On the effectiveness of phrase distance measures on separability and cohesion of meaning: A multilingual review

Anonymous ACL submission

Abstract

This paper presents an automated method for evaluating phrase distance measures based on cohesion and diffusion measurements, eliminating the need for direct human judgment. The evaluation involves five homegrown datasets, each consisting of 200 headlines or abstracts from news articles, subdivided into 20 sets. Two datasets are in Arabic, while others include news in French, German, and English. Each set contains 10 texts with shared meaning but different cohesion, and diffusion is modeled by distances between articles with different meanings. The benchmark for evaluating phrase distance measures combines Silhouette Index properties with the mean of Pearson Correlations over distance matrix pairs.

Our findings reveal that Yule distance with binary embeddings consistently surpasses other measures. Phrase distance performance remains steady across languages, tokenizers and sentences' lengths.

1 Introduction

With the rise of the transformer model (Vaswani et al., 2023) and the recent prominence of OpenAI's ChatGPT model (Wu et al., 2023), interest in large language models (LLMs) modeling and applications have surged to unprecedented levels. Phrase Distance Measures (PDM)s, which measure the distance in meaning between two sentences or paragraphs, are pivotal are indispensable for evaluating LLMs (Lai et al., 2023). They assist in comparing expert-known true answers with those from chat-enabled LLMs, which highlights the need to understand the deviation from a known truth. Solutions, relying on context-aware ChatBots with architectures built on vector databases encoding domain-specific contexts (Mansurova et al., 2023; Yager, 2023; Neumann et al., 2023), necessitate effective and rapid PDMs to locate relevant contexts in response to user queries. The performance of PDMs themselves is an active field of research and discussion and several new PDMs have been developed (Zhao et al., 2019; Zhang et al., 2020; Rei et al., 2020; Sellam et al., 2020; Yuan et al., 2021) and studied for explainability (Leiter et al., 2022).

In most cases PDMs themselves are evaluated by human judgement. In this work we propose assessing distance measures by their ability to cluster similar content (cohesion) and differentiate disparate content (diffusion). We apply this method on phrase distances inspired by the work in bioinformatics (Haschka et al., 2021), effectively scoring commonly used PDMs and investigate the effectiveness of different PDMs using five hand-designed datasets in four languages: Arabic, English, French, and German. These datasets, constructed from news articles scraped from various outlets, consist of 20 articles which shares the same meaning but expressed in distinct styles. Human selection ensures that the 10 different texts for the same article meet this requirement.

Herein we present a multidimensional study, varying tokenization, embedding, and distance measures on word embeddings to identify the optimal phrase distance measure. Across all datasets and languages, the Yule distance, with a simple binary word embedding vector, consistently yields the most promising results in practical contexts.

2 Background and Related Work

A comprehensive review spanning over 15 years in the development of PDMs underscores the significance of consistency and highlights the challenges in reporting machine learning model performance. This complexity inherently complicates the interpretation of model performance reports (Blagec et al., 2022). The paper emphasizes that ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005) stand out

as the most commonly reported evaluation metrics (Blagec et al., 2022).

Critiques of these metrics often center around their correlation with human judgment, with discussions acknowledging the limitations of n-gram correlations (Reiter, 2018), such as those found in BLEU (Papineni et al., 2002) or various ROUGE (Lin, 2004) variants. Interestingly, these discussions suggest that metrics based on F_1-measure yield superior performance (Lavie et al., 2004). Many advanced PDMs have seen their light in recent years (Zhao et al., 2019; Zhang et al., 2020; Rei et al., 2020; Sellam et al., 2020; Yuan et al., 2021) but rely on computationally intensive tasks and the community is further reluctant to adopt them (Leiter et al., 2022). Due to the multilingual character of this study, we were unable to assess all of them comprehensively. Nevertheless, we utilized the multilingual Bert variant (Devlin et al., 2019) together with the Bert score (Zhang et al., 2020). The expansion of the method shown herein to other variants remains trivial if the goal is a single language comparison.

Although previous studies have utilized scrapped news datasets for phrase distance metric evaluation (Agirre et al., 2016), this paper introduces a straightforward and established method in bioinformatics (Haschka et al., 2021) to systematically assess PDMs used in the field of LLM/AI research. This approach aims to provide a more nuanced understanding of the performance and limitations of phrase distance measures.

