UDC 004.93

WELER: A COMPLEX METRIC FOR TEXT QUALITY ASSESSMENT

Dumyn A. R. – Post-graduate student at Lviv Polytechnic National University, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0003-2111-2899.

Shakhovska N. B. – Dr. Sc., Professor, Rector at Lviv Polytechnic National University, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0002-6875-8534.

ABSTRACT

Context. Assessing text quality is essential for reliable AI that processes language. In ASR, it reflects how faithfully speech becomes text; in OCR, how accurately images yield text; and in NLP, how correct and coherent outputs are.

Objective. The goal of the work is the creation of a complex metric for text quality assessment.

Method. Classic metrics WER and CER are narrow: they capture only lexical edits, weigh all changes equally, ignore context and semantics, and often skip punctuation and case, masking readability issues and error types. We propose WELER, a hybrid metric that blends weighted WER and CER with a semantic component based on contextual embeddings to measure meaning preservation. Weights can be set manually or learned (e.g., via PCA), adapting the metric to ASR, OCR, or NLP tasks. Key challenges include computational cost, choosing optimal weights through correlation with human judgments, and the need for high-quality reference data. Proposed WELER metric integrates accurate word and character level error counting, using Levenshtein distance as a basis, with advanced semantic similarity methods based on contextual embeddings. This allows WELER to take into account not only what was incorrectly recognized, but also how much this error affects the meaning and understanding of the text. The inclusion of self-adjusting weights depending on the text category is a key feature of WELER, which allows adapting the metric to the specific requirements of different applications and domains, prioritizing those aspects of quality that are most critical for a particular task.

Results. Proposed WELER metric is an alternative solution in this direction. It integrates accurate word and character level error counting, using Levenshtein distance as a basis, with advanced semantic similarity methods based on contextual embeddings.

Conclusions. WELER, like all metrics based on reference data, relies on accurate and consistent human-verified transcriptions. Errors in the reference data can affect the accuracy of the assessment. Therefore, for complex metrics, the quality and representativeness of these data are especially important, since semantic and weighted errors are much more sensitive to the quality of the annotation than simple word counts.

KEYWORDS: snatural language processing, automatic speech recognition, text quality assessment, WER, CER, WELER.

ABBREVIATIONS

ASR is automatic speech recognition;
OCR is optical character recognition;
NLP is natural language processing;
WER is word error rate;
CER is character error rate;
NLI is Natural Language Inference;
WELER is weight-based error rate.

NOMENCLATURE

S is a number of substitutions;
D is a number of deletions (word omitted);
I is a number of insertions (extra word);
N is a number of words in the template;
S_c is a number of character substitutions;
D_c is a number of omitted characters;
I_c is a number of extra characters;
N_{char} is a total number of characters in the reference text;
S_{word} is a number of word-level substitutions obtained from the Levenshtein alignment;
D_{word} is a number of word-level deletions obtained from the Levenshtein alignment;
I_{word} is a number of word-level insertions obtained from the Levenshtein alignment;
N_{word} is a total number of words in the reference text;
w_S is an adjustable weighting factors for S_{word};
w_D is an adjustable weighting factors for D_{word};
w_L is an adjustable weighting factors for $\underline{I_{word}}$;
S_{char} is a numbers of character-level substitutions;
D_{char} is a numbers of character-level deletions;
I_{char} is a numbers of character-level insertions;
W_{wer} is a weighted word error rate;
W_{cer} is a weighted error rate in symbols;
$SemErr$ is a semantic error indicator;
α is a normalized W_{wer};
β is a normalized W_{cer};
γ is a normalized $SerErr$;
Emb_{inut} is a vector of embeddings of the reference text;
Emb_{output} is a vector of recognized text embeddings.

INTRODUCTION

Text quality assessment plays a crucial role in functionality of systems that work with text data. Especially artificial intelligence. In areas such as ASR, OCR, and NLP, accurate text quality assessment is fundamental to determining the effectiveness of the system. For example, in ASR, it determines how accurately spoken words are converted to text, and in OCR, the correctness of text extraction from images. In NLP tasks related to text generation, processing, and quality assessment helps determine the correctness and coherence of the obtained result [1].

Traditionally, WER and CER metrics have been widely used to assess text quality. These metrics are quantitative measures of how well extracted data matches a reference, usually expressed as a percentage [1]. WER and CER are derived from the Levenshtein distance, which defines the minimum number of editing operations (replacements, deletions, or insertions) required to transform one sequence into another, WER works at the word level, counting errors resulting from word substitutions,

DOI 10.15588/1607-3274-2026-1-7

deletions, or insertions. CER, on the other hand, focuses on character-level accuracy, counting similar errors for individual characters. These metrics have become the standards for evaluating the performance of ASR and OCR models, due to their relative ease of calculation and straightforward interpretation.

However, despite their popularity, WER and CER have significant limitations that hinder their ability to fully reflect text quality. They mainly measure lexical accuracy at a superficial level, without taking into account semantic similarity, word importance, grammatical correctness, or the impact of punctuation errors [2]. For example, WER penalizes minor spelling errors as well as errors that completely change the meaning of a sentence, leading to discrepancies between automatic evaluation and human perception of text quality.

Due to these restrictions, there was a need to develop a more comprehensive Evaluation metric. Such metrics should go beyond simply counting errors at a superficial level, integrating a deeper understanding of semantics and context, and allowing for differential weighting of different types of errors according to their impact on the overall quality and understandability of the text.

Object of the research: automated text quality assessment in AI systems (ASR, OCR, and NLP) based on reference transcriptions/texts.

Subject of the research: the hybrid WELER metric – its components (weighted WER, weighted CER, and a semantic error indicator from contextual embeddings), normalization choices, and data-driven weighting (e.g., PCA-based) for different task categories.

This study aims to develop a hybrid metric for assessing the quality of generated text based on the use of basic metrics and taking into account semantic and weighted approaches. The development of such a hybrid metric reflects the growth of the assessment of artificial intelligence systems, moving from a purely quantitative error count to a more qualitative, human-perceived understanding of "correctness". This transition is an intermediate stage for creating more reliable and user-oriented artificial intelligence systems.

1 PROBLEM STATEMENT

We study the problem of learning a composite error metric (WELER) for text sequences. For each example we have a reference text R, a system output H (from ASR/OCR/NLP), domain metadata m (e.g., language, task type, noise), and a human quality score y in [0,1]. From (H, R) we compute three normalized component errors in [0,1]: a weighted word-level edit error, a weighted character-level edit error, and a semantic difference (for instance, one minus cosine similarity of sentence embeddings or an NLI-based score). WELER is a single score that combines these components with non-negative weights that sum to one; the weights may be fixed or predicted from m so the metric adapts to the domain. Lower WELER means fewer errors (you may invert it if you prefer "higher is better").

The goal is to learn the weights (and, if needed, normalization/adaptation parameters) so that 1 − WELER matches the human scores y as closely as possible on the dataset, typically by minimizing an average loss such as MAE or MSE, or by maximizing rank agreement. We require monotonicity (if any component error increases, WELER cannot decrease) and, optionally, stability across groups (e.g., languages). For decision making, a threshold on 1 − WELER can be chosen to label outputs as acceptable or not, tuned to maximize a target metric like F1.

