Tokengram_F, a fast and accurate token-based chrF++ derivative

Sören Dréano
ML-Labs
Dublin City University
soren.dreano2@mail.dcu.ie

Derek Molloy
School of Electronic Engineering
Dublin City University
derek.molloy@dcu.ie

Noel Murphy
School of Electronic Engineering
Dublin City University
noel.murphy@dcu.ie

Abstract

The tokengram_F metric presented in this paper is a novel approach to evaluating machine translation that has been submitted as part of the WMT23 challenge. It offers a new perspective on evaluating machine translation that takes advantage of modern tokenization algorithms to provide a more natural representation of the language in comparison to word *n*-grams.

Tokengram_F is an F-score-based evaluation metric for Machine Translation that is heavily inspired by chrF++ and can act as a more accurate replacement. By replacing word *n*-grams with *n*-grams obtained from tokenization algorithms, tokengram_F captures similarities between words sharing the same semantic roots.

While requiring minimal training based on an open corpus of monolingual datasets, the tokengram_F metric proposed still retains excellent performance that is comparable to more computationally expensive metrics. The tokengram_F metric demonstrates its versatility by showing satisfactory results, even when a tokenizer for a specific language is not available. In such cases, the tokenizer of a related language can be used instead, highlighting the adaptability of the tokengram_F metric to less commonly-used languages.

1 Introduction

Machine Translation (MT) is a subdomain of Neural Language Processing (NLP) that is focused on the translation of one natural language to another, with the aim of producing natural-sounding sentences. To evaluate the quality of algorithm-generated translations, evaluation metrics provide quantitative scores to objectively assess the accuracy of the model. Machine-generated translations are compared to human-generated translations in different ways depending on the evaluation metric. In recent years, machine translation has seen a great deal of progress in terms of accuracy and fluency. However, there is still a need for more robust evaluation metrics that can effectively measure the quality of machine-generated translations.

Popular metrics in MT include BLEU (Papineni et al., 2002), which measures the overlap between sequences of words in the reference and generated texts; chrF (Popović, 2015), which is an F1 score at the character-level; chrF++ (Popović, 2017), which extends chrF with word *n*-grams; TER (Snover et al., 2006), which counts the erroneously-aligned words between the reference and the generation, and METEOR (Banerjee and Lavie, 2005), which takes into account synonyms and stems.

More recent metrics rely on neural network architectures, such as COMET (Rei et al., 2020) and MS-COMET (Kocmi et al., 2022). By using the similarity of vector representations of the generated translation and the reference translation, they provide state-of-the-art machine evaluation of translations at the cost of being expensive to train and compute.

Since its third instance in 2006, the Workshop on Statistical Machine Translation (WMT) has released an evaluation task to compare metrics each year. Generated translations are usually ranked by humans, and the correlation coefficient between the human-performed ranking and the evaluation metric-performed ranking determines the quality of the metric.

2 Tokengram_F

2.1 Tokenization

As text cannot be directly processed by machine learning algorithms, it first has to be converted into a numerical representation. Tokenization splits the text into smaller character sequences, including but not limited to phonemes, syllables, letters, words or base pairs, collectively named tokens. The tokenization process also often consists of adding special tokens, such as the unknown <unk> token to represent never-seen characters or the padding

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 730–737
December 6–7, 2023. ©2023 Association for Computational Linguistics

<pad> token to pad the sentence to a fixed length. Each token can be converted to and from a unique identifier, which is usually an integer between 0 and the maximum vocabulary size minus one.

2.2 chrF++

An *n*-gram refers to a consecutive series of *n* tokens that are extracted from a given corpus of text or speech, with these units of text being defined based on the particular context of the application. A character-gram, or unigram, is a token that contains exactly one character, while a word-gram contains an entire word. chrF++ is an F-score using both word-grams and character-grams to compare the generated translation to the reference translation. The general formula is

$$ngrF\beta = (1+\beta^2)\frac{(ngrP \times ngrF)}{(\beta^2 \times ngrP + ngrR)}$$

where β determines the weight of the recall as discussed in Section 3.4.3.

2.3 Modern tokenization algorithms

Subword-based tokenization divides words depending on their number of occurrences in the training data. Subwords can be combined to represent less frequent words or even words that were not present in the training data. For instance, in cases where a word such as “decaying” is absent from the vocabulary, an English tokenizer may represent it by combining the “decay” and “ing” tokens.

Byte Pair Encoding (BPE) (Sennrich et al., 2016), had been used for data compression long before it was ever applied in NLP-related tasks. After first counting the frequency of each unique word in the training data, BPE merges frequent occurrences of subword pairs until it reaches the desired vocabulary size.

