i+1}, w_{i+2}\}$. In case that the tri-gram is not found, bi-grams back off is used and uni-gram back off is used when a bi-gram is not found. The sentence that gives the highest probability is considered as the correct one indicating that the confusion set member in that sentence is a better choice and hence more likely to be the correct word.

For example, in the sentence ‘وبعد ذلك تم عرض فيلم قصير عن المركز ثم بدأت دفعة الحوار بين الجلسات’ the word ‘قصير’ is the target word. The probability is calculated for three tri-grams:

عرض فيلم قصير
فيلم قصير عن
قصير عن المركز

The same is done for the other member of the confusion set ‘صغير’, the sentence that gives the highest probability is considered the correct one and the confusion set member is considered the best replacement.

6.5 Experimental Results

In this section, the experimental results using the above two techniques are presented.

Our Al-Riyadh newspaper corpus mentioned in chapter 4 is used in our experiments. Only sentences that contain words from the confusion sets are extracted and divided into 70% for training and the remaining for testing. Statistics for confusion set sentences are shown in detail in Table 6.1 above. Note that some sentences for some confusion sets are

reduced because they have many occurrences. For instance, the words of the confusion set { 'كثير', 'كبير' } occurred in the corpus 32,569 times but only 9,233 of sentences that contain words from this confusion set are used. Moreover we encountered some short sentences of length three or less, these sentences seem to be sub titles within the articles. These sentences are also excluded from the sentences used in the experiments.

6.5.1 Word Co-occurrence

For word co-occurrence, we experimented with window sizes of (2, 3, 5, 7, 12, and 20) where k is half the window size. Table 6.2 shows the results of each window. Each row in the table shows the results for one confusion set, it shows the number of context words for each word in that confusion set in the training phase. It also shows the correction accuracy for each window size. There are three rows for each confusion set, in the first we consider all the surrounding words, in the second the function words are ignored, while in the third row we follow the standard deviation approach by (C. Ben Othmane Zribi and M. Ben Ahmed 2012). The results show that the best average accuracy was achieved when $k = 3$. This confirms the results of (A. Golding 1995) and contradicts with the conclusion of (C. Ben Othmane Zribi and M. Ben Ahmed 2012) that the longer the context the better the correction accuracy. We repeated the experiments ignoring the stop (function) words (the results of the second row). The results show that ignoring the function words does not improve the correction rate. This may be due to the nature of Arabic language and the way in which words co-occur with certain function words (i.e. some words are recognized easier if a specific function word precedes or comes after them). Our results for $k = 3$ are better than the results obtained in (C. Ben Othmane Zribi and M. Ben Ahmed 2012) for all window sizes. Table 6.3 compares between the baseline

method and the word co-occurrence method for a window of ± 3 , the table shows that the co-occurrence method on the average of (91.5%) is better than the baseline method (76.6%). Table 6.4 shows the confusion matrix of the words in the confusion sets using the word co-occurrence method for the same window size.

Golding (A. Golding 1995) pruned context words that have insufficient information, by ignoring context words that occur less than 10 times within the context of the confusion sets. We pruned such words from the context obtained in the training phase but the accuracy rate dropped, we then ignored words that occurred 10 and 5 times but the accuracy always got worse as the number goes larger i.e. 10 in this case. We think that each word in the context is useful in the discrimination process.

Table 6.2: Context words method for different values of k using whole words, ignoring stop word and standard deviation used by Bin Othman.

| Confusion Set ($w_1 - w_2$) | # of co-occurrences $w_1 - w_2$ | $w \pm 2$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 3$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 5$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 7$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 12$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 20$ Accur acy% |
|-------------------------------|------------------------------------|----------------------------|------------------------------------|----------------------------|------------------------------------|----------------------------|------------------------------------|----------------------------|------------------------------------|-----------------------------|------------------------------------|-----------------------------|
| كثير - كبير | 12553-4731 | 82.1 | 18091-7422 | 80.7 | 26369-11808 | 78.2 | 32590-15230 | 77.9 | 43603-21413 | 77.3 | 54476-27597 | 77.3 |
| | 12053-4454 | 80.4 | 17489-7038 | 78.9 | 25651-11320 | 78.0 | 31805-14680 | 77.9 | 42708-20757 | 77.8 | 53490-26860 | 77.6 |
| | 12553-4731 | 74.9 | 18091-7422 | 70.2 | 26369-11808 | 69.5 | 32590-15230 | 70.5 | 43603-21413 | 72.7 | 54476-27597 | 70.1 |
| قصير - صغير | 1296-900 | 89.5 | 1951-1466 | 88.7 | 3080-2403 | 84.2 | 4013-3171 | 79.7 | 5819-4745 | 74.4 | 7722-6430 | 73.3 |
| | 1131-758 | 85.7 | 1753-1290 | 84.6 | 2823-2179 | 82.0 | 3704-2921 | 77.6 | 5443-4439 | 73.9 | 7284-6080 | 72.2 |
| | 1296-900 | 89.6 | 1951-1466 | 87.6 | 3080-2403 | 84.9 | 4013-3171 | 81.0 | 5819-4745 | 77.8 | 7722-6430 | 78.3 |
| اثاث - اساس | 3073-249 | 96.6 | 4601-373 | 97.6 | 6983-610 | 97.5 | 8864-818 | 97.2 | 12355-1234 | 97 | 1729-16004 | 96.8 |
| | 2834-208 | 97.3 | 4297-310 | 96.6 | 6584-519 | 97.4 | 8443-696 | 96.8 | 11831-1065 | 97 | 15401-1544 | 97 |
| | 3073-249 | 87.7 | 4601-373 | 82.0 | 6983-610 | 71.6 | 8864-818 | 72.1 | 12355-1234 | 71.6 | 1729-16004 | 72.6 |
| إمارة - عمارة | 355-1079 | 95.8 | 562-1804 | 95.1 | 937-2986 | 93.5 | 1279-3987 | 92.9 | 1965-5999 | 92.5 | 2786-8187 | 90.5 |
| | 291-966 | 95.8 | 482-1656 | 95.5 | 820-2790 | 94.1 | 1136-3755 | 93.3 | 1786-5706 | 91.3 | 2565-7836 | 90.1 |
| | 355-1079 | 94.7 | 562-1804 | 93.7 | 937-2986 | 94.3 | 1279-3987 | 91.3 | 1965-5999 | 89.9 | 2786-8187 | 91.3 |
| سمن - ثمن | 1300-68 | 97.3 | 1932-99 | 98.4 | 2976-159 | 98.9 | 3860-218 | 99.2 | 5616-332 | 98.9 | 7689-470 | 98.9 |
| | 1158-50 | 99.5 | 1753-74 | 98.9 | 2739-124 | 98.7 | 3590-168 | 98.7 | 5285-267 | 98.9 | 7299-387 | 99.2 |
| | 1300-68 | 98.9 | 1932-99 | 99.2 | 2979-159 | 99.5 | 3860-218 | 99.2 | 5616-332 | 98.4 | 7689-470 | 96.5 |
| الإمارة - العمارة | 371-890 | 87.3 | 599-1319 | 87.3 | 1019-2030 | 92.3 | 1381-2618 | 92.3 | 2090-3770 | 90.5 | 2848-5021 | 89.1 |
| | 321-786 | 91.6 | 539-1200 | 93.3 | 925-1872 | 93.3 | 1268-2431 | 93.3 | 1935-3540 | 91.6 | 2659-4755 | 90.1 |
| | 371-890 | 91.6 | 599-1319 | 90.9 | 1019-2030 | 89.1 | 1381-2618 | 88.0 | 2090-3770 | 87.3 | 2848-5021 | 88 |