2 REVIEW OF THE LITERATURE

The issue of developing additional metrics for assessing the quality of work of artificial intelligence methods with the tasks of natural language processing, optical text recognition [3], audio-to-text conversion [3], plagiarism detection [4], etc., is widely studied in academic circles. In particular, the authors [5] propose the *H_eval* metric, which combines semantic correctness with the traditional WER error. This metric works much faster than BERTScore and has a strong correlation with NLP tasks. In [6], the metric was introduced SemDist is the distance between the embedding vectors for the reference and generated texts obtained through RoBERTa. Semantic evaluation has been shown to be more relevant for natural language understanding tasks than simple WER.

SeMaScore metric [7], developed for ASR, integrates traditional error metrics with a robust measure of semantic similarity. It then computes segment scores using contextual embeddings. and cosine similarity. This approach demonstrates how traditional methods can serve as the basis for new, more sophisticated estimates. In the work [8] proposed a metric that combines NLI score, semantic and phonetic similarity. The proposed metric achieves a correlation of with human intelligibility judgments on data with language features (dysarthria and dysphonia discourse). Article [9] proposes the BERTScore metric, which calculates the similarity between texts based on contextual embeddings and shows that this approach correlates better with human evaluation than traditional n-gram metrics.

WER [10] is one of the most common metrics for evaluating text recognition accuracy. It measures the percentage of incorrect words in the generated text (hypothesis) compared to the reference text [1]. The formula for calculating WER is given below

$$WER = \frac{S + D + I}{N}. \qquad (1)$$

It is important to note that the WER value can exceed 1 or 100%, especially in cases where the number of insertions significantly exceeds the number of words in the reference text. Lower WER values indicate higher accuracy. WER is a particularly valuable tool for evaluating ASR and OCR performance, particularly in scenarios where the emphasis is on free-form text recognition. This

© Dumyn A. R., Shakhovska N. B., 2026
DOI 10.15588/1607-3274-2026-1-7

includes applications such as document digitization, handwriting transcription, multilingual text recognition, book and manuscript archiving, and automating the transcription of meeting notes or legal documents.

Before calculating the WER, a text pre-processing process known as normalization is usually applied to ensure a fair comparison. This involves converting all text to lowercase, removing punctuation, and standardizing numbers (e.g., "5 доларів". instead "5 $") and expansion of abbreviations.

WER is based on the Levenshtein distance, which works at the word level. This relationship means that WER, as a metric, is derived from an algorithm that counts the minimum number of edit operations to transform one sequence of words into another. This use of Levenshtein distance for WER results in all word-level errors (substitutions, deletions, insertions) receiving the same weight. For example, the substitution of the word "плисти" to "плести", which is a homophone or "кит" to "кат", which is a completely different word, will have the same error weight, despite their different impact on semantic meaning. This insensitivity to semantic nuances is a direct consequence of the underlying algorithm and constitutes a significant limitation.

Furthermore, the need for extensive text normalization before calculating WER, which includes removing punctuation and ignoring capitalization, indicates that WER in its standard form is not a holistic measure of text quality. Rather, it measures lexical relevance after preprocessing. This means that important aspects of text quality, such as formatting and grammatical correctness, which are often removed during normalization, are effectively ignored by standard WER. This highlights the need for a new metric that could explicitly include these aspects or allow for their weighted inclusion.

CER [10] is another key metric for assessing recognition accuracy, which, unlike WER, focuses on accuracy at the character level. CER measures how many characters in the source data differ from the reference data, taking into account substitutions, deletions, and insertions of characters relative to the total number of characters in the reference data. The formula for CER is identical to the WER formula, but applied to characters, and is given below

$$CER = \frac{S_C + D_C + I_C}{D_C} . \quad (2)$$

CER provides a more detailed error assessment than WER. For example, minor typographical errors, such as "опрацюваня" instead of "опрацювання", will result in a full word substitution error in WER, but only a single character substitution error in CER. This allows for a more accurate assessment of systems where even small errors can have significant consequences, such as in coding, formal documents, or specialized terminology.

CER is particularly useful in scenarios where word boundaries may be absent, such as numeric data, alphanumeric codes, or where accurate character recognition is critical, such as serial numbers, financial data, passport numbers. It is also applicable for languages that do not have clear spaces between words, such as Chinese.

Although CER offers a more granular error estimate, it still has a fundamental limitation of WER – the lack of semantic understanding. Even a single character error can significantly change the meaning of a word or sentence, which CER alone does not capture. This means that while CER is more accurate in indicating where an error has occurred, it does not assess the impact of that error on the value, which is a critical aspect of the quality of the estimate.

The usefulness of CER in specific domains, such as numeric data, codes, or languages without clear word boundaries, suggests that a truly robust text quality metric should be adaptive or configurable to prioritize different levels of error (symbolic vs. lexical) depending on the application requirements. This suggests that the coefficients for WER/CER in the hybrid metric will allow it to be adapted to these diverse needs.

Levenshtein distance [12], also known as edit distance, is a metric that quantifies the difference between two strings. It calculates the minimum number of character-level editing operations (insertions, deletions, or substitutions) required to transform one string into another. The algorithm was proposed by Vladimir Levenshtein in 1965, and it quickly became the basis for tasks such as spell checking, DNA sequencing, and duplicate text detection. The implementation of Levenshtein distance is based on dynamic programming.

As already mentioned, the Levenshtein distance is the main algorithm for calculating WER and CER. It is used to align the recognized text with the reference text and identify minimal editing operations at the word or character level.

There are variants of edit distance that extend the basic Levenshtein distance. For example, the Damerau-Levenshtein distance includes transpositions, i.e., the rearrangement of two adjacent characters, as a single editing operation, which allows for a more accurate representation of some types of errors, such as those that occur in typing. Weighted edit distance allows for different weights to be assigned to insertion, deletion, and replacement operations. These extensions demonstrate that even within the edit distance paradigm, researchers have recognized that not all errors are equally important, which in turn creates the prerequisites for the application of additional weights.

Levenshtein distance is purely lexical a comparison metric that is not intended for semantic or contextual understanding. Its applications, such as spell checking or plagiarism detection, are primarily concerned with string similarity rather than semantic equivalence. However, despite this, Levenshtein distance can serve as a structural framework for integrating higher-level semantic information. For example, in metrics such as SeMaScore, Levenshtein distance is used for initial segment alignment. This allows for identifying relevant parts of the text, even if

DOI 10.15588/1607-3274-2026-1-7

they contain errors, and then applying semantic comparison to these aligned parts. This approach is a way of combining lexical and semantic evaluation.

To overcome the limitations of traditional metrics such as WER and CER, researchers have developed more sophisticated approaches that take semantic meaning into account and allow for weighting of different types of errors [13].

Semantic similarity metrics move away from simply counting lexical errors and focus on preserving the meaning of the text. They use contextual embeddings to represent words and sentences in a multidimensional space where semantically similar texts are located closer together. Cosine similarity is commonly used to measure this proximity.

BERTScor metric uses contextual embeddings from pre-trained language models such as BERT or RoBERTa to measure semantic similarity between texts. It calculates precision, completeness, and F1-measure by aligning tokens based on vector similarity. BERTScore correlates better with human judgment than traditional n-gram metrics such as BLEU, ROUGE, because it is able to recognize when different words or phrases convey the same meaning, even if their surface forms are different. This is a direct solution to the problem of lack of semantic understanding in WER/CER [14].

