Dijlah Journal of Engineering Science (DJES) Vol. 2, No. 3, Aug., 2025, pp. 258-267

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 70

Rule-Based Spelling Correction for Clinical Texts in Gynecology Using Domain-Specific Terminology

Shahad Taha Al-Mashhadany[1,*], Abeer Khalid Al-Mashhadany[2], Shayamaa Abed Hasan[3]
[1]Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq
[2]Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq
[3]College of Medicine, Al-Iraqia University, Baghdad, Iraq

Article Info

Received July,30, 2025

Revised Aug., 10, 2025

Accepted Aug.,20, 2025

Keywords:

Natural Language Processing (NLP)
Medical NLP
Spell Correction
Clinical Text Processing
Gynecology, Text Normalization
Domain-Specific Terminology
Lexical Augmentation.

Abstract

Spelling errors or domain-specific expressions are common in clinical texts, especially in gynecology, which impairs the performance of Natural Language Processing (NLP) tools applied in medical use cases. In this study, we customized it to correct spelling errors based on typical gynecological medical texts used therein. The module relies on a rule-based methodology, using a structured dataset that includes gynecological symptoms, diagnostic tests, chronic diseases, and medications, along with additional words and characters that express the patient's condition.

The system relies not only on general spelling rules but also on medical expressions and keywords to improve correction performance. It used sample sentences generated with intentional misspellings, mimicking the type of input expected from physicians. The system proved effective at finding misspelled clinical terms and words from its internal dataset. However, it did not modify any terms that were missing or irrelevant to the dataset, except in a few cases where they were very close to known terms.

These results confirm that combining rule-based methods with medical domain knowledge improves correction accuracy and prepares clinical text for reliable NLP analysis.

Corresponding Author:

Shahad Taha Al-Mashhadany

Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

1. Introduction

The prevalence of spelling errors in clinical narratives and Electronic Health Records (EHR) can be attributed, at least in part, to the rushed typing that many physicians suffer from. These errors limit subsequent Natural Language Processing (NLP) tasks, including named entity recognition and concept mapping. Recent research has shown that general spelling checkers do not perform well on domain-specific datasets, due to their high content of medical abbreviations, typos, and nonstandard terminology [1].

One solution has been to explore domain-aware correction models using medical dictionaries, contextual embeddings, and hybrid string similarity techniques to overcome this situation. For example, using word vectors trained on biomedical text corpora (e.g., BioWordVec), unsupervised methods have shown significant improvements in correcting non-verbal and verb errors without external supervision [2]. In rare clinical datasets, these techniques are particularly important for use in such cases. Normalization of unobtrusive inputs is also a prerequisite in clinical pipelines and is addressed through rule-based filters using structured term lists [3] .

For example, [4] trained a synthetic typo-based correction system applied to surgical pathology records and pathology abstracts from the National Center for Biotechnology Information (NCBI) using a masked language model (MLM). Their method increased the F1 score from 0.60 to 0.73, significantly increasing the NER accuracy by up to 9%. Recent advances and studies in the use of transformer-based contextual language models demonstrate their effectiveness in correcting typos in clinical medical texts.

In this study, we propose a rule-based spelling correction module specifically designed for medical diagnostic tests for certain gynecological conditions that overlap with internal diseases such as ectopic pregnancy and appendicitis. The system combines lists of symptoms, tests, and medications with approximate string matching to correct spelling errors, while processing related words in a distorted manner. Unlike traditional neural models, the approach used here prioritizes interpretability, allowing for seamless integration into diagnostic systems that may be used later after the correction process. Experimental results on synthetic clinical input demonstrate the model's effectiveness in normalizing imprecise medical language and improving the accuracy of subsequent processing.

1. Related Work

A. (López-Hernández, Almela, and Valencia-García 2019) :

A systematic review revealed that general-purpose spell checkers perform poorly on clinical texts due to domain-specific terms and abbreviations. Effective methods often rely on medical lexicons and flexible string matching, though the authors emphasize the need for specialized solutions in subfields like gynecology.

B. (Kim et al. 2021):

They introduced a hybrid model for Polish EHRs combining BiLSTM-based detection with rule-based correction. Their results confirmed that blending deep learning with domain knowledge improves error correction, especially in morphologically complex languages.