| Confusion Set ($w_1 - w_2$) | # of co-occurrences $w_1 - w_2$ | $w \pm 2$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 3$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 5$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 7$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 12$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 20$ Accuracy% |
|-------------------------------|---------------------------------|---------------------|---------------------------------|---------------------|---------------------------------|---------------------|---------------------------------|---------------------|---------------------------------|----------------------|---------------------------------|----------------------|
| ثورة - ثروة | 708-430 | 78.7 | 1043-656 | 81.3 | 1593-1079 | 83.0 | 1441-2095 | 80.4 | 3134-2246 | 74.8 | 4337-3153 | 80.0 |
| | 600-361 | 77.4 | 903-567 | 82.2 | 1422-950 | 80.7 | 1886-1285 | 75.7 | 2883-2039 | 76.1 | 4035-2911 | 79.1 |
| | 708-430 | 75.7 | 1043-656 | 77.4 | 1593-1079 | 76.1 | 1441-2095 | 71.3 | 3134-2246 | 64.4 | 4337-3153 | 65.2 |
| الحرق - الأرق | 406-162 | 85.9 | 568-247 | 85.9 | 882-378 | 85.9 | 1127-468 | 85.9 | 1593-663 | 86.6 | 2028-866 | 89.3 |
| | 327-132 | 82.6 | 461-208 | 85.2 | 751-321 | 87.9 | 976-400 | 85.2 | 1405-570 | 89.3 | 1716-755 | 91.3 |
| | 406-162 | 85.1 | 568-247 | 83.8 | 882-378 | 84.5 | 1127-468 | 80.4 | 1593-663 | 84.4 | 2028-866 | 89.7 |
| مصر - مضر | 556-1434 | 84.5 | 773-2102 | 87.0 | 1138-3098 | 88.3 | 1464-3939 | 88.3 | 2086-5640 | 89.5 | 2774-7455 | 87.9 |
| | 499-1290 | 86.8 | 700-1925 | 88.3 | 1026-2867 | 89.0 | 1330-3680 | 87.7 | 1914-5320 | 88.8 | 2568-7093 | 87.7 |
| | 556-1434 | 84.7 | 773-2102 | 81.1 | 1138-3098 | 77.2 | 1464-3939 | 76.7 | 2086-5640 | 77.6 | 2774-7455 | 75.3 |
| اسعار - اشعار | 415-4339 | 98.3 | 662- 6558 | 98.1 | 1072-9733 | 97.8 | 1454-12163 | 97.4 | 2194-16585 | 97.3 | 2991-21052 | 97.0 |
| | 345-4064 | 98.3 | 567 - 6203 | 97.8 | 943 - 9306 | 97.6 | 1301- 11681 | 97.4 | 1998 - 16019 | 97.2 | 2757 - 20426 | 97.1 |
| | 415-4339 | 97.6 | 662-6558 | 95.2 | 1072-9733 | 92.2 | 1454-12163 | 88.9 | 2194-16585 | 86.2 | 2991-21052 | 86.8 |
| القرعة - القرحة | 818-214 | 91.6 | 1203-316 | 93.5 | 1804-475 | 93.5 | 2273-615 | 95.0 | 3223-890 | 96.9 | 4253-1174 | 97.3 |
| | 722-166 | 95.8 | 1076-259 | 95.4 | 1634-399 | 96.6 | 2081-522 | 96.9 | 2989-773 | 98.5 | 3980-1024 | 97.7 |
| | 818-214 | 91.9 | 1203-316 | 94.2 | 1804-475 | 94.2 | 2273-615 | 93.1 | 3223-890 | 92.3 | 4253-1174 | 93.1 |
| الرحم - الرسم | 1471- 737 | 95.3 | 2132-1147 | 96.4 | 3178-1846 | 97.7 | 4048-2414 | 97.8 | 5618-3534 | 97.9 | 7214- 4683 | 97.5 |
| | 1322-643 | 96.1 | 1947- 1017 | 97.6 | 2937 - 1683 | 98.2 | 3772 - 2215 | 98.5 | 5282 - 3292 | 97.9 | 6830 - 4391 | 98.2 |
| | 1471-737 | 93.6 | 2132-1147 | 91.9 | 3178-1846 | 91.4 | 4048-2414 | 96.0 | 5618-3534 | 92.35 | 7214- 4683 | 95.0 |

| Confusion Set ($w_1 - w_2$) | # of co-occurrences $w_1 - w_2$ | $w \pm 2$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 3$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 5$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 7$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 12$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 20$ Accur acy% |
|-------------------------------|---------------------------------|----------------------|---------------------------------|----------------------|---------------------------------|----------------------|---------------------------------|----------------------|---------------------------------|-----------------------|---------------------------------|-----------------------|
| حال - مال | 7204-1625 | 93.8 | 10744-2582 | 92.2 | 15935-4254 | 91 | 19940-5609 | 90.3 | 27378-8094 | 89.8 | 35033-10652 | 89.5 |
| | 6868-1440 | 91.7 | 10313-2355 | 91.2 | 15388-3952 | 90.4 | 19322-5272 | 89.63 | 26635-7688 | 89.4 | 34209-10185 | 89.3 |
| | 7204-1625 | 76.0 | 10744-2582 | 60.0 | 15935-4254 | 47.2 | 19940-5609 | 44.4 | 27378-8094 | 42.4 | 35033-10652 | 42.0 |
| يفرق - يغرق | 111-244 | 92.8 | 157-416 | 89.9 | 250-700 | 81.2 | 327-954 | 79.7 | 524-1466 | 78.3 | 732-2100 | 76.8 |
| | 73-187 | 50.7 | 107-331 | 56.5 | 186-580 | 69.6 | 252-814 | 75.4 | 423-1273 | 75.4 | 609-1872 | 79.7 |
| | 111-244 | 89.9 | 157-416 | 84.1 | 250-700 | 82.6 | 327-954 | 85.5 | 524-1466 | 81.2 | 732-2100 | 79.7 |
| عسل - غسل | 599-406 | 92.5 | 978-607 | 94.0 | 1686-984 | 94 | 2231-1323 | 91.5 | 3291-2004 | 91.5 | 4338-2707 | 91.1 |
| | 515-343 | 93.2 | 874-523 | 95.0 | 1540-867 | 93.2 | 2062-1185 | 92.5 | 3071-1820 | 90.8 | 4083-2481 | 90.75 |
| | 599-406 | 92.9 | 978-607 | 94.7 | 1686-984 | 93.6 | 2231-1323 | 92.9 | 3291-2004 | 94.7 | 4338-2707 | 91.1 |
| تصير - تسيير | 1214-80 | 94.6 | 1887-116 | 96.5 | 3219-185 | 96.5 | 4370-242 | 96.5 | 6475-347 | 96.5 | 8860-443 | 96.5 |
| | 1062-58 | 96.0 | 1665-85 | 96.5 | 2941-143 | 97.0 | 4040-188 | 97.0 | 6080-267 | 97.0 | 8395-343 | 97.3 |
| | 1214-80 | 96.2 | 1887-116 | 97.0 | 3219-185 | 97.0 | 4370-242 | 97.0 | 6475-347 | 96.0 | 8860-443 | 96.2 |
| اصرار - اسرار | 435-684 | 85.6 | 643-1075 | 85.6 | 1051-1780 | 86.4 | 1392-2404 | 81.2 | 2079-3619 | 80.8 | 2918-5009 | 97.2 |
| | 363-579 | 85.6 | 545-945 | 82.8 | 907-1598 | 80.8 | 1223-2186 | 81.2 | 1857-3351 | 80.0 | 2643-4697 | 78.0 |
| | 435-684 | 81.2 | 643-1075 | 68.8 | 1051-1780 | 59.6 | 1392-2404 | 55.6 | 2079-3619 | 58.0 | 2918-5009 | 61.6 |
| الظهور - الزهور | 464-968 | 88.1 | 699-1507 | 89.0 | 1116-2477 | 89.6 | 1442-3255 | 91.4 | 2092-4846 | 91.7 | 2821-6583 | 91.7 |
| | 402-845 | 91.7 | 620-1339 | 90.2 | 1014-2252 | 92.0 | 1314-3003 | 91.4 | 1923-4522 | 94.2 | 2618-6201 | 91.7 |
| | 464-968 | 86.9 | 699-1507 | 84.1 | 1116-2477 | 82.3 | 1442-3255 | 79.8 | 2092-4846 | 78.9 | 2821-6583 | 81.0 |

