Delving into Evaluation Metrics for Generation: A Thorough Assessment of How Metrics Generalize to Rephrasing Across Languages

Yixuan Wang
New York University
grace621524@gmail.com

Qingyan Chen
Tufts University
tchen4256@gmail.com

Duygu Ataman
New York University
ataman@nyu.edu

Abstract

Language generation has been an important task in natural language processing (NLP) with increasing variety of applications especially in the recent years. The evaluation of generative language models typically rely on automatic heuristics which search for overlaps over word or phrase level patterns in generated outputs and traditionally some hand-crafted reference sentences in the given language ranging in the forms from sentences to entire documents. Language, on the other hand, is productive by nature, which means the same concept can be expressed potentially in many different lexical or phrasal forms, making the assessment of generated outputs a very difficult one. Many studies have indicated potential hazards related to the prominent choice of heuristics matching generated language to selected references and the limitations raised by this setting in developing robust generative models. This paper undertakes an in-depth analysis of evaluation metrics used for generative models, specifically investigating their responsiveness to various syntactic structures, and how these characteristics vary across languages with different morphosyntactic typologies. Preliminary findings indicate that while certain metrics exhibit robustness in particular linguistic contexts, a discernible variance emerges in their performance across distinct syntactic forms. Through this exploration, we highlight the imperative need for more nuanced and encompassing evaluation strategies in generative models, advocating for metrics that are sensitive to the multifaceted nature of languages.

1 Introduction

In the context of Natural Language Processing (NLP), evaluating generative models typically refers to a two-fold process: while the generated output should first of all be a grammatically and semantically plausible utterance in the target language, it should also fulfil in form or meaning the requirements of a specific task the system is built for. For instance machine translation model output is typically assessed based on how well the system output can represent the meaning of a sentence in another language, while outputs of summarization or question answering systems should be conveying factual information about a given context representing information. The evaluation at hand can then seek to gauge the accuracy, fluency, and appropriateness of the output for the given application at the same time.

While a through and accurate evaluation of any NLP system should eventually involve human assessment, due to time and cost considerations, a prominent approach especially during system development typically relies on automatic heuristics which can provide costless reinforcement on the sufficiency or efficacy of the model settings or resources used in system development. Automatic evaluation metrics are generally designed with the principle of comparing the similarity of system output to a gold-standard utterance presenting an example of an accurate system output, by relying on the rate of common words (Papineni et al., 2002; Doddington, 2002). However, such metrics tend to fall back significantly when the output happens to contain a rephrased version of the context due to stylistic or syntactic variations in the generative process. Many languages with rich morphology not only can change in form at the subword level through inflectional or derivational transformations, one can also observe free word order where the same phrase can be written as a combination of the words in many different orders, and still convey the same meaning. In such cases, word-level metrics are known to fail to capture accurate evaluations (Culy and Riehemann, 2003; Callison-Burch et al., 2006; Birch et al., 2010; Mathur et al., 2020). Alternatively, (Popović, 2015) proposed n-gram match-

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 23–31
November 1, 2023 ©2023 Association for Computational Linguistics

ing at the character level, which has been more appropriate for the evaluation in morphologically-rich languages. However, matching based approaches still might miss semantic nuances in the generated language. Recent studies proposed the alternative approach to use vector similarity in distributed representations (Zhang et al., 2019). This method provides a better semantic notion over simple word matching heuristics, yet there is not a well-established understanding on the robustness of pre-trained language representations and how well they may generalize across languages and domains.

While valuable, each metric has its challenges, especially given the intricate tapestry of global languages. Previous work has compared the performance of evaluation metrics in different tasks (Liu et al., 2016; Shen et al., 2022; Moghe et al., 2022), however, a task-agnostic analysis that focuses on providing insight on the assessment of generalization capability in generative language models and its measurement across languages with different syntactic typology has never been performed. Our study embarks on an extensive examination of evaluation metrics within a linguistic framework where our objective is to understand how these metrics perform in capturing the essence of rephrased language and generalize across diverse syntactic structures and linguistic complexities. For this purpose, we select four prominent automatic evaluation metrics representative of a different approach to evaluation metric formulation: BLEU (Papineni et al., 2002), chrf (Popović, 2015), NIST (Doddington, 2002) and BERTScore (Zhang et al., 2019) and use these metrics to compute the similarity across collections of sentences that are paraphrases of each other, in 71 different languages from 12 distinct language families, and measure how different linguistic features affect the applicability of each metric in similarity detection across paraphrased language. Our study aims to extend the understanding of evaluation metric performance and highlights potential gaps and areas for further research in considering the future of generative models and how they can be better developed to capture linguistic nuances. Through this endeavor, we aim to refine the evaluation process for generative models across multiple languages and promote the study of generative models in potentially many new under-studied languages.