While semantic metrics offer significant benefits, they also have their challenges. Generating contextual embeddings can be computationally intensive, especially for large datasets. Furthermore, the reliance on deep learning models can make them less interpretable than traditional metrics. This suggests a trade-off between the enrichment of the score and its practical applicability and interpretability.

The concept of weighting different types of errors is not new. As early as 1990, Hunt proposed a weighted WER, where substitutions were given a weight of 1 and deletions and insertions were given a weight of 0.5 [15]. This early example shows the recognition that not all errors are of equal severity.

Composite metrics combine multiple evaluation metrics into a single score, often using weighted averages. This approach allows for a more holistic assessment by balancing different aspects of performance. For example, in studies evaluating the user experience of chatbots, composite metrics integrate usability, engagement, and error rate. Weights for such metrics can be derived empirically, for example, using principal component analysis (PCA), based on empirical patterns in user interaction.

Using methods such as PCA to determine weights provides a scientifically sound approach to assigning importance to different components of a metric, moving beyond arbitrary assignments. If a composite metric can balance usability, engagement, and error rate for chatbots, a similar framework can be applied to text quality assessment. This allows for the flexibility to tailor the metric using weights to meet the specific priorities of different ASR, NLP, or OCR use cases. For example, character accuracy might be a priority for serial number recognition in banking, while word accuracy might be a priority for meeting transcription, and semantic relevance might be a priority for conversational AI.

Despite their widespread use, traditional error rate metrics have several critical limitations that reduce their ability to provide a complete and accurate assessment of text quality [10]. Using WER, CER, and Levenshtein distance alone or in simple combinations has significant limitations. In particular, these metrics are characterized by a lack of semantic sensitivity, since WER and CER treat all errors the same, regardless of their impact on meaning. For example, replacing "погода" with "погоди" may be only a single word error, but completely change the meaning of the sentence. WER does not distinguish "Я люблю фрукт" from "Я люблю фрукти".

Another disadvantage is word order sensitivity, as WER and CER do not take word order into account, which can be critical for NLP tasks. For example, "собака вкусила хлопчика" and "хлопчика вкусила собака" will have a high WER, even though they have not different contexts.

In particular, traditional error rate metrics are insensitive to the semantic meaning and importance of words, WER and CER treat all errors equally, regardless of their impact on the meaning or importance of the word, or ignore specialized terminology. For example, a typographical error such as "опроцювання" instead of "опрацювання" has the same WER as "день" instead of "пень", despite their radically different semantic consequences. Similarly, "самоповага" has the same WER as "само повага". This means that the metrics do not distinguish between critical errors that change meaning from minor errors that do not affect comprehension. The main reason for this is that the underlying Levenshtein distance algorithm on which WER and CER are based assigns the same weight to each editing operation, resulting in a uniform weighting of all types of errors. This equal weighting makes WER and CER poor indicators for human perception of quality, as people implicitly weigh errors differently depending on their impact.

Standard WER calculations often normalize text by removing punctuation and ignoring capitalization, thereby ignoring these critical aspects of text quality and readability. Although punctuation and capitalization are important for readability, WER does not take this information into account. Grammatical errors are also not typically directly evaluated by WER/CER. This means that if punctuation is removed during normalization, WER cannot account for punctuation errors, even if they are important for a particular application.

Problem of text normalization and different lengths transcriptions is another limitation of WER and CER. Inconsistent text normalization, for example, "5 доларів" instead "5 $" may artificially inflate the WER. In addition, WER depends on the length of the transcription, as longer texts have more room for errors. The WER value may exceed 1 or 100 %, which may be counterintuitive to the "error rate".

© Dumyn A. R., Shakhovska N. B., 2026
DOI 10.15588/1607-3274-2026-1-7

3 MATERIALS AND METHODS

To overcome the limitations of traditional metrics and integrate the advantages of semantic and weighted approaches, a new hybrid metric is proposed, namely an improved error rate based on the Levenshtein distance and the use of WELER weights.

WELER is designed as a multi-layered approach that provides a comprehensive assessment of text quality. It integrates quantitative error metrics WER and CER with qualitative semantic understanding within customizable and weighted structure.

The proposed metric uses the Levenshtein distance for reliable alignment at both the word and character levels. This will ensure error counting and will serve as a structural basis for higher-level analysis.

To provide semantic and contextual understanding, WELER includes a semantic similarity component to assess the conveyed meaning, which is a major limitation of traditional WER/CER. This component assesses how well the generated text preserves the intended meaning even if the lexical forms differ.

Entering weighting factors will allow users to prioritize different aspects of text quality, such as character accuracy, word accuracy, semantic relevance, depending on their specific application and domain requirements. This solves the problem of "evenly weighting" errors in WER/CER. The weights in this approach can be either manually entered by the user or calculated using the principal component analysis method or based on empirical patterns in user interaction. The goal of using the principal component analysis method is to obtain a weighted combination of error rate metrics, where the weights will be determined automatically, without manual adjustment.

WELER is a composite measure, potentially expressed as a weighted sum of normalized error components or a weighted average of the accuracy components. The goal is a single, interpretable measure that reflects the overall quality of a text.

Weighted word error rate W_{wer} includes differentiated penalties for substitutions, deletions, and insertions at the word level

$$W_{wer} = \frac{w_S \cdot S_{word} + w_D \cdot D_{word} + w_I \cdot I_{word}}{N_{word}}. \qquad (3)$$

Similarly to W_{WER}, but at the symbol level, (4) is calculated. This will allow get a more detailed error assessment

$$W_{cer} = \frac{w_S \cdot S_{char} + w_D \cdot D_{char} + w_I \cdot I_{char}}{N_{char}}. \qquad (4)$$

The third indicator used in the calculation of *WELER* is semantic error indicator *SemErr*. It is obtained from a semantic similarity metric, for example, BERTScore, a modified component of SeMaScore or similar components (depending on the specifics of the task, language, etc.), normalized to represent "error" rather than similarity (6). To calculate the semantic error indicator, it is advisable to use the semantic distance mechanism. Semantic distance is defined as the distance between pairs of reference and hypothetical texts in the space of embeddings at the sentence level, usually using models such as RoBERTa, and cosine similarity (5)

$$SemDist\,[E,H] = 1 - \frac{e \cdot h}{\|e\| \cdot \|h\|}, \qquad (5)$$

where *E* and *H* are vector representations of reference and hypothetical text

$$SemErr = \frac{1 - \dfrac{Emb_{input} \cdot Emb_{output}}{\|Emb_{input}\| \cdot \|Emb_{output}\|}}{2}, \qquad (6)$$

where Emb_{inut} – vector of embeddings of the reference text; Emb_{output} – vector of recognized text embeddings.

The general WELER formula is to use a weighted combination of the above components (7).

$$WELER = \alpha \cdot W_{wer} + \beta \cdot W_{cer} + \gamma \cdot SemErr, \qquad (7)$$

where α, β, γ are global weighting factors, the sum of which is 1, allowing to prioritize word accuracy, character accuracy or semantic correspondence. These factors can be adjusted depending on the specific of application.

Table 1 lists the components of the proposed hybrid metric.

Table 1 – WELER components and weighing scheme

Metric component	Granularity	Customizable weighting factors	Purpose/Benefit
Weighted word error rate W_{wer}	Word level	w_S, w_D, w_I	Lexical accuracy, flexible penalization of different types of word errors
Weighted error rate in symbols W_{cer}	Character level	w_S, w_D, w_I	Detailed error detection, sensitivity to small typos, importance for numerical data.
Semantic error indicator *SemErr*	Semantic Sentence Level	In the internal scales of the model	Preserving meaning, understanding context, taking into account the impact of low-level errors on semantics.
General *WELER*	General	α,β,γ	Comprehensive quality assessment, setting priorities between lexical accuracy and semantic fidelity.

© Dumyn A. R., Shakhovska N. B., 2026
DOI 10.15588/1607-3274-2026-1-7

WELER value ranges from 0 to 1, where 0 is a complete match between the reference and generated text.

The weights α, β, γ in this approach can be either manually entered by the user depending on the task at hand, or calculated using the principal component analysis method. For example, for optical serial number recognition tasks, the coefficient β will be high to prioritize character accuracy. For general voice recognition, α and γ can be balanced to take into account both word accuracy and meaning preservation.

Below is a method for calculating weights using the principal component analysis (PCA) method. The purpose of PCA is to obtain a weighted combination of the above metrics, where the weights will be determined automatically. The principal components are the directions with the greatest variance in the data. Accordingly, this approach allows us to understand which metrics contribute the most to the variability, determines the relative weight of each metric. PCA also allows us to take into account the task category, for example, recognition of short commands, dialogues, long texts, technical documents, etc., and to distribute weights for WER, CER and semantic similarity relative to the category, since the task category can significantly affect the distribution of weights.

The first step for the principal component analysis method is to construct an observation matrix (8) based on data for N examples and d metrics, in this case d=3, since 3 metrics are used.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} \end{bmatrix}, \tag{8}$$

where $x_{i1} - W_{wer}; x_{i2} - W_{cer}; x_{i3} - SemErr.$

Normalization can be performed to avoid the impact of differences in metric scales (9)

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, j = 1,2,3. \tag{9}$$

The next step is to construct a covariance matrix of size, which describes how the values of *WER*, *CER* change. and *SemErr* together. Next, the direct PCA task is performed, that is, the search for eigenvalues and vectors (10) is performed:

$$\sum v = \lambda \cdot v. \tag{10}$$

The next stage is the interpretation of the weights based on the first principal component, i.e. the eigenvector (11)

$$W_1 = (W_{WER}, W_{CER}, W_{SemErr}), \tag{11}$$

where W_{WER}, W_{CER}, W_{SemErr} are values corresponding to the largest λ.

Obtained weights after normalization can be interpreted as the relative importance of metrics in the overall indicator, i.e.

$$\alpha = |w_{wer}|, \beta = |w_{cer}|, \gamma = |w_{SemErr}|, \alpha + \beta + \gamma = 1 .$$

Since there are different areas of application of the proposed hybrid metric in relation to the task, it is advisable to introduce a task category that will affect the automatic calculation of weight coefficients. Let k be given task categories, for example, teams, dialogs, documents. For each category, it is worth building your own matrix (12):

$$X^{(k)} = \{(WER_i, CER_i, SemErr_i) \mid task_i = k\}. \tag{12}$$

Then, for each such matrix, it is necessary to calculate its covariance matrix, the eigenvalues and vectors of which are represented as (13)

$$\sum^{(k)} v^{(k)} = \lambda^{(k)} \cdot v^{(k)}. \tag{13}$$

Accordingly, for each category, its own weight vector (7) will be obtained

$$W^{(k)} = (W_{WER}^{(k)}, W_{CER}^{(k)}, W_{SemErr}^{(k)}). \tag{14}$$

Another way to factor task categories into the weighting is to factorize with the category, which means that the total weights are composed of a global part and a category correction. Then the weight vector is computed on all data, and the correction is computed as the difference between the local and global weights. This approach allows us to see how the category shifts the importance of the metric, for example, for the command category the correction value for CER will be greater than 0, since symbols are the most important.

Using task categories when calculating weight coefficients allows you to obtain a flexible integral metric that can automatically adapt to the type of task.

The choice of normalization strategy should depend on the specific requirements of the application. For example, if punctuation errors are important for a given task, then punctuation should not be removed during preprocessing. This emphasizes that the decision to include or exclude certain types of errors through normalization directly affects what WELER measures as “quality”. This requires that the WELER coefficients be flexible enough to reflect these preprocessing choices.

WELER, by integrating semantic understanding, allows us to solve specific challenges that are difficult for traditional metrics. In particular, the proposed metric helps in working with homophones and similar – sound-

DOI 10.15588/1607-3274-2026-1-7

ing words. The semantic component can help distinguish lexically similar but semantically different words that WER by itself would mark equally. This allows the metric to reflect the real impact of such errors on understanding.

Another task that the proposed metric will help solve is working with ambiguous word boundaries, when For languages that do not have clear spaces between words, such as Chinese, the CER component and its underlying character-level alignment based on Levenshtein distance become especially valuable.

In automatic audio recognition tasks, audio quality affects error rates, but the semantic component of WELER can help assess whether meaning of the text even under increased lexical errors caused by noise. This allows W ELER to go beyond simple detection what errors have occurred, to understanding how much those errors affect understanding. This shift in focus from purely technical accuracy to a more user-oriented definition of quality is a significant step.

4 EXPERIMENTS

Table 2 shows examples of using *WELER* with weighting coefficients α=0.3, β=0.3, γ=0.4.

The table shows the calculated metrics *WER*, *CER*, *SemErr* and the calculated value of the proposed metric WELER. Row 4 shows the use of synonyms and paraphrasing, where the metrics WER and CER show a poor error result of 0.667 and 0.5758, since for these metrics the words. “Машина” and “Авто” are two completely different words, which confirms the unsuitability of these metrics for working with synonyms, giving a high error rate. The semantic distance for the considered sentence is low and is 0.0108, since the meaning of the sentence is perfectly preserved. Accordingly, WELER has a value of 0.377 and demonstrates a significantly better, lower, error result than WER/CER.

Row 5 shows complete sentence divergence and loss of context. WELER combines high scores of all three metrics and produces a high overall error. This demonstrates that WELER does not simply underestimate the scores, the proposed metric responds adequately when the meaning of the text is truly lost.

Another example of reordering and the use of synonyms in the text is shown in row 10. The word order is completely reversed, so *WER* and *CER* record a large error, almost 0.8571 for *WER*. WELER takes into account the high structural errors with *WER*/*CER*, but balances them with an almost perfect semantic match. The result of 0.486e is a much more adequate assessment of the text, the WELER value is almost twice as good as the *WER* value.

