Security of Infocommunication Systems and Internet of Things

2024 Vol 2, No 1

https://doi.org/10.31861/sisiot2024.1.01009

Received 29 July 2024; revised 07 August 2024; accepted 16 August 2024; published 30 August 2024

Technologies Overview for Typo Segregation

Oleksandr Skliarov and Ganna Zavolodko*

Department of Multimedia and Internet Technologies and Systems, National Technical University 'Kharkiv Polytechnic Institute', Kharkiv, Ukraine

*Corresponding author (E-mail: anna.zavolodko@khpi.edu.ua)

ABSTRACT The article focuses particularly on the difference between typos (accidental mechanical errors) and spelling or conceptual errors that arise from insufficient knowledge of language rules. Modern typo detection methods are analyzed, highlighting the advantages and disadvantages of each. The Levenshtein method is one of the most common algorithms for detecting and correcting errors in text. It effectively identifies and corrects errors in short words where the number of operations to convert the erroneous word to the correct one is small. However, this method does not consider the context in which the word is used, which can lead to incorrect corrections. The keyboard layout-based typo detection method analyzes probable errors that can occur due to the proximity of keys on the keyboard. It is simple to implement and integrate into existing spell-checking systems but does not consider the context of word usage. The contextual analysis method for typo detection relies on using contextual information to identify and correct errors in text, requiring significant computational resources and a large, diverse corpus of texts for effective model training. Deep models, such as BERT or GPT, consider the context of entire sentences or even larger text blocks, allowing for high accuracy in typo detection but require significant computational resources for training and inference, as well as large volumes of high-quality data for training. Machine learning methods, such as n-grams and Bayesian classifiers, show significant potential due to their simplicity and efficiency but may not account for complex dependencies between words and context, reducing their accuracy. The study highlights the importance of accurate error detection in student assessment systems, where typos can affect final grades and the relevance of answers.

KEYWORDS typo, spelling error, typo detection methods, automation of typo correction.

I. INTRODUCTION

In today's world, information spreads at an incredible speed, so the importance of accuracy and correctness of texts becomes more and more obvious. Typos and errors in the text can significantly affect the perception of information, its credibility and the reputation of the author or publication.

The topic of finding and separating typos also plays an important role in the system of assessing students' knowledge and determining their academic progress and level of preparation. One of the most common assessment methods is testing, which allows you to quickly and objectively determine the level of students' knowledge. However, in the process of answering tests, students often make typos and mistakes that can affect the final grades and their relevance. Therefore, it is very important to be able to find and distinguish typos from errors, analyzing the nature of their origin and the impact on the student's answer. Automatic typing detection helps to quickly and accurately evaluate students' written work, focusing on their knowledge and not on random errors. In turn, clearly distinguishing errors allows students to better understand their weaknesses and work on their correction [1].

First, it is worth understanding the difference between typos and errors. Distinguishing between a typo (random error) and an error (spelling or conceptual error) involves understanding the root causes and nature of each error [2].

A typo (random error) is a minor, random error that occurs while typing or entering text. Usually contains incorrect characters due to accidental pressing or "stuck" keys, the cause can also be fast typing.

Typos can be detected by the following signs:

- typos are often the result of mechanical errors, such as pressing the wrong key, double entering or skipping a key;
- typos usually refer to characters that are located next to each other on the keyboard or are the result of normal hand movements, so the printed word is very close to the correct form;
- in context, the incorrect word may not fit logically, but is often similar to the intended word in terms of spelling or phonetics;
- linked terms are a common mistake in a technical text context because in this context there are many method or class names that contain linked terms, such as "LinkedList" or "connectToServer". Sometimes these words have a "camel" case, and sometimes users simply skip spaces between words.

An error (spelling or conceptual error) is a misuse, wording, grammatical, spelling, syntactic, or stylistic inaccuracy that results from a lack of knowledge or understanding rather than a mechanical error.

Errors are characterized by:

- errors occur due to misunderstanding, incorrect memorization or gaps in knowledge, that means they are cognitive in nature;
- an incorrect word may phonetically or semantically differ from the intended one;
- errors may follow certain patterns, such as general spelling rules or language-specific phonetic errors.