| Confusion Set ($w_1 - w_2$) | # of co-occurrences $w_1 - w_2$ | $w \pm 2$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 3$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 5$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 7$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 12$ Accur acy% | # of co-occurrences $w_1 - w_2$ | $w \pm 20$ Accur acy% |
|-------------------------------|---------------------------------|----------------------|---------------------------------|----------------------|---------------------------------|----------------------|---------------------------------|----------------------|---------------------------------|-----------------------|---------------------------------|-----------------------|
| حاوية - هاوية | 50-258 | 93.6 | 75-372 | 93.6 | 124-573 | 93.6 | 171-745 | 92.6 | 275-1125 | 92.6 | 407-1568 | 91.5 |
| | 41-209 | 90.4 | 61-302 | 91.5 | 99-486 | 93.6 | 137-646 | 96.8 | 22-997 | 94.7 | 336-1423 | 93.6 |
| | 50-258 | 91.5 | 75-372 | 91.5 | 124-573 | 92.6 | 171-745 | 90.4 | 275-1125 | 90.4 | 407-1568 | 90.4 |
| شراب - شراب | 73-277 | 93.8 | 109-396 | 92.3 | 176-621 | 84.6 | 242-823 | 84.6 | 382-1229 | 86.2 | 561-1675 | 86.2 |
| | 48-231 | 89.2 | 75-330 | 89.2 | 123-530 | 86.2 | 180-714 | 86.2 | 298-1094 | 87.7 | 450-1507 | 87.7 |
| | 73-277 | 95.4 | 109-396 | 90.8 | 176-621 | 87.7 | 242-823 | 86.2 | 382-1229 | 84.6 | 561-1675 | 81.5 |
| مصير - مسير | 45-919 | 95.0 | 67-1351 | 94.1 | 102-2096 | 96.0 | 136-2757 | 96.5 | 212-4048 | 96.0 | 313-5550 | 96.0 |
| | 36-804 | 95.0 | 51-1195 | 95.5 | 80-1889 | 95.0 | 105-2522 | 95.0 | 167-3756 | 95.5 | 253-5204 | 95.0 |
| | 45-919 | 96.5 | 67-1351 | 96.0 | 102-2096 | 96.0 | 136-2757 | 96.0 | 212-4048 | 96.0 | 313-5550 | 95.5 |
| مهنة - محنة | 74-1257 | 97.1 | 107-1904 | 98.4 | 174-2971 | 98.1 | 232-3849 | 98.4 | 362-5604 | 98.1 | 519-7587 | 98.1 |
| | 52-1102 | 99.0 | 79-1710 | 97.8 | 138-2729 | 98.7 | 184-3566 | 98.7 | 299-5259 | 98.4 | 443-7192 | 98.1 |
| | 74-1257 | 98.4 | 107-1904 | 98.4 | 174-2971 | 98.1 | 232-3849 | 98.1 | 362-5604 | 98.1 | 519-7587 | 98.1 |
| يصب - يسب | 15-549 | 99.4 | 22-913 | 98.8 | 35-1625 | 99.4 | 44-2249 | 99.4 | 65-3443 | 99.4 | 97-4850 | 99.4 |
| | 7-460 | 99.4 | 13-796 | 99.4 | 24-1459 | 99.4 | 28-2050 | 99.4 | 42-3194 | 99.4 | 66-4554 | 99.4 |
| | 15-549 | 99.4 | 22-913 | 99.4 | 35-1625 | 99.4 | 44-2249 | 99.4 | 65-3443 | 99.4 | 97-4850 | 99.4 |
| علم - ألم | 1319-2335 | 84.8 | 1970-3558 | 84.3 | 3071-5554 | 85.2 | 3957-7196 | 84.1 | 5540-10312 | 85.2 | 7197-13700 | 84.4 |
| | 1155-2096 | 84.9 | 1763-3259 | 83.8 | 2821-5190 | 84.2 | 3661-6777 | 84.0 | 5173-9821 | 85.6 | 6780-13145 | 84.7 |
| | 1319-2335 | 82.9 | 1970-3558 | 82.1 | 3071-5554 | 77.5 | 3957-7196 | 75.1 | 5540-10312 | 74.7 | 7197-13700 | 76.4 |