Table 2 – Examples of using WELER [17]

#	Reference text	Generated text	WER	CER	SemErr	WELER
1	Привіт, світ! Як у вас справи?	Привіт, світ! Як у вас справи?	0	0	0	0
2	Зараз чудова погода для прогулянки.	Зараз чудова пригода для прогулянки.	0.2	0.0571	0.0771	0.108
3	Я люблю програмування на Python	Я люблю програмування на Pyton.	0.2	0.0645	0.0685	0.1068
4	Машина їде дуже швидко по дорозі.	Авто рухається швидко по дорозі.	0.6667	0.5758	0.0108	0.377
5	Українська мова дуже мелодійна	Китайська кухня дуже смачна.	0.75	0.5483	0.2993	0.5092
6	Привіт, як справи? Усе добре!	Привіт як справи Усе добре	0.6	0.1034	0.0356	0.2253
7	Він пішов до школи сьогодні.	Він пішли до школи сьогодні.	0.2	0.0714	0.0004	0.0816
8	Котику-муркотику з м’яким животиком.	Де сметанка що була тут ще зранку.	1	0.8889	0.2604	0.6708
9	Адреса: вул. Свободи, 10	Адреса: вул. Свободи, 1О	0.25	0.0417	0.0159	0.0939
10	Автомобіль швидко рухався дорогою до міста Львів.	До міста Львів дорогою швидко рухалося авто.	0.8571	0.7347	0.0222	0.4864
11	Розробка штучного інтелекту змінює світ.	Розробка ШІ змінює світ.	0.4	0.45	0.0623	0.2799
12	Щоб встигнути потрібно плисти за течією до сходу сонця.	Щоб встигнути потрібно плести кошики до сходу сонця.	0.3333	0.1818	0.1538	0.2161
13	Плисти вперед до сходу сонця.	Плести кошики до сходу сонця.	0.4	0.2414	0.1596	0.2562

Figure 1 shows a distribution diagram of four speech recognition quality assessment metrics. It is clear from the diagram that *WER* has a fairly wide range from 0 to almost 1. The median for *WER* is around 0.5, which means that half of the examples have errors in around 50% of the words. The *CER* metric also has a wide range of, but a lower median of around 0.25, so the system performs better at the symbol level than at the word level. The diagram also illustrates that there are a significant number of examples with very large errors. The *SemErr* metric shows that most of the values are very low, 0–0.2, meaning that the meaning of the sentences is mostly preserved even

DOI 10.15588/1607-3274-2026-1-7

with errors in the words. There are a few outliers down to around 0.6.

Table 2 – Examples of using WELER

#	Reference text	Generated text	WER	CER	SemErr	WELER
1	Привіт, світ! Як у вас справи?	Привіт, світ! Як у вас справи?	0	0	0	0
2	Зараз чудова погода для прогулянки.	Зараз чудова пригода для прогулянки.	0.2	0.0571	0.0771	0.108
3	Я люблю програмування на Python	Я люблю програмування на Pyton.	0.2	0.0645	0.0685	0.1068
4	Машина їде дуже швидко по дорозі.	Авто рухається швидко по дорозі.	0.6667	0.5758	0.0108	0.377
5	Українська мова дуже мелодійна	Китайська кухня дуже смачна.	0.75	0.5483	0.2993	0.5092
6	Привіт, як справи? Усе добре!	Привіт як справи Усе добре	0.6	0.1034	0.0356	0.2253
7	Він пішов до школи сьогодні.	Він пішли до школи сьогодні.	0.2	0.0714	0.0004	0.0816
8	Котику-муркотику з м’яким животиком.	Де сметанка що була тут ще зранку.	1	0.8889	0.2604	0.6708
9	Адреса: вул. Свободи, 10	Адреса: вул. Свободи, 1O	0.25	0.0417	0.0159	0.0939
10	Автомобіль швидко рухався дорогою до міста Львів.	До міста Львів дорогою швидко рухалося авто.	0.8571	0.7347	0.0222	0.4864
11	Розробка штучного інтелекту змінює світ.	Розробка ШІ змінює світ.	0.4	0.45	0.0623	0.2799
12	Щоб встигнути потрібно плисти за течією до сходу сонця.	Щоб встигнути потрібно плести кошики до сходу сонця.	0.3333	0.1818	0.1538	0.2161
13	Плисти вперед до сходу сонця.	Плести кошики до сходу сонця.	0.4	0.2414	0.1596	0.2562

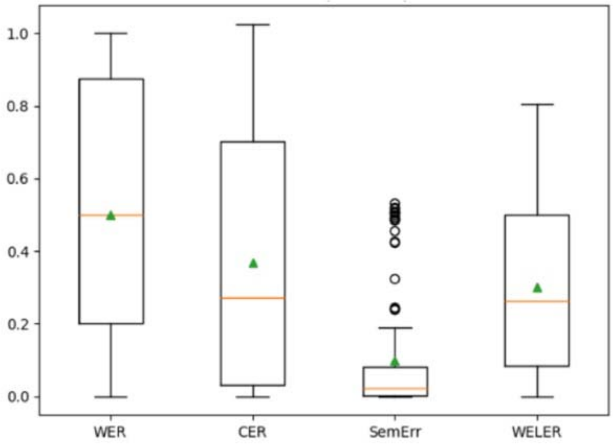

Figure 1 – Boxplot of the distribution of four speech recognition quality assessment metrics

The proposed WELER metric exhibits a smaller spread than *WER* and *CER*, namely 0–0.8, with a median of about 0.3. The proposed metric balances between formal and semantic errors. In general, we can conclude that at the symbol level the system makes fewer errors than at the word level, and semantics is mostly preserved even when there are spelling or lexical errors. The combined WELER metric gives a more balanced assessment than WER or CER separately. Figures 2 and 3 show graphs of pairwise dependencies between the studied metrics.

The analysis of the relationship between the traditional recognition metrics *WER*, *CER*, the proposed WELER metric and semantic errors *SemErr* demonstrates different levels of correlation. In particular, the comparison of *WER* and *SemErr* shows that even with high values of errors at the word level of 0.8–1.0, the *SemErr* indicator often remains low, which indicates the possibility of preserving the meaning despite numerous orthographic or syntactic deviations. The same is true for *CER*, where, although at high values of 0.8–1.0 there are cases of significant semantic distortions >0.4, most of the data is concentrated in the range of low values *SemErr* <0.1. This allows us to conclude that errors at the symbol level do not always lead to the destruction of meaning, but their excessive number significantly reduces semantic accuracy. The closest relationship with *SemErr* is revealed by the combined WELER metric; a smooth increase in semantic errors is observed with an increase in its value, which confirms the higher correspondence of WELER to the level of content preservation compared to *WER* and *CER*.