II. TECHNOLOGIES FOR TYPO SEGREGATION

Since typos separation and search have been researched for a long time, today there are various methods for their implementation. Next, we will consider the most popular methods and give a comparative description of their disadvantages and advantages. Today, one of the most common methods is the Levenshtein method.

The Levenshtein method, also known as the Levenshtein distance or editorial distance, is one of the most common algorithms for detecting and correcting errors in text. This method measures the minimum number of single-character operations (insertions, deletions, substitutions) required to transform one string of characters into another [3-4].

The process of calculating the Levenshtein distance is carried out using a matrix, where each cell represents the cost of converting a substring of one row into a substring of another. The first row and first column of the matrix are filled with the indices of the row characters, and the rest of the matrix is filled by calculating the minimum cost of the operations (insertion, deletion, replacement) required to match the characters.

The Levenshtein method effectively detects and corrects errors in short words, where the number of operations to convert an erroneous word into a correct one is small. For example, correcting "katt" to "cat" requires one substitution and one deletion, giving a Levenshtein distance of 2.

Overall, the Levenshtein method remains one of the most common and widely used algorithms for finding typos due to its simplicity of implementation and ability to effectively correct mechanical errors in texts.

Method of searching for typos considering the layout of the keyboard is based on the analysis of possible errors that may occur due to the close position of the keys on the keyboard. This is an approach that aims to detect mechanical typos that often occur during rapid typing. The essence of the method is to build a matrix that displays the distance between each pair of keys on the keyboard. Proximity is defined as the number of adjacent keys that can be mistakenly pressed while typing. Based on this matrix, a list of possible error options is created for each word in the text, considering the proximity of the keys. Then these variants are checked against the dictionary, and if the variant is a dictionary word, it can be considered as a potential fix.

TABLE 1. Analysis of the advantages and disadvantages of the Levenstein method.

Criterion	Detailed description
Advantages	Levenshtein's algorithm is fairly simple to understand and implement, which makes it popular in many spell-checking programs. The method works well for detecting typos in short words, where the distance between the wrong word and the correct word is small.
Disadvantages	Levenstein's method does not consider the context in which the word is used, which can lead to incorrect correction. For example, in the sentence "I went to the see", the word "see" is not incorrect in terms of editorial distance, but it is grammatically incorrect. In the processing of long words or large texts, the algorithm can become resource-intensive, since it needs to calculate the distances between many pairs of words.

TABLE 2. Analysis of the advantages and disadvantages of the keyboard layout method.

Criterion	Detailed description
Advantages	Analysis of the proximity of keys on the keyboard allows you to effectively detect errors that occur due to accidental pressing of adjacent keys. For example, letters "р" and "п" are located next to each other on the Ukrainian keyboard, so they are often confused. The method is quite simple to implement and integrate into already existing spell checking systems. It can be applied as an additional layer of verification to the main algorithms. Since the method is based on simple key location comparison operations, it is very fast and computationally efficient. Low cost of implementation: Unlike machine learning-based methods, this approach does not require training on large data sets, which reduces the cost of implementation and maintenance.
Disadvantages	The method does not consider the context in which the word is used. This means that it may not detect or misinterpret errors that depend on the meaning of a word in a particular sentence. For example, errors like "buy" instead of "bought" may not be detected by the method. This method is not able to detect errors that arise due to ignorance of the rules of spelling or grammar, that is, cognitive errors are not considered by this method. Different language and regional keyboard layouts can make this method difficult to use. For example, the layout for the Ukrainian language differs from English, which requires a separate setting for each language. The method can offer corrections for words that are correct but look like potential misspellings due to the proximity of the keys. This can lead to an excessive number of false corrections. Because it needs to calculate the distances between many pairs of words.