| Confusion Set ($w_1 - w_2$) | # of co-occurrences $w_1 - w_2$ | $w \pm 2$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 3$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 5$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 7$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 12$ Accuracy% | # of co-occurrences $w_1 - w_2$ | $w \pm 20$ Accuracy% |
|-------------------------------|---------------------------------|---------------------|---------------------------------|---------------------|---------------------------------|---------------------|---------------------------------|---------------------|---------------------------------|----------------------|---------------------------------|----------------------|
| الشعر - الشعر | 2798-2023 | 91.8 | 4072-2915 | 94.0 | 5975-4294 | 95.8 | 7522-5398 | 96.3 | 10441-7288 | 96.8 | 13478-9126 | 96.5 |
| | 2524-1844 | 93.4 | 3749-2694 | 94.4 | 5571-4006 | 96.2 | 7087-566 | 96.6 | 9937-6884 | 97.2 | 12911-8682 | 96.6 |
| | 2798-2023 | 87.4 | 4072-2915 | 87.8 | 5975-4294 | 88.5 | 7522-5398 | 88.1 | 10441-7288 | 90.1 | 13478-9126 | 91.8 |
| الارض - العرض | 3301-3592 | 84.8 | 4911-5290 | 85.5 | 7605-7853 | 84.7 | 9739-9912 | 84.9 | 13678-13701 | 83.8 | 17913-17640 | 84.0 |
| | 3050-3328 | 83.5 | 4602-4990 | 84.1 | 7206-7499 | 83.9 | 9283-9506 | 84.5 | 13141-13220 | 84.1 | 17298-17084 | 84.0 |
| | 3301-3592 | 81.8 | 4911-5290 | 79.7 | 7605-7853 | 76.9 | 9739-9912 | 78.8 | 13678-13701 | 80.5 | 17913-17640 | 81.6 |
| السكر - الشكر | 1969-1497 | 96.8 | 2970-2406 | 97.8 | 4590-3910 | 98.3 | 5898-5144 | 98.1 | 8149-7629 | 97.6 | 10404-10451 | 97.1 |
| | 1774-1336 | 96.1 | 2717-2191 | 97.2 | 4273-3631 | 97.8 | 5540-4818 | 97.6 | 7722-7221 | 97.2 | 9923-9983 | 96.8 |
| | 1969-1497 | 95.6 | 2970-2406 | 93.3 | 4590-3910 | 90.5 | 5898-5144 | 89.0 | 8149-7629 | 89.4 | 10404-10451 | 90.6 |
| سلاسة ثلاثة | 198-1539 | 97.0 | 279-2198 | 97.6 | 447-3492 | 98.3 | 608-4722 | 98.0 | 939-7127 | 97.3 | 1297-9755 | 97.1 |
| | 163-1410 | 97.4 | 234-2036 | 97.7 | 380-3256 | 97.9 | 530-4453 | 97.7 | 829-6793 | 97.3 | 1167-9336 | 97.0 |
| | 198-1539 | 97.0 | 279-2198 | 96.8 | 447-3492 | 96.7 | 608-4722 | 96.4 | 939-7127 | 93.6 | 1297-9755 | 90.3 |
| Average | | | | | | | | | | | | |
| with stop words | | 91.5 | | 91.5 | | 91.1 | | 90.7 | | 90.3 | | 90.3 |
| without stop words | | 91.0 | | 90.1 | | 90.8 | | 90.5 | | 90.4 | | 90.1 |
| Bin Othman | | 86.3 | | 81.7 | | 77.6 | | 73.6 | | 76.1 | | 76.2 |

Table 6.3: Comparison between the baseline method and word co-occurrence method with a window size of ± 3 .

| Confusion Set | Baseline Accuracy % | Word Co-occurrence $k = 3$ Accuracy % |
|-------------------|---------------------|---------------------------------------|
| كبير – كثير | 78.8 | 80.7 |
| صغير – قصير | 54.1 | 88.7 |
| أساس – أثاث | 96.9 | 97.6 |
| عمارة – إمارة | 81.2 | 95.1 |
| ثمن – سمن | 98.7 | 98.4 |
| العمارة – الإمارة | 74.3 | 87.3 |
| ثروة – ثورة | 57.4 | 81.3 |
| الأرق – الحرق | 79.2 | 85.9 |
| مضر – مصر | 74.9 | 87.0 |
| أشعار – أسعار | 96.1 | 98.1 |
| القرعة – القرحة | 77.9 | 93.5 |
| الرحم – الرسم | 62.2 | 96.4 |
| حال – مال | 78.8 | 92.2 |
| يغرق – يفرق | 87.3 | 89.9 |
| غسل – عسل | 70.5 | 94.0 |
| تصير – تسير | 95.4 | 96.5 |
| إصرار – إصرار | 66.8 | 85.6 |
| الزهور – الظهور | 75.2 | 89.0 |
| هاوية – حاوية | 90.4 | 93.6 |
| سراب – شراب | 83.1 | 92.3 |
| مسير – مصير | 95.6 | 94.1 |
| محنة – مهنة | 98.1 | 98.4 |
| يسبب – يصب | 99.4 | 98.8 |
| ألم – علم | 63.7 | 84.3 |
| السعر – الشعر | 65.8 | 94.0 |
| الأرض – العرض | 59.6 | 85.5 |
| السكر – الشكر | 50.5 | 97.8 |
| ثلاثة – سلاسة | 96.2 | 97.6 |
| Average | 76.6 | 91.5 |

Table 6.4: Confusion matrix between words in the confusion sets using context words method for $k = 3$.

| Confusion set | Word | | | Accuracy% | Total accuracy% |
|-------------------|---------|---------|---------|-----------|-----------------|
| كبير – كثير | | كبير | كثير | | 80.0 |
| | كبير | 1901 | 272 | 87.5 | |
| | كثير | 282 | 315 | 52.8 | |
| صغير – قصير | | صغير | قصير | | 85.9 |
| | صغير | 244 | 31 | 88.7 | |
| | قصير | 44 | 213 | 82.9 | |
| أساس – أثاث | | أساس | أثاث | | 97.6 |
| | أساس | 1060 | 22 | 98 | |
| | أثاث | 4 | 12 | 75.0 | |
| عمارة – إمارة | | إمارة | عمارة | | 95 |
| | إمارة | 406 | 21 | 95.1 | |
| | عمارة | 4 | 74 | 94.9 | |
| ثمن – سمن | | ثمن | سمن | | 98.4 |
| | ثمن | 361 | 3 | 99.2 | |
| | سمن | 3 | 2 | 40 | |
| ثروة – ثورة | | ثورة | ثروة | | 81.3 |
| | ثورة | 70 | 15 | 82.4 | |
| | ثروة | 28 | 117 | 80.7 | |
| العمارة – الإمارة | | الإمارة | العمارة | | 87.3 |
| | الإمارة | 192 | 17 | 91.9 | |
| | العمارة | 19 | 56 | 74.7 | |
| الأرق – الحرق | | الأرق | الحرق | | 85.9 |
| | الأرق | 113 | 16 | 87.6 | |
| | الحرق | 5 | 15 | 75 | |
| مضر – مصر | | مصر | مضر | | 87 |
| | مصر | 388 | 40 | 90.7 | |
| | مضر | 33 | 101 | 75.4 | |

| | | | | | | |
|-----------------|--|--------|--------|-----|------|------|
| أشعار – أسعار | | أسعار | أشعار | | 98.1 | |
| | | أسعار | 2796 | 50 | | 98.2 |
| | | أشعار | 6 | 64 | | 91.4 |
| القرعة – القرحة | | القرعة | القرحة | | 93.5 | |
| | | القرعة | 198 | 11 | | 94.7 |
| | | القرحة | 6 | 47 | | 88.7 |
| الرحم – الرسم | | الرحم | الرسم | | 96.4 | |
| | | الرحم | 501 | 19 | | 96.3 |
| | | الرسم | 11 | 292 | | 96.4 |
| حال – مال | | حال | مال | | 92.2 | |
| | | حال | 3170 | 191 | | 94.3 |
| | | مال | 135 | 697 | | 83.8 |
| يفرق – يفرق | | يفرق | يفرق | | 89.9 | |
| | | يفرق | 11 | 3 | | 78.6 |
| | | يفرق | 4 | 51 | | 92.7 |
| غسل – غسل | | غسل | عسل | | 94 | |
| | | غسل | 189 | 8 | | 95.9 |
| | | عسل | 9 | 75 | | 89.3 |
| تصير – تسيير | | تسيير | تصير | | 96.5 | |
| | | تسيير | 353 | 11 | | 97 |
| | | تصير | 2 | 5 | | 71.4 |
| إصرار – إصرار | | إصرار | إصرار | | 85.6 | |
| | | إصرار | 153 | 22 | | 87.4 |
| | | إصرار | 14 | 61 | | 81.3 |
| الزهور – الظهور | | الظهور | الزهور | | 89 | |
| | | الظهور | 232 | 22 | | 91.3 |
| | | الزهور | 14 | 59 | | 80.8 |
| هاوية – حاوية | | حاوية | هاوية | | 93.6 | |
| | | حاوية | 83 | 4 | | 95.4 |
| | | هاوية | 2 | 5 | | 71.4 |
| سراب – شراب | | شراب | سراب | | 92.3 | |
| | | شراب | 51 | 2 | | 96.2 |