Instead of starting from a small vocabulary representing the set of unique words, and growing in size from there (as in BPE), Unigram (Kudo, 2018) initialises its base vocabulary to a large number of symbols and then trims it down to the desired size. It is analogous to factor analysis, as at each step it calculates the loss of information that would be induced by removing each token, and then erases the less important ones from its vocabulary.

The sentence “The kingly sovereign governs” becomes:

```
words:  "The"  "kingly"  "sovereign"  "governs"
tokens:  "The"  "king"  "ly"  "sovereign"  "govern"  "s"
```

2.4 Replacing word-grams

When using word-grams for scoring, each word is compared regardless of its proximity to other words. For example, the words “say” and “saying” share a common root but this link would be lost when using word-grams.

In this work, the authors claim that modern tokenization algorithms can be used instead of word-grams to split the text in a more natural manner that reflects the structure of each language. Tokengram_F is an evaluation metric derived from chrF++ that replaces the use of word-grams by tokens learned either by Unigram or by BPE.

3 Methodology

3.1 Framework

SentencePiece (Kudo and Richardson, 2018) is a fast data-driven text tokenizer and detokenizer implementing the Unigram algorithm. The vocabulary size (number of individual tokens) needs to be provided before training. The minimum vocabulary size would consist of the number of special tokens and individual characters of the alphabet for each language. A large vocabulary size might lead to overfitting and a reduction in the effectiveness of the model, given that some parameters will be dedicated to rare words.

3.2 Training

The tokengram_F score uses the same 3-letter ISO-639-2 language code as the Tatoeba dataset, while the WMT tasks rely on the 2-letter ISO-639-1 language code. The website of the Library of Congress (Library of Congress, 2017) was used for conversions between the two norms.

The Tatoeba Translation Challenge (Tiedemann, 2020) is an initiative that aims to evaluate the effectiveness of MT systems on a large, diverse, and high-quality parallel corpus. While the main training data relies on OPUS (Tiedemann, 2012), which provides open-source sentence-aligned text corpora to support data-driven NLP, Tatoeba also provides monolingual datasets extracted from CirrusSearch Wikimedia dumps (Foundation, 2023).

Out of the 279 different languages available, 240 had a sufficiently large corpus to be included in the work described in this paper.

3.3 Exceptions

As the tokengram_F metric is dependent on the utilization of the Tatoeba monolingual datasets for tokenizer training, adaptations were necessary to accommodate languages that are not represented within this dataset.

3.3.1 Livonian:

While there is no dataset in the Tatoeba Translation Challenge to train a Livonian tokenizer, the Latvian tokenizer produced satisfactory results and was utilised as a substitute, highlighting the versatility of the tokengram_F metric and its ability to accommodate languages that are less frequently used.

3.3.2 Serbian and Indonesian:

The Tatoeba challenge does not offer monolingual datasets for neither Serbian nor Indonesian. Nevertheless, the Tatoeba Wikimedia data, which are appropriate for tokenizer training and available for both the Serbian and Indonesian languages, were employed as a substitute.

3.4 Tokengram_F parameters

3.4.1 Tokenization algorithm:

While Tokengram_F can be used with any tokenization algorithm, this study examined both BPE and Unigram.

3.4.2 *n*-gram length:

This parameter determines the number of items in the reference that will be compared with each item in the source sentence, to assess the degree of correspondence of the two sentences.

Previous work (Popović, 2015, 2017) has indicated that for chrF++ there is no necessity to set the maximum word *n*-gram length beyond N=6.

3.4.3 Beta:

In this metric, the relative importance of precision and recall in the evaluation metric is determined by the β parameter. When beta is equal to 1.0, precision and recall have equal importance, while when beta is equal to 3.0, recall is three times more significant than precision. Previous research (Popović, 2015) has evaluated two beta values, 1.0 and 3.0, with the latter being considered “the most promising variant” due to the higher correlations it obtained.

3.4.4 Vocabulary size:

The goal of the present study is to mitigate the effect of infrequent words on the accuracy of the tokengram_F metric. To investigate the influence of vocabulary size on the performance of the metric, three tokenizers were trained for each language using vocabulary sizes of 16,000, 32,000, and 50,000 tokens. As the average vocabulary size tends to decrease from one year to the next (Libovický, 2021), wider vocabularies have not been examined.

3.5 Optimal parameters

3.5.1 Finetuning

The optimal parameters were determined based on achieving the highest average correlation among segments or systems across three datasets: WMT20 (Mathur et al., 2020), WMT21 (Freitag et al., 2021), and WMT22 (Freitag et al., 2022).

Initially, the *n*-gram length was assessed at values of 3, 6, and 9, while maintaining a vocabulary size of 50,000. Consistent with the findings of the original paper, a *n*-gram length of 6 demonstrated the strongest correlation as shown in Table 1, and thus it was chosen for subsequent evaluations.