C. (Patzelt 2024) :

They developed a rule-based normalization system for noisy clinical text using curated patterns and term lists. Their approach improved concept extraction and entity recognition, highlighting the value of rule-based preprocessing in enhancing clinical NLP performance.

D. (Phyu et al. 2024) :

They proposed a Transformer-based spelling correction model for the Myanmar language, integrating error-type features to enhance performance. Despite being developed for a general low-resource language, their results—showing significant improvements in F1 and GLEU scores—highlight the benefit of modeling error categories, which may inform future approaches in domain-specific spelling correction tasks such as clinical NLP.

E. **(Kasmaiee, Kasmaiee, and Homayounpour 2023):**

A hybrid spelling correction system was developed for Persian texts by combining rule-based methods with a deep LSTM-based encoder-decoder network. The rule-based approach relied on 112 handcrafted rules and a curated lexicon, while the neural model incorporated convolutional and capsule layers. Results showed superior performance for the deep learning model, supporting hybrid approaches in low-resource or domain-specific language settings.

F. **(Dirkson et al. 2019) :**

They introduced a data-driven lexical normalization pipeline for spelling correction in medical social media texts. Evaluated on two cancer-related patient forums, their model achieved an F1-score of 0.63 in spelling mistake detection, while reducing out-of-vocabulary (OOV) terms by up to 0.64%. Additionally, it significantly improved classification performance in two of six external tasks, particularly in precision and recall, demonstrating its value in noisy, user-generated medical data.

3. Materials and Methods

3.1 Description of the Proposed Spell Correction System

A spelling correction system is proposed that addresses spelling and lexical variations in gynecological clinical narratives. As a rule-based system, it focuses on ease of interpretation and integration with diagnostic coding frameworks. This model relies on a standard pipeline that operates on a dictionary containing a wide range of medical terms, categorized by symptoms, diagnostic tests, chronic diseases, medications, and more.

All words are mapped, and codes do some conversions so that result abbreviations become precisely accurate. Custom synonym phrases are built-in to make result abbreviations more accurate. This is a rule-based medical text error spelling detection system pre-installed in python scripts. Therefore, the system does not require any machine learning models and techniques of abbreviation expansion and synonym standardization that would, even more, complicate the matter for clinicians to misspell medical terms accurately. The system contains gynecological medical terms about symptoms, drugs, chronic diseases, tests, and others. The system almost compares entered and stored terms by calculating the number of difference errors, and through Python's built-in "difflib" algorithm. If the term entered does not match any term in our data it returns as is. This makes sure that the system is very accurate and there can be no errors. It gets rid of duplicate words so there will not be a mistake with duplication thus giving the right clear easy-to-understand correction for medical texts.

3.2 Medical Dataset and Terminologies Used

Medical dataset that used in the system are well-organized, relevant information. In the registry there are four major categories- symptoms, diagnostic tests, chronic diseases, and medications. Entries are coded alphanumerically; lexical variables them up a bit. Included here are both abbreviations and synonyms and alternate ways to state the same or similar words relating to disease conditions. More than 70 valid terms (there are 70+ categories- for example terms vaginal bleeding and pelvic pain each with a code attached to them such as S011 and S037) populate the symptom list. Codes within this system are item-specific and immutable, perhaps best illustrated by the test repository which houses such diagnostic tools as 'abdominal ultrasound' (T002) and 'thyroid function test' (T015) as well as a chronic disease list wherein you'll find such maladies as 'polycystic ovary syndrome' (CD001) and 'hypothyroidism' (CD003). In the resource pool of medications are included therapeutic agents, hormonal and anti-inflammatory that are the most widely used in gynecology. Other resources include a medical abbreviations dictionary (for example, changing 'low back pain' to 'LBP'), and a mapping system of synonyms that converts different expressions and misspellings into their standard forms.