| | | | | | | |
|---------------|------|-------|-------|-----|------|------|
| | سراب | 3 | 9 | 75 | | |
| مسير - مصير | | مصير | مسير | | 94.1 | |
| | | مصير | 188 | 7 | | 96.4 |
| | | مسير | 5 | 2 | | 28.6 |
| محنة - مهنة | | مهنة | محنة | | 98.4 | |
| | | مهنة | 307 | 3 | | 99 |
| | | محنة | 2 | 3 | | 60 |
| يسب - يصب | | يصب | يسب | | 98.8 | |
| | | يصب | 159 | 1 | | 99.4 |
| | | يسب | 1 | 0 | | 0 |
| ألم - علم | | علم | ألم | | 84.3 | |
| | | علم | 570 | 79 | | 87.8 |
| | | ألم | 82 | 292 | | 78.1 |
| السعر - الشعر | | السعر | الشعر | | 94 | |
| | | السعر | 1087 | 57 | | 95 |
| | | الشعر | 47 | 533 | | 91.9 |
| الأرض - العرض | | العرض | الأرض | | 85.5 | |
| | | العرض | 1189 | 184 | | 86.6 |
| | | الأرض | 140 | 718 | | 83.7 |
| السكر - الشكر | | السكر | الشكر | | 97.8 | |
| | | السكر | 729 | 9 | | 98.8 |
| | | الشكر | 23 | 727 | | 96.9 |
| ثلاثة - سلاسة | | ثلاثة | سلاسة | | 97.6 | |
| | | ثلاثة | 633 | 14 | | 97.8 |
| | | سلاسة | 2 | 11 | | 84.6 |

6.5.2 N-gram Language Models

Unlike the unsupervised method, the n-gram models in supervised method are built using only the sentences that contain the words from the confusion sets. The seventy percent training sentences for each of the twenty eight confusion sets are used to build the language models. Table 6.5 shows the statistics of the language models for the N-gram supervised method.

Table 6.5: Statistics of the language models for the training sentences in the supervised method

| No. of words | Uni-grams | Bi-grams | Tri-grams |
|---------------------|------------------|-----------------|------------------|
| 3,131,258 | 138,108 | 1,425,641 | 2,395,324 |

The experiments were run on the remaining thirty percent sentences of each of the confusion sets. The steps in testing are explained in detail in section 6.4. Table 6.6 shows the comparison between the baseline and the N-gram methods.

The results show that the n-gram language models scores an average of 95.9% accuracy compared with an average accuracy of 76.6% for the baseline method.

We ran other experiments using separate language models built for each confusion set training sentences; we refer to it as Separate LMs. The same procedure is applied to the test sentences as only the language models for that confusion set is used. The average accuracy obtained from the Separate LMs is 94.7%. Table 6.7 shows the results for the Separate LMs and compares them with the results obtained by the other techniques. This indicates that there is no advantage of using separate language models.

Table 6.6: Comparison between the baseline and the N-gram methods

| Confusion Set | Baseline Accuracy % | N-Gram Accuracy % | No. of test sentences |
|-------------------|---------------------|-------------------|-----------------------|
| كبير – كثير | 78.8 | 97.1 | 2770 |
| صغير – قصير | 54.1 | 93.4 | 532 |
| أساس – أثاث | 96.9 | 98.6 | 1100 |
| عمارة – إمارة | 81.2 | 97.6 | 505 |
| ثمن – سمن | 98.7 | 99.5 | 369 |
| العمارة – الإمارة | 74.3 | 90.5 | 284 |
| ثروة – ثورة | 57.4 | 80.4 | 230 |
| الأرق – الحرق | 79.2 | 87.1 | 149 |
| مضر – مصر | 74.9 | 92.5 | 562 |
| أشعار – أسعار | 96.1 | 99.3 | 2916 |
| القرعة – القرحة | 77.9 | 86.8 | 262 |
| الرحم – الرسم | 62.2 | 96.6 | 823 |
| حال – مال | 78.8 | 98.5 | 4193 |
| يغرق – يفرق | 87.3 | 87.0 | 69 |
| غسل – عسل | 70.5 | 94.7 | 281 |
| تسير – تصير | 95.4 | 95.7 | 351 |
| إصرار – إصرار | 66.8 | 85.2 | 250 |
| الزهور – الظهور | 75.2 | 89.0 | 327 |
| هاوية – حاوية | 90.4 | 94.7 | 94 |
| سراب – شراب | 83.1 | 89.2 | 65 |
| مسير – مصير | 95.6 | 97.0 | 202 |
| محنة – مهنة | 98.1 | 98.7 | 315 |
| يسب – يصب | 99.4 | 99.4 | 161 |
| ألم – علم | 63.7 | 92.1 | 1023 |
| السعر – الشعر | 65.8 | 92.3 | 1724 |
| الأرض – العرض | 59.6 | 92.3 | 2231 |
| السكر – الشكر | 50.5 | 97.6 | 1488 |
| ثلاثة – سلاسة | 96.2 | 99.6 | 660 |
| Average | 76.6 | 95.9 | |

Table 6.7: Comparison between the baseline, the separate LMs and the N-gram methods.

| Confusion Set | Baseline Accuracy % | N-Gram Accuracy % | Separate LMs Accuracy % | No. of test sentences |
|-------------------|---------------------|-------------------|-------------------------|-----------------------|
| كبير – كثير | 78.8 | 97.1 | 95.3 | 2770 |
| صغير – قصير | 54.1 | 93.4 | 92.8 | 532 |
| أساس – أثاث | 96.9 | 98.6 | 98.9 | 1100 |
| عمارة – إمارة | 81.2 | 97.6 | 96.4 | 505 |
| ثمن – سمن | 98.7 | 99.5 | 99.5 | 369 |
| العمارة – الإمارة | 74.3 | 90.5 | 91.6 | 284 |
| ثروة – ثورة | 57.4 | 80.4 | 83.5 | 230 |
| الأرق – الحرق | 79.2 | 87.1 | 88.4 | 149 |
| مضر – مصر | 74.9 | 92.5 | 91.1 | 562 |
| أشعار – أسعار | 96.1 | 99.3 | 99.3 | 2916 |
| القرعة – القرحة | 77.9 | 86.8 | 92.6 | 262 |
| الرحم – الرسم | 62.2 | 96.6 | 94.9 | 823 |
| حال – مال | 78.8 | 98.5 | 98.6 | 4193 |
| يغرق – يفرق | 87.3 | 87.0 | 84.1 | 69 |
| غسل – عسل | 70.5 | 94.7 | 93.6 | 281 |
| تسير – تصير | 95.4 | 95.7 | 96.2 | 351 |
| إصرار – إصرار | 66.8 | 85.2 | 88.8 | 250 |
| الزهور – الظهور | 75.2 | 89.0 | 91.1 | 327 |
| هاوية – حاوية | 90.4 | 94.7 | 90.4 | 94 |
| سراب – شراب | 83.1 | 89.2 | 84.6 | 65 |
| مسير – مصير | 95.6 | 97.0 | 96.5 | 202 |
| محنة – مهنة | 98.1 | 98.7 | 53.0 | 315 |
| يسبب – يصب | 99.4 | 99.4 | 99.4 | 161 |
| ألم – علم | 63.7 | 92.1 | 91.7 | 1023 |
| السعر – الشعر | 65.8 | 92.3 | 91.3 | 1724 |
| الأرض – العرض | 59.6 | 92.3 | 89.7 | 2231 |
| السكر – الشكر | 50.5 | 97.6 | 97.0 | 1488 |
| ثلاثة – سلاسة | 96.2 | 99.6 | 97.3 | 660 |
| Average | 76.6 | 95.9 | 94.7 | |