3 Methodology

The methodology employed in this study revolves around assessing the effectiveness of various PDMs in capturing the cohesion and diffusion dynamics within an expert-curated dataset. In section **??**, we define our distance nomenclature while 3.2 elaborates on the construction of the dataset which features clusters of phrases sharing identical meanings, strategically distant from clusters conveying different meanings. As per the definition in equation (3), a PDM is a composite of tokenization algorithms (section 3.3), embeddings (section 3.4), and vector or n-gram distances (section 3.5). Our approach involves systematic variations in all three components, a detailed exposition of which follows below.

The efficacy of distinguishing these clusters, thereby assessing how closely phrases with similar meanings align and how distinctly they stand from phrases with different meanings, is quantified through two performance indices. These indices are based on the Silhouette method and a Pearson Correlation of pairwise distance matrix elements concerning an optimal distance matrix, elucidated in section 3.6.

3.1 Formulation

3.2 Datasets

Five datasets were curated each comprising 20 distinct news contents or meanings. For every news content, 10 diverse news outlets that articulated the same information in varying were identified. Consequently, each dataset encompasses a total of 200 items. The selection of languages differed for each dataset, including English, French, German, and two datasets in Arabic. This deliberate multilingual approach, along with variations in word and sentence lengths across datasets, was adopted to test the robustness of the evaluation across diverse data sets as shown in Figure 1. The data collection spanned news sources such as public crime reports, tabloid press, technology updates, and financial news, deliberately excluding religious and extreme political content. This careful selection ensures a balanced and representative evaluation of various phrase distance measures across a broad spectrum of data.

3.3 Tokenizers and Tokenizer Training

For the training of tokenizers, we leveraged a subset of 50,000 articles from the wiki40b dataset (Guo et al., 2020) for each language: German, French, English, and Arabic. The Hugging Face tokenizers library (Wolf et al., 2020) facilitated the training process. The trained tokenizers include *Byte-Pair Encoding* (Sennrich et al., 2016), *Unigram* (Kudo, 2018), *WordPiece* (Devlin et al., 2019), and *WordLevel* tokenization. A vocabulary size of 32,000 was selected for each tokenizer, aligning with the input size of common large language models (LLMs) (OpenAI et al., 2023; Touvron et al., 2023).

Post-training, we applied these tokenizers to encode the 50,000 articles they were trained on. Notably, we observed that the number of tokens generated by these algorithms is language-dependent. Table 1 provides insights into the token generation. This language-dependent tokenization variation underscores the importance of considering linguistic

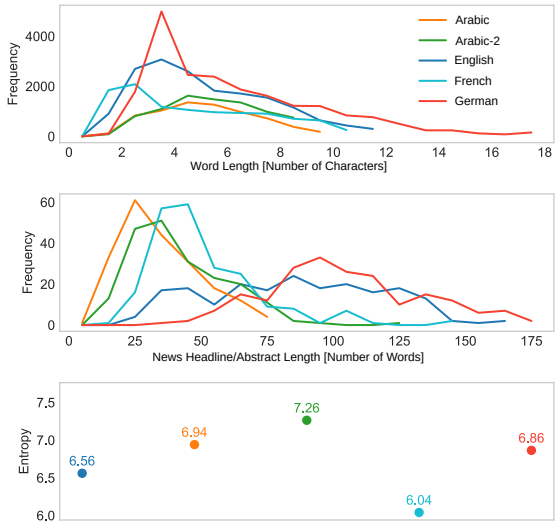

Figure 1: The diversity of the 5 datasets in 4 languages scrapped from various news sources. Entropy is here calculated word-wise: $H(\text{ds}) = -\sum_{i=1}^{n} P(w_i) \log_2 P(w_i)$ where $P(w_i)$ is the probability to find the word w_i out of n words in the dataset ds.

nuances when applying tokenization in multilingual contexts.