2 Evaluation Metrics for Language Generation

In this study, we focus on sentence-level generation and adopt four commonly used evaluation metrics developed for the automatic evaluation of machine translation. Here we briefly define the formulation of each method.

2.1 BLEU (Bilingual Evaluation Understudy)

Introduced by Papineni et al. (2002), BLEU was one of the first automated metrics comparing machine-generated translations to human reference translations. The BLEU score, typically between 0 (worst) and 1 (best), is given by:

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

where:

- BP is the brevity penalty,
- w_n are the weights for each n-gram (usually set to 1/N),
- p_n is the precision for the n-th n-gram.

Usually, if a candidate sentence is shorter, the n-gram tends to get a higher score. The brevity penalty helps control this effect by scaling the frequency over the sentence length.

$$BP = \begin{cases} 1 & \text{if c > r} \\ \exp^{(1-r/c)} & \text{if c} \leq \text{r} \end{cases} \quad (2)$$

The second term in Eq. 2.1 ensures all n-grams's weights be uniformly distributed. Since the overall accuracy decreases with the increase of n-gram, the general n-gram is taken as 4-gram.

2.2 chrF

Building on BLEU's success, chrF is a metric that assesses n-gram similarity at the character level, intuitively more suitable for the evaluation of morphologically-rich languages. The overall CHR-F score is the weighted harmonic mean of the F-scores for each n-gram size. The weights are determined by the frequency of each n-gram size in the reference text.

$$\text{chrF} = 2 \times \frac{P \times R}{P + R} \quad (3)$$

where:

- P is character-level precision,
- R is character-level recall.

Unlike BLEU, the metric is not sensitive to the position of the n-grams in the sentence, making it a more flexible and robust metric.

2.3 NIST

Developed by the National Institute of Standards and Technology[1], NIST improves upon BLEU's formulation, with an emphasis on rewarding rare n-grams. The NIST score is given by:

$$\text{NIST} = \frac{\sum_{n=1}^{N} w_n \log p_n}{\sum_{n=1}^{N} w_n} \tag{4}$$

where:

- w_n are the weights for each n-gram, which are adjusted based on the informativeness of the n-gram,
- p_n is the precision for the n-th n-gram.

2.4 BERTScore

A more contemporary metric, BERTScore, taps into BERT's contextual embeddings to determine text quality. The similarity between a system output and reference sentence is computed as:

$$P_{\text{score}} = \frac{1}{N_P} \sum_{i=1}^{N_P} \max_{j=1}^{N_R} \cos(e_{P_i}, e_{R_j})$$

$$R_{\text{score}} = \frac{1}{N_R} \sum_{j=1}^{N_R} \max_{i=1}^{N_P} \cos(e_{P_i}, e_{R_j})$$

$$F1_{\text{score}} = \frac{2 \cdot P_{\text{score}} \cdot R_{\text{score}}}{P_{\text{score}} + R_{\text{score}}}$$

where:

- N_P and N_R are the number of tokens in P and R, respectively.
- e_{P_i} and e_{R_j} are the BERT embeddings of the i-th token in P and the j-th token in R, respectively.
- cos denotes the cosine similarity between two vectors.

[1]https://www.nist.gov

3 Experimental Methodology

Metrics have undeniably evolved over time, mirroring the advancements in generative models. The above metrics represent this transformation, showcasing the progression from rudimentary n-gram matching to nuanced evaluations via deep learning embeddings. a desired property in each generative language model is to be able to produce plausible language in as many stylistic or syntactic variations the language allows. In order to assess how sensitive each metric is to generalization in the subword or phrase level syntactic structures, i.e. rephrasing, we design a set of experiments that compute similarity between paraphrased utterances in different languages.

By the nature of their design, some metrics may be able to capture certain typological forms and patterns better than others, and thus correlate better with languages with those features. In order to test how each metric may suit better capturing grammatical generalization in different languages, we perform an in-depth analysis over the similarity scores and how well they correlate with different types of linguistic features.

