# Automatic Spelling Correction based on n-Gram Model

S. M. El Atawy
Dept. of Computer Science
Faculty of Specific Education
Damietta University, Egypt

A. Abd ElGhany
Dept. of Computer
Damietta University
Egypt

## ABSTRACT
A spell checker is a basic requirement for any language to be digitized. It is a software that detects and corrects errors in a particular language. This paper proposes a model to spell error detection and auto-correction that is based on n-gram technique and it is applied in error detection and correction in English as a global language. The proposed model provides correction suggestions by selecting the most suitable suggestions from a list of corrective suggestions based on lexical resources and n-gram statistics. It depends on a lexicon of Microsoft words. The evaluation of the proposed model uses English standard datasets of misspelled words. Error detection, automatic error correction, and replacement are the main features of the proposed model. The results of the experiment reached approximately 93% of accuracy and acted similarly to Microsoft Word as well as outperformed both of Aspell and Google.

## General Terms
Spelling correction, N-gram.

## Keywords
N-gram - Spelling correction - Misspelling detection - Spell checker - Information retrieval.

## 1. INTRODUCTION
A very important reason for considering English as a global language is that the world's knowledge is usually preserved in English [1]. Recently and with the spread of global English as an essential tool for trade, worldwide exchange and communication, more interest has been focused on the problems and needs of Arab learners studying English [2]. Arab learners of' English encounter problems in both speaking and writing [3][4][5]. A number of studies discuss that many Arabic students face difficulties in learning [6][7].

English at various levels and with different skills (e.g. McCardle and Hoff, [8], Abdul Haq [9], Hoffman [10]). Students need reading and writing skills during their studies and graduation, which helps them to be competitive in the working world. Therefore, spelling instruction must include compensatory skills that make the Arab student write in correct English writing. Spell checkers which will enable them to compensate for spelling weaknesses. Spellcheckers are able to provide the target word for misspellings due to keyboarding and spelling rule application errors [11]. The problem with the development of technologies and algorithms the correct words become automatically in the digitized texts a continual research challenge. There are good reasons for the continuing efforts in this field to develop of applications possible [12]. Spell checking was dating back to the 1960s especially in the works of Damerau (1964) [13]. Spellchecking is the task of predicting which words in a document are misspelled [14]. Spell checking are very important for a number of computer application such as text processors, web browsers, search engines, and others [15]. There are two types of spell checker, error detection and error correction. In this paper, we are designed, implemented and evaluated an end-to-end system that performs spellchecking and auto correction.

This paper is organized as follows: Section 2 illustrates types of spelling error and some of related works. Section 3 explains the system description. Section 4 presents the results and evaluation. Finally, Section 5 includes conclusion and future work.

## 2. RELATED WORKS
Spell checking techniques have been substantially, such as error detection & correction. Two commonly approaches for error detection are dictionary lookup and n-gram analysis. Most spell checkers methods described in the literature, use dictionaries as a list of correct spellings that help algorithms to find target words [16]. Many different solutions have been proposed such as Gökhan Dalkılıç and Yalçın Çebi [17] suggested method based on n-gram analysis to find incorrectly spelled words in a mass of text. The first step to use n-gram is to determine the language specific n-gram using a corpus. But a corpus cannot be big enough to find all the possible word n-gram. Back-off smoothing method is one of the methods to estimate the frequency of the unknown n-gram in a corpus. If a non-existent n-gram is found the word is determined as a misspelling. A dictionary is a lexical source that contains list of correct words a particular language. dictionary-based methods (de Amorim, 2009), still have a performance limitation because of their intrinsic architecture, one common alternative to this performance limitation is the use of dictionaries organized as Finite State Automata (FSA).