6.5.3 Combining Methods

We combined the word co-occurrence method and the N-gram language models method in our experiments. The combination method checks the decisions made by the two methods, if they agree on a decision, either correct or incorrect, this decision is considered as the decision of the combined method. If the two methods do not agree on a decision, the method uses the difference of the probabilities for each method to decide which decision to choose, the one with the highest difference probability will be considered as the taken decision. The difference probability for each method is shown in Equations 6.3.

$$Difference = \frac{Probability_{word1} - Probability_{word2}}{Probability_{word1}} \quad (6.3)$$

where: $Probability_{word1} \geq Probability_{word2}$

The results of the combining methods with comparison the other methods results are shown in Table 6.7. In some confusion sets, the combined method scored a better accuracy rate than the other method. However, with average accuracy rate of 95.9%, the N-gram language method scored the best results among all methods presented in this chapter. Table 6.8

Table 6.7: Comparing the baseline, context-words, N-gram, and combined methods.

| Confusion Set | Baseline Accuracy % | Word Co-occurrence Accuracy % | N-Gram Accuracy % | Combined Accuracy % |
|-------------------|---------------------|-------------------------------|-------------------|---------------------|
| كبير – كثير | 78.8 | 80.7 | 97.1 | 92.2 |
| صغير – قصير | 54.1 | 88.7 | 92.5 | 91.3 |
| أساس – أثاث | 96.9 | 97.6 | 98.6 | 98.5 |
| عمارة – إمارة | 81.2 | 95.1 | 96.8 | 97.3 |
| ثمن – سمن | 98.7 | 98.4 | 99.5 | 99.7 |
| العمارة – الإمارة | 74.3 | 87.3 | 90.5 | 91.6 |
| ثروة – ثورة | 57.4 | 81.3 | 70.4 | 87.0 |
| الأرق – الحرق | 79.2 | 85.9 | 77.6 | 87.1 |
| مضر – مصر | 74.9 | 87.0 | 82.0 | 92.2 |
| أشعار – أسعار | 96.1 | 98.1 | 99.3 | 99.1 |
| القرعة – القرحة | 77.9 | 93.5 | 86.8 | 94.2 |
| الرحم – الرسم | 62.2 | 96.4 | 89.2 | 97.8 |
| حال – مال | 78.8 | 92.2 | 97.0 | 97.1 |
| يغرق – يفرق | 87.3 | 89.9 | 81.2 | 84.1 |
| غسل – غسل | 70.5 | 94.0 | 84.3 | 95.7 |
| تسير – تصير | 95.4 | 96.5 | 95.7 | 96.2 |
| إصرار – إصرار | 66.8 | 85.6 | 85.2 | 86.0 |
| الزهور – الظهور | 75.2 | 89.0 | 87.8 | 89.0 |
| هاوية – حاوية | 90.4 | 93.6 | 91.5 | 94.7 |
| سراب – شراب | 83.1 | 92.3 | 89.2 | 89.3 |
| مسير – مصير | 95.6 | 94.1 | 96.5 | 97.0 |
| محنة – مهنة | 98.1 | 98.4 | 98.7 | 99.1 |
| يسبب – يصب | 99.4 | 98.8 | 99.4 | 99.4 |
| ألم – علم | 63.7 | 84.3 | 90.2 | 91.5 |
| السعر – الشعر | 65.8 | 94.0 | 90.8 | 95.2 |
| الأرض – العرض | 59.6 | 85.5 | 88.8 | 92.3 |
| السكر – الشكر | 50.5 | 97.8 | 97.6 | 98.5 |
| ثلاثة – سلاسة | 96.2 | 97.6 | 99.6 | 97.4 |
| Average | 76.6 | 91.5 | 95.9 | 95.4 |

We reduced the error rate in the combination method by rejecting the unmatched decisions made by the two techniques. In other words, if the two methods agree on a decision, this decision is considered as the decision of the combined method, otherwise we reject the decision. Previously the error rate was 4.6% without using the rejection scheme. Although the accuracy rate reduced, however, after applying the rejection scheme the error rate dropped to 1.8% , that is about 61% of the combined method errors has been reduced using the combination with rejection. Table 6.8 shows the reduction of error rate using the rejection scheme.

Table 6.8: Reducing error rate in the combination method using rejection.

| Confusion Set | Without rejection | | Using rejection | | |
|-------------------|-------------------|--------------|-----------------|--------------|-------------|
| | Accuracy % | Error rate % | Accuracy % | Error rate % | Rejection% |
| كبير – كثير | 92.2 | 7.8 | 78.0 | 2.0 | 20 |
| صغير – قصير | 91.3 | 8.7 | 83.1 | 3.8 | 13.2 |
| أساس – أثاث | 98.5 | 1.5 | 97.2 | 0.9 | 1.9 |
| عمارة – إمارة | 97.3 | 2.7 | 93.7 | 1.0 | 5.3 |
| ثمن – سمن | 99.7 | 0.3 | 98.1 | 0.3 | 1.6 |
| العمارة – الإمارة | 91.6 | 8.4 | 56.0 | 3.2 | 40.8 |
| ثروة – ثورة | 87.0 | 13.0 | 71.3 | 9.1 | 19.6 |
| الأرق – الحرق | 87.1 | 12.9 | 82.3 | 8.8 | 8.8 |
| مضر – مصر | 92.2 | 7.8 | 83.3 | 3.6 | 13.2 |
| أشعار – أسعار | 99.1 | 0.9 | 97.8 | 0.5 | 1.8 |
| القرعة – القرحة | 94.2 | 5.8 | 89.5 | 2.0 | 8.6 |
| الرحم – الرسم | 97.8 | 2.2 | 94.4 | 1.5 | 4.1 |
| حال – مال | 97.1 | 2.9 | 91.7 | 1.0 | 7.3 |
| يغرق – يفرق | 84.1 | 15.9 | 81.2 | 4.3 | 14.5 |
| غسل – عسل | 95.7 | 4.3 | 89.7 | 1.0 | 9.3 |
| تسير – تصير | 96.2 | 3.8 | 95.2 | 3.0 | 1.9 |
| إصرار – إصرار | 86.0 | 14.0 | 56.0 | 6.0 | 38.0 |
| الزهور – الظهور | 89.0 | 11.0 | 82.6 | 4.6 | 12.8 |
| هاوية – حاوية | 94.7 | 5.3 | 92.6 | 4.3 | 3.2 |
| سراب – شراب | 89.3 | 10.7 | 80.0 | 1.5 | 18.5 |
| مسير – مصير | 97.0 | 3.0 | 94.0 | 3.0 | 3.0 |
| محنة – مهنة | 99.1 | 0.9 | 97.8 | 0.6 | 1.6 |
| يسب – يصب | 99.4 | 0.6 | 99.0 | 0.5 | 0.5 |
| ألم – علم | 91.5 | 8.5 | 79.7 | 3.3 | 17.0 |
| الشعر – الشعر | 95.2 | 4.8 | 88.6 | 2.3 | 9.1 |
| الأرض – العرض | 92.3 | 7.7 | 80.9 | 3.1 | 16.0 |
| السكر – الشكر | 98.5 | 1.5 | 95.2 | 0.4 | 4.4 |
| ثلاثة – سلاسة | 97.4 | 2.6 | 70 | 0.5 | 29.5 |
| Average | 95.4 | 4.6 | 87.8 | 1.8 | 10.4 |