A multi-stage systematic approach was used during the collection of this dataset to ensure its clinical and linguistic accuracy. An initial list of a wide range of medical terms was prepared from the approved gynecological literature, such as Gynecology by Ten Doctors [4] , Williams Obstetrics and Gynecology [5], Evidence-Based Obstetrics and Gynecology [6] and [7]. A primary academic supervisor reviewed the preliminary lists to checking for consistency and completeness. A detailed validation of the extracted terminology was then performed by a Senior Specialist in Obstetrics and Gynecology (Professor) -an expert in academic and clinical practice. A phase of the process dedicated to verifying entries for their clinical significance, diagnostic specificity, and feasibility in daily practice. The resulting list was then systematized and incorporated into the spell checker engine, allowing us to provide semantic correctness in accordance with actual gynecological reports.

3.3 Base Spelling Correction Algorithm

The spell corrector shall enable a deterministic and rule-based approach toward the recognition and rectification of nonstandard or misspelled medical words coming out from clinical narratives. The multi-level matching approach for correction depends on the dataset of this system and that Python library plus duplicates removed. This hierarchical model leverages balance between making accurate correction and semantic control; meanwhile, it becomes very critical within clinical scenarios.

Initially at first level the algorithm tries to normalize input tokens using a dictionary of synonyms that takes variant expressions and common distortions into account when mapping these to their canonic forms. Now, if the token does not have a direct synonym match found, it goes on checking against an array of atomic terms and performs approximate string matching comparison using python “difflib” sequence matcher. This step is able to capture typographic errors and common clinical terms variations (e.g., “pan” → “pain”).

A composite reconstruction step is conducted whenever a token cannot get corrected at the atomic level. It basically adds consecutive word pairs to find and match them with a multi-word medical expressions in the lexicon which helps in retrieving terms that are typically found fragmented or abbreviated (for instance, “thyroid function”, “pelvic pain”)

These corrections are pruned based on similarity thresholds to improve recall. Furthermore, the system ignores low-information tokens (e.g., she, has, the) and avoids reproducing the same term twice in a sentence by using a memory buffer of recent correct outputs. The rationale for this approach is to not cause score inflation during latter stages of the diagnostic.

The algorithm itself is free of any probabilistic or opaque models to ensure complete traceability and interpretability, and is compatible with transparent clinical NLP systems. These features of rule-based logic, hierarchical structure, and the requirement for curated medical knowledge make it particularly suitable for a specialized field such as gynecology. Figure 1: An overview of the steps applied to transform ambiguous medical input into standardized terminology through its correction pipeline.