FSA are especially interesting for morphologically rich languages such as Hungarian, Finnish, and Turkish. One example of a study for spell checking that organized the dictionaries as FSA is [18] Hulden (2009) presented algorithm for finding approximate matches of a string in a finite-state automaton, given some metric of similarity such as minimum edit distance. The algorithm can use a variety of metrics for determining the distance between two words; and points out that finding the closest match between word and a large list of words, is an extremely demanding task. V. Ramaswamy and H. A. Girijamma [19] presented a way to convert finite automaton to fuzzy automaton as fuzzy automaton is better than finite automaton for strings comparison when individual levels of similarity for particular pairs of symbols or sequences of symbols are defined. A finite automaton is useful in defining whether a given string is accepted or not whereas fuzzy automaton determines the extent to which the string is accepted. The method presented serves as an alternative to the FSA-based dictionaries that reduce the number of distances that have to be calculated for each misspelling and therefore improving processing speed.

According to the related works, there are two types of spelling errors: cognitive errors and typographic errors [20].

1) Typographic errors: A study by Damerau [21] shows that 80% of the typographic errors fall into one of the following four categories:

   a. Single letter inserting; e.g. typing computer for ccomputer.

   b. Single letter deleting, e.g. typing computer for cmputer.

   c. Single letter substituting, e.g. typing computer for compoter.

   d. Transposition of two adjacent letters, e.g. typing computer for cumpoter.

2) Cognitive errors: these errors occur when the correct spellings of the word are not known. In this type, the pronunciation of misspelled word is the same or similar to the pronunciation of the intended correct word. e.g. "peace" for piece [22].

# 3. THE METHODS

## 3.1 The description of the method

There are many algorithmic techniques for detecting and correcting spelling errors in text [23]. Error Correction Approaches like Neural Based [24], Levenshtein Edit Distance [25], Similarity Keys [22], Rule-Based [26], Probabilistic [22], and N-gram [27]. Shannon discussed the idea of using n-gram in language processing [28]. After this first work, the idea of using n-gram has been applied to many problems such as speech recognition, translated word, correction word, prediction and spelling correction. This technique, purely statistical, does not require any knowledge of the document language. Another advantage of the n-gram is the automatic capture of the most frequent roots [28]. N-gram can be used in two ways, either without a dictionary or together with a dictionary [20]. N-gram is used without a dictionary, this way employs to find in which position in the incorrect word the error occurs. If there is a special way to change the incorrect word so that it contains only correct n-gram, there is as correction. The performance of this way is low; but it is a simple way and does not require a dictionary [30]. In the other way, together with a dictionary, n-gram is used to define the distance between words, but the words are always checked versus the dictionary. This needs many ways, e.g. analysis how many n-gram the misspelled word and a dictionary word have common, weighted by the length of words [31].

The proposed method based on the n-gram model. It can detect the correction suggestions by giving weights to a list of scope correction candidates, based on n-gram statistics and lexical resources.

Lexical resources present linguistic information about words of natural languages. This information can be presented in data structures, from plain lists to complicated with many types of linguistic information and relations associated with the entries stored in the resource [23] [32]. Lexical resources are used for language and knowledge engineering. It plays a role in preparing, processing and managing the information needed by computers and humans [33] [34] [35]. We propose to use dictionaries of Microsoft Word program due to it is main dictionaries contains the most common words. It covers verbs, nouns, adverbs and adjectives, but may not include certain names, technical terms, abbreviations, or specialized capitalization. In our purposed method, we use the words provided from this resource to correct the misspelled word. Thus, it used to extract all words contained in it with all its linguistic relationships. Then the proposed method automatically replaces the selected suggestion in the input text. N-gram probability is applied to detect suggestion of the error words. The proposed method is programmed and evaluated on "Matlab" program.

## 3.2 Compute the similarity

There are several approaches based on similarity key, minimum edit distance, neural networks, and n-gram. The probability is proposed to perform the task of error correction. N-gram is used to comparing string letters [36] [37]. It is language independent, this technique only compares the letters of words regardless of the language used, it computes the similarity between two strings by counting the number of similar n-gram they share. Based on the similarity coefficient is compute the more similar n-gram between two strings exist the more similar they are [29]. In general, approach the similarity coefficient δ is performed by Equation (1).

where s1 and s2 are the n-gram sets for two words s1 and s2 which they compared. | s1 ∩ s2 | indicates the number of similar n-gram in s1and s2, and | s1 ∪ s2 | indicates the number of unique n grams in the union of s1 and s2.