DOI 10.15588/1607-3274-2026-1-7

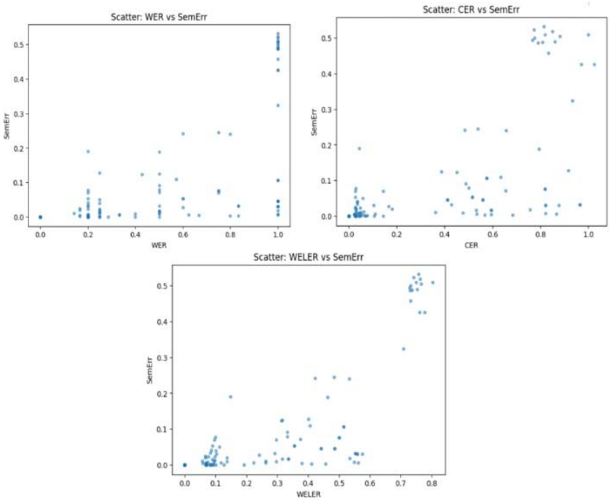

Figure 2 – Pairwise dependencies between SemErr and other metrics

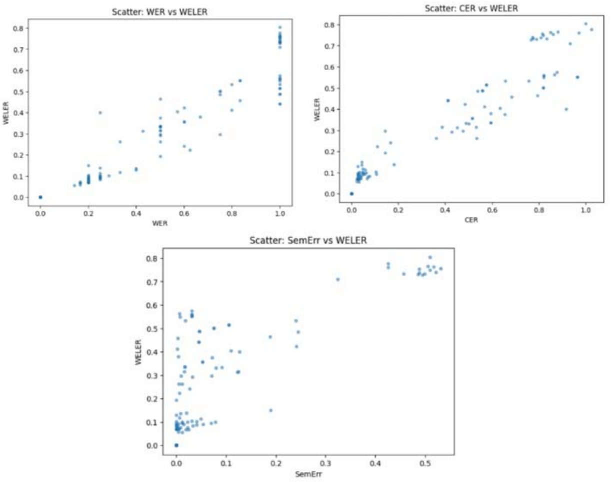

Figure 3 – Pairwise dependencies between WELER and other metrics

Analysis of the relationships between different metrics shows that WELER demonstrates a high correlation with both *WER* and CER, in the first case, an almost linear increase in WELER with an increase in *WER* is observed, which indicates the inclusion of *WER* as a key component of the combined metric; in the second case, a similar pattern is recorded, which confirms the integration of *CER* into the WELER structure. At the same time, the relationship between *SemErr* and WELER is less unambiguous, most examples are characterized by a low level of semantic errors 0–0.1 with a wide range of WELER values, however, there are also cases where a high *SemErr* indicator 0.4–0.5 is accompanied by an increased WELER. This indicates that WELER is able to respond to semantic distortions, but its value can remain high even with preserved content.

It can be argued that *WER* and *CER* exhibit an almost linear relationship with WELER, while *SemErr* correlates with them less strongly, since *WER* and *CER* capture formal errors rather than substantive ones. This confirms the ability of WELER to display more balanced results, reflecting at the same time the nature of errors at the form level and partially taking into account semantic accuracy.

© Dumyn A. R., Shakhovska N. B., 2026
DOI 10.15588/1607-3274-2026-1-7

In order to test the approach to automatically determining weight coefficients using the principal components method, a dataset was created consisting of two categories: “text” (fictional sentences) and “number” (serial numbers). After applying the algorithm to the values of the *WER*, *CER*, and *SemErr* metrics within each category, automatically determined weight coefficients were obtained, the results of which are shown in Table 3.

Table 3 – Automatically determined weights using the principal component method

Category	*WER*	*CER*	*SemDist*
number	0.3826	0.37718	0.2402
text	0.3470	0.3483	0.3046

Table 4 shows a comparison of WELER values calculated using manually entered weighting coefficients (α=0.3, β=0.3, γ=0.4) and weighting coefficients determined by applying the principal component method (α=0.347, β=0.3483, γ=0.3046).

Table 4 – Examples of calculating the WELER metric with manually selected coefficients and using PCA [17]

#	Reference text	Generated text	WELER	WELER (PCA)
1	Привіт, світ! Як у вас справи?	Привіт, світ! Як у вас справи?	0	0
2	Зараз чудова погода для прогулянки.	Зараз чудова пригода для прогулянки.	0.108018	0.112825
3	Я люблю програмування на Python	Я люблю програмування на Pyton.	0.106754	0.112746
4	Машина їде дуже швидко по дорозі.	Авто рухається швидко по дорозі.	0.377028	0.435193
5	Українська мова дуже мелодійна	Китайська кухня дуже смачна.	0.509243	0.542481
6	Привіт, як справи? Усе добре!	Привіт як справи Усе добре	0.225291	0.255112
7	Він пішов до школи сьогодні.	Він пішли до школи сьогодні.	0.081602	0.09442
8	Котику-муркотику з м’яким животиком.	Де сметанка що була тут ще зранку.	0.670843	0.736009
9	Адреса: вул. Свободи, 10	Адреса: вул. Свободи, 1O	0.093865	0.10612
10	Автомобіль швидко рухався дорогою до міста Львів.	До міста Львів дорогою швидко рухалося авто.	0.486422	0.560139
11	Розробка штучного інтелекту змінює світ.	Розробка ШІ змінює світ.	0.279902	0.314533
12	Щоб встигнути потрібно плисти за течією до сходу сонця.	Щоб встигнути потрібно плести кошики до сходу сонця.	0.216085	0.225878
13	Плисти вперед до сходу сонця.	Плести кошики до сходу сонця.	0.25624	0.271503

WELER calculated using the principal component method differs only in scaling, the values are slightly shifted, but the general approach of the balanced metric value is preserved.

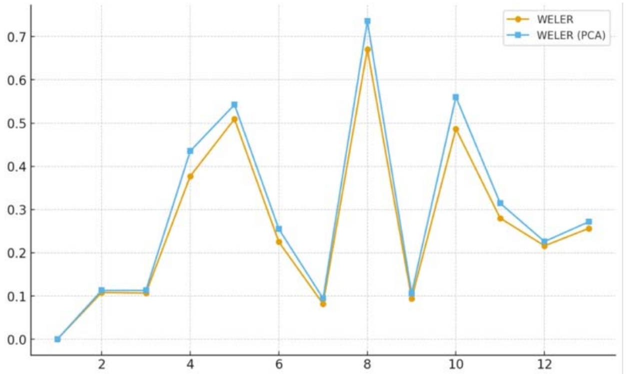

Figure 4 – Pairwise dependencies between WELER and other metrics

As can be seen from Figure 4 and Table 4, in most cases the WELER and WELER (PCA) values demonstrate almost parallel dynamics, however, the application of the principal components method somewhat enhances the differences between individual examples. The highest values, approximately 0.5–0.5, are recorded in examples No. 5, No. 8 and No. 10, where significant semantic distortion of the text was observed. The lowest values, approximately 0.0–0.1, are characteristic of examples No. 1, No. 2, No. 3, No. 7 and No. 9, where deviations were limited to minor spelling or grammatical errors. The PCA method shows higher sensitivity to semantic differences, which is especially noticeable in example No. 4, where the difference between the values 0.38 and 0.43 makes the deviation more pronounced.

5 RESULTS

As a result, WELER, compared to the classic WER and CER, provides a more balanced assessment of text quality, as it penalizes significant semantic deviations but mitigates the penalty for minor spelling or stylistic differences, which better reflects the real impact of errors on comprehension.

Proposed WELER metric has the potential to improve text quality assessment in a wide range of artificial intelligence tasks, providing a more tailored assessment.

For automatic speech recognition, WELER can provide a more granular analysis of speech-to-text systems. It is able to distinguish between minor phonetic recognition errors and critical semantic errors. The metric can be configured to prioritize verbatim reproduction for legal transcriptions (high α and β) or semantic understanding for conversational AI (high γ). This flexibility in tuning the coefficients is a major advantage of WELER for a variety of applications.

For natural language processing tasks such as text generalization or machine translation, WELER can evaluate both lexical accuracy, such as correct word choice, and semantic preservation, ensuring the accuracy of form and meaning of the generated text. The coefficients can be adjusted to penalize grammatical errors more severely if fluency is a top priority.