Subsequently, the beta values of 2, 3, and 4 were examined specifically for an *n*-gram length N of 6. The best overall correlation is obtained with Unigram and β=3.0.

Table 3 presents the results obtained with a vocabulary size of 32,000. As with the previous vocabulary size, the choice of the tokenization algorithm only slightly affects the results. A β of 3.0 or 4.0 seems to give the best results.

As shown in Table 2, the vocabulary size of 16,000, which was the smallest size examined, exhibits generally weaker correlations compared to larger sizes, thus precluding the exploration of smaller sizes.

3.5.2 Results

Despite the marginal disparity, when the results are not rounded, the optimal parameters for tokengram_F are found to be a vocabulary size of 50,000, unigram tokenization, an *n*-gram length N=6, and a β value of 3.0.

3.6 Source code

The source code of tokengram_F is available at https://github.com/SorenDreano/tokengram_F.

4 Conclusion

Instead of using word *n*-grams as a basis for comparing a generated translation with a reference translation, tokengram_F utilises contemporary tokenization algorithms to accomplish this task. As a result, words that share common roots are deemed similar, regardless of whether they are exact matches or not.

The results obtained from evaluating the tokengram_F metric on the WMT20, WMT21, and WMT22 datasets indicate that the use of tokens generated through the SentencePiece framework leads to improved performance compared to the use of traditional word-grams in the chrF++ metric. Full results are displayed in Appendix A. The tokengram_F metric is a simple and efficient method for obtaining a reasonable correlation with human rankings, with the added benefit of requiring minimal training time to be applied to new languages.

In the segment-level task of the WMT22 edition, tokengram_F managed to obtain better overall correlations than any other metric that could provide results for all language pairs, including ones that require extensive neural networks to operate. With the exception of two tasks, tokengram_F outperformed both chrF and chrF++ metrics.

In conclusion, the tokengram_F metric is presented as a promising alternative for evaluating the quality of machine translations, as it offers a simple and efficient solution with above-average performance compared to other models. The findings of this study provide strong evidence of the potential of the tokengram_F metric as a valuable evaluation tool for machine translation. Its combination of simplicity, efficiency, and adaptability make it an attractive alternative to existing metrics and a promising direction for future research in the field.

5 Further work

The optimal number of tokens in a tokenizer may vary depending on the language. Subsequent research could concentrate on determining the most suitable vocabulary size per language.

The majority of the tokenizers were trained using the MonoLinguage Datasets from the Tatoeba Challenge, which are based on data from the Wikimedia Foundation. It remains possible that alternate data sources may produce varying results.

6 Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

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.

Wikimedia Foundation. 2023. Wikimedia downloads.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu - neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jindřich Libovický. 2021. Jindřich's blog – machine translation weekly 86: The wisdom of the wmt crowd. Online, Accessed: 07.02. 2023.

Library of Congress. 2017. Codes for the representation of names of languages.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

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.

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, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *CoRR*, abs/2009.09025.

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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge - realistic data sets for low resource and multilingual MT. *CoRR*, abs/2010.06354.

A Appendix. Tables

Table 1: Correlations over all metrics depending of the hyperparameters with a vocabulary size of 50,000

Vocabulary	50000									
Algorithm	BPE					Unigram				
Beta	4	3			2	4	3			2
n-gram	6	9	6	3	6	6	9	6	3	6
System WMT20	0.875	0.871	0.876	0.871	0.877	0.876	0.871	0.876	0.871	0.877
System WMT21	0.715	0.715	0.716	0.713	0.717	0.716	0.715	0.717	0.714	0.718
System WMT22	0.837	0.831	0.834	0.834	0.829	0.836	0.831	0.835	0.834	0.830
Segment WMT20	0.277	0.274	0.277	0.279	0.277	0.276	0.273	0.277	0.278	0.277
Segment WMT21	0.158	0.155	0.158	0.166	0.157	0.160	0.159	0.158	0.170	0.152
Segment WMT22	0.398	0.393	0.398	0.399	0.400	0.398	0.393	0.399	0.397	0.400
Average	0.543	0.540	**0.544**	0.543	0.543	**0.544**	0.540	**0.544**	0.543	0.542

Table 2: Correlations over all metrics depending of the hyperparameters with a vocabulary size of 16,000

Vocabulary size	16000					
Tokenization algorithm	BPE			Unigram		
Beta	4	3	2	4	3	2
n-gram length	6					
System WMT20	0.875	0.876	0.877	0.876	0.877	0.877
System WMT21	0.715	0.716	0.717	0.716	0.717	0.718
System WMT22	0.839	0.837	0.83	0.839	0.837	0.83
Segment WMT20	0.278	0.278	0.277	0.277	0.277	0.277
Segment WMT21	0.152	0.153	0.157	0.146	0.15	0.153
Segment WM22	0.398	0.399	0.4	0.398	0.4	0.402
Average	0.543	0.543	0.543	0.542	0.543	0.543