This method is quite effective for detecting typos, since mechanical errors often occur due to pressing adjacent keys. For example, in the layout of the English keyboard, the letters "d" and "f" are located next to each other, so they can often be confused. The method is simple to implement and integrate into already existing spell-checking systems and is also fast and efficient in terms of computing resources. However, it has certain limitations, such as the inability to consider the context of word use, which can lead to incorrect correction of errors that depend on the meaning of the word in a particular sentence. Also, different language and regional keyboard layouts can make this method difficult to use, as each language requires a separate setting. Despite these shortcomings, the keyboard layout-based typing search method remains a useful tool for improving the accuracy of text documents.

Method of contextual analysis for the separation of typos is based on the use of contextual information to detect and correct errors in the text. This includes analyzing surrounding words, phrases, and even sentences to determine whether a word is correct or a potential typo.

The principles of contextual analysis for typo detection are based on using information from the surrounding text to accurately identify and correct errors. This method consists in considering the context, which means analyzing the words, phrases and sentences surrounding the potential typo. The basic idea is that words are not used in isolation, and their meaning and correctness can be determined by considering the words that are next to them [5].

Contextual analysis uses various patterns, such as collocations, which determine the frequency of co-occurrence of certain words, and grammatical structures, which help identify inconsistencies in sentence structure. For example, if the word "error" occurs more often with the word "occurred" than with the word "lipstick", then the system can determine that "lipstick" in this context is a typo. Context analysis methods also include language models such as n-gram models, which analyze sequences of words to estimate the probability of certain combinations, and neural network-based models such as recurrent neural networks (RNNs) or transformers (e.g. BERT, GPT). These models can consider long-term dependencies in the text, which allows more accurate analysis of the context.

TABLE 3. Analysis of the advantages and disadvantages of the method of contextual analysis.

Criterion	Detailed description
Advantages	Usage of context significantly increases the accuracy of error detection and correction compared to methods that analyze only individual words. The method allows detecting not only typos, but also grammatical errors, which makes it more universal.
Disadvantages	Context analysis requires significant computing resources, especially when using neural networks. Effective training of models requires a large and diverse corpus of texts. Insufficient quantity or poor quality of data can reduce the accuracy of the analysis.

In addition, contextual analysis uses statistical techniques, such as the Bayesian approach, to estimate the probability that a word is a correct or incorrect spelling based on the context. This approach helps to increase the accuracy of error detection and correction, making the method more universal.

In general, the method of contextual analysis significantly increases the accuracy of detecting and correcting typos compared to methods that analyze only individual words. However, this approach requires significant computing resources, especially when using neural networks, and depends on the quality of the training data.

Deep learning and machine learning methods play a key role in finding and correcting typos in texts due to their ability to analyze large amounts of data and take context into account. Deep neural networks use transformer architectures to understand the context of an entire sentence or even larger blocks of text. This allows them to achieve high accuracy in detecting typos, as they can consider the words on either side of the erroneous word, greatly improving the understanding of the text [6].

Deep models are trained on large volumes of data, allowing them to independently learn complex patterns and dependencies in text.

For example, the GPT-3 model has been trained on terabytes of text data from the Internet, giving it the ability to process different types of text data, including unstructured text, news, and social media posts.

Machine learning has become an integral part of many technological solutions used for data processing and analysis. Machine learning techniques such as n-grams and Bayesian classifiers show significant potential due to their simplicity and efficiency. At the same time, they have their advantages and disadvantages, which affect their use in various scenarios.

The advantages of machine learning methods, such as

TABLE 4. Analysis of advantages and disadvantages of deep learning methods.

Criterion	Detailed description
Advantages	Deep models, such as BERT or GPT, take into account the context of the entire sentence or even larger text blocks, which allows for high accuracy in detecting misprints. Deep models can work with different types of textual data, including unstructured texts, news, social media, etc., and adapt to different writing styles and text genres.
Disadvantages	Deep models require significant computational resources for training and inference, which can be expensive and technically challenging. To achieve high accuracy, models require large amounts of high-quality data for training.

TABEL 5. Analysis of advantages and disadvantages of machines learning methods.