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

This chapter summarizes our major contributions in this thesis. The goal of this research is to design and implement a prototype for automatic context sensitive spell checking and correction of Arabic text. This chapter also discusses the limitations of our work with possible enhancements and future research directions.

7.1 Conclusions

Spell checkers are important tools for document preparation, word processing, searching and document retrieval. Spelling errors that result in a real word in the dictionary cannot be detected by conventional spell checkers. In this thesis, we designed and implemented different techniques for spell checking and correction that are able to detect and correct real-word errors in Arabic text automatically.

A corpus of Arabic text was collected from different resources. This corpus was used to build different language models for spell checking and correction. Different dictionaries with different sizes were also built using the collected corpus. In addition, we collected confusion sets from different resource, like OCR misrecognized words and others from the most common mistakes committed by non-native Arabic speakers.

We implemented an unsupervised model for real-word error detection and correction for Arabic text in which N-gram language models are used. The unsupervised model uses the probabilities from the language models to detect as well as to correct real-word errors. Language models can handle different types of errors and not only restricted to specific

types (semantic or syntactic) of errors or predefined sets of errors (confusion sets). However, the technique requires large memory size to store the language models.

Different supervised models were also implemented that use confusion sets to detect and correct real-word errors. A window-based technique was used to estimate the probabilities of the context words of the confusion sets. N-gram language models were also used to detect real-word errors by examining the sequences of n words. The same language model was also used to choose the best correction for the detected errors.

The experimental results of the techniques show promising correction accuracy, especially the supervised methods. However, it was not possible to compare our results with other published works as there is no benchmarking dataset for real-word errors correction for Arabic text.

The unsupervised approach discussed in chapter 5 is a general method that can detect and correct different kinds of errors and not only restricted to a set of predefined errors. The method was able to detect 60.5% to 76.2% of all real-word errors, with different threshold values, using the n-gram probability information. 56% to 70.7% of the detected real-word errors have been correctly amended. Unfortunately, precision rate is too much low. Only 7.8% to 12.3 of total detected errors are rightly detected, the others were false positives. Published work suffers from the same problem of low precision rates.

Supervised learning methods using confusion sets scored better accuracy than the unsupervised method. These methods have several advantages. They can handle errors caused by common confused words in an easy way, simply by considering the predefined confusion sets. They are also not restricted to spelling variations with only one character

difference. In addition, they require less memory space. The downside with supervised methods, however, is that all the confusion sets must be defined in advance. These methods can only detect specific errors that are predefined ahead of time in a form of confusion sets. Thus the word will be checked for being an error or not only if it is a member of any of the predefined confusion sets.

We therefore see supervised methods based on confusion sets as complementary to the unsupervised method based on language models and vice versa. Each method is appropriate for a particular type of error. We believe that a spell checker that uses both methods will be a comprehensive spell checker and will be capable of detecting and correcting errors efficiently.

7.2 Future Directions

The methods used in this thesis have some limitations that need to be enhanced and problems that need to be resolved. In this section, we are suggesting some solutions that could be used to improve the methods' performance.

- To resolve the problem of false positives, the system may be made interactive; the detected errors are flagged as suspicious words with their possible corrections. Then it is the user's job to recognize whether the original word or one of its candidate suggestions is what was intended.
- A good lemmatizer could be an enhancement that may resolve the problem of false positives as most of them has the same lemma. We think if a good lemmatizer is involved, it would be able to check whether the suspicious word has the same lemma

with the suggested candidate correction. If so, we can easily ignore these suspicious words, however this might still raise some false negatives.

- The use of Part-Of-Speech (POS) n-grams and possibly mixing them with the words n-grams is expected to reduce the false positives and may solve data sparseness problem. However, POS n-grams may not solve the problem of proper nouns false positives as nouns have the same POS tag.
- The language models require large memory size, which limits the practicality of the system especially when the language models are very large. This problem could be resolved by making the system online. The system might be installed in a dedicated server so that the user can access the spell checker via an online service. The language models, dictionary and all other data would be stored in the server. We think this solution may solve the memory space problem.
- We think a spell checker will perform better when it is targeted to a specific subject. Language models can be built using a corpus with specific topics. For example, when building a medical context sensitive spell checker, the vocabulary would be smaller and therefore the n-gram language models would be smaller as well. Only words that are likely to be used by doctors could then be suggested.
- Supervised methods can be improved by increasing the number of confusion sets that cover most of the errors. Clearly, not all errors are covered by the stated confusion sets in this thesis. We created confusion sets that are driven from the dictionary that groups words with a minimum edit distance of one. The number of sets was huge and the sets have large number of words (sometimes it exceeds 30 words in a set). As a future work, we plan to create more realistic confusion sets that contain words driven

from the dictionary with minimum edit distance of one given that the different characters are neighbors in the keyboard. This reduces the number of words in each confusion set. The sets will also be more realistic as a user may type one word for another, in the same set, because of keyboard slips.

The methods of detecting and correcting real-word spelling errors that we have presented in this thesis need to be integrated with a conventional spell checker for non-word errors. Integrating the presented methods with a conventional spell checker with a suitable user interface for a word processor results in a spell checker that can be tested on more realistic data.