3.1 Data

The experiment uses data from the TaPaCo Dataset (Scherrer, 2020), which is a multilingual paraphrase corpus extracted from the Tatoeba platform[2], an online platform that collects translations via crowd-sourcing that allows the public mass to provide translations and annotations to sentences. The TaPaCo dataset is built by matching sentences within the Tatoeba database via context automatically based on the multilingual pivoting approach introduced by Lewis and Steedman (2013). The matched sentences are organized in sets with verified non-trivial accuracy of between 50 to 75 percent. The database consists of roughly 1.9 million sentences, with a range of 200 to 250,000 sentences in each language. Of the 73 languages in the TaPaCo dataset, 42 are languages from the Indo-European language family group, the remaining 31 are composed of languages from various families such as Afro-Asiatic, Austronesian, Sino-Tibetan, Turkic, Uralic, and other constructed languages. Only the paraphrased sentences from the TaPaCo dataset are used in the experiment to compute the metric scores for each language. Any annotations

[2]https://tatoeba.org

of the sentences are stripped from the data when computing the metric scores from the sentences.

3.2 Metrics

In our experiments, we use the nltk (Natural Language Toolkit) version 3.7 for calculating BLEU, chrF and NIST scores.

Typological feature data for 73 languages were surveyed from the URIEL database (Littell et al., 2017) that contains a collection of language typology data via the lang2vec[3] library. This database was initially developed as part of DARPA's (Defence Advanced Research Project Agency's Low Resource Language for Emergent Incidents project) LORELEI project to develop tools for automated human language technology for low resource languages. For our examination, we select five categories of language typological features:

1. geography ("geo") – Geographic distances between languages on the globe

2. syntax average ("syntax_average") – an average score representing the distinctness of the paradigms observed in a given language in terms of syntax

3. phonology average ("phonology_average") – an average score representing speech sounds production rules of a language

4. inventory average ("inventory_average") – an average score representing features related to phonetic inventories or the lexical patterns of a language

5. learned ("learned") – a learned predictive feature dataset used for typological predictions

Feature datum of the 71 languages selected corresponding to the overlapping languages between the TaPaCo dataset and the feature data for languages available in the lang2vec database are surveyed for this experiment. Each set of the typological feature data is given as a single high-dimensional vector that represents the feature datum of the language in question in numerical values. Some represent the presence or absence of certain features in the language. Thus, the average feature score of languages cannot be collected trivially by taking means of the independent numerical scores. In order to preserve data, these high dimensional feature vectors are transformed into one-dimensional vectors with one point for each language using PCA (Principal Component Analysis) (Bro and Smilde, 2014) to be compared with the metric scores computed using the TaPaCo data. To collect the metric scores on the TaPaCo dataset, sentences within the same paraphrased group in the same language are split off into pairs in order to compute their metric scores. A mean average of the scores from then sentence pairs in each language is taken to represent the language's score evaluated by a particular metric. Finally, to examine the correlation relation between the typological features of a language and the evaluation metric performances on the language as a whole, Pearson's correlation coefficient was computed between each different metric score and the average transformed typological feature. Figures 1 to 5 illustrate how typological features are distributed across language families in linguistic features, such as syntax, phonology, inventory, geology, etc.

[3]https://github.com/antonisa/lang2vec

4 Results

The metric scores graph (Fig. 6) presents the distribution of all metric scores computed over paraphrases and organized by language family:

- Constructed Languages: toki(Toki Pona), tlh(Klingon; tlhIngan-Hol), vo(Volapük), jbo(Lojban)

- Afro-Asiatic: ar(Arabic), ber(Berber), he(Hebrew), kab(Kabyle)

- Austroasian: id(Indonesian), tl(Tagalog), war(Waray), Creolecbk(Chavacano)

- Indo-European: af(Afrikaans), be(Belarusian), bg(Bulgarian), bn(Bengali), br(Breton), ca(Catalan), cs(Czech), da(Danish), de(German), el(Greek), en(English), eo(Esperanto), es(Spanish), fr(French), gl(Galician), gos(Gronings), hi(Hindi), hr(Croatian), hy(Armenian), io(Ido), is(Icelandic), it(Italian), kw(Cornish), la(Latin), lfn(Lingua Franca Nova), lt(Lithuanian), mk(Macedonian), mr(Marathi), nb(Norwegian Bokmål), nds(Low German), nl(Dutch), orv(Old Russian), pes(Iranian Persian), pl(Polish), pt(Portuguese), ro(Romanian), ru(Russian), sl(Slovenian), sr(Serbian), sv(Swedish), uk(Ukrainian), ur(Urdu)