3.4 Embeddings

In this section, we explore four distinct types of embeddings tailored to various distance measures. Certain vector distance measures, such as the L_p-norm or cosine distance (equation (3)), necessitate vectors in $\vec{v} \in \mathbb{R}^n$ form. Therefore, we employed a simple embedding technique for each dataset item, mapping them into a vector with a dimensionality equal to the dictionary size. This involved counting the occurrences of each token and placing them at the corresponding index.

Other distance measures, like Jaccard or Yule distance, are designed for binary vectors $\vec{v} \in \{0, 1\}^n$. Here, we straightforwardly set the index for a specific token in the vector to 1 if the token occurred at least once.

In addition to these fundamental embeddings, we introduced two advanced embedding methods:

1. **Singular Value Decomposition based embeddings:**
   - Utilizing the trained tokenizers detailed in Section 3.3, we encoded 50,000 Wikipedia articles corresponding to the language of the tokenizer.
   - For each article, we created a vector with dimensions matching the vocabulary size (32,000), indicating the token counts.
   - Performing a singular value decomposition on these encoded datasets, we retained the right-hand vectors corresponding to the top 100 singular values.
   - Embeddings were then generated by projecting the encoded data onto these right-hand vectors, resulting in embeddings $\vec{v}_{\text{SVD}} \in \mathbb{R}^{100}$.
   - The singular value decomposition aimed to enhance the initial embeddings by filtering variance, yielding more suitable embeddings.
2. **Bert model embeddings:**
   - We generated embeddings based on the Bert model (Devlin et al., 2019).
   - This embedding approach has the advantage of distinguishing phrases by meaning, capturing nuances like the difference between "I went home and cooked food for my family" and "I stayed home and cooked food for my family."
   - These embeddings were directly built from the coefficients found in the transformer architecture of a Bert model.

These diverse embedding strategies allow us to capture different aspects of linguistic information, enabling a comprehensive evaluation of phrase distance measures.

3.5 Distance Measures

This study defines a PDM as the image,

$$D : p_1 \times p_2 \to d, \tag{1}$$

of the operator D applied to p_1 and p_2 representing the character strings of two texts of arbitrary length. p_1 and p_2 are as such elements of a field of texts, and $d \in \mathbb{R}$ represents the similarity in meaning between the phrases p_1 and p_2.

We propose an optimal distance measure, $D_{opt}(p_1, p_2)$ defined as:

$$D_{\text{opt}}(p_1, p_2) = \begin{cases} 0 \text{ if } p_1, p_2 \in a \\ 1 \text{ if } p_1 \in a, p_2 \in b \end{cases}, \tag{2}$$

where a and b are sets of 10 news articles that share the same *meaning* but are written by different outlets in varying style. The Silhouette Index

and the mean of Pearson correlations, calculated under different distance measures D are employed as benchmarks. Optimal cases yield indices close to 1, while the worst cases result in -1 (Silhouette Index) or 0 (mean of Pearson correlations). This article presents a dataset and strategy for evaluating distance measures, addressing both cohesion and diffusion. A PDM is considered effective if it identifies articles with the same meaning as close (cohesion) and those with different meanings as distant (diffusion).

To compare two phrases they are transformed into vectors $\vec{v} \in \mathbb{R}^n$. This vector representation, known as an embedding enables the application of a vector distance measure:

$$\begin{aligned} T(p_1) &= \vec{v}_{p_1}, \\ T(p_2) &= \vec{v}_{p_2}, \\ V(\vec{v}_{p_1}, \vec{v}_{p_2}) &= d, \end{aligned} \tag{3}$$

where $T(p)$ is a composition of text tokenization and embedding algorithms.

The comprehensive D involves multiple operations, encompassing tokenization and embedding, ultimately leading to the formation of V. Furthermore, V can be applied to both real $\vec{v} \in \mathbb{R}^n$ and binary $\vec{v} \in 0, 1^n$ embeddings. Additionally, there exist PDMs that operate directly on tokens, exemplified by widely used paraphrasing distances like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). This nuanced approach allows for a multifaceted exploration of various linguistic aspects within the scope of phrase distance measures.