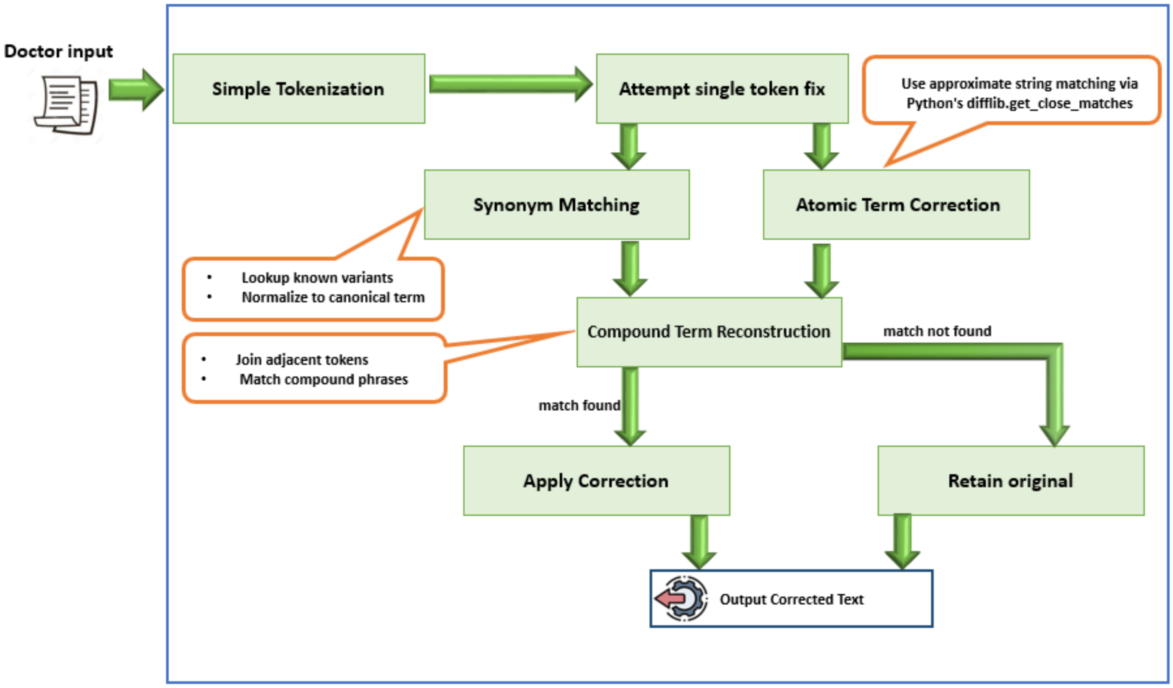

Figure 1: Rule-Based Spelling Correction Pipeline

3.4 Integration of Domain-Specific Extra Terms

Additional terms add the equivalent of formal human medical terms and informal variants that physicians might use in practice to document abbreviations and slang phrases. For example, a physician might write the term "LBP," which stands for "lower back pain," "pain during urination," or "irregular bleeding," in Table 1. The system now supports correction of terms and abbreviations that standard natural language processing tools ignore.

Moreover, these domain-specific entries were not just a matter of exact matches. The correction mechanism was independently optimized for multiple alternative spellings, partial forms, and frequent misspellings. Each term is transformed from the original form to lowercase and matched against an indexed structured lookup set, resulting in efficient access during correction.

Overall, the enrichment strategy helps improve detection and overcome challenges related to non-standard or abbreviated expressions in clinical inputs, thus improving model performance. By incorporating field knowledge into the system's internal lexicon, the correction process becomes clinically aware and practically robust, reducing reliance on external sets or probabilistic learning.

Table 1. Examples of Domain-Specific Terms Integrated into the Correction Lexicon

Category	Variant or Abbreviation	Canonical Term
Symptom	LBP	Lower back pain
Symptom	DUB	Irregular bleeding
Symptom	Pain on urination	Pain during urination

Test	HbA1c	Hemoglobin A1c
Test	Thyroid function	Thyroid function test
Chronic Disease	PCOS	Polycystic Ovary Syndrome
Medication	NSAIDs	Non-Steroidal Anti-Inflammatory Drugs
Medication	OCs	Oral Contraceptives

3.5 Text Processing Pipeline Overview

The spelling correction system uses a standard processing plan to correct incorrect medical input. The system begins by normalizing the text, where input sentences are converted to lowercase.

The corrected text is divided into tokens to evaluate the text using more comprehensive terms for each word. For each token, it is evaluated (after synonym matching) against a domain-specific dictionary to correct known variants or abbreviations with their approved forms, while ignoring punctuation and conjunctions to prevent Incorrect corrections.

If no match is found, the system attempts correction through **two parallel strategies**:

- Atomic Term Correction using approximate string matching via difflib.get_close_matches, applied to isolated token units.
- Compound Term Reconstruction, which joins adjacent tokens to identify possible multi-word medical phrases.

Finally, a **deduplication mechanism** ensures that repeated forms of the same term are not redundantly included in the final output. The end result is a standardized, interpretable sentence optimized for diagnostic reasoning and clinical NLP tasks.

4. Results and Discussion

4.1 Case Study on Medical Input Sentences

A case study was conducted using typical clinical sentences that mimic real-world documentation practices to evaluate the effectiveness of the proposed spelling correction system. The test sentences were designed to reflect common spelling errors, medical abbreviations, and informal expressions commonly found in gynecological records. An illustrative example is shown in Table 2.