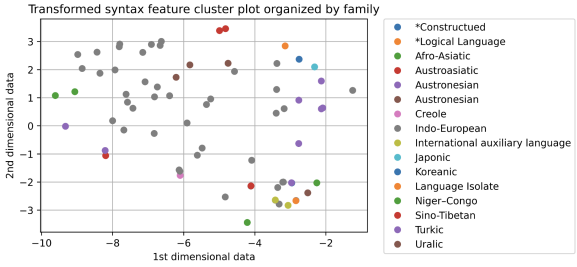

Figure 1: Scatter Cluster Plot of Syntactic Information of languages, grouped by language family

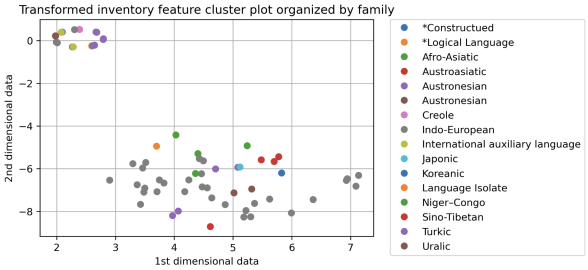

Figure 2: Scatter Cluster Plot of Inventory Information of languages, grouped by language family

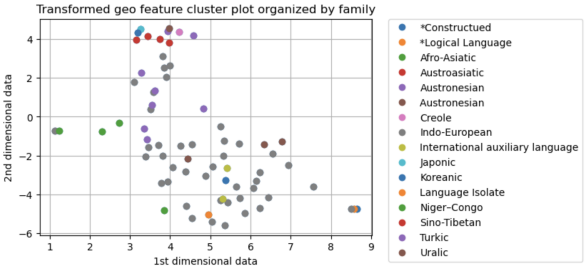

Figure 3: Scatter Cluster Plot of Geology Information of languages, grouped by language family

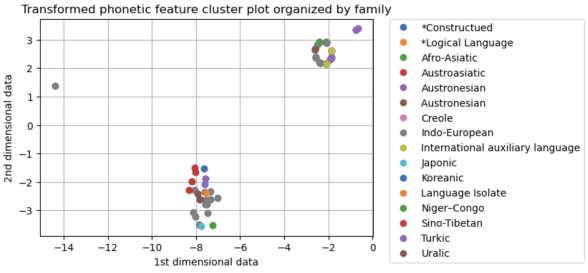

Figure 4: Scatter Cluster Plot of Phonetic Information of languages, grouped by language family

- International auxiliary language: ia(Interlingua), ie(Interlingue)
- Japonic: ja(Japanese)
- Koreanic: ko(Korean)
- Language Isolate: eu(Basque)
- Niger–Congo: rn(Kirundi)

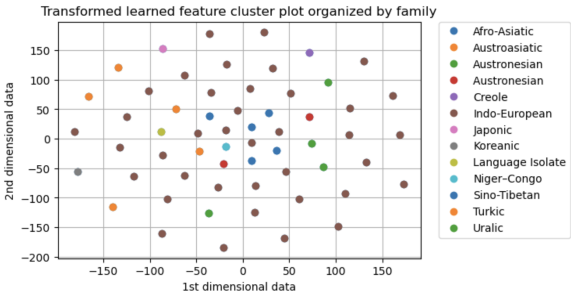

Figure 5: Scatter Cluster Plot of Learned Information of languages, grouped by language family

- Sino-Tibetan: cmn(Mandarin Chinese), wuu(Wu Chinese), yue(Yue Chinese)
- Turkic: az(Azerbaijani), ota(Turkish, Ottoman), tk(Turkmen), tr(Turkish), tt(Tatar), ug(Uyghur)
- Uralic: et(Estonian), fi(Finnish), hu(Hungarian)

On average, we observe the highest BLEU scores are computed in the Creole language with 0.3692516 points followed by the Indo-European language family, with an average of 0.2861689. Average BLEU scores for Japonic, Koreanic, Niger-Congo, Afro-Asiatic, Turkic and Uralic languages are much lower with scores ranging as 0.14158, 0.196627, 0.177759, 0.0.23553, 0.22831 and 0.2283065, respectively. For these language groups with relatively complex morphology, we observe the chrF scores, on the other hand, to be much higher on average, with scores of 0.43876 in Turkic, 0.50198 in Uralic, 0.48934 in Afro-Asiatic and 0.50487 in Niger-Congo languages. In Japonic, Koreanic and Sino-Tibetan languages, the scores are relatively low, with 0.38995, 0.35384 and 0.30561 respectively, indicating neither n-gram matching based metric are able to capture the rephrasing in example sentences.