The similarity coefficient for the misspelled word "camputer" and the correct word "computer" using an n-gram with n = 2 (bi-gram) shown in Table 1 (as an example), as well as, Figure 1 Illustrates the implementation of n-gram.

**Table 1: An example of Calculating the bigrams similarity coefficient between two words**

| **bi-grams** | **Computer** | **Computer** |
|---|---|---|
| Co | 1 | - |
| Om | 1 | - |
| Mp | 1 | 1 |
| Pu | 1 | 1 |
| Ut | 1 | 1 |
| Te | 1 | 1 |
| Er | 1 | 1 |
| Ca | - | 1 |
| Am | - | 1 |
| Similarity coefficient | **5/9 = 0.55** | |

The user selects the wrong word from error list and the proposed method performs the n-gram for this word by comparing it with each word in the dictionary and gives the words in the suggestion list with similarity coefficient (δ)=1. Then system again selects the word from suggestion list and replaces it in input text. The system

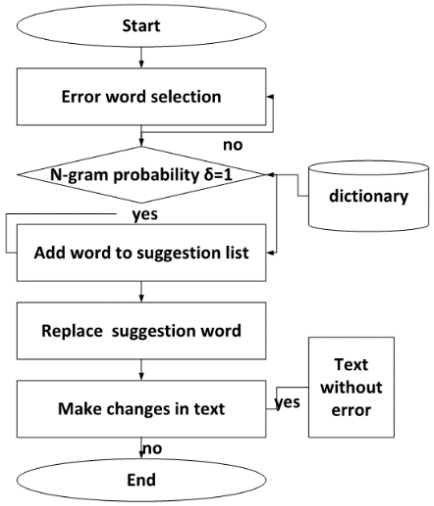


**Figure 1. The sequence of n-gram method in the proposed method.**

## 3.3 The graphical user interface (GUI) of the proposed method

The GUI of the proposed method (spellchecker) appears whenever the application is stared which contains the text area, the user can enter the input text to spell checking. There are some buttons on the main window such as Spell Check. It can be clicked when user wants to check the text entered in the text area for spell check of errors. When user clicks on this button, the system begins process the word then appear the list of word suggestion in text area, the system shows the correct word in anther text area and display the complete correction of entered text in anther text area. There are second button labeled "Delete all" to delete the entered text to perform a new process.

## 4. RESULTS AND EVALUATION

Some experiments are done to evaluate the automatic spelling correction as following, we care about evaluating the quality of the proposed suggestion. To achieve this, the evaluation was done on the whole English commonly misspelled word list provided in [38]. This list, contains about 350 words, and we used a list of 3900 words in English common misspellings. This list of common misspellings is appeared at a table consisting of two columns. The first one shows the correct spelling of the word, and the second the misspelled word. We divided evaluate into two phases. The first evaluation was done on the whole English commonly misspelled word list. In this evaluation, only considered the correction words which were rated as best correction word, if the second word would have been the correct suggestion, this was count as error correction. First used all misspelled words of the list, using the bi-gram case and just correct the first suggestion. The proposed method corrected 354 misspelled words (98%) and failed for 6 misspelled words (2%). Afterwards, we tested the system again by creating intentional errors, these errors were divided into eight categories as flows:

- Single letter inserting to the word
- Single letter deleting from the word
- Single letter substituting in the word
- Transposition of two adjacent letters
- Differ in one character from the original word.
- Differ in one letter removed or added, plus one letter different.
- Differ in repeated characters removed or add and 1 character different
- Differ in having two consecutive letters exchanged and one character different.

A test set of 2800 spelling errors has been created, the spellchecker is corrected 2380 spelling error words (85%) and failed in 420 words with misspellings (15%). When used more than standard such as insert or deleted or substituting or transposition more than two letter, this showed low efficiency and performance. Spellchecker succeed on 2100 words (75%) and failed in 700 words (25%) as shown in Table 2.