The following vector distance measures were applied on real, $\vec{v} \in \mathbb{R}^n$, embeddings:

- L_p-norms for $p = \{1, 2\}$: These norms are for the $p = 1$ case further known as Manhattan for the $p = 2$ case Euclidean distance. They are defined by the equation:

$$L_p = \left[\sum_{i=1}^{n} |v_{1,i} - v_{2,i}|^p \right]^{\frac{1}{p}} \tag{4}$$

- The cosine distance: This distance is built from the angle between the two vectors:

$$V_{\cos} = \frac{\langle v_1 | v_2 \rangle}{\sqrt{\langle v_1 | v_1 \rangle} \sqrt{\langle v_2 | v_2 \rangle}}, \tag{5}$$

with $\langle v_1 | v_2 \rangle$ denoting the inner product between v_1 and v_2.

It is essential to highlight that, given the high dimensionality and the well-documented curse of dimensionality, particularly with naive embeddings, the anticipated superior performance of the cosine distance over the L_1-norm is acknowledged (Sohangir and Wang, 2017). Additionally, an expected advantage of the L_1-norm over the L_2-norm is recognized (Aggarwal et al., 2001).

Further, the following binary vector distance measures were evaluated:

- Jaccard distance:

$$V_{\text{Jac}}(v_1, v_2) = \frac{|v_1 \cup v_2| - |v_1 \cap v_2|}{|v_1 \cup v_2|}, \tag{6}$$

where $|v_1 \cup v_2|$ (union) represents the number of vector elements that are 1 in v_1 or v_2, and $|v_1 \cap v_2|$ (intersection) represents the number of elements that are in both v_1 and v_2 1.

- Yule distance:

$$\begin{aligned} \xi &= |\{i : i \in v_1, i \notin v_2\}|, \\ \eta &= |\{i : i \notin v_1, i \in v_2\}|, \\ \zeta &= |\{i : i \in v_1, v_2\}|, \\ \tau &= |\{i : i \notin v_2, v_2\}|, \\ d(x, y) &= \frac{2\xi\eta}{\zeta\tau}. \end{aligned} \tag{7}$$

The F_1-score served as an additional distance metric in our study, but with a distinct approach. Unlike the typical binary embeddings, we applied it to real $\vec{v} \in \mathbb{R}^n$ embeddings, utilizing a weighted average, as implemented in scikit-learn (Pedregosa et al., 2011).

In addition to paraphrasing distances working with real and binary embeddings, the effectiveness of paraphrasing distance metrics that operate on n-grams—recurrent sequences of tokens within a phrase is investigated. To explore this, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) scores, all of which rely on n-grams were used.

3.6 Evaluation Coefficients/Benchmark Indices

In evaluating phrase distance measures D, two distinct performance indicators were used with our curated dataset outlined in Section 3.2. These benchmarks enable a comprehensive assessment of a phrase distance measure's performance.

1. **Silhouette Index:** The Silhouette Index (Rousseeuw, 1987) was implemented as follows, considering a pair of different news articles a and b:

$$k_{j,a} = \frac{1}{n_a}\sum_{i=1}^{n_a} D(p_a(i), p_a(j)), \quad (8)$$

$$k_{j,b} = \frac{1}{n_b}\sum_{i=1}^{n_b} D(p_b(i), p_a(j)), \quad (9)$$

$$s_j = \begin{cases} k_{j,a} < k_{j,b}: & 1 - \frac{k_{j,a}}{k_{j,b}} \\ k_{j,a} = k_{j,b}: & 0 \\ k_{j,a} > k_{j,b}: & \frac{k_{j,b}}{k_{j,a}} - 1 \end{cases} \quad (10)$$

$$S = \frac{1}{n_a}\sum_{j=1}^{n_a} s_j, \quad (11)$$

where $p_{a,b}(i)$ is the i-th element news articles with, a the same meaning or, b a different meaning. The Silhouette varies between $-1 \leq S \leq +1$. It is negative if the distances between the articles with the same meaning would be spread out and less clustered together than news articles with a different meaning. More interesting are of course cases where articles of the same meaning cluster together and articles of different meanings are more distant from each other. In this instance, the Silhouette index results in a positive value. Our rationale is based on the understanding that an optimal distance measure would produce Silhouette scores approaching +1.