NIST scores are also highest for the Indo-European languages with an average of 0.80087 and AustroAsiatic languages with an average score of 0.70759, however, the scores relatively remain high for morphologically-rich languages, such as in Afro-Asiatic family, the average NIST score is 0.71675, followed by 0.64827 in Turkic, 0.69190 in Uralic languages, indicating a general improvement for better balancing the more frequent and rare n-gram statistics. In Koreanic and Niger-Congo the scores are very low, with 0.47798 and 0.37228, respectively.

Finally, the distributed semantic similarity score BERTScore obtains the overall, with an average of 0.8684 in Indo-European, ranging to slightly different values in different language families with 0.8268 in Turkic, 0.85171 in Uralic, 0.87922 in Afro-Asiatic, 0.84763 in Niger-Congo, 0.87247 in Koreanic and 0.86648 in Japonic languages, suggesting to be the most applicable metric across languages with varying typological characteristics.

We further explore the details of how each metric respond to different linguistic aspects of language by analyzing the correlation between evaluation metric scores and various linguistic typological features. Our analysis yields a spectrum of results that underscore the intricacies of language generation evaluation. Considering the provided correlation coefficients:

Syntactic Average:

- **BLEU, chrF, and NIST:** exhibited negative correlations with syntactic construction, implying that as syntactic complexity increases, thet fall back in capturing similarities in the outputs and reference language utterances. This might hint that these metrics struggle to capture syntactic nuances, or the general process of rephrasing that we explicitly integrate in our experimental setting, which is not surprising due to their heavy relying on ordered sequential patterns.

- **BERTScore:** exhibits a positive correlation suggesting its potential aptness in gauging syntactic richness or its increased robustness to languages with complex syntactic patterns.

Geography:

- **BLEU and BERTScore:** Both metrics indicate a relationship between geographical distances and their evaluation scores, possibly hinting at regional linguistic patterns that these metrics are sensitive to. These results are in line with the metric scores in Figure 6 and how they show clear differences across language families from different geographical locations in the distribution of either metric.

- **chrF and NIST:** Negative correlations may imply a diminished sensitivity or lack of significance related to geographical linguistic nuances or a different type of sensitivity to regional patterns.

Inventory:

- Most metrics showed a negative inclination with the exception of NIST, which had a very marginal positive correlation. This could signify a negative relationship between the effect of phonetic inventories to the specific task of similarity in case of varied syntactic expression. Notably, BERTScore's significantly negative score could highlight a potential shortfall of n-gram based methods being able to capture lexical variety and how it may reflect in the generated language.

Phonology:

- **BLEU, CHAR-F, and NIST:** leaning negative, suggest that traditional metrics might not be fully equipped to capture the richness of speech sound production rules.

- **BERTScore:** Moves in a positive direction, suggesting that embedding-based metrics like BERTScore might offer a new perspective to represent cross-lingual distributed information.

Learned:

- We find mixed results with learned linguistic feature representations. Our findings indicate that the sensitivity of metrics to learned predictive feature datasets is varied. BLEU, CHAR-F, and NIST have negative correlations, in contrast with BERTScore which has positive correlation, emphasizing the potential alignment of data-driven approaches in their distributed nature of information.

In sum, while some evaluation metrics manifest robustness in certain linguistic dimensions, clear disparities are evident across different syntactic and typological realms. Our findings propose significant differences in applicability of certain evaluation metrics to sets of language families with general typological differences in their syntactic characteristics. We find n-gram based metrics like BLEU to be very limited in applicability to relatively simple syntactic constructions observed in Indo-European languages, however, generally failing to provide any informative score in majority of language families with the common characteristic of complex morphosyntactic properties. Although chrF was developed in a way to cope with this limitation, we still fail to find it robust enough to

