Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd

ISSN,e-ISSN: 2548-1304
Volume:IDAP-2023, pp:44-51, 2023
https://doi.org/10.53070/bbd.1350547
Research Paper

The Role of Phonological Errors in Evaluation Metrics

Fonolojik Hataların Değerlendirme Metriklerindeki Rolü

Ayşegül Çağlı[1], Vakkas Karakurt[1], Kürşat Edabalı Yıldırım[1]
Fatih Soygazi[1], Yılmaz Kılıçaslan[1]

[1]Bilgisayar Mühendisliği Bölümü, Aydın Adnan Menderes Üniversitesi, Aydın, Türkiye

(201805049@stu.adu.edu.tr, 201805008@stu.adu.edu.tr, 221805117@stu.adu.edu.tr, fatih.soygazi@adu.edu.tr, yilmaz.kilicaslan@adu.edu.tr)

Received:Aug.26,2023 Accepted:Aug.26,2023 Published:Oct.18,2023

Özetçe— Son yıllarda, Doğal Dil İşleme (DDİ), özellikle metin özeti oluşturma ve makine çevirisi alanlarında yoğun bir araştırma artışı yaşamıştır. ROUGE ve BLEU gibi değerlendirme metrikleri, N-gram temelli yaklaşımlar kullanılarak metinlerin kalitesini değerlendirmek için yaygın olarak kullanılmaktadır. Ancak, bu metrikler özellikle sosyal medya platformlarından elde edilen verilere uygulandığında, sesbilgisel hataların yaygınlığı nedeniyle zorlanmaktadır. Bu çalışma, sesbilgisel hataların kaynaklarını ve frekansını belirlemeye odaklanmakta ve bu hataları dikkate almalı mı sorusuna cevap niteliği taşımaktadır. Bu konuyla ilgili olarak sesbilgisel hataların sık görüldüğü bir platform olan Twitter'dan veri toplanmış ve incelenmiştir. Ayrıca mevcut literatür de gözden geçirilmiştir. Makale, Levenshtein ve Damerau-Levenshtein gibi düzenleme mesafesi algoritmalarını mevcut metriklere entegre ederek onları geliştirmeyi önermektedir. Sesbilgisel hataları değerlendirmelere dahil ederek, DDİ ve makine çevirisi alanlarında doğruluk ve güvenilirliği artırmayı hedeflemektedir. Bu çalışmanın nihai amacı, bu alanlarda daha hassas ve güvenilir değerlendirme metrikleri oluşmasına katkı sağlamaktır.

Anahtar Kelimeler : *Doğal Dil İşleme, Fonolojik Hatalar, Rouge, Makine Çevirisi, Değerlendirme Metrikleri, Düzeltme Uzaklığı Metrikleri.*

Abstract—In recent years, Natural Language Processing (NLP) has seen a surge in research, particularly in the areas of text summarization and machine translation. Evaluation metrics like ROUGE and BLEU have been widely used to assess the quality of texts using N-gram based approaches. However, these metrics often struggle when applied to data sourced from the internet, such as social media platforms, due to the prevalence of phonological errors. This study focuses on identifying the sources and frequency of phonological errors while addressing the question of whether they should be considered or not. Data from Twitter, a platform known for phonological errors, was collected, and studied, along with existing literature on the subject. The article proposes enhancing existing metrics by integrating edit distance algorithms like Levenshtein or Damerau-Levenshtein. By considering phonological errors in evaluations, this approach aims to improve accuracy and reliability in the NLP and machine translation domains. The ultimate goal of this study is to contribute to more sensitive and reliable evaluation metrics in these fields.

Keywords: *Natural Language Processing, Phonological Errors, Rouge, Machine Translation, Evaluation Metrics, Edit Distance Metrics.*

1.Introduction

In recent years, there has been a significant surge in research efforts within the field of Natural Language Processing (NLP). A large portion of these endeavours has been focused on text summarization and machine translation. Consequently, various metrics have been developed to assess and compare the quality of texts. Among

the most widely utilized metrics are ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy), both of which employ an N-gram-based approach for evaluation. These metrics are primarily applied for evaluating text summarization. N-gram-based approach involves analysing the frequency and similarity of consecutive sequences of "n" words or characters to assess the quality of text or translation.

Our observation indicates that these metrics often prove insufficient when applied to datasets gathered from the internet, especially from social media platforms where phonological errors are widespread. The presence of such errors makes it challenging to achieve the same level of accuracy and reliability in evaluations. Consequently, these metrics tend to produce more accurate results when applied to clean and error-free data sources such as articles and books. Considering this challenge, our primary focus has been to identify the source of phonological errors and determine the frequency at which individuals make such mistakes. Subsequently, we plan to enhance the performance of the algorithm by introducing a penalty rate that is inversely proportional to the number of mistakes made since detecting a low rated error is much harder than detecting a higher rated error. By doing so, we aim to improve the overall quality of the algorithm's output. We have found that these errors generally stem from two main causes: a lack of knowledge about grammar and spelling rules and typographical mistakes arising from keyboard usage. For this reason, we collected data from Twitter, where phonological errors are prevalent, and conducted studies on our custom dataset. Additionally, some data is collected from existing literature on this subject to demonstrate the significance of considering phonological errors in evaluations.

In this article, we propose the incorporation of additional edit distance algorithms, such as Levenshtein and Damerau-Levenshtein, alongside existing metrics to enhance their performance. This integration is expected to yield better and improved results with the currently available metrics. We delve into the importance of considering phonological errors when evaluating results generated by NLP tasks. The ultimate aim of this study is to contribute towards making evaluation metrics in the NLP.

2. Literature Review

Despite the limited availability of data for machine translation and text summarization in Turkish, we managed to find some research conducted on non-native Turkish speakers to determine the types and frequencies of phonological errors. One such study conducted by Uzdu Yıldız, F., & Çetin, B. (2020) focuses on analyzing the Turkish writing of non-native individuals. The participants in this study came from diverse backgrounds and diverse countries.

ROUGE and BLEU are commonly used metrics that rely on N-gram-based analysis. That makes them most effective when applied to clean data sources like books or articles. Notably, some noteworthy studies conducted by Baykara, B. and Güngör, T. (2023), Chin-Yew Lin (2004), and Yvette Graham (2015) underscore the prevalence of these metrics in the field of text summarization. However, our hypothesis was that these metrics might have limitations. To bolster this argument, studies by Schluter, Natalie (2017) and Liu, Feifan, and Liu, Yang (2008) highlight specific shortcomings of these metrics.

In our continued research aimed at determining the appropriate algorithm for word comparison, we came across a study conducted by Sağlam, B. & Özek, F. (2023). This research delves into phonetic distinctions between Azerbaijani, Turkish, and Turkmen Turkish. It utilizes the Levenshtein Distance Algorithm to quantify the linguistic divergence among these languages. Through this examination, we observed that a distance algorithm can indeed effectively compare and measure differences between distinct languages.

In additional research, we encountered another study conducted by Santoso, Puji et al. (2019). This study focuses on comparing the Levenshtein Algorithm with the Damerau-Levenshtein Algorithm, an improved version of the Levenshtein Algorithm, for measuring word distances in Indonesian fairy tales. It demonstrates that employing the Damerau-Levenshtein Algorithm produces superior results compared to the Levenshtein Algorithm. Relevant research employing these algorithms includes studies by L. Yujian and L. Bo (2007), as well as Youness Chaabi and Fadoua Ataa Allah (2022).

In the realm of machine translation, there exist two primary approaches: statistical and rule based. Since the statistical approach demands more extensive data and resources, in contrast to the rule-based approach, our ultimate focus lies in determining whether phonological errors can be effectively corrected through the rule-based method. In pursuit of this question, a noteworthy thesis study conducted by Çalış, T. (2017) emerges. This study compares Google Translate with the syntactic transfer method for translation. The findings highlight that employing the syntactic transfer method yields superior results compared to solely relying on statistical approach of Google Translate. Moreover, the study suggests that combining these two approaches can lead to even more

significant improvements in translation quality, aligning well with our ultimate intention of achieving better results in addressing phonological errors.

3. Examples of Phonological Errors

There are few errors that can occur in natural language. These can be categorized as morphological, syntactic, semantic, and phonological errors. Morphological errors involve mistakes in the use of suffixes. Syntactic errors arise from the improper placement of words within a sentence. Semantic errors disrupt the meaning of a sentence.

The topic of this article, phonological errors, stems from spelling mistakes. There can be two reasons for phonological errors: either due to a lack of knowledge or as a result of typographical errors while using a keyboard. The reason why the most preferred metrics are not applied to the level of phonological errors might be their typical use in extracting article or book summaries. However, this situation leads to the inability to extract data from large datasets on social media platforms like Twitter. To understand the importance of phonological errors in Turkish, let's provide a few examples of phonological errors:

Table 1. Examples of phonological errors both in Turkish and English.

	Word (Turkish)	Word (English)
Normal Form	muayene	examination
Duplication Error	muayenee	eexamination
Deletion Error	muayne	exmination
Substitution Error	muatene	ezamination
Addition Error	muayenet	exambination
Transposition Error	muayeen	examinatino

Duplication error: Occurs when some letters are repeated unnecessarily in a word.

Deletion error: Involves the omission of some letters from a word.

Substitution error: Refers to the act of writing a different letter instead of the correct one. Often caused by keyboard mistakes.

Addition error: Involves adding an extra letter to a word that is not normally present.

Transposition error: Occurs when two letters in a word are accidentally switched or transposed.

4. Unveiling the Impact of Phonological Errors

Whether phonological errors are worth considering or not, we conducted a comprehensive search for previous studies and conducted our own research. We then compared the collected data and arrived at a conclusion.

4.1 Collected Data from Related Works

The study conducted by Uzdu Yıldız, F., & Çetin, B. (2020) based on people from various backgrounds and different countries. To prevent any confusion, in their study, lexical errors correspond to semantic errors. The Table 2 is directly from the article written by Uzdu Yıldız, F., & Çetin, B. (2020).

Table 2. Average error rates of non-native Turkish learners (Uzdu Yıldız, F., & Çetin, B. (2020)).

Level	Grammar	Syntax	Lexical	Spelling and punctuation
Average	33.98	11.97	13.93	40.91

We conducted calculations on average data to determine penalty rates for error detection, aiming to improve our algorithm in future studies. Detecting lower-rated errors proves to be more challenging than identifying higher-rated errors. Thus, we made the decision to impose more substantial penalties on the algorithm as the error rate decreases. The corresponding penalty rates are presented in Table 3:

Table 3. Penalty rates based on the data of non-native people.

Syntax error penalty rate	≈ 39.93%
Lexical (Semantic) error penalty rate	≈ 34.31%
Grammar error penalty rate	≈ 14.07%
Spelling error penalty rate	≈ 11.69%

4.2 Twitter Research

While Uzdu Yıldız, F., & Çetin, B. (2020)'s study indicates that spelling errors are the second most common type, we decided to conduct our own research on Turkish people to verify this claim. To achieve this, we collected our own data from the social media platform Twitter, as it contains numerous instances of mistyped words and other types of errors.

We get 38,750 random words from Twitter (April 9th, 2021) and compared their grammatical and spelling error rates. The results are presented in Table 4.

Table 4. The Twitter data presents error rates among Turkish users.

	Count	Rate
Total word	38750	-
Total error	2558	-
Spelling error	1599	62.51%
Grammar error	959	37.49%

4.3 Comparing Data

According to the study conducted by Uzdu Yıldız, F., & Çetin, B. (2020), the spelling error rate among non-native individuals is 40.91% of the total errors, while our study on native speakers found a spelling error rate of 62.51% of the total errors.

Based on the data collected from Twitter, it is evident that even native Turkish writers tend to make spelling errors more than any other types of errors. Our findings align with the study conducted on non-native individuals, leading us to reaffirm our earlier conclusion: it is necessary to account for phonological errors.

After deciding to consider phonological errors, we need to determine the most suitable algorithm for enhancing the current metrics.

5. The Algorithms for Handling Phonological Errors

There are numerous edit-based and token-based algorithms that can be applied to evaluate various metrics across different fields. These algorithms have proven to be useful in many domains, such as the medical field, where the Smith-Waterman algorithm is employed for local sequence alignment. Similarly, in business applications, the use of MLIPNS would be considered a wise choice. The range of applications is extensive, but when it comes to correcting phonological mistakes, two prominent options stand out: the Levenshtein Algorithm and the Damerau-Levenshtein Algorithm.

5.1 Levenshtein and Damerau-Levenshtein Algorithms

The Levenshtein algorithm, also known as the Levenshtein distance or edit distance, is a string metric used to quantify the dissimilarity between two sequences, typically words or strings. It provides a way to calculate the minimum number of single-character operations required to transform one sequence into another. These operations can be insertions, deletions, or substitutions of individual characters.

The Damerau-Levenshtein algorithm is an extension of the Levenshtein algorithm. This algorithm would be described as a string metric that calculates the minimum number of single-character operations required to transform one string into another. These operations include insertions, deletions, substitutions, and transpositions of adjacent characters. It was proposed by Fred J. Damerau in 1964 and later refined by Vladimir Levenshtein in 1965.

Table 5 illustrates the Levenshtein and Damerau-Levenshtein distances between two words.

Table 5. Comparing Levenshtein and Damerau-Levenshtein Algorithms.

Index	Correct Word	Wrong Word	Levenshtein Distance	Damerau-Levenshtein Distance
1	**Kitap**	**Ktiap**	**2**	**1**
2	Merhaba	Merhbaa	2	1
3	Programcılık	Prgortamcılk	4	3
4	Okul	Ookul	1	1
5	Evim	Evimm	1	1
6	Kitaplık	Kitaplkı	2	1
7	Yemek	Yemel	1	1
8	**Telefon**	**Telefmo**	**2**	**2**
9	Bilgisayar	Biglisaayr	4	2
10	Televizyon	Televizyn	1	1
11	Paylaşımcı	Direksiyon	10	10

As observed in Table 5, the Damerau-Levenshtein distance, which considers the possibility of transposition, is always less than or equal to the Levenshtein distance. Let's examine the following examples from the table:

(1) Kitap – Ktiap: To calculate the Levenshtein distance, we need to first remove the "i" in the word "Ktiap" and then add an "i" between "K" and "t." However, for Damerau-Levenshtein distance, it only requires one calculation, which is the transposition of the letters "i" and "t".

(8) Telefon – Telefmo: To compute the Levenshtein distance, we must remove the letter "m" and add the letter "n" at the end, which requires two operations. Similarly, for the Damerau-Levenshtein distance, transposition cannot be applied here since the letter "m" does not exist in the original word. Therefore, the calculations for both Levenshtein distance and Damerau-Levenshtein distance are the same, involving deletion and addition operations.

As demonstrated in Table 5, the Damerau-Levenshtein algorithm proves to be more sensitive when errors stem from phonological mistakes. Its inclusion of transposition probability enables more efficient detection of keyboard errors. Hence, the Damerau-Levenshtein algorithm is better equipped to handle and correct phonological errors effectively.

6. Proposed Enhancement of Evaluation Metrics

We hypothesized that combining two metrics could improve the quality of summarized or translated text. To validate this hypothesis, we chose to implement the Damerau-Levenshtein distance algorithm, which accounts for

letter transpositions, into one of the most used evaluation metrics, ROUGE, and evaluating whether our initial assumption held true.

ROUGE-1 measures the number of matching one-grams between the model-generated text and a human-produced reference.

6.1 Default ROUGE-1 Metric

Consider the reference R and the candidate summary C:

R: “Bugün ktiap okudum”

C: “Ben bugün kitap okudum”

ROUGE-1 precision can be computed as the ratio of the number of unigrams in C that appear also in R, (that are the words “bugün” and “okudum”), over the numbers of unigrams in C.

ROUGE-1 precision = 2/4 = 0.5

ROUGE-1 recall can be computed as the ratio of the number of unigrams in R that appear also in C (that are the words “bugün”, “okudum”), over the number of unigrams in R.

ROUGE-1 recall = 2/3 = 0.66

ROUGE Calculation (F1 Score) = 2 * (precision * recall) / (precision + recall)

Default ROUGE-1 F1 Score = 2 * (0.33) / (1.16) = 0.57

6.2 Enhancing ROUGE-1 Metric with Damerau-Levenshtein Algorithm

R: “Bugün ktiap okudum”

C: “Ben bugün kitap okudum”

Damerau-Levenshtein distance (D-L distance) of “kitap” and “ktiap” is 1. (Transposition of “i” and “t”)

Normalized D-L distance (adjusting distance between 0 and 1) can be computed as the ratio of D-L distance over the length of the longest one between two words (here, “kitap” and “kitpa” has the same length, which is five).

Normalized D-L Distance = 1/5 = 0.2

The precision of ROUGE-1 can be calculated by taking the count of unigrams in C that also appear in R (specifically, the words "bugün" and "okudum"), and then adding the normalized Damerau-Levenshtein similarity (where 1 minus the normalized D-L distance is used), and finally dividing this sum by the total count of unigrams in C.

Normalized D-L Similarity = 1 - 0.2 = 0.8

ROUGE-1 precision = (2 + 0.8) / 4 = 0.7

ROUGE-1 recall can be computed by taking the number of unigrams in R that appear also in C (that are the words “bugün”, “okudum”), and then adding the normalized Damerau-Levenshtein similarity, and finally dividing this sum by the number of unigrams in R.

ROUGE-1 recall = (2 + 0.8) / 3 = 0.93

Enhanced ROUGE-1 F1 Score = 2*(0.7 * 0.93) / (0.7 + 0.93) = 0.8

While the default ROUGE metric indicates a 57% similarity between these two sentences, our improved ROUGE metric, which incorporates the Damerau-Levenshtein algorithm, demonstrates an 80% similarity. Given that the reference and candidate sentences are very close in terms of similarity, the fact that the enhanced metric yields a higher similarity percentage compared to the default one holds significant value for tasks such as summarization and machine translation, particularly when dealing with datasets containing typographical errors

Table 6. Comparing default ROUGE-1 and Enhanced ROUGE-1 Scores.

Index	Reference Text	Candidate Text	ROUGE-1 Score	Enhanced ROUGE-1 Score
1	Hava ***cok*** güzel	Hava çok güzel	0.66	0.87
2	Kitap ***duygu*** anlıyor	Kitap duygularını anlıyor	0.66	0.85
3	Tavuk döner tercih ederim	Tavuk döner tercih ederim	1	1
4	Geç ***kalmmak icin*** uyanıyorum	Geç kalmamak için uyanıyorum	0.5	0.96
5	***Kumanadyı kaybettm***	Kumandayı kaybettim	0	0.89
6	Internet sorunları ***yaşyorum***	Internet sorunları yaşıyorum	0.66	0.96
7	Hangi filmi ***izliyeceme*** karar veremedim	Hangi filmi izleyeceğime karar veremedim	0.8	0.95
8	Yeni menüsünü inceledim	Bugün hava kötü	0	0
9	***Sopr yaparkene dkkatli olamak önemldir***	Spor yaparken dikkatli olmak önemlidir	0	0.85
10	***Seayahat*** etmeyi ***terçih edreim***	Seyahat etmeyi tercih ederim	0.25	0.89