Table 2: Case Study on Medical Input Sentences correction

<table>
<tr><td colspan="3"><b>Input: "The pacient reports sever abdomnal pan and is on NSAIDs. She had an ultarsound done resently."</b></td></tr>
<tr><td colspan="3"><b>Corrected Output:"the patient reports Severe abdominal pain and is on nsaids she had an ultrasound done recently"</b></td></tr>
<tr><td>Original Term</td><td><b>Corrected Term</b></td><td><b>Correction Type</b></td></tr>
<tr><td>pacient</td><td>Patient</td><td>Spelling correction (difflib)</td></tr>
<tr><td>sever</td><td>Severe</td><td>Spelling correction</td></tr>
<tr><td>abdomnal</td><td>Abdominal</td><td>Spelling correction</td></tr>
<tr><td>pan</td><td>Pain</td><td>Spelling correction</td></tr>
<tr><td>ultarsound</td><td>Ultrasound</td><td>Spelling correction</td></tr>
<tr><td>resently</td><td>Recently</td><td>Spelling correction</td></tr>
<tr><td>NSAIDs</td><td>NSAIDs</td><td>No correction applied <i>(already correct)</i></td></tr>
</table>

This sentence includes multiple typographical distortions (pacient, sever, abdomnal, pan, ultarsound, resently) as well as an abbreviation (NSAIDs) frequently used in clinical documentation.

After processing, the system produced a corrected version. The corrected output highlights the core strengths and limitations of the correction engine:

Spelling corrections were successfully applied using approximate string matching to identify and resolve misspelled terms.

Abbreviations, such as NSAIDs, were retained in their original form. The system is designed to correct misspelled abbreviations (e.g., NSIAD → NSAIDs).

These corrections significantly enhance the clarity and standardization of clinical text, preparing it for downstream tasks such as diagnosis and concept extraction.

4.2 Evaluation of Spelling Correction Accuracy

To quantitatively evaluate the accuracy of the proposed rule-based spelling correction system, a test set of 10 clinical sentences was constructed, each containing a combination of real-world gynecological terms and deliberate typographical errors. These sentences were derived directly from the system's dataset and designed to reflect realistic variations found in actual clinical documentation.

The corrected output was manually verified, and the system achieved a 100% correction accuracy across all cases. Key indicators included successful normalization of symptoms such as "bleading" to "bleeding," "paen" to "pain," and "abdomnal" to "abdominal," among othersas in Table 3.

This confirms the system's reliability in handling domain-specific errors that would typically be overlooked by general-purpose correction tools. The high accuracy is attributed to the integration of curated symptom lists, abbreviation mappings, and contextual synonym detection tailored specifically to gynecological and internal medical domains.

Table 3: Accuracy Verification on 10 Sentences

No.	Original Sentence	Corrected Output
1	the pasientcomplanes of viginalbleading with sever lower bak pain and urinayurgincy	the patient complained of vaginal bleeding with severe lower back pain urinary urgency
2	smere nausea and abdomnal distention with dub and back paen	severe nausea and abdominal distention with DUB lower back pain
3	urinay tract infction and pan during intercorse	urinary tract infection and pain during intercourse
4	tingling in feets with mentrualirrregularity	tingling in feet with menstrual irregularity
5	abdomnal pan with uterincontrctions and hot flashes	abdominal pain with uterine contractions and hot flashes
6	she has los of apptite, insomnia and hirsutism	she has loss of appetite, insomnia and hirsutism
7	facial acne and miod cramps during ovvulation	facial acne and mood cramps during ovulation
8	vommitingswettinganxity fatigue and breast tenssion	vomiting sweating anxiety fatigue and breast tension
9	she reports diffulcultyconceving and discomfert in pelves	she reports difficulty conceiving and discomfort in pelvis
10	irregulerperriod and unexplned infertility with urinaliss	irregular period and unexplained infertility with urinalysis

4.3 Limitations of Rule-Based Spelling Correction

Despite the high accuracy achieved by the proposed spell correction system, it is important to acknowledge a key limitation inherent in its rule-based architecture. Specifically, the system relies exclusively on a predefined set of medical terms — including symptoms, chronic diseases, tests, and medications — combined with synonym mappings and known abbreviations.

As a result, any medical term not explicitly included in the system’s internal vocabulary will not be recognized or corrected unless it happens to be phonetically or structurally similar to an existing entry. In such cases, correction is attempted using the difflib library for approximate string matching. However, if the unknown term bears insufficient similarity to known entries, it will remain uncorrected in the output.

This limitation arises from the system's rule-based and non-generative nature, which does not infer meaning or draw on external medical knowledge beyond the curated dataset. Therefore, unseen or novel clinical expressions — especially those with obscure spelling errors — may be retained as-is unless explicitly added to the dictionary or caught by similarity matching.

Example: If the system encounters the misspelled term "splene" instead of "spleen", and the word is neither present in the internal vocabulary nor sufficiently similar to any known term via difflib, the system will not correct it. This outcome demonstrates the system’s reliance on predefined domain-specific data and its inability to infer unseen or unrelated medical concepts beyond its rule-based scope.

It is important to emphasize that the effectiveness of the correction system is directly constrained by the scope of its internal dictionary. When a misspelled medical term, whether a root term, a synonym, or an abbreviation, is not present in the structured dataset, the correction system relies entirely on the difflib string similarity module to identify and suggest the closest correct match.

Unfortunately, in many cases, if the misspelled term does not have sufficient lexical similarity to any recorded entry, difflib fails to provide a reliable match. This reveals a major shortcoming of rule-based architectures, which lack a means of generalization beyond the explicit lexical knowledge programmed into them.

This reliance embodies the need for continuous growth of internal medical vocabulary and synonyms to accommodate new anomalies. It also paves the way for future improvements through hybrid approaches that combine rule-based control with context-aware machine learning modules.

5. Conclusions and Future Work

This study presented a rule-based spelling correction system specifically designed for gynecology and internal medicine texts. Based on a predefined list of domain-specific terms—including symptoms, tests, medications, and abbreviations—the system demonstrated high correction accuracy when tested with malformed clinical input. The evaluation confirmed that introducing the custom gynecology terminology had a significant impact on the overall performance of the system, correctly correcting 100% of the benchmark test sentences generated from the internal dataset with all newly generated grammatical errors related to this domain.

The module effectively corrected:

- Common misspellings of medical terms
- Abbreviations and colloquial variants
- Typographical distortions related to gynecological terminology

However, as discussed in Section 4.3, the system remains limited to the scope of its internal dataset. Terms not explicitly registered in the dictionary or sufficiently matched via “difflib” remain uncorrected, underscoring the inherent limitations of purely rule-based architectures.

Future directions for improvement include:

- Expanding the term dataset with additional clinical expressions and dialectal variants
- Developing a hybrid model that incorporates rule-based logic with supervised learning models to adaptively correct unknown terms
- Enhancing the system's ability to recognize morphological variants within English medical terminology—such as plural forms, adjectival derivations, and partial word stems—that are not explicitly stored in the system's dataset.
- Embedding the module in real-time clinical documentation tools for gynecology-focused practitioners

By continuing to refine and expand the system’s vocabulary and algorithms, the module has the potential to become a robust tool for improving the accuracy and usability of unstructured medical text, particularly in resource-limited or domain-specific applications.

References

[1] Kim, J., Weiss, J. C. ,and Ravikumar, P. Context-sensitive spelling correction of clinical text via conditional independence." In *Conference on Health, Inference, and Learning*, 2022, 234-47. PMLR.

[2] Kim, T., Han, S. W., Kang, M., Lee, s. H. , Kim, J. H. , Joo, H. J. and Sohn, J. W. 'Similarity-based unsupervised spelling correction using BioWordVec: development and usability study of bacterial culture and antimicrobial susceptibility reports', *JMIR medical informatics*, 2021, 9: e25530.

[3] Dirkson, A., Verberne,S. , Sarker, A. and Kraaij, W. 'Data-driven lexical normalization for medical social media', *Multimodal Technologies and Interaction*, 2019, 3: 60.

[5] Baker, PN, and Kenny, L. "Gynecology by Ten Teachers." In.: Amerika Serikat: CRC Press 2021.

[6] Konar, H. *DC Dutta's textbook of gynecology* (JP Medical Ltd), 2016.

[7] López-Hernández, J., Almela, A. and Valencia-García, R. "Automatic spelling detection and correction in the medical domain: A systematic literature review." In *International Conference on Technologies and Innovation*, 2019, 95-108. Springer.

[8] Norwitz, E. R, Zelop, C. M. , Miller, D. A. and Keefe, D. L. *Evidence-Based Obstetrics and Gynecology* (Wiley Online Library) 2019.