Criterion	Detailed description
Advantages	Many machine learning techniques, such as n-grams or Bayesian classifiers, are relatively simple to implement and fast to execute. For example, n-gram models can quickly estimate the probability of word combinations without significant computing resources. Machine learning methods usually require less computing resources compared to deep models. For example, Bayesian models can run on ordinary computers without the need for high-performance equipment. Machine learning methods are often more interpretable, allowing for a better understanding of how and why certain predictions were made. For example, logistic regression allows you to clearly see the weights of each parameter.
Disadvantages	Simpler models may not consider the complex dependencies between words and context, which reduces their accuracy. For example, n-gram models cannot consider long contexts and complex grammatical constructions. Machine learning methods often require more manual work to configure and optimize. For example, creating and configuring features for classic machine learning models can be a time-consuming process. As with deep learning, machine learning methods require high-quality data to train on, although their volumes may be smaller. For example, incorrect or incomplete data can significantly reduce the performance of a model.

ease of implementation, speed of execution and interpretability of results, make them attractive for many tasks that do not require complex computing resources. For example, n-gram models and Bayesian classifiers can run on ordinary computers and provide sufficient accuracy in simple tasks. However, their drawbacks, including their limited ability to account for complex contextual dependencies and the need for significant manual tuning, reduce their performance compared to deep models. Despite this, traditional machine learning techniques remain important tools for analyzing textual data, especially in cases where resources are limited, and model transparency is a key factor.

III. MODERN RESEARCH

Teachers often encounter typographical errors when creating exam tests, which can affect the quality of the assessment. Traditional proofreading methods are time-consuming and often less efficient, especially given the large volume of documents. The article [7] discusses a study aimed at developing a spelling correction application using the Damerau-Levenshtein distance method to help teachers in detecting and correcting typos and errors in test scenarios of exams. The application provides word suggestions for unrealistic word errors and can handle different types of documents, increasing the efficiency and accuracy of the teacher's proofreading tasks.

The program uses the Damerau-Levenshtein distance, which improves upon the basic Levenshtein distance by adding transpose operations for more efficient detection and correction of typos. The system supports various input formats, including direct text input, file uploads, and document processing.

The developed application significantly increases the effectiveness of checking test exam scripts by automating the detection and correction of typos. The Damerau-Levenstein distance method has proven to be effective in handling different types of errors, providing high accuracy and convenient suggestions for correction. Future improvements may include reducing processing time and expanding the dictionary to include more words.

The study introduces MeDict [8], a health dictionary application designed to improve word search performance by correcting typos using the Damerau-Levenshtein Distance Algorithm. The app eliminates common mechanical and knowledge-based input errors by providing optimized search suggestions to improve user experience and information retrieval in medical terminology.

The app was tested with a set of medical terms that are commonly prone to typos. The Damerau-Levenshtein distance algorithm successfully provided accurate word suggestions, greatly improving search performance. The application was evaluated using the Technology Acceptance Model (TAM). 86.2% of users agreed that the app is useful and 86.9% find it easy to use.

MeDict effectively solves the problem of typos when searching for medical terms, making it a valuable tool for medical students and professionals. The app's implementation of the Damerau-Levenshtein distance algorithm provides accurate and efficient search results, improving the overall user experience.

This research [9] focuses on the development of an automatic error correction method for writing in English using deep neural network methods. The goal is to address grammatical errors in students' writing in English by creating a model that can automatically detect and correct these errors. The paper highlights the integration of statistical learning with deep learning using models such as Seq2Seq with attention and the transformer model.

The integration of Seq2Seq Attention models, transformer networks, and innovative learning strategies such as curriculum learning, and MASS led to the development of a robust automatic error correction system. This system not only helps to check and correct English writing, but also supports students' autonomous learning by removing the limitations of traditional correction methods.

The research [10] is devoted to the correction of context-dependent spelling and typographical errors in English documents using deep learning methods. The focus is on typographical errors, which are often caused by incorrect keystrokes. The proposed approach uses deep learning models, including autoregressive (AR) and auto-encoding (AE) language models, to efficiently detect and correct these errors.