**Table 2: The Results of comparison between bi-gram and tri-gram in data set (2800 words)**

| | **Bigram** | **more than two letter(trigram)** |
|---|---|---|
| Correct | 2380 (85%) | 2100 (75%) |
| Wrong | 420(15%) | 700 (25%) |

In another evaluation stage, we randomly selected a set of only 150 misspelled words obtained from Wikipedia [38] and not the whole list. All error types and starting letters of the words were taken into account. We compared the Spellchecker with Microsoft Word, Google, and Aspell. Microsoft Word provides a list of suggestion corrections we took just the first suggestion from the list assuming that the first suggestion is the most likely one proposed by the algorithm. Google provided only one correction suggestion. Whereas, Aspell provides a list of suggestion corrections we took just the first suggestion. Table 3 show that the Spellchecker finds the correct spelling for 135 words (90%). In comparison, Google can correct 125 words (83%) words, Aspell can correct 116 words (77%) while Microsoft Word can correct 135 words (90%) as shown in Table 3.

**Table 3: The results of comparison between the proposed method, Google, Microsoft Word, and Aspell**

| | **Spell checker** | **Google** | **Microsoft Word** | **Aspell** |
|---|---|---|---|---|
| Correct | 135 (90%) | 125 (83%) | 135 (90%) | 116 (77%) |
| Wrong | 15 (10%) | 25(17%) | 15(10%) | 34(23%) |

The results of the Spellchecker were similar to the Microsoft Word 2013 results, but Microsoft Word provides to the user a list of suggestion which the user can choice from it, while the Spellchecker has the ability to automatic correction of words without user choice from a list of suggestion. This is an advantage in the Spellchecker, saving user's time at choose from a list of suggestion.

**Table 4. A sample of the results of the error corrections between all methods.**

| **Spelling Error** | **Correct Spelling** | **Our Spellchecker** | **Google** | **Microsoft word** | **Aspell** |
|---|---|---|---|---|---|
| Accetable | Acceptable | acceptable | acceptable | acceptable | acceptable |
| Ahainst | Against | Against | Against | Against | gainst |
| anthr | Anther | anther | anthropologie | Anther | anthr- |

| | | | | | |
|---|---|---|---|---|---|
| alleuance | Allegiance | Alliance | Alliance | Alliance | Allegiance |
| Awfull | Awful | Awful | Awfully | Awful | Awful |
| bncsuse | Because | because | - | Because | bonasus |
| Bevoming | Becoming | becoming | Becoming | becoming | Becoming |
| Bginning | Beginning | beginning | Beginning | beginning | Beginning |
| Behieve | Believe | Beehive | Beehive | Beehive | beehive |
| bellweather | Bellwether | bellwether | bellwether | bellwether | Bellwether |
| Bouy | Buoy | Buoy | Bouygues | Buoy | bout |
| bussness | Business | business | Business | business | butchness |
| commited | Committed | committed | committed | committed | committed |
| Concede | Concede | concede | conceding | concede | conched |
| Collt | Coolly | Colt | Colt | Colt | coll |
| Dacquiri | Daiquiri | daiquiri | Daiquiri | Daiquiri | daiquiri |
| Definite | Definite | definite | Definitely | Definite | definite |
| Experidnce | Experience | experience | experience | experience | Experience |
| etrame | Extreme | extreme | Etrade | Extreme | ihram |
| fasonating | Fascinating | assonating | fascinating | assonating | fascinating |
| Firey | Fiery | Fiery | Fiery | Fiery | fire |
| Foreig | Foreign | foreign | Foreign | Foreign | foreign |
| Guaratee | Guarantee | guarantee | guarantee | guarantee | orante |
| Guidace | Guidance | guidance | Guidance | guidance | guidance |
| Haras | Harass | harass | harassment | Harass | harass |
| Heighth | Height | height | Height | Height | height |
| Inoculae | Inoculate | inoculate | inoculated | inoculate | inoculate |
| Intelligene | Intelligence | intelligence | intelligence | intelligence | Intelligence |
| Jewellery | Jewelry | jewelry | Jewellery | Jewelry | jewellery |
| Judgement | Judgment | judgment | Judgemental | judgment | judgment |
| Kerel | Kernel | kernel | Kernel | Kernel | carnal |
| lisence | License | silence | License | Silence | licence |
| Licence | License | license | Licence | License | license |
| Liggtenin | Lightnin | lightning | Liggteni | lightning | lightning |
| g | g | | ng | | |
| Loese | Lose | Lose | Lose | Lose | leese |
| Medeval | Medieval | medieval | Medieval | medieval | medieval |
| Momento | Memento | memento | Momento | memento | momento |
| Necessary | Necessary | necessary | necessary | necessary | necessary |
| Neice | Niece | Niece | Niece | Niece | nice |
| Niehbor | Neighbor | neighbor | Neighbor | neighbor | neighbor |
| Tjranny | Tyranny | tyranny | -------- | Tyranny | tyranny |