Table 3: Correlations over all metrics depending of the hyperparameters with a vocabulary size of 32,000

Vocabulary size	32000					
Tokenization algorithm	BPE			Unigram		
Beta	4	3	2	4	3	2
n-gram length	6					
System WMT20	0.875	0.876	0.877	0.876	0.876	0.877
System WMT21	0.715	0.716	0.717	0.716	0.717	0.718
System WMT22	0.837	0.835	0.83	0.837	0.835	0.83
Segment WMT20	0.277	0.277	0.277	0.277	0.277	0.277
Segment WMT21	0.157	0.158	0.157	0.158	0.158	0.153
Segment WM22	0.397	0.399	0.400	0.399	0.401	0.402
Average	0.543	**0.544**	0.543	**0.544**	**0.544**	0.543

Table 4: Tokengram_F results on the WMT20 dataset compared to chrF++

Language pair	tokengram_F system r	chrF++ system r	tokengram_F segment τ	chrF++ segment τ
en-cs	**0.865**	0.833	**0.485**	0.478
en-de	**0.961**	0.958	**0.371**	0.367
en-ru	**0.981**	0.952	**0.162**	0.156
en-ta	0.941	**0.956**	**0.590**	0.579
en-zh	0.851	**0.983**	**0.403**	0.388
en-ja	**0.949**	0.328	**0.521**	0.506
en-pl	**0.958**	0.315	**0.256**	0.255
en-iu	**0.433**	0.338	**0.340**	0.338
cs-en	**0.872**	0.844	**0.095**	0.09
de-en	0.997	**0.998**	**0.440**	0.435
pl-en	0.508	**0.970**	0.032	**0.034**
ta-en	**0.957**	0.522	0.184	**0.186**
km-en	**0.984**	0.965	**0.281**	0.275
ps-en	0.894	**0.964**	0.143	**0.145**
ja-en	**0.972**	0.763	**0.251**	0.245
ru-en	0.921	**0.977**	**0.055**	0.054
zh-en	**0.960**	0.841	**0.130**	**0.130**
iu-en	**0.765**	0.726	0.242	**0.246**

Table 5: Tokengram_F results on the WMT21 dataset compared to chrF (chrF was used in place of chrF++ as chrF++ results were not reported)

Language pair	tokengram_F system r	chrF system r	tokengram_F segment τ	chrF segment τ
en-cs	**0.978**	0.970	**0.549**	0.531
en-zh	**0.625**	0.549	**0.121**	0.092
en-ha	**0.760**	0.748	0.185	**0.186**
en-ja	**0.967**	0.966	**0.384**	0.371
en-ru	**0.756**	0.943	**0.214**	0.201
en-de	**0.842**	0.831	**0.448**	0.098
cs-en	**0.562**	**0.562**	**-0.052**	-0.053
zh-en	0.269	**0.723**	**0.395**	-0.035
ha-en	0.921	**0.924**	**0.021**	**0.021**
ja-en	0.823	**0.831**	**0.006**	0.005
ru-en	0.579	**0.593**	**-0.123**	-0.126
de-en	**0.424**	0.357	**-0.151**	-0.162
fr-de	**0.655**	0.646	0.049	**0.054**
de-fr	**0.504**	0.498	**0.111**	0.110
bn-hi	**0.949**	0.941	**0.079**	0.071
hi-bn	**0.877**	0.872	**0.335**	0.327
xh-zu	**0.999**	0.998	**0.306**	0.301
zu-xh	0.997	**0.999**	0.529	**0.530**

Table 6: Tokengram_F results on the WMT22 dataset compared to chrF++ (chrF was used in place of chrF++ as chrF++ results were not reported)

Language pair	tokengram_F system r	chrF system r	tokengram_F segment τ	chrF segment τ
en-cs	0.602	**0.689**	0.077	**0.147**
en-zh	**0.248**	0.210	-0.044	**0.051**
en-hr	0.899	**0.920**	**0.274**	0.185
en-ja	0.927	**0.931**	**0.241**	0.142
en-liv	**0.989**	0.988	**0.370**	0.101
en-ru	**0.852**	0.813	**0.659**	0.153
en-uk	0.869	**0.895**	**0.178**	0.177
en-de	0.799	**0.811**	**1.000**	0.085
liv-en	**0.985**	0.969	**0.500**	0.184
zh-en	0.787	**0.881**	**0.415**	0.071
sah-ruh	**1.000**	**1.000**	**0.856**	0.430
uk-cs	0.971	**0.979**	**0.350**	0.171
cs-uk	0.921	**0.927**	**0.311**	0.195