Experiments conducted as part of the study demonstrate the effectiveness of various deep learning models in correcting context-dependent spelling errors:

- embedding-based correction: Models such as GloVe and fastText have shown significant improvement in detecting and correcting errors based on word embedding techniques;
- AR and AE models: AE language models such as BERT and RoBERTa have outperformed AR models in terms of correction accuracy due to their ability to use bidirectional contextual information.

The study highlights the superiority of deep learning approaches, particularly AE models, in correcting context-sensitive spelling errors in English documents. The models' ability to understand and use context makes them very effective at solving typos.

Natural Language Processing (NLP) technology offers innovative solutions to improve human-computer interaction, including error correction in English text. This research paper [11] by Juan Long explores the intelligent correction of words and grammatical errors in student English essays, presenting a model that balances mathematical and statistical methods with technological solutions.

The model effectively corrects non-verbal errors (such as insertion, loss, replacement, and substitution) and grammatical errors (such as singular-plural mismatches, subject-predicate mismatches, and modal verb errors) using a combination of statistical models and grammar rules. The obtained data indicate a high level of accuracy in correcting both non-verbal and grammatical errors.

The study demonstrates the effectiveness of using NLP technologies to automatically correct grammatical errors in English essays. By combining statistical models with grammar rules, the proposed system effectively detects and corrects a wide range of errors, thus improving the quality of writing in English for non-native speakers. The results indicate significant potential for the application of such technologies in educational institutions, providing valuable tools for both students and teachers.

The objective of analyzing these articles was to understand and demonstrate how the typo segregation methods mentioned above find their application in existing solutions and to highlight that typo detection is a modern problem with many possible approaches to solve it. In conclusion, these studies demonstrate the significant potential of advanced deep learning algorithms and models in automating and improving error correction processes in various fields. The integration of technologies such as the Damerau-Levenshtein distance method, deep learning models, and NLP offers accurate, efficient solutions that improve the quality of written evaluations and specialized terminology retrieval.

IV. CONCLUSION

In today's world, where the importance of accuracy and correctness of texts increases every day, the difference between accidental typos and spelling or conceptual errors must be clearly defined to ensure the relevance of information and correct understanding of the text.

Searching for typos and errors in text is a complex task, for the solution of which various methods can be used, including the Levenshtein method, contextual analysis, machine learning models, use of dictionaries and analysis of frequent error patterns. Now, there is no method that would be ideally suited for solving this type of problem, due to various shortcomings. Therefore, the integration and combination of several different methods and approaches, from simple algorithms to advanced deep learning models, allows you to create effective systems for automatic detection and correction of typos, which significantly improves the quality of texts and the objectivity of assessing students' knowledge.

AUTHOR CONTRIBUTIONS

O.S., G.Z. – formal analysis; O.S., G.Z. – conceptualization, methodology; O.S. – investigation; O.S. – writing-original draft preparation, writing-review and editing; G.Z. – supervision, validation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article.

REFERENCES

[1] A. A. Khansir and F. Pakdel, "Place of error correction in English language teaching," *Educational Process: International Journal*, vol. 7, no. 3, pp. 189-199, 2018.

[2] D. Hládek, J. Staš, and M. Pleva, "Survey of automatic spelling correction," *Electronics*, vol. 9, no. 1670, 2020.

[3] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171-176, 1964.

[4] Y. Korolekh and G. Zavolodko, "Enhancing digital search: Synergizing the Levenshtein algorithm with NLP techniques," in *IX International Scientific and Practical Conference "Scientific Problems and Options for Their Solution,"* Bucharest, Romania, Feb. 7-9, 2024, International Scientific Unity, pp. 60-64.

[5] D. Ittner and H. Baird, "Programmable contextual analysis," in *Document Analysis Systems*, A. Spitz and A. Dengel, Eds. Singapore: World Scientific, 1995, pp. 76-92.

[6] E. Puerto, J. Aguilar, and A. Pinto, "Automatic spell-checking system for Spanish based on the Ar2p neural network model," *Computers*, vol. 13, no. 3, p. 76, 2024.