References

- Al-Sulaiti, L. 2004. "Designing and Developing a Corpus of Contemporary Arabic." *Master's thesis, School of Computing in the University of Leeds.*
- Alkanhal, Mohamed I., Mohamed A. Al-Badrashiny, Mansour M. Alghamdi, and Abdulaziz O. Al-Qabbany. 2012. "Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions." *IEEE Transactions on Audio, Speech, and Language Processing* 20(7): 2111–2122.
- Bergsma, Shane, Dekang Lin, and Randy Goebel. 2008. "Web-Scale N-gram Models for Lexical Disambiguation." In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, , p. 1507–1512.
- Brill, Eric, and Robert C. Moore. 2000. "An improved error model for noisy channel spelling correction." *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*: 286–293.
- Church, Kenneth W, and William A Gale. 1991. "Probability scoring for spelling correction." *Statistics and Computing* 1(2): 93–103.
- Damerau, Fred J. 1964. "A technique for computer detection and correction of spelling errors." *Communications of the ACM* 7(3): 171–176.
- Flexner, B. 1983. "Random House unabridged dictionary." (2nd ed.). *New York: Random House.*
- Fossati, Davide, and Barbara Di Eugenio. 2007. "A Mixed Trigrams Approach for Context Sensitive Spell Checking" ed. Alexander Gelbukh. *Computational Linguistics and Intelligent Text Processing* 4394: 623–633.
- Golding, A. 1995. "A Bayesian Hybrid Method for Context-Sensitive Spelling Correction." *Proceedings of the 3rd Workshop on Very Large Corpora, Boston, MA*: 39–53.
- Golding, A, and Y Schabes. 1996. "Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction." *Proceedings of the 34th annual meeting on Association for Computational Linguistics*: 71–78.
- Golding, Andrew R, and Dan Roth. 1996. "Applying Winnow to Context-Sensitive Spelling Correction." *Machine Learning, arXiv preprint cmp-lg/9607024*: 182–190.
- Haddad, B, and M Yaseen. 2007. "Detection and Correction of Non-Words in Arabic: A Hybrid Approach." *International Journal of Computer Processing of Oriental Languages (IJCPOL)* 20(4): 237–257.

- Hassan, A, H Hassan, and S Noeman. 2008. "Language Independent Text Correction using Finite State Automata." *Proceedings of the 2008 International Joint Conference on Natural Language Processing (IJCNLP)*.
- Hirst, G., and A. Budanitski. 2001. "Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion." *Department of Computer Science, Toronto, Ontario, Canada*.
- Hirst, Graeme, and Alexander Budanitsky. 2005. "Correcting real-word spelling errors by restoring lexical cohesion." *Natural Language Engineering* 11(1): 87–111.
- Islam, A., and D. Inkpen. 2011. "An Unsupervised Approach to Detecting and Correcting Errors in Short Texts." *Computational Linguistics* 1.1: 1–38.
- Islam, Aminul, and Diana Inkpen. 2009. "Real-word spelling correction using Google Web IT 3-grams." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3 - EMNLP '09* 3(August): 1241.
- Jiang, Jay J, and David W. Conrath. 1997. "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy." *Proceedings of International Conference Research on Computational Linguistics (Rocling X)*.
- Kernighan, Mark D, Kenneth W Church, and William A Gale. 1990. "A spelling correction program based on a noisy channel model." In *Proceedings of the 13th conference on Computational linguistics*, Association for Computational Linguistics, p. 205–210.
- Kukich, Karen. 1992. "Technique for automatically correcting words in text." *ACM Computing Surveys* 24(4): 377–439.
- L. Cherry, and N. Macdonalil. 1983. "The Writer's Workbench software Byte." 241–248.
- Lehal, Gurpreet Singh. 2007. "Design and Implementation of Punjabi Spell Checker." *International Journal of Systemics, Cybernetics and Informatics*: 70–75.
- Levenshtein, V. 1966. "Binary Codes Capable of Correcting Deletions and Insertions and Reversals." *Soviet Physics Doklady* 10(8): 707–710.
- Liang, Hsuan Lorraine. 2008. Science "Spell checkers and correctors: A unified treatment." Doctoral dissertation, University of Pretoria, South Africa.
- Mahdi, Adnan. 2012. "Spell Checking and Correction for Arabic Text Recognition." *Master's thesis, KFUPM University, Department of Information & Computer Science*.
- Mays, Eric, Fred J. Damerau, and Robert L. Mercer. 1991. "Context based spelling correction." *Information Processing & Management* 27(5): 517–522.

- Mitton, R. 1987. "Spelling checkers, spelling correctors and the misspellings of poor spellers." *Information Processing* 23(5): 495–505.
- Ben Othmane Z C Ben Fraj F, Ben Ahmed M. 2005. "A Multi-Agent System for Detecting and Correcting 'Hidden' Spelling Errors in Arabic Texts." In *Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science NLUCS*, ed. Bernadette Sharp. INSTICC Press, p. 149–154.
- Ben Othmane Zribi, C., and M. Ben Ahmed. 2012. "Detection of semantic errors in Arabic texts." *Artificial Intelligence* 1: 1–16.
- Ben Othmane Zribi, C., Hanene Mejri, and M. Ben Ahmed. 2010. "Combining Methods for Detecting and Correcting Semantic Hidden Errors in Arabic Texts." *Computational Linguistics and Intelligent Text Processing*: 634–645.
- Ben Othmane Zribi, Chiraz, and Mohamed Ben Ahmed. 2003. "Efficient automatic correction of misspelled Arabic words based on contextual information, Lecture Notes in Computer Science." *Springer* 2773: 770–777.
- Pedler, Jennifer. 2007. "Computer Correction of Real-word Spelling Errors in Dyslexic Text." *PhD Thesis University of Birkbeck, London*.
- Shalan, K, R Aref, and A Fahmy. 2010. "An Approach for Analyzing and Correcting Spelling Errors for Non-native Arabic learners." In *the Proceedings of The 7th International Conference on Informatics and Systems, INFOS2010*, Cairo, p. 53–59.
- St-Onge, D. 1995. "Detecting and correcting malapropisms with lexical chains." *Master's thesis, University of Toronto, Computer Science Department*: Also published as Technical Report CSRI–319.
- Stolcke, A. 2002. "SRILM — An Extensible Language Modeling Toolkit." *Intl. Conf. on Spoken Language Processing*.
- Thorsten, Brants, and Franz Alex. 2006. *Web IT 5-gram corpus version 1.1*.
- . 2009. *Web IT 5-gram, 10 European languages version 1*. Philadelphia.
- Verberne, Suzan. 2002. "Context-sensitive spell checking based on word trigram probabilities Context-sensitive spell checking based on word trigram probabilities." *Master's thesis, University of Nijmegen, February-August*.
- Wagner, Robert A, and Michael J Fischer. 1974. "The String-to-String Correction Problem." *Journal of the ACM* 21(1): 168–173.
- Wilcox-O'Hearn, L Amber, G Hirst, and A Budanitsky. 2008. "Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer

- model” ed. A Gelbukh. *Computational Linguistics and Intelligent Text Processing* 4919(2000): 605–616.
- Wing, A M, and A D Baddeley. 1980. “Spelling Errors in Handwriting: A Corpus and Distributional Analysis.” In *Cognitive Processes in Spelling*, ed. Uta Frith. Academic Press, p. 251–283.
- Young, Charlene W, Caroline M Eastman, and Robert L Oakman. 1991. “An analysis of ill-formed input in natural language queries to document retrieval systems.” *Information Processing & Management* 27(6): 615–622.
- Zamora, E. 1981. “The use of trigram analysis for spelling error detection.” *Information Processing & Management* 17(6): 305–316.

Vitae

Name : Majed Mohammed Abdulqader Al-Jefri

Nationality : Yemeni

Date of Birth :8/12/1983

Email : majfbi@gmail.com

Address : Aden, Yemen

Academic Background : Received Bachelor of Science (B.Sc) in Computer
Science from Al-Ahgaff University in 2006 with GPA 4.49 out of 5