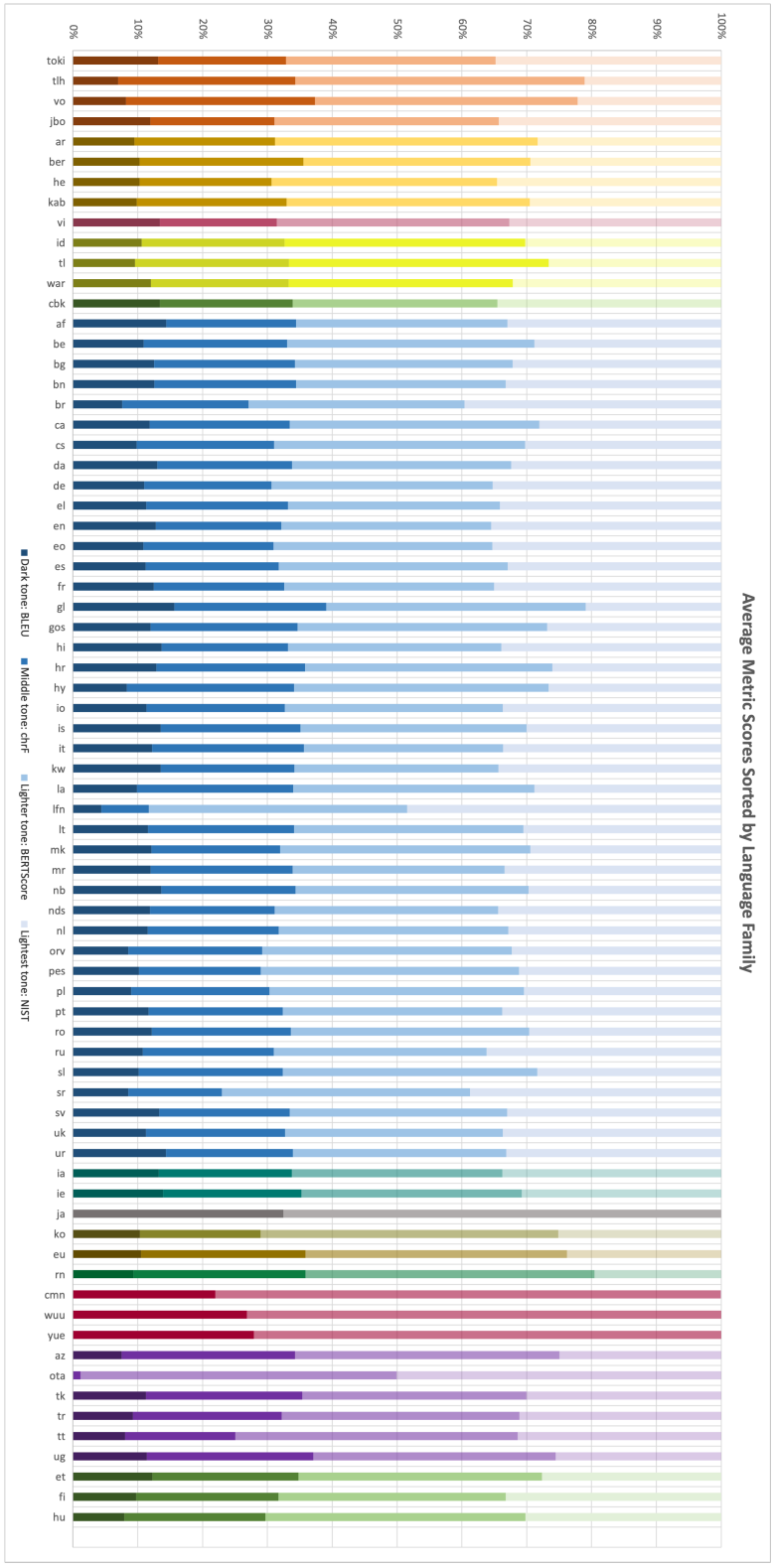

Figure 6: The fluctuation of average scores of different languages computed using different metrics. Each language family is represented with a different color (Constructed: Maroon, Afro-Asiatic: Orange, Austroasiatic: Pink, Austronesian: Lemon, Creole: Pine, Indo-European: Blue, International Auxiliary Language: Teal, Japonic: Grey, Koreanic: Crocodile, Language Isolate: Brown, Niger-Congo: Emerald, Sino-Tibetan: Crimson, Turkic: Purple, Uralic: Olive). The evaluation results in each language are presented using the metric scores BLEU (darkest tone), chrF (middle tone) BERTScore (lighter tone) and NIST (lighest tone), respectively.
**NIST scores are scaled to the range of 0 to 1 using the formula: $scaled\_score = \frac{NIST\_score - min\_NIST}{max\_NIST - min\_NIST}$ with $max\_NIST = 1.900$ and $min\_NIST = 0$.

	Syntactic	Geography	Inventory	Phonology	Learned
BLEU	-0.18217	0.22827	-0.09892	-0.18603	-0.13972
NIST	-0.16295	-0.35622	0.02184	-0.14601	-0.18125
chrF	-0.18701	-0.13178	-0.12033	-0.18603	-0.0402
BERTScore	0.35427	0.17401	-0.30546	0.16748	0.16193

Table 1: Correlation results between each metric and typological feature

be applicable to different language families, but, a better alternative in a subset of agglutinative languages like Turkic and Uralic language families. A not well-adopted metric in the recent years, NIST had shown interestingly robust performance across languages supported by a more balanced formulation in n-gram statistics, as indicated in its ability to perform relatively well in the evaluation of language generated in sparse languages. The distributed space similarity metric BERTScore had in overall the best results in being able to capture syntactic, semantic and phonological information across languages much better compared to all other surface-level heuristics. We remain to future work how well it generalizes across languages and domains with limited data available to build pre-trained representations.

The insights gleaned underscore the imperative for a multifaceted, holistic approach to evaluation, one that is attuned not only to textual fidelity but also to the vast tapestry of linguistic features that define our global languages. Future endeavors in the realm of NLP should prioritize the development and refinement of evaluation metrics that genuinely reflect the richness of human languages.

5 Conclusion

This paper provided an analytic study on the evaluation of language generation and how optimal evaluation measures can be developed in a task-agnostic way that can generalize well across different rephrasing choices that are common in natural language. In order to provide insight on the applicability of commonly used evaluation metrics for language generation, we performed extensive experiments on multilingual paraphrase collections and measured the robustness and efficacy of each metric in capturing syntactic variations across languages with varying syntactic typology. Our findings confirm the general fallback of surface level matching based heuristics in both applicability and accuracy across languages with different characteristics, and suggest the future of evaluation in language generation lies in the direction of pre-trained language representation. We hope our study helps better understand how more robust evaluation metrics can be developed, eventually promoting more studies in the development of generative models in many under-studied language families.

Limitations

In spite of the task-agnostic evaluation setting adopted in our study, it's worth discussing potential limitations on the applicability of our findings when deployed in specific generative tasks or domains. Our study mainly aims to inspire a more general approach to the design of evaluation of language generation, with a focus on linguistic typology and how syntactic characteristics may affect the efficacy of evaluation metrics of different nature. In this scope, we adopt two major types of approaches to metric formulation, surface level heuristics and distributed semantic space similarity comparison. There may exist additional metrics not in the scope of this project, which we leave the reader to experiment with in similar settings. In this context, we do not strongly suggest the adoption of a particular metric, but generally aim to provide a novel perspective on different language families and how their typological characteristics should be considered in metric design. Eventual deployment of a particular metric in a given task may yield additional insight on another level that may not have been captured in our specific experimental design. We invite all readers to beware again the nature of controlled scientific methodology and how each experimental setting is refined to verify a particular scope and hypothesis.

Acknowledgements

We express our gratitude to the anonymous reviewers whose invaluable feedback sharpened our work. We are especially indebted to Professor Ataman for her expert guidance and steadfast support throughout this endeavor. We extend our heartfelt thanks to New York University, particularly the Dean's

Undergraduate Research Fund (DURF) and the NYU Courant Institute of Mathematical Sciences' Pathways to AI Program, for their crucial financial backing.

References

Alexandra Birch, Miles Osborne, and Phil Blunsom. 2010. Metrics for mt evaluation: evaluating reordering. *Machine Translation*, 24:15–26.

Rasmus Bro and Age K Smilde. 2014. Principal component analysis. *Analytical methods*, 6(9):2812–2831.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Chris Culy and Susanne Z Riehemann. 2003. The limits of n-gram translation evaluation metrics. In *Proceedings of Machine Translation Summit IX: Papers*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Mike Lewis and Mark Steedman. 2013. Unsupervised induction of cross-lingual semantic relations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 681–692.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2022. Extrinsic evaluation of machine translation metrics. *arXiv e-prints*, pages arXiv–2212.

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.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.