The Silhouette Index, the arithmetic mean of Silhouettes over each set of articles of the same meanings to characterize a phrase distance measures effectiveness with a dataset was also used. The closer the Silhouette Index is to +1 the more effective the phrase distance measure is in clustering together articles of the same meaning and differentiating them from articles of a different meaning.

$$SI = \frac{\sum_{\text{pairs}} S(c_i, c_j)}{n_{\text{pairs}}}. \quad (12)$$

A higher SI indicates greater effectiveness in clustering together articles of the same meaning and differentiating them from those with different meanings.

2. **Pearson Correlation Index** We defined a Pearson Correlation Index (PI) for pairwise distance matrices, starting with the definition of a distance matrix M:

$$M_{i,j} = \begin{cases} 1: & (p_i \in a) \wedge (p_j \in b) \\ 0: & \text{otherwise} \end{cases}, \quad (13)$$

and a distance matrix K under a given phrase distance measure D:

$$K_{i,j} = D(p_i, p_j). \quad (14)$$

The Pearson correlation for a single set pair of news articles a and b is then calculated as:

$$P = \sum_{\forall i,j \in a,b} \frac{M_{i,j}K_{i,j}}{\sqrt{\left(M_{i,j} - \bar{M}\right)^2}\sqrt{\left(K_{i,j} - \bar{K}\right)^2}}, \quad (15)$$

where $\bar{M}$ and $\bar{K}$ are the means of the elements in matrices M and K. The overall Pearson Correlation Index for a dataset is given by:

$$PI = \sum_{\text{Pairs of Sets}} P_{\text{Pair}}. \quad (16)$$

Under optimal conditions, a phrase distance D that correctly separates phrases of the same meaning and diffuses phrases of different meanings is expected to yield a PI close to 1.

By combining these indices with our datasets and the outlined equations, we possess the necessary tools to comprehensively evaluate the performance of a phrase distance measure.

4 Experimental Results

Our experimental findings demonstrate consistent results across languages and datasets, with only a minor discrepancy observed in the Pearson Correlation Index (PI) for the Yule distance in the Arabic dataset. Despite significant variations in the number of token generations, particularly notable in Arabic compared to other languages, the impact on the effectiveness of phrase distance measures remained limited.

Our key observation is that the selection of a well-behaved phrase distance measure, capable of identifying a suitable vector distance, holds greater significance than the choice of tokenizer. Notably, binary distance measures, lacking specialized embeddings, consistently outperformed specialized paraphrasing distances such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or METEOR (Banerjee and Lavie, 2005) in practical scenarios.

The ability to effectively cluster together phrases with the same meaning and distinguish those with different meanings is vividly illustrated in Figure 2.

These results emphasize the practical utility of binary distance measures and sheds light on their superior performance compared to their specialized counterparts in various linguistic contexts. The effectiveness of a PDM in capturing the nuances of meaning appears to be more closely tied to its inherent properties than the specifics of the tokenization process.

4.1 Differences in Languages and Token Generation

Despite training tokenizers for English, French, German, and Arabic on datasets of approximately the same size, the number of tokens generated varies significantly across languages, as illustrated in Table 1. Notably, the Arabic language exhibits nearly 10 times the token generation compared to the other languages. This discrepancy can be attributed to Arabic being considered a low-resource language (Alyafeai et al., 2021; Mofijul Islam et al., 2022).

However, our analysis indicates that the increased number of tokens in Arabic does not substantially impact the efficacy of downstream metrics. This observation suggests that, despite language-specific variations in token generation, the performance of PDMs remains robust and comparable across diverse linguistic contexts. The resilience of these measures underscores their versatility in handling variations in tokenization outputs, offering consistent performance across languages.

4.2 Effects of Embeddings

Our investigation extended beyond the impact of tokenizers to explore the influence of embeddings on the performance of phrase distance measures. As detailed in Section 3.4,the generic embeddings—vectors were generated with a dimensionality that matched the tokenizer's vocabulary size—by counting the occurrences of tokens in a given phrase. Additionally, the Singular Value Decomposition (SVD) computed an orthonormal basis, aiming to emphasize tokens with high variance in the dataset. This transformation not only enhances the importance of specific tokens but also reduces dimensionality, mitigating the curse of dimensionality and potentially improving vector distance measures.