For OCR tasks, the application of WELER can provide a more robust evaluation of systems, especially for documents where both character-level accuracy is important, such as serial numbers, financial data, and word-level accuracy, such as free text fields. The semantic component can even assess whether contextual meaning to the data obtained. Providing a more comprehensive and customizable metrics WELER can facilitate better decision-making in the development and deployment of models by allowing engineers to optimize systems for specific real-world performance criteria rather than a generic, potentially misleading error rate. This can lead to significant cost savings by reducing the need for extensive manual post-processing and quality control in document digitization, transcription, and data entry processes. It can also be used to assess the similarity of texts and translation quality with reference translations with corresponding weight changes for these tasks.

It can also be used to assess the similarity of texts and translation quality with reference translations with corresponding weight changes for these tasks.

6 DISCUSSION

WELER is the first step towards more comprehensive metrics for assessing text quality. Future research will focus on aspects of improving dynamic weighting. In particular, it is worth continuing to investigate methods for automatically adjusting WELER coefficients based on the domain of the input text, its complexity, or perceived criticality, for example, taking into account the thematic focus of the texts (medical and everyday conversation, etc.). This task can be solved by using machine learning approaches to learn optimal weighting coefficients.

Another important area of further research is integration with metrics for detecting grammatical errors and punctuation, in particular including established GEC metrics, e.g. M2, ERRANT, GLEU, SARI, and punctuation error metrics, e.g. FER, NER for punctuation, to the WELER framework can contribute to a truly holistic assessment of linguistic quality. Although the integration of these metrics is important for certain tasks, it will significantly increase the complexity of WELER and create noise for speech recognition tasks, since such tasks mostly do not have punctuation.

Researching how WELER and its components work in different languages, especially those with complex morphological structures or without explicit word boundaries, and adapting weighting schemes accordingly is another important direction.

CONCLUSIONS

Text quality assessment is a cornerstone for the development and improvement of artificial intelligence systems in such critical areas as ASR, NLP and OCR. Although traditional metrics such as WER and CER are widely used and based on the robust concept of Levenshtein distance,

© Dumyn A. R., Shakhovska N. B., 2026
DOI 10.15588/1607-3274-2026-1-7

their insensitivity to semantic meaning, grammatical correctness, and punctuation features limits their ability to reflect the true quality of text from a human perceptual perspective. This discrepancy highlights the urgent need to develop more comprehensive and adaptive assessment tools.

Proposed WELER metric is an alternative solution in this direction. It integrates accurate word and character level error counting, using Levenshtein distance as a basis, with advanced semantic similarity methods based on contextual embeddings. This allows WELER to take into account not only what was incorrectly recognized, but also how much this error affects the meaning and understanding of the text. The inclusion of self-adjusting weights depending on the text category is a key feature of WELER, which allows adapting the metric to the specific requirements of different applications and domains, prioritizing those aspects of quality that are most critical for a particular task.

However, WELER, like all metrics based on reference data, relies on accurate and consistent human-verified transcriptions. Errors in the reference data can affect the accuracy of the assessment. Therefore, for complex metrics, the quality and representativeness of these data are especially important, since semantic and weighted errors are much more sensitive to the quality of the annotation than simple word counts. The implementation of WELER, while promising significant benefits, requires careful consideration of issues such as computational complexity, empirical determination of optimal weights, and ensuring high quality of reference data. Future research could extend WELER to include dynamic weight adjustment, integration with grammatical and punctuation error metrics, and exploration of LLM-based scoring capabilities, which could lead to further developments in the field of text quality assessment. Ultimately, WELER represents a robust and flexible framework for more accurate and holistic text quality assessment, which will contribute to the development of more efficient and user-friendly AI systems.

ACKNOWLEDGEMENTS

The authors would like to thank the industry researchers, authors of works in references, and the anonymous reviewers. This research received no external funding.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: equally significant, evenly proportional.

Data availability: The manuscript has no associated data except in the example within itself.

Software availability: The manuscript has no specific associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies to create the submitted work. Artificial intelligence was used for translation and grammar checks.

REFERENCES

1. Hamed I. Benchmarking Evaluation Metrics for Code-Switching Automatic Speech Recognition, *2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023 : proceedings*. [Piscataway], IEEE, 2023, pp. 999–1005. DOI: 10.1109/SLT54892.2023.10023181
2. Measure and improve speech accuracy, *Cloud Speech-to-Text Documentation*. Available at: https://cloud.google.com/speech-to-text/docs/speech-accuracy (accessed: 22 July 2025).
3. Dumyn A., Fedushko S., Syerov Y. Review of Automatic Speech Recognition Systems for Ukrainian and English Language, *Data-Centric Business and Applications : proceedings*. Cham, Springer, 2024. (Lecture Notes on Data Engineering and Communications Technologies, Vol. 212).
4. Shakhovska N., Shvorob I. The method for detecting plagiarism in a collection of documents, *2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, Ukraine, 2015 : proceedings*. [Piscataway], IEEE, 2015, p. 142–145. DOI: 10.1109/STC-CSIT.2015.7325453
5. Sasindran Z., Yelchuri H., Prabhakar T. V., Rao S. A new hybrid evaluation metric for automatic speech recognition tasks, *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) : proceedings.* [Piscataway], IEEE, 2023, pp. 1–7. DOI: 10.48550/arXiv.2211.01722
6. Kim S., Arora A., Le D., Yeh C.-F., Fuegen C., Kalinli O., Seltzer M. L. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding, *arXiv preprint arXiv:2104.02138,* 2021. Link: https://arxiv.org/abs/2104.02138
7. Sasindran Z., Yelchuri H., Prabhakar T. V. SeMaScore: a new evaluation metric for automatic speech recognition tasks, *arXiv preprint arXiv:2401.07506,* 2024. Link: https://arxiv.org/abs/2401.07506
8. Phukon B., Zheng X., Hasegawa-Johnson M. Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches, *arXiv preprint arXiv:2506.16528*, 2025. Link: https://arxiv.org/abs/2506.16528
9. Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating text generation with BERT, *arXiv preprint arXiv:1904.09675,* 2019. Link: https://arxiv.org/abs/1904.09675
10. James J., Gopinath D. P. Advocating character error rate for multilingual ASR evaluation, *arXiv preprint arXiv:2410.07400,* 2024. Link: https://arxiv.org/abs/2410.07400
11. Van Schaik T., Pugh B. A field guide to automatic evaluation of LLM-generated summaries, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, ACM, 2024, pp. 2832–2836.

DOI 10.15588/1607-3274-2026-1-7

12. Arockiya Jerson J., Preethi N. An analysis of Levenshtein distance using dynamic programming method, *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (ICMISC 2022).* Singapore, Springer Nature Singapore, 2023, pp. 525–532.
13. Greenacre M., Groenen P. J., Hastie T., d'Enza A. I., Markos A., Tuzhilina E. Principal component analysis, *Nature Reviews Methods Primers,* Vol. 2, № 1, Article 100.
14. Measuring the Accuracy of Automatic Speech Recognition Solutions, *arXiv.* Available at: https://arxiv.org/html/2408.16287v1 (accessed: 22 July 2025).
15. Hunt M. A. Word Errors and the Significance of Weighted Accuracy Measures, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 1990.
16. Neudecker C., Baierer K., Gerber M., Clausner C., Antonacopoulos A., Pletschacher S. A survey of OCR evaluation tools and metrics, *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing.* New York, ACM, 2021, pp. 13–18.
17. Dumyn A.R. Hibrydna metryka otsinky yakosti tekstu na osnovi kontekstnoho zvazhuvannya, *Tavriys'kyy naukovyy visnyk. SeriyaL Tekhnichni nauky*, 2025, №4, ch. 1, pp. 85-93

Received 20.10.2025.
Accepted 13.01.2026.
Published 27.03.2026.