Table 6 presents the scores for ROUGE-1 and the improved ROUGE-1 that uses the Damerau-Levenshtein method on several sentences. Words with errors are highlighted in bold and written in italics for easy identification. Once more, the enhanced metric demonstrates significantly higher similarity scores, particularly in cases where mistakes are attributed to phonological errors.

In subsequent studies, if it were integrated into a word-based system, achieving a higher similarity score when comparing two distinct yet highly similar words would be more advantageous. This enhancement would enable the system to identify wrong words more effectively, subsequently leading to the discovery of the correct word through the utilization of this elevated similarity score.

7. Conclusion

This paper has highlighted the significance of considering phonological errors in Natural Language Processing (NLP) and machine translation evaluations, particularly when dealing with data sourced from social media platforms like Twitter. Numerous studies on this subject consistently indicate that phonological errors rank among the most common mistakes made by both native and non-native writers.

To address this issue, we proposed the integration of edit distance metrics, specifically the Damerau-Levenshtein algorithm alongside existing metrics, such as ROUGE and BLEU, to improve their performance. Subsequently, we demonstrated that this improved metric yields more realistic results.

Overall, this paper aims to contribute to the advancement of evaluation metrics in NLP and machine translation domains by addressing the challenges posed by phonological errors. By improving the accuracy and reliability of evaluation metrics, we can enhance the quality of text summarization and machine translation outputs. Ultimately, the goal is to foster more precise and dependable NLP applications and language translation tools in the future.

8. References

Uzdu Yıldız, F., & Çetin, B. (2020). Errors in written expressions of learners of Turkish as a foreign language: A systematic review. Journal of Language and Linguistic Studies, 16(2), 612-625. Doi: 10.17263/jlls.759261

Sağlam, B. & Özek, F. (2023). Levenshtein Uzaklık Algoritmasına Göre Azerbaycan, Türkiye ve Türkmen Türkçeleri Arasındaki Fonetik Uzaklık. Asya Studies-Academic Social Studies / Akademik Sosyal Araştırmalar, 7(Special Issue / Özel Sayı 3), 45-64.

Çalış, T. Sözdizimsel Aktarıma Dayalı Makale Çevirisi Yüksek Lisans Tezi, Trakya Üniversitesi, 2017

Stanley, Theban & Hacioglu, Kadri. (2011). Statistical Machine Translation Framework for Modeling Phonological Errors in Computer Assisted Pronunciation Training System.

L. Yujian and L. Bo, (2007) "A Normalized Levenshtein Distance Metric," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1091-1095, doi: 10.1109/TPAMI.2007.1078.

Santoso, Puji, et al. (2019) “Damerau levenshtein distance for indonesian spelling correction,” J. Inform 13.2: 11.

Youness Chaabi, Fadoua Ataa Allah, (2022), “Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram,” Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 8, Part B, Pages 6116-6124, ISSN 1319-1578.

Schluter, Natalie. (2017). The limits of automatic summarisation according to ROUGE. 41-45. 10.18653/v1/E17-2007.

Liu, Feifan & Liu, Yang. (2008). Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries.. 201-204. 201-204. 10.3115/1557690.1557747.

Baykara, B., Güngör, T. (2023). Morphosyntactic Evaluation for Text Summarization in Morphologically Rich Languages: A Case Study for Turkish. In: Métais, E., Meziane, F., Sugumaran, V., Manning, W., Reiff-Marganiec, S. (eds) Natural Language Processing and Information Systems. NLDB 2023. Lecture Notes in Computer Science, vol 13913. Springer, Cham.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.