Subsequently, we utilized Bert embeddings, constructed from coefficients within the Bert model's neural network. Given that these embeddings generate vectors in $\mathbb{R}^n$ and not binary vectors, and considering the inapplicability of n-gram counting and similarity metrics such as BLEU and ROUGE, we confined the embeddings comparison to L_p and cosine vector distances.

Our observations indicate a modest improvement in scores with different embeddings. However, it is noteworthy that distance measures paired with basic binary embeddings consistently outperformed those utilizing SVD or Bert-style embeddings. The superiority of primitive binary embeddings in conjunction with various distance measures underscores their robustness and efficiency in capturing meaningful linguistic nuances across different datasets and languages.

4.3 Effectiveness of Different Distance Measures

Our findings highlight notable disparities in the effectiveness of various distance measures. n-gram counting measures, such as BLEU and ROUGE, exhibit suboptimal performance compared to classical metrics. Even within the ROUGE score, experimentation with different n-gram sizes reveals a diminishing performance trend with higher n. In contrast, the cosine distance, when coupled with straightforward embeddings, consistently outperforms these n-gram-based metrics. L_p-distances display weaker performance than cosine distance across the board.

It is to be notes that all metrics were surpassed by the Yule distance, which consistently yields Silhouette scores and Pearson Correlation indices of pairwise distance matrices in the 0.5-0.9 ranges.

5 Limitations

While our study focuses on the cohesion of similarity and separability of different news headlines and abstracts, assessed through the Silhouette Index (12) and a mean of Pearson Correlation over the matrix elements of pairwise distance matrices (16), it does not delve into deep linguistic features beyond the cohesion and diffusion approach.

Despite the dataset's diversity, encompassing a Semitic language, two Indo-Germanic, and one Romance language, limitations arise from the absence of languages with vastly different grammar and script, such as Mandarin, Chinese, or Hindi.