УДК 004.93

WELER: КОМПЛЕКСНИЙ ПОКАЗНИК ДЛЯ ОЦІНКИ ЯКОСТІ ТЕКСТУ

Думин А. Р. – аспірант кафедри систем штучного інтелекту Національного університету «Львівська політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0003-2111-2899.

Шаховська Н. Б. – д-р техн. наук, професор, ректор Національного університету «Львівська політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0002-6875-8534.

АНОТАЦІЯ

Актуальність. Оцінка якості тексту є важливою для надійного штучного інтелекту, який обробляє мову. В ASR вона відображає, наскільки точно мовлення стає текстом; в OCR – наскільки точно зображення перетворюють текст; а в NLP – наскільки правильними та зв'язними є виходи.

Мета. Метою роботи є створення складної метрики для оцінки якості тексту.

Метод. Класичні метрики WER та CER є вузькими: вони фіксують лише лексичні редагування, однаково зважують усі зміни, ігнорують контекст та семантику, і часто пропускають пунктуацію та регістр, маскуючи проблеми читабельності та типи помилок. апропонована метрика WELER інтегрує точний підрахунок помилок на рівні слів та символів, використовуючи відстань Левенштейна як основу, з передовими методами семантичної подібності, заснованими на контекстному вбудовуванні. Це дозволяє WELER враховувати не лише те, що було неправильно розпізнано, але й те, наскільки ця помилка впливає на значення та розуміння тексту. Включення самоналаштовуваних ваг залежно від категорії тексту є ключовою особливістю WELER, яка дозволяє адаптувати метрику до конкретних вимог різних застосувань та областей, надаючи пріоритет тим аспектам якості, які є найбільш критичними для конкретного завдання.

Результати. Метрика WELER пропонується як ефективний підхід до оцінювання якості тексту. Її концептуальна основа полягає в інтеграції традиційного підрахунку помилок на словесному та символьному рівнях, заснованого на відстані Левенштейна, із сучасними методами оцінювання семантичної подібності, що використовують контекстуальні векторні подання. Такий підхід забезпечує більш комплексне відображення впливу помилок на змістову цілісність та інтерпретованість результатного тексту.

Висновки. WELER, як і всі метрики, засновані на довідкових даних, спирається на точні та послідовні транскрипції, перевірені людиною. Помилки в довідкових даних можуть впливати на точність оцінки. Тому для складних метрик якість та репрезентативність цих даних є особливо важливими, оскільки семантичні та зважені помилки набагато чутливіші до якості анотації, ніж проста кількість слів.

КЛЮЧОВІ СЛОВА: обробка природної мови, оцінка якості тексту, WER, CER, WELER.

ЛІТЕРАТУРА

1. Hamed I. Benchmarking Evaluation Metrics for Code-Switching Automatic Speech Recognition / I. Hamed // 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023 : proceedings. – [Piscataway]: IEEE, 2023. – P. 999–1005. DOI: 10.1109/SLT54892.2023.10023181
2. Measure and improve speech accuracy // Cloud Speech-to-Text Documentation. – Available at: https://cloud.google.com/speech-to-text/docs/speech-accuracy (accessed: 22 July 2025).
3. Dumyn A. Review of Automatic Speech Recognition Systems for Ukrainian and English Language / A. Dumyn, S. Fedushko, Y. Syerov // Data-Centric Business and Applications : proceedings. – Cham : Springer, 2024. – (Lecture Notes on Data Engineering and Communications Technologies, Vol. 212).
4. Shakhovska N. The method for detecting plagiarism in a collection of documents / N. Shakhovska, I. Shvorob // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, Ukraine, 2015 : proceedings. – [Piscata-

DOI 10.15588/1607-3274-2026-1-7

way] : IEEE, 2015. – P. 142–145. DOI: 10.1109/STC-CSIT.2015.7325453

5. A new hybrid evaluation metric for automatic speech recognition tasks / [Z. Sasindran, H. Yelchuri, T. V. Prabhakar, S. Rao] // 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) : proceedings. – [Piscataway] : IEEE, 2023. – P. 1–7. DOI: 10.48550/arXiv.2211.01722
6. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding / [S. Kim, A. Arora, D. Le et al.] // arXiv preprint arXiv:2104.02138. – 2021. Link: https://arxiv.org/abs/2104.02138
7. Sasindran Z. SeMaScore: a new evaluation metric for automatic speech recognition tasks / Z. Sasindran, H. Yelchuri, T. V. Prabhakar // arXiv preprint arXiv:2401.07506. – 2024. Link: https://arxiv.org/abs/2401.07506
8. Phukon B. Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches / B. Phukon, X. Zheng, M. Hasegawa-Johnson // arXiv preprint arXiv:2506.16528. – 2025. Link: https://arxiv.org/abs/2506.16528
9. BERTScore: Evaluating text generation with BERT / [T. Zhang, V. Kishore, F. Wu et al.] // arXiv preprint arXiv:1904.09675. – 2019. Link: https://arxiv.org/abs/1904.09675
10. James J. Advocating character error rate for multilingual ASR evaluation / J. James, D. P. Gopinath // arXiv preprint arXiv:2410.07400. – 2024. Link: https://arxiv.org/abs/2410.07400
11. Van Schaik T. A field guide to automatic evaluation of LLM-generated summaries / T. Van Schaik, B. Pugh // Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York : ACM, 2024. – P. 2832–2836.
12. Arockiya Jerson J. An analysis of Levenshtein distance using dynamic programming method / J. Arockiya Jerson, N. Preethi // Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (ICMISC 2022). – Singapore : Springer Nature Singapore, 2023. – P. 525–532.
13. Principal component analysis / [M. Greenacre, P. J. Groenen, T. Hastie et al.] // Nature Reviews Methods Primers. – Vol. 2, № 1. – Article 100.
14. Measuring the Accuracy of Automatic Speech Recognition Solutions // arXiv. – Available at: https://arxiv.org/html/2408.16287v1 (accessed: 22 July 2025).
15. Hunt M. A. Word Errors and the Significance of Weighted Accuracy Measures / M. A. Hunt // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 1990.
16. A survey of OCR evaluation tools and metrics / [C. Neudecker, K. Baierer, M. Gerber et al.] // Proceedings of the 6th International Workshop on Historical Document Imaging and Processing. – New York : ACM, 2021. – P. 13–18.
17. Dumyn A. R. Hibrydna metryka otsinky yakosti tekstu na osnovi kontekstnoho zvazhuvannya / A. R. Dumyn // Tavriys’kyy naukovyy visnyk. SeriyaL Tekhnichni nauky. – 2025. – №4, ch. 1 – P. 85–93

© Dumyn A. R., Shakhovska N. B., 2026
DOI 10.15588/1607-3274-2026-1-7