## 5. CONCLUSIONS

This paper proposed a language-independent spellchecker that is based on n-gram techniques. It is used in detecting and correcting spell errors. The main features of the proposed model can be summarized in giving the suggestions for detected errors and providing the correction automatically using the first suggestion. Furthermore, the proposed model is evaluated using English standard data sets of misspelled words. The results of the proposed spellchecker were similar to the results of Microsoft Word, while it outperforms the two industrial applications of Aspell and Google in first order ranking of suggestion. In future, we intend to improve the accuracy of the proposed model and apply it in Arabic text.

## 6. REFERENCES

[1] Intakhab Alam Khan: Learning difficulties in English: Diagnosis and pedagogy in Saudi Arabia, Educational Research (ISSN: 2141-5161) Vol. 2(7) pp. 1248-1257 July 2011.

[2] Baheej, Kassem: Difficulties that Arab Students Face in Learning English and the Importance of the Writing Skill Acquisition, PHD, Moldova State University, Sep. 2014.

[3] Rababah, Ghaleb: Communication Problems Facing Arab Learners of English, ERIC Processing and Reference Facility,2002.

[4] Refaat, M. M., Ewees, A. A., Eisa, M. M., & Ab Sallam, A. Automated assessment of students Arabic free-text answers. Int. J. Cooperative Inform Syst., 12, 2012. 213-222.

[5] Ewees, A. A., Eisa, M., & Refaat, M. M. Comparison of cosine similarity and k-NN for automated essays scoring. cognitive processing, 3(12). 2014.

[6] Arafa, M. N., Elbarougy, R., Ewees, A. A., & Behery, G. M. A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation. International Journal of Image, Graphics & Signal Processing, 10(4). 2018.

[7] Bialy, Asmaa Awad, A A. Ewees, & A F ElGamal. A Proposed Method for Summarizing Arabic Single Document. International Journal of Computer Applications 180(34). 2018. 9-14.

[8] McCardle, P and E. Hoff. Childhood bilingualism: research on infancy through school age. Clevedon: Multilingual Matters, 2006.

[9] Abdul Haq, F. An Analysis of Syntactic Errors in the Composition of Jordanian Secondary Students.

Unpublished MA Thesis. Jordan. Yarmouk University. 1982.

[10] Hoffman, Charlotte. "Towards a description of trilingual competence." International Journal of Bilingualism 2001. pages 1-17.

[11] Donna J. Montgomery et.al.: The Effectiveness of Word Processor Spell Checker Programs to Produce 'Thrget Words for Misspellings Generated by Students with Learning Disabilities, Journal of SpecialEducation Technology. p.p 27- 42, 16(2), Spring, 2001.

[12] K. Kukich: Techniques for automatically correcting words in text, ACM Computing Surveys, 24(4), 377–439, 1992.

[13] Fred J. Damerau: "technique for computer detection and correction of spelling errors". Communications of the ACM, Volume 7 Issue 3, March 1964 pp:171–176.

[14] Mitton. Ordering the suggestions of a spellchecker without using context. Natural Language Engineering, 15(2):173–192, 2009.

[15] Whitelaw, B. Hutchinson, G. Chung, and G. Ellis: Using the web for language independent spellchecking and auto correction. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP2009), pages 890–899, Singapore, 2009.

[16] Renato Cordeiro de Amorim and Marcos Zampieri: Effective Spell Checking Methods Using Clustering Algorithms, Proceedings of Recent Advances in Natural Language Processing, pages 172–178, Hissar, Bulgaria, 7-13 September 2013.

[17] Gökhan Dalkılıç and Yalçın Çebi: Turkish Spelling Error Detection and Correction by Using Word N-gram, IEEE, 2009.

[18] M. Hulden: Fast approximate string matching with finite automata. Procesamiento del Lenguaje Natural, 43:57–64, 2009.

[19] V. Ramaswamy and H. A. Girijamma: Conversion of Finite Automata to Fuzzy Automata for String Comparison, International Journal of Computer Applications (0975 – 8887) Volume 37– No.8, January 2012.

[20] Neha Gupta and Pratistha Mathur: Spell Checking Techniques in NLP: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 12, December 2012.

[21] J DamerauA: technique for computer detection and correction of spelling error", Communication ACM, 1964.

[22] Rakesh Kumar, Minu Bala, Kumar Sourabh: A study of spell checking techniques for Indian Languages, JK Research Journal in Mathematics and Computer Sciences, Vol. (1) No. (1) March 2018.

[23] Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger: Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness, Polibits (40) 2009.

[24] V. J. Hodge and J. Austin, "A comparison of standard spell checking algorithms and novel binary neural approach," IEEE Trans. Know. Dat. Eng., Vol. 15:5, pp. 1073-1081, 2003.

[25] R. A. Wagner and M. J. Fisher, "The string to string correction problem," Journal of Assoc. Comp. Mach., 21(1):168-173, 1974.

[26] E. J. Yannakoudakis and D. Fawthrop, "An intelligent spelling error corrector," Information Processing and Management, 19:1, 101-108, 1983.

[27] V.Gupta M. Lennig P. Mermelstein, "A Language Model in a Large-Vocabulary Speech Recognition System," in Computer Speech & Language Volume 6, Issue 4, October 1992, Pages 331-344.

[28] C. E. Shannon: "Prediction and entropy of printed English," Bell Sys. Tec. J. (30):50–64, 1951.

[29]Abdelbiifkahmoun and Zakarta Ftbcrrichi: Experimenting N-gram in Text Categorization, The International Arab Journal of Information Technology vol. 4. No. 4. October 2007.

[30] Baljeet Kaur: Review On Error Detection and Error Correction Techniques in NLP, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014.

[31] Hema P. H, Sunitha C: Spell Checker for Non-Word Error Detection: Survey, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015.

[32] Eisa, M. M., Ewees, A. A., Refaat, M. M., & Elgamal, A. F. Effective medical image retrieval technique based on texture features. International Journal of Intelligent Computing and Information Science, 13(2). 2013. 19-33

[33] Wim Peters, "Lexical Resources," NLP group, Dept. of Comp. Sc., Uni. of Sheffield, 2001.

[34] Atta E. ElAlfi, Moahmed M. ElBasuony and S. M. ElAtawy. Intelligent Arabic text to Arabic Sign Language Translation for Easy Deaf Communication. International Journal of Computer Applications 92(8). 2014. 22-29

[35] Atta E. ElAlfi and EL S. M. Atawy. Intelligent Arabic Sign Language to Arabic text Translation for Easy Deaf Communication. International Journal of Computer Applications 180(41). 2018. 19-26

[36] Hall, P.; Dowling, G. (1980). Approximate String Matching. Computing Surveys 12(4), pages 381–402.

[37] James L. Peterson: "Computer Programs for Detecting and Correcting Spelling Errors", Communications of the ACM, Volume 23 Number 12, December 198.

[38] Wikipedia, Commonly misspelled English words, https://en.wikipedia.org/wiki/Commonly_misspelled_English_words#Unlimited_misspellings, 2018.