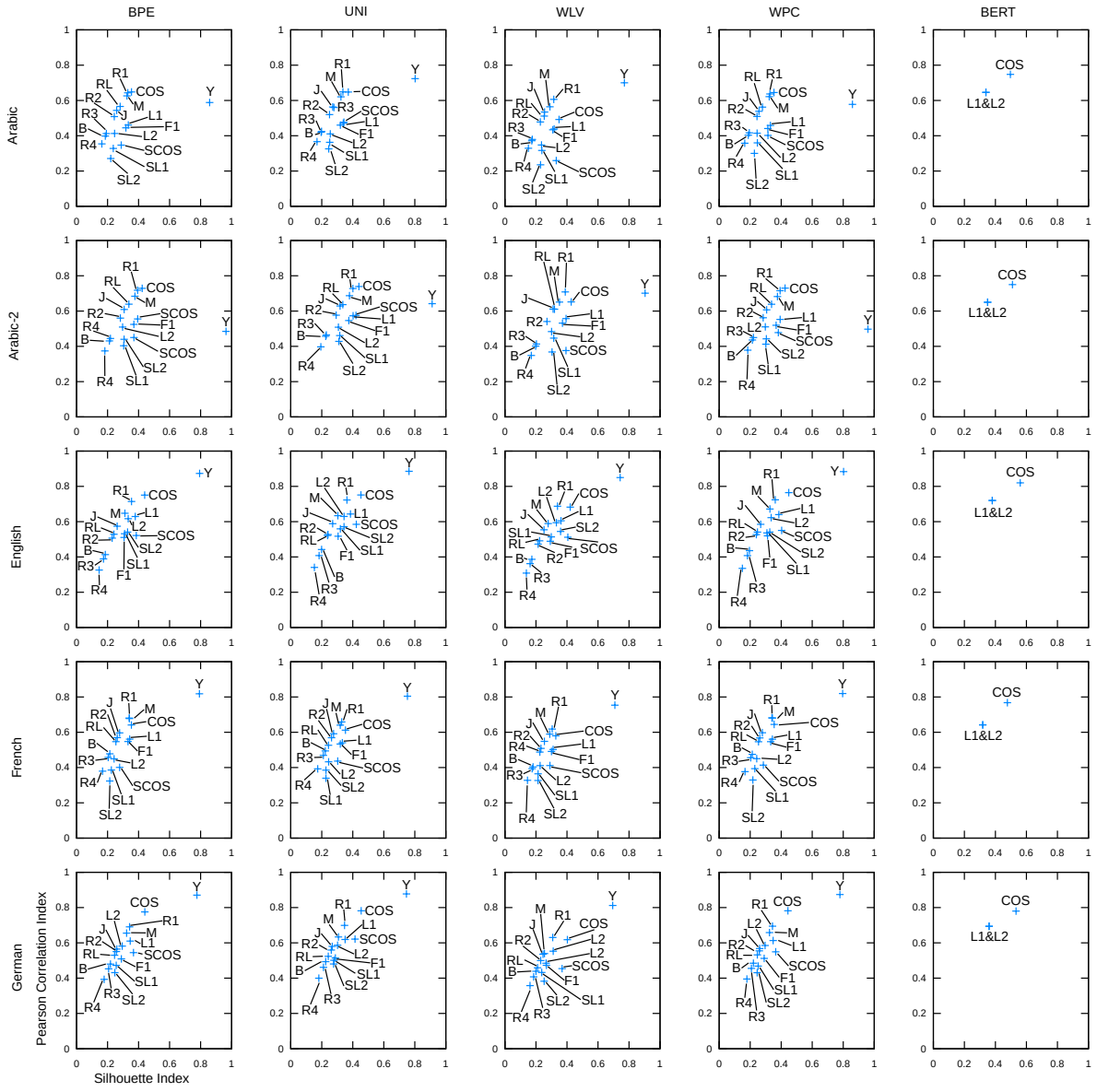

Figure 2: The performance of PDMs according to SI (12) and PI (16) indices. PDMs found in the upper right corners perform inherently better at cohesion of phrases of similar *meaning* and diffusion of phrases with different *meaning* than distance measures in the lower left corners.
PDMs are annotated as follows: **L1**: L_1-distance, **L2**: L_2-distance, **COS**: Cosine Distance, **SL1**: L_1-distance on SVD embeddings, **SL2**: L_2-distance on SVD embeddings, **R1**. . .**4/RL** Rouge 1. . .4/L variant scores, **B**: Bleu score, **M**: Meteor score, **J**: Jaccard distance, **Y**: Yule distance.
Columns outline the different tokenizers used: [BPE] Byte-Pair-Encoding (Sennrich et al., 2016), [UNI] Unigram (Kudo, 2018), [WLV] by WordLevel tokenization, and [WPC] by WordPiece tokenization (Devlin et al., 2019). Further, Bert Embeddings were used with the WordPiece implementation used by Bert.
Rows outline the different datasets: Each dataset contains 20 different meanings with 10 similar news articles per meaning, totaling 200 items.

Language	Words	Bytes	BPE	UNI	WLV	WPC
English	24.7M	159M	37.3M	36.2M	31.3M	39.7M
German	19.0M	144M	48.7M	60.0M	31.6M	51.5M
French	19.1M	129M	64.2M	70.7M	49.4M	67.0M
Arabic	15.8M	172M	554.6M	554.9M	295.0M	577.7M

Table 1: Number of tokens generated from different tokenizers for 50 000 articles in the given languages of the wiki40b dataset.
[Words] outlines the number of words found in the 50 000 articles. [Bytes] corresponds to the size in bytes of the 50 000 articles. The columns outline the number of tokens generated by the [BPE] Byte-Pair-Encoding (Sennrich et al., 2016), [UNI] Unigram (Kudo, 2018), [WLV] by WordLevel tokenization, and [WPC] by WordPiece tokenization (Devlin et al., 2019)

While results between Arabic and the other languages appear similar, subtle variations, such as the slight degradation of the effectiveness of the Yule distance measured by the Pearson Index (16), suggest challenges in extrapolating our findings to linguistically distinct languages.

Additionally, the exclusion of several novel phrase distance measures stems from underlying models that either lacked the resources for retraining or did not support the multilingual nature of this study.