[7] V. C. Mawardi, F. Augusfian, J. Pragantha, and S. Bressan, "Spelling correction application with Damerau-Levenshtein distance to help teachers examine typographical error in exam test scripts," *E3S Web Conf.*, vol. 188, p. 00027, Sep.

2020, doi: 10.1051/e3sconf/202018800027.

[8] W. Clarissa and F. P. Putri, "MeDict: Health dictionary application using Damerau-Levenshtein distance algorithm," *IJNMT (International J. New Media Technol.)*, vol. 7, no. 2, pp. 98-101, 2020, doi: 10.31937/ijnmt.v7i2.1654.

[9] L. Cheng, P. Ben, and Y. Qiao, "Research on automatic error correction method in English writing based on deep neural network," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2709255, 2022.

[10] J.-H. Lee, M. Kim, and H.-C. Kwon, "Deep learning-based context-sensitive spelling typing error correction," *IEEE Access*, vol. 8, pp. 152565-152578, 2020.

[11] J. Long, "A grammatical error correction model for English essay words in colleges using natural language processing," *Mobile Information Systems*, vol. 2022, no. 5, pp. 1-9, Jul. 2022.

Oleksandr Skliarov

Oleksandr, 24 years old, Master's in the field of "Information Technologies" with a specialty in "Computer Science" at National Technical University "Kharkiv Polytechnic Institute", co-founder of ReMnemo

ORCID ID: 0009-0006-7232-6319

Ganna Zavolodko

Ganna, 46 years old, Ph.D., associate professor at National Technical University "Kharkiv Polytechnic Institute", IEEESenior; CEO, co-founder of ReMnemo

ORCID ID: 0000-0003-0000-8910

Огляд технологій відокремлення одруківок

Олександр Скляров, Ганна Заволодько*

Кафедра Мультимедійних та Інтернет-технологій і систем, Національний Технічний Університет «Харківський Політехнічний Інститут», Харків, Україна

*Автор-кореспондент (Електронна адреса: anna.zavolodko@khpi.edu.ua)

АНОТАЦІЯ У статті особлива увага приділяється різниці між одруківками (випадковими механічними помилками) та орфографічними або концептуальними помилками, які виникають через недостатнє знання мовних правил. Проаналізовані сучасні методи виявлення одруківок, виявлені переваги та недоліки кожного з них. Метод Левенштейна є одним із найпоширеніших алгоритмів для виявлення та виправлення помилок у тексті, який ефективно виявляє та виправляє помилки в коротких словах, де кількість операцій для перетворення помилкового слова в правильне невелика. Проте цей метод не враховує контекст використання слова, що може призводити до неправильного виправлення. Метод пошуку одруківок з урахуванням розкладки клавіатури базується на аналізі ймовірних помилок, які можуть виникати через близьке розташування клавіш на клавіатурі та є простим для реалізації та інтеграції у вже існуючі системи перевірки правопису, але не враховує контекст використання слова. Метод контекстуального аналізу для відокремлення одруківок базується на використанні контекстної інформації для виявлення та виправлення помилок у тексті вимагає значних обчислювальних ресурсів і потребує великого та різноманітного корпусу текстів для ефективного навчання моделей. Глибокі моделі, такі як BERT або GPT, враховують контекст цілих речень або навіть більших текстових блоків, забезпечуючи високу точність виявлення друкарських помилок, але вимагають значних обчислювальних ресурсів для навчання та висновків, а також великих обсягів високоякісних даних для навчання. Методи машинного навчання, такі як n-grams та Байєсівські класифікатори, демонструють значний потенціал завдяки своїй простоті та ефективності проте вони можуть не враховувати складні залежності між словами та контекстом, що знижує їхню точність. Дослідження показує важливість точного виявлення таких помилок у системі оцінювання знань студентів, де одруківки можуть впливати на підсумкові оцінки та релевантність відповідей.

КЛЮЧОВІ СЛОВА одруківка, орфографічна помилка, методи відокремлення одруківок, автоматизація відокремлення одруківок.