6 Discussion

This study introduced a method for evaluating the effectiveness of phrase distance measures in discerning phrases of the same meaning from those with different meanings. This method, based on a carefully curated dataset, eliminates the need for further human intervention inspired by the work of (Haschka et al., 2021) in bioinformatics. By employing the Silhouette Index and Pearson Correlation of Distance Matrices, our method provides a robust and automated means of assessing diverse phrase distance measures.

Our results challenge the conventional wisdom in the field of Large Language Models (LLMs), revealing that straightforward embeddings and distance measures can outperform widely used metrics such as BLEU and ROUGE. Importantly, these findings hold across varied datasets, showcasing independence from the choice of tokenizers, languages, and phrase lengths.

Notably, for phrase distance measures with non-binary embeddings, the cosine distance emerges as a preferred choice. However, when utilizing binary embeddings, the Yule distance consistently outperforms other distance measures. This outcome has significant implications for the implementation of vector databases. If future databases store encoded phrases as binary data, it could streamline data query and retrieval processes, potentially achieving efficiency gains through Binary Operations and Single Instruction Multiple Data (SIMD) mechanisms.

To further explore the impact of different embeddings, we generated SVD-based embeddings and commonly used Bert embeddings. While our results indicate a favorable effect of Bert embeddings, SVD-based embeddings did not yield similar improvements. Nevertheless, it is noteworthy that the Yule distance continues to outperform Bert embeddings in conjunction with classical distance measures, emphasizing the robustness and efficacy of the Yule distance across various embedding types.

7 Conclusion

In conclusion, this study introduces an automated methodology for evaluating phrase distance metrics, with a particular focus on the cohesion and diffusion dynamics within phrases of similar or distinct meanings. Employing diverse datasets spanning multiple languages, our comprehensive evaluation of various distance measures underscores the consistent superiority of the Yule distance, especially when coupled with binary embeddings. This observed performance extends across linguistic variations, demonstrating language and length independence in our findings.

Furthermore, our exploration into the impact of different embeddings reveals the notable efficacy of binary embeddings, particularly when employed in conjunction with the Yule distance. The results underscore the practical implications of optimizing phrase distance measures, especially in the context of large language models. This work provides valuable insights into refining the performance of such

measures across varied linguistic scenarios.

The significance of our research lies in its contribution to the evolving landscape of natural language processing, where robust and efficient PDMs are essential. By presenting a nuanced understanding of the effectiveness of diverse metrics and embeddings, this study serves as a foundation for future advancements in the optimization and application of phrase distance measures within the realm of large language models.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Zaid Alyafeai, Maged S. Al-shaibani, Mustafa Ghaleb, and Irfan Ahmad. 2021. Evaluating various tokenizers for arabic text classification.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald. 2022. A global analysis of metrics used for measuring performance in natural language processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Thomas Haschka, Jean Baptiste Morlot, Leopold Carron, and Julien Mozziconacci. 2021. Improving distance measures between genomic tracks with mutual proximity. *Briefings in Bioinformatics*, 22(6):bbab266.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *CoRR*, abs/1804.10959.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. 2004. The significance of recall in automatic metrics for mt evaluation. In *Machine Translation: From Real Users to Research*, pages 134–143, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation.

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.

Aigerim Mansurova, Aliya Nugumanova, and Zhansaya Makhambetova. 2023. Development of a question answering chatbot for blockchain domain. *Scientific Journal of Astana IT University*, 15(15):27–40.

Md Mofijul Islam, Gustavo Aguilar, Pragaash Ponnusamy, Clint Solomon Mathialagan, Chengyuan Ma, and Chenlei Guo. 2022. A vocabulary-free multilingual neural tokenizer for end-to-end task learning. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 91–99, Dublin, Ireland. Association for Computational Linguistics.

Alexander Tobias Neumann, Michal Slupczynski, Yue Yin, Chenyang Li, and Stefan Decker. 2023. Citation recommendation chatbot for professional communities. In *Collaboration Technologies and Social Computing*, pages 52–67, Cham. Springer Nature Switzerland.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess,

Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sahar Sohangir and Dingding Wang. 2017. Improved sqrt-cosine similarity measurement. *Journal of Big Data*, 4(1):25.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,

Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Kevin G. Yager. 2023. Domain-specific chatbots for science using embeddings. *Digital Discovery*, 2:1850–1861.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance.