 doi:10.20944/preprints202201.0018.v1

Guidance to Pre-tokeniztion for SacreBLEU: Meta-Evaluation in Korean

Ahrii Kim[1] and **Jinhyun Kim**[2]
Kakao Enterprise
235 Pangyoeok-ro Bundang-gu Seongnam-si,
Gyeonggi-do, Republic of Korea
ria.i, rob.k@kakaoenterprise.com

Abstract

SacreBLEU, by incorporating a text normalizing step in the pipeline, has been well-received as an automatic evaluation metric in recent years. With agglutinative languages such as Korean, however, the metric cannot provide a conceivable result without the help of customized pre-tokenization.

In this regard, this paper endeavors to examine the influence of diversified pre-tokenization schemes –word, morpheme, character, and subword– on the aforementioned metric by performing a meta-evaluation with manually-constructed into-Korean human evaluation data.

Our empirical study demonstrates that the correlation of SacreBLEU (to human judgment) fluctuates consistently by the token type. The reliability of the metric even deteriorates due to some tokenization, and MeCab is not an exception. Guiding through the proper usage of tokenizer for each metric, we stress the significance of a character level and the insignificance of a Jamo level in MT evaluation.

Link to our code is available at http://github.com/ko_sacrebleu

1 Introduction

For almost two decades, BLEU (Papineni et al., 2002) has played a vital role in Machine Translation (MT) evaluation as an all-time favored automatic metric, whether we actually "like" it or not. Marie et al. (2021) statistically backed up such a trend, reporting that in the past decade (2010-2020), about 98.8% of research papers submitted in ACL under the title of MT regarded BLEU as a prime evaluation metric. Although we kept getting stern warnings against its use (Tan et al. 2015; Callison-Burch et al. 2006), 89% of new cutting-edge metrics that exhibited better correlation with human judgment (than other metrics including BLEU) were never deployed in actual researches to date (Marie et al., 2021). The research community opted for stabilizing it instead of exploring new ones, and the best alternative seemed to be SacreBLEU (Post, 2018).

The biggest strength of SacreBLEU was that it reduced the influence of pre-processing scheme on the score computation that could have otherwise fluctuated upon any minor changes such as a type of tokenizers, a split of compound nouns, use of unknown tokens for rare words or casing (Post, 2018). By embracing the text normalizing step in the architecture, SacreBLEU could provide more trustworthy evaluation scores.

While it was gaining weight in the literature, the trust issue was still prominent in agglutinate languages such as Korean. Languages of such typology by design required language-specific tokenization to consider semantic features, which was absent from the pipeline of SacreBLEU. The rule of thumb was to use MeCab-ko[1] beforehand as directed in Workshop on Asian Translation (Nakazawa et al., 2020), but its correlation to human judgment in MT evaluation was not officially confirmed.

In the context where Korean is not capable of taking advantage of SacreBLEU's protective layer, we shed light on the influence of a varied pre-tokenization pipeline on the given automatic metric that features three lexical-similarity-based types: BLEU, TER (Snover et al., 2006), and ChrF (Popović, 2015). At the same time, we share empirical lessons for SacreBLEU when applying it in the Korean language in MT evaluation, some of which are as follows.

At the segment level:

1. The pre-tokenization enhances the credibility of BLEU and TER in almost all cases.

2. MeCab is beneficial to BLEU and TER, but

[1] https://bitbucket.org/eunjeon/mecab-ko

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2022 doi:10.20944/preprints202201.0018.v1

Level	Denominationß	Particle	Ending	Affix	Example
Word	Eojeol	X	X	X	혜미가, 동화를, 읽었다
	Word	O	X	X	혜미, -가, 동화, -를, 읽었다
	Word	O	O	X	혜미, -가, 동화, -를, 읽, -었다
Morpheme	Morpheme	O	O	O	혜미, -가, 동화, -를, 읽, -었, -다
Syllable	Eumjeol	-	-	-	혜, -미, -가, 동, -화, -를, 읽, -었, -다
Consonant & Vowel	Jamo	-	-	-	ㅎ, -ㅖ, ㅁ, -ㅣ, ㄱ, -ㅏ, ㄷ, -ㅗ, -ㅇ, ㅎ, -ㅘ, ㄹ, -ㅡ, -ㄹ, ㅇ, -ㅣ, -ㄹㄱ, ㅇ, -ㅓ, -ㅆ, ㄷ, -ㅏ

Table 1: Definition of a Korean word (Nam et al., 2019) in comparison to other meta-levels.

when it comes to ChrF, there is a good chance that it damages its reliability.

3. The influence of the subword level is insignificant. In the worst case, the segmentation at this level could even degrade the credibility of ChrF and TER.

4. The segmentation at character level could substitute other token units, especially in BLEU and ChrF.

At the corpus level:

1. The metric score can be inflated up to twice when tokenized.

2. The superiority among systems measured by SacreBLEU does not comply with human judgment. The mean value of the segment-level scores can be a safe option at this moment.

2 Background

2.1 Word Segmentation

> A single distinct meaningful element of speech or writing, [...] and *typically shown with a space on either side* when written or printed.[2]

The general definition of word, as shown above, conjectures that it is separated by space. Such assumption, however, is arguable in Korean, whose word does not always come with a space on either side.

As displayed in Table 1, there are three approaches in defining a word: comprehensive, compromising, and analytic. They diverge on the endowment of qualification as an independent element to three components – post-positional particle, ending, and affix (Nam et al., 2019). Following the comprehensive standpoint, what is typically understood as a word in Western languages is equivalent to *Eojeol* in Korean. The compromising perspective sees that endings and affixes are not independent while the analytic approach recognizes the self-reliance of the endings. That much active discussion is possible with the morpheme boundary, which makes the Korean word decomposition complex.

In this paper, we describe two peculiar aspects of the Korean letter (or *Hangeul*). In the first place, **a character has a sub-layer**. The word *read*, for instance, is composed of four characters: r-e-a-d. The equivalent Korean word 읽 in Table 1 is also a character, but at the same time it is a combination of two consonants (ㅇ , ㄺ) and one vowel (ㅣ). We call this sub-layer *Jamo* (ㅇ - ㅣ - ㄺ).

Secondly, **Jamo are position-wise**; they are situated in a fixed position of *Choseong* (initial, ㅇ), *Jungseong* (middle, ㅣ), and *Jongseong* (final, ㄺ), respectively. Some affixes or morphemes take the form of Jongseong, making a diversified token scenario between the morpheme and Jamo level possible. The elaborate example is given in Table 2.

2.2 Token Level

We define four meta-levels of segmentation for our experiment: word, morpheme, character, and subword. As discussed in Section 2.1, particle (or *Josa*), ending, and affix are the fork of a road to the classification of the tokens.

- **Word**: We adopt the comprehensive perspective. Hence, this token level is conceptually equivalent to Eojeol, which does not consider particles, ending or affixes as an independent element.

- **Morpheme**: This token level considers particles, endings, and affixes independent. The level of segmentation varies from tokenizer to tokenizer.

[2] https://www.oxfordlearnersdictionaries.com

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2022 doi:10.20944/preprints202201.0018.v1

Word		모델	레웅	데임	은	아직	그	누구도	시도한	적	없는	방식으로	캣워크를	활보했다
Morpheme			레 웅	데 임				누구 도	시도	하 ㄴ		없 는	방식 으로	캣워크 를	활보 했다
				데 이 ㅁ				누 구도						캣 워크 를	활보 했 다
														캐 엇 워크 를	활 보 했 다
															활보 하 었 다
															활보 하 았 다
Character		모 델	레 웅	데 임	은	아 직	그	누 구	도	시 도	한	적	없 는	방 식 으 로	캣 워 크 를	활 보 했 다
Subword	choseong	ㅁ ㄷ	ㄹ ㅇ	ㄷ ㅇ	ㅇ	ㅇ ㅈ	ㄱ	ㄴ ㄱ	ㄷ	ㅅ ㄷ	ㅎ	ㅈ	ㅇ ㄴ	ㅂ ㅅ	ㅇ ㄹ	ㅋ ㅇ ㅋ	ㄹ	ㅎ	ㅂ	ㅎ	ㄷ
	jungseong	ㅗ ㅔ	ㅔ ㅛ	ㅔ ㅣ	ㅡ	ㅏ ㅣ	ㅡ	ㅜ ㅜ	ㅗ	ㅣ ㅗ	ㅏ	ㅓ	ㅓ ㅡ	ㅏ ㅣ	ㅡ ㅗ	ㅐ ㅟ	ㅡ	ㅡ	ㅏ	ㅛ	ㅐ	ㅏ
	jongseong	ㄹ	ㅐ	ㅁ	ㄴ	ㄱ					ㄴ	ㄱ	ㅂㅅ ㄴ	ㅇ ㄱ	ㅅ			ㄹ	ㄹ		ㅆ	

Table 2: All possible tokenization schemes with the tokenizers applied in this study.

- **Character**: This token level is morphologically similar to but semantically different from Eumjeol or syllable. The segmentation strategy of this level does not require a tokenizer.
- **Subword**: This token level is the smallest token unit in Korean. A character is decomposed into Jamo, and each consonant and vowel is considered a token.[3]

2.3 Tokenizer

The meta-level tokens come into shape with the help of tokenizers in some cases. We implement seven tokenizers on the morpheme level – Kkma, Hannanum, Komoran, Okt and MeCab from KoNLPy (Park and Cho, 2014), Kiwi (Korean Intelligent Word Identifier)[4] and machine-learning-based Khaiii (Kakao Hangul Analyzer III)[5]– and two tokenizers –SentencePiece (SPM) (Kudo and Richardson, 2018) and Jamo[6]– on the subword level.

2.3.1 Tag Set

Most Korean morphological analyzers have their roots in the 21st Century Sejong Project launched in 1998 intending to build a national framework for large-scale Korean corpora (21st Sejong Project, 1999). The tokenizers feature a different number of tag sets derived from the Sejong tag sets, as described in Appendix A.

The prototype tag set is preserved in Komoran and similarly in MeCab and Khaiii. The tokenizer with the most diminutive tag set is Kkma with 56 tags. It can provide a detailed analysis of endings (Eomi). The reverse case is Okt, a tokenizer targeted for Twitter, with 19 tags. Woo and Jung (2019) report its outstanding performance with typos, emojis, and punctuations. Hannanum also features a small-sized tag set (22 tags). Its exceptionally compressed division of particle tags is noticeable, among others. As expected, the central divergence of the tag sets is observed in particles, endings, and affixes.

2.3.2 Tokenization Scenario

As displayed in Table 2, the exemplary sentence is extracted from our data set to show a various tokenization scenario. It accumulates all possible scenarios of the meta-levels studied in our experiment.

The instance shows that the segmentation is most diversified in verbs (활보했다) with nine possibilities. It is also intriguing that some tokenizers consider Jongseong as an independent token (하, -ㄴ). Such case is Hannanum, Kkma, Komoran, Khaiii and Kiwi.

3 Experiment

3.1 Experiment Setup

As Korean evaluation data is rarely available, we organized a human evaluation of four commercial NMT systems for the English-to-Korean translation direction with Direct Assessment (DA), the conventional human metric at Conference on Machine Translation (Barrault et al., 2020). Consequently, a machine evaluation with BLEU, TER, and ChrF of SacreBLEU is performed. With the resources at hand, the correlation between the two evaluation results is computed at the segment and corpus level.

3.1.1 Dataset

- **Source Test Set**: The original English texts are borrowed from WMT 2020 English III-type test set, composed of 2,048 sentences (61 documents) with segment split. The segment-split format allows them the freedom of translating beyond a sentence level (Barrault et al., 2020).

[3] Although SPM is considered as a subword tokenizer in general, such a concept does not coincide with Korean linguistics. Hence, while its output takes the shape of morpheme, we categorize SentencePiece in the subword level.

[4] https://github.com/bab2min/Kiwi

[5] https://github.com/kakao/khaiii

[6] https://github.com/JDongian/python-jamo

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2022 doi:10.20944/preprints202201.0018.v1

- **Reference Translation**: The Korean reference translation is created by a group of professional translators. The translators are advised not to post-edit MT. To guarantee the highest translation quality, we double-check the final revision. The revision is implemented only if the sentence is semantically erroneous.

- **System Translation**: We employ four online MT models including our own –*Kakao i*[7]–. They are denominated as Sys_A, Sys_B, Sys_P and Sys_Q in order to keep their anonymity for legal reason. The system translations are obtained on July 21, 2021.

In terms of normalizing data, errors in the source test sets and their subsequent impact on the system translations (Kim et al., 2021) remain undealt with. Only some minor technical issues, i.e. different single quotations (‘ and ’), are normalized.

3.1.2 Human Evaluation

DA is a metric where an evaluator scores each sentence on a continuous scale of 0-100 in the category of Adequacy and Fluency. We hire 25 professional translators and assign each person a HIT of more or less 300 translated sentences with the context of the documents maintained. They are advised to consider the context when making a judgment. They are allowed to reverse their previous decisions within a document.

They are either holders of a master’s degree in interpretation and translation in the English-Korean language pair or freelance translators with a minimum of two years of experience. In light of the fact that all participants are new to MT evaluation, we provide a detailed guideline for the experiment.

One judgment per system translation is gathered, amounting to 16,384 (8,192 of Adequacy and Fluency) evaluation data. The judgment on Fluency is only utilized as supplementary information.

3.1.3 Quality Control

Out of the 8,192 Adequacy judgments, the first ten judgments of each evaluator are considered an adjusting step, and so, they are removed. The scores are then normalized with judge-wise Z-scores. With them, Inter-Quartile Range (IQR) is computed as in Equation 1, where Q_1 and Q_3 signify the first and third quartile values. The outlier x belongs to the range that meets the condition.

[7] https://www.translate.kakao.com

Having removed 5.67% of the data, we base our observation on 7,727 judgments.

$$x < Q_1 - 1.5 \cdot (Q_3 - Q_1)$$

or

$$x > Q_3 + 1.5 \cdot (Q_3 - Q_1) \quad (1)$$

3.1.4 Computation

The hypothesis and reference translations are tokenized by the aforementioned 11 token units without applying any additional normalization. Consequently, the scores of the automatic metrics are computed, and their Pearson’s correlation coefficient r to the the human Adequacy judgment are measured as such:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H}) \cdot (M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \quad (2)$$

H and M refer to the machine and human DA score, respectively, and $\overline{H}$ and $\overline{M}$, their mean values. The Pearson’s r measures the linear relationship between them.

During the process, some of the issues have concerned us:

- **Do we adjust n-gram parameters?**
  By default, the BLEU score is a geometric mean of four-grams. As the token unit is divergent, on the one hand, we attempt to avoid a circumstance where any tokenizer benefits from the n-gram parameter. On the other hand, the default n-grams of ChrF at the word level are zero. To make the consequence of the token unit visible and compatible, we acquire in advance the best-correlated n-grams of each token unit to the human judgment. The n-grams of the token unit, therefore, is specified together with the outcome.

- **TER scores over 1.0?**
  Theoretically speaking, $TER = 1.0$ represents a perfect match between reference and hypothesis sentences. However, we detect that when a hypothesis is too short for its reference, the (sentence-level) TER score exceeds 1.0. Such cases are avoided by being normalized to 1.0 afterward.

- **Is there a tie rank?**
  We run a Wilcoxon ranksum test to identify

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2022 doi:10.20944/preprints202201.0018.v1

	BLEU								TER								ChrF							
n	*500*	*1000*	*1500*	*2000*	*2500*	*3000*	*4000*	*5000*	*500*	*1000*	*1500*	*2000*	*2500*	*3000*	*4000*	*5000*	*500*	*1000*	*1500*	*2000*	*2500*	*3000*	*4000*	*5000*
Komoran	0.623	1.776	0.399	1.523	1.745	0.868	**0.011**	0.914	**3.95**	2.712	2.139	2.439	3.385	3.777	3.575	3.47	1.138*	0.286	**0.005**	**0.071**	**0.005**	**0.068**	2.045	1.214
Khaiii	**0.392**	1.985	0.426	**0.386**	1.007	0.978	0.058	0.390	3.012	3.368	1.993	2.759	3.725	3.668	3.689	3.154	0.864	0.259	0.068	0.177	0.056	0.084	**0.754**	1.262
Kiwi	1.763	1.868	**0.155**	1.582	0.565	**0.267**	0.113	0.450	3.571	3.311	2.798	2.3	4.482	4.039	3.165	3.981	1.119	**0.066**	0.322	0.149	0.561	0.16	0.902	**0.484**
Kkma	1.061	1.368	0.839	1.499	2.738	1.8	0.721	0.720	3.073	2.692	3.671	2.888	3.854	**4.48**	**4.666**	4.517	0.705	0.815	0.495	1.23	0.493	0.499	1.082	3.443
Character	0.421	**0.345**	0.247	2.19	**0.275**	0.448	0.957	**0.166**	2.493	2.632	2.953	3.281	1.947	3.08	2.792	3.057	0.301	0.321	0.913	0.584	0.834	1.776	1.546	0.656
MeCab	0.736	0.488	1.421	0.752	0.525	0.529	1.206	2.309	3.275	2.303	**3.753**	2.593	**4.666**	3.305	4.569	2.535	0.766	0.598	0.603	0.842	0.65	2.648	1.424	1.358
SPM	1.164	0.957	1.572	2.905	1.32	2.052	1.624	1.753	3.148	**3.5**	3.447	2.686	3.44	4.203	4.153	**4.706**	0.539	0.773	0.832	1.035	0.612	3.496	1.659	1.753
Jamo	1.053	0.411	0.317	0.635	0.728	2.061	2.294	1.186	1.149	2.271	2.022	0.51*	0.851	1.059	0.357*	1.304	**0.27**	0.867	2.94	3.471*	3.501*	1.222	1.351	4.714*
Okt	1.824	2.438	3.432	1.227	3.678	4.104	3.488	3.992	2.235	1.324	1.498	**3.452**	2.865	1.818	2.699	3.093	0.786	0.875	0.58	2.394	1.959	3.608	3.289	1.281
Hannanum	2.221	4.063	3.736	4.096	5.474*	4.507	4.691	4.976	0.975	0.842	0.289	0.954	1.625	0.699	0.699	0.067	0.771	1.665*	1.934*	3.16	1.319	4.212*	5.342*	3.684
None	4.24*	4.43*	5.124*	5.412*	5.078	5.93*	5.65*	5.984*	0.164*	0.062*	0.173*	0.629	0.492*	0.213*	0.663	0.041*	0.311	1.016	1.787	1.483	2.927	2.344	3.659	2.585

Table 3: Average segment-wise ranks of the token unit for SacreBLEU resampled by n samples. The highest ranking is in bold while the lowest is marked with asterisk (*).

BLEU			TER		ChrF		
	r	*n*		*r*		*r*	*char_n*	*word_n*
Jamo	0.3079	5	Kkma	0.2949	Komoran	0.3304	3	1
Kiwi	0.3073	2	Kiwi	0.2933	Khaiii	0.3301	3	1
Komoran	0.3062	2	Mecab	0.2912	Kiwi	0.3295	3	1
Character	0.3054	2	Komoran	0.2898	Kkma	0.3263	3	0
Khaiii	0.3051	2	Spm	0.2894	Mecab	0.3230	3	1
Kkma	0.3019	2	Khaiii	0.2884	Character	0.3230	3	2
Mecab	0.3017	2	Character	0.2860	Spm	0.3222	3	1
Spm	0.3008	2	Okt	0.2776	Okt	0.3208	3	1
Okt	0.2843	2	Jamo	0.2765	Hannanum	0.3200	3	0
Hannanum	0.2663	2	Hannanum	0.2503	Word	0.3199	3	0
Word	0.2498	1	Word	0.2383	Jamo	0.3160	5	0

Figure 1: Result of Pearson correlation r with ranking clusters by Wilcoxon ranksum test at the segment level. The token units in the identical background color are considered tie. The adjustment of n-grams is specified in the n columns.

the statistical significance between the ranks. First, we set a cluster boundary with the p-value based on the assumption that two token types whose p-value is less than 0.05 ($p < 0.05$) are statistically different. After counting such cases for all combinations, those with the same number of counts are considered a tie.

- **Is the sample size enough?**
  To yield a credible result, we apply Bootstrap Resampling suggested by Koehn (2004). Out of 7,727 sentences, different blocks of sub-samples are extracted in a binary round (N out of 7,727 and M out of N) on a random basis. Iterating I times, M out of M samples are randomly selected to print the final result. Koehn (2004) reported that they reached a 95% confidence level with 394 samples (N) and near 100% with 3,000 samples when assessing MT systems with BLEU. While we follow the precedence by setting up the parameters similarly as $M = 6000, N = 3000, I = 1000$, we provide additional results with variations in N and I in Table 3.

3.2 Experiment Result

BLEU, TER, and ChrF scores and their correlation results are analyzed at the segment/corpus level. While finding the best pre-tokenization scheme is intriguing, our primary focus is to examine that SacreBLEU results are susceptible to the pre-tokenization scenarios.

3.2.1 Segment Level

Table 1 shows the Pearson correlation of the token units clustered by Wilcoxon ranksum test. The result of Bootstrap Resampling is given in Table 3.

BLEU. Without resampling, the result shows that the highest correlation is observed in a broad spectrum of token units: Jamo, Kiwi, Komoran and Character. The lowest correlation is when pre-tokenization is absent, which is consistently witnessed in the resampled outcome, except for one instance of Hannanum ($n = 2500$). It is safe to conclude, therefore, that any tokenizer can enhance the credibility of this metric, but Hannanum is not an option. Besides, we also report the moderate impact of MeCab.

TER. Before resampling, the morpheme level (Kiwi and Kkma) goes best with this metric, followed by MeCab. Considering the lowest correlation of the word level, it is a reasonable guess that pre-tokenization is essential in this metric. Such trend is still valid when the data is resampled, aside from the two negative cases in Jamo. Among the tokenizers, the positive influence of Kkma is noteworthy as the sample size grows. Moreover, MeCab seems to be a good fit for this metric in Korean. The least recommendable token unit is Jamo, not only because its correlation is low on average but also the increased token size is markedly disadvantageous to the computational cost of this metric.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2022 doi:10.20944/preprints202201.0018.v1

ChrF. With the 7,727 data set, the optimal token unit for this metric is morpheme (Komoran and Khaiii). Komoran is remarkably well-fitted to this metric even when the data is resampled. However, the most differential aspect of this metric is that tokenization can be harmful. For instance, in a pilot study we find that the best-correlated word n-grams for Kkma and Hannanum are zero, meaning that it is best when they are not taken into consideration. All but the character level have a history of deteriorating the correlation of the metric. In that respect, Hannanum and Jamo produce the most unstable result. It draws our attention that when the sample size is small (n=500), most of the pre-tokenization worsens its correlation, and MeCab is not an exception.

All in all, we conclude that at the segment level, the correlation of the three metrics fluctuates by the pre-tokenization scheme, but there is no direct relationship between the token shape and the correlation of the metrics. Nevertheless, if we are to draw a meaningful conclusion from the result, any pre-tokenization is better than the word level in BLEU and TER, but in ChrF the token unit should be carefully selected. The combination of MeCab and BLEU/TER stands out, while the popular tokenizer can detriment the credibility of ChrF.

Unlike our expectations, the subword level does not serve as a dependable token unit for the Korean MT evaluation. Instead, there is a good chance that it harms the correlation of ChrF and TER. Furthermore, the expanded vocab size increases the computational cost exponentially. As a substitute, we highlight the effectiveness of the character-level segmentation, which guarantees a fast deployment without any tokenizer and, at the same time, is proven to be as reliable as or often better than MeCab in all metrics.

3.2.2 Corpus Level

Table 4 - 6 show human DA scores and SacreBLEU scores of four online systems in relation to the tokenization scheme. The noticeable finding is that in all three cases the spectrum of the score is substantial. The highest-ranked system in the human z-score (Sys_A) obtained a 28.09 BLEU score without segmentation, but 48.71 when with the character-level tokenization. Likewise, the most overestimated version of TER is before tokenization (82.33 - 89.69), as expected, while the most underestimated version of score is on the Jamo level (51.96 - 54.69). In ChrF, the range mentioned above is from 42.74 - 45.72 on the None level to 51.19 - 53.80 on the Jamo level. We, thus, confirm the possibility that the absolute score of SacreBLEU can be doubled just by selecting a different tokenizer.

While so, the more severe problem is that the ranks by score, irrelevant from the pre-tokenization typology, do not comply with human perception. The human average scores place the systems in the order of $[Sys_A = 1, Sys_B = 2, Sys_P = 3, Sys_Q = 4]$, but almost all automatic scores position them as $[Sys_A = 2, Sys_B = 1, Sys_P = 3, Sys_Q = 4]$. Moreover, in many cases the BLEU scores are prone to rank them as $[Sys_A = 2, Sys_B = 1, Sys_P = 4, Sys_Q = 3]$ except when tokenized by MeCab, Kiwi and Khaiii. Such an unreliable performance of this metric can be derived from either the small number of systems or the existence of outlier systems in the comparison (Mathur et al., 2020). We raise the issue of a questionable SacreBLEU score at the corpus level, leaving its verification to our future work.

Figure 2 shows the Pearson correlation of SacreBLEU when with the pre-tokenization. The most faithful score is achieved with Khaiii and Kiwi in all three metrics. The correlation of MeCab is also striking. Despite their discernible performance on this level, however, we reiterate that none of the options represent the system rankings as humans perceive. In this respect, we propose the mean value of segment-wise SacreBLEU score as a substitute.

4 Related Works

Recently, a word segmentation got the limelight with the outstanding achievement of subword-level pipelines such as SPM or Byte-Pair Encoding (BPE) (Sennrich et al., 2015) in many NLP tasks (Zhang et al. 2015;). In MT, in specific, the segmentation is tightly entangled with the translation quality mainly due to the handling of unseen vocabulary. In that respect, many studies observed that identifying the best tokenization is language-dependant.

Huck et al. (2017) discovered that their model displayed the highest performance when BPE was coupled with a suffix split in German. In a similar manner, Lee et al. (2016) suggested that their fully character-level NMT model outperformed BPE models, especially in the Finnish-English pair. Domingo et al. (2018) demonstrated that when five languages were under study, no single best tok-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2022 doi:10.20944/preprints202201.0018.v1

	Ave.↑	Ave. z	None	Okt	Mecab	Komoran	Kkma	Kiwi	Khaiii	Hannanum	Character	Spm	Jamo
Sys_A	**68.783**	**0.203**	28.099	33.398	38.341	40.275	40.986	41.022	40.005	36.939	48.712	41.015	48.467
Sys_B	67.160	0.112	**28.932**	**34.351**	**39.185**	**41.007**	**41.920**	**41.997**	**40.881**	**37.793**	**49.553**	**41.948**	**49.188**
Sys_P	64.688	0.027	23.941	30.415	35.605	36.621	37.236	38.458	37.034	32.902	45.924	37.213	45.098
Sys_Q	57.734	-0.220	25.941	31.382	35.602	37.304	38.063	38.138	36.939	34.058	47.096	38.155	46.602

Table 4: The variation of BLEU score of the four MT systems by token type along with human DA scores and their z-scores. The best scores are in bold.

	Ave. ↑	Ave. z	None	Okt	Mecab	Komoran	Kkma	Kiwi	Khaiii	Hannanum	Character	Spm	Jamo
Sys_A	**68.783**	**0.203**	82.811	68.223	64.142	63.041	62.253	62.352	63.412	67.833	57.718	62.391	52.932
Sys_B	67.160	0.112	**82.334**	**67.332**	**63.519**	**62.585**	**61.545**	**61.649**	**62.867**	**67.249**	**56.364**	**61.083**	**51.962**
Sys_P	64.688	0.027	89.652	69.882	64.898	64.859	63.479	62.983	64.346	71.199	62.163	65.914	54.063
Sys_Q	57.734	-0.220	86.699	70.356	66.611	65.641	64.751	64.758	66.126	71.199	59.771	64.767	54.697

Table 5: The variation of TER score of the four MT systems by token type along with human DA scores and their z-scores. The best scores are in bold.

	Ave. ↑	Ave. z	None	Okt	Mecab	Komoran	Kkma	Kiwi	Khaiii	Hannanum	Character	Spm	Jamo
Sys_A	**68.783**	**0.203**	44.897	46.508	47.544	48.904	46.326	49.299	48.763	46.019	47.887	47.932	53.140
Sys_B	67.160	0.112	**45.725**	**47.345**	**48.370**	**49.635**	**47.131**	**50.096**	**49.560**	**46.826**	**48.707**	**48.807**	**53.807**
Sys_P	64.688	0.027	42.742	44.171	45.342	46.182	43.796	47.017	46.354	43.401	45.699	45.357	51.198
Sys_Q	57.734	-0.220	43.505	45.134	46.031	47.166	44.639	47.557	47.011	44.378	46.533	44.378	51.775

Table 6: The variation of ChrF score of the four MT systems by token type along with human DA scores and their z-scores. The best scores are in bold.

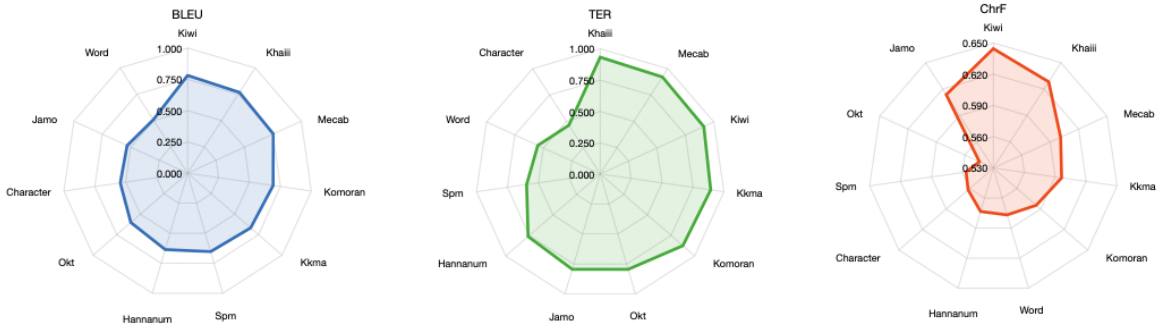

Figure 2: Result of Pearson correlation r at the corpus level.

enizer could lead to a more refined translation quality for all languages. They made a remark that such phenomenon was striking in morphologically rich languages such as Japanese.

In relation to Korean in this regard, the concept of detailed segmentation intrigued many researchers in NLP in general (Park et al. 2018; Kim et al. 2020; Yongseok and Lee 2020; Park et al. 2020), for a Korean word had a large volume of affixes and morphemes. In MT in specific, various combinations of token units were suggested. Among them, Park et al. (2019) stated that when their NMT model was trained with a dataset of subtitles tokenized with SPM Unigram after removing postpositional particles, it obtained a higher BLEU score than when with simple BPE.

While they mentioned that smaller token unit was not always an answer in the case of Korean, recent studies paid attention to a smallest unit, the Jamo. Moon and Okazaki (2020) introduced Jamo-Pair Encoding, combining Jamo with BPE. Eo et al. (2021) utilized Jamo from a functional viewpoint by regarding Choseong and Jungseong as one token and leaving Jongseong as another. They demonstrated that the model with such segmentation outperformed the model of Park et al. (2019).

We differ from the aforementioned studies in that we explore the impact of tokenization on the MT evaluation. As the gold standard in this field is human, we prioritize the examination of the pattern of its impact on SacreBLEU, instead of discovering a superior token unit. Assuming that the matching system of N-grams in BLEU and ChrF or edit distance in TER is exceptionally vulnerable to the lexical shape of tokens, we observe how such changes are in line with human perception. We believe it is a pitfall to the MT evaluation of synthetic languages[8] such as Korean.

5 Conclusion

On the condition that pre-tokenization is obligatory for the agglutinative language such as Korean when computing automatic evaluation metrics, we endeavor to demonstrate the influence of dynamic token units on the credibility of SacreBLEU, including BLEU, TER, and ChrF at both the segment and corpus level in the into-Korean translation direction.

For the meta-evaluation, we perform a human evaluation with 25 professional translators on system translations produced by four NMT models. When the Pearson correlation is measured, the result shows that the impact of token type differs from metric and its computation level. At the segment level, we report that any pre-tokenization enhances the correlation of BLEU and TER while it should be carefully selected in ChrF. At the corpus level, ranking by the SacreBLEU scores turns out to be inaccurate regardless of the pre-tokenization scheme.

Contrary to our expectation, the diminutive segmentation by the subword level shows signs of ineffectiveness. Instead, we put an emphasis on the role of the character level.

[8] It refers to a language group whose word is composed of an exceptional number of morphemes.

Acknowledgements

Special thanks to our team members for their thoughtful comments and healthy discussions.

References

The 21st Sejong Project. 1999. Construction of korean basic data (academic service report).

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

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, Trento, Italy. Association for Computational Linguistics.

Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2018. How much does tokenization affect neural machine translation? *CoRR*, abs/1812.08621.

Sugyeong Eo, Chanjun Park, Hyeonseok Moon, and Heuiseok Lim. 2021. Research on subword tokenization of korean neural machine translation and proposal for tokenization method to separate jongsung from syllables. *Journal of the Korea Convergence Society*, 12(3):1–7.

Matthias Huck, Simon Riess, and Alexander M. Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *WMT*.

Ahrii Kim, Yunju Bak, Jimin Sun, Sungwon Lyu, and Changmin Lee. 2021. The suboptimal wmt test sets and their impact on human parity. *Preprints*.

Hwichan Kim, Tosho Hirasawa, and Mamoru Komachi. 2020. Zero-shot North Korean to English neural machine translation by character tokenization and phoneme decomposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 72–78, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395. Association for Computational Linguistics.

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.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. *CoRR*, abs/2106.15195.

Nitika Mathur, Tim Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: reevaluating the evaluation of automatic machine translation evaluation metrics. *CoRR*, abs/2006.06264.

Sangwhan Moon and Naoaki Okazaki. 2020. Jamo pair encoding: Subcharacter representation-based extreme Korean vocabulary compression for efficient subword tokenization. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3490–3497, Marseille, France. European Language Resources Association.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Suzhou, China. Association for Computational Linguistics.

Gisim Nam, Yeonggeun Ko, Hyunkyung Yu, and Hyeongyong Choi. 2019. *Korean standard grammar (표준 국어문법론)*. Hankook Munhwasa, Korea.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Chanjun Park, Gyeongmin Kim, and Heuiseok Lim. 2019. Parallel corpus filtering and korean-optimized subword tokenization for machine translation. *Annual Conference on Human and Language Technology*, pages 221–224.

Eunjeong L. Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In *Proceedings of the 26th Annual Conference on Human Cognitive Language Technology*, Chuncheon, Korea.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various korean NLP tasks. *CoRR*, abs/2010.02534.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. Subword-level word vector representations for Korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2429–2438, Melbourne, Australia. 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Liling Tan, Jon Dehdari, and Josef van Genabith. 2015. An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan. Workshop on Asian Translation.

Kyungjin Woo and Suhyeon Jung. 2019. Comparison of korean morphology analyzers according to the types of sentence. *Proceedings of the Korean Information Science Society Conference*, pages 1388–1390.

Choi Yongseok and Kongjoo Lee. 2020. Performance analysis of korean morphological analyzer based on transformer and bert. *Journal of KIISE*, 47(8):730–741.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2022 doi:10.20944/preprints202201.0018.v1

A Appendix

Category			Sejong	Okt	Komoran	MeCab-ko	Kkma	Hannanum	Khaiii	Kiwi
# of tags			*42*	*19*	*42*	*43*	*56*	*22*	*46*	*47*
Substantive	**noun**	**general**	NNG	Noun	NNG	NNG	NNG	NC	NNG	NNG
		proper	NNP		NNP	NNP	NNP	NQ	NNP	NNP
		dependent	NNB		NNB	NNB	NNB	NB	NNB	NNB
		unit				NNBC	NNM			
	pronoun		NP		NP	NP	NP	NP	NP	NP
	numeral		NR		NR	NR	NR	NN	NR	NR
Predicate	**verb**		VV	Verb	VV	VV	VV	PV	VV	VV
	adjective		VA	Adjective	VA	VA	VA	PA	VA	VA
	auxiliary		VX	-	VX	VX	VXV	PX	VX	VX
							VXA			
	copula	**positive**	VCP	-	VCP	VCP	VCP	-	VCP	VCP
		negative	VCN	-	VCN	VCN	VCN	-	VCN	VCN
Modifier	**article**	**determiner**	MM	Determiner	MM	MM	MDT	MM	MM	MM
		numeral					MDN			
	adverb	**general**	MAG	Adverb	MAG	MAG	MAG	MA	MAG	MAG
		connective	MAJ	Conjunction	MAJ	MAJ	MAC		MAJ	MAJ
Interjection	**interjection**		IC	Exclamation	IC	IC	IC	II	IC	IC
Post-positional Particle	**case-marking**	**subjective**	JKS	Josa	JKS	JKS	JKS	JC	JKS	JKS
		complement	JKC		JKC	JKC	JKC		JKC	JKC
		adnominal	JKG		JKG	JKG	JKG		JKG	JKG
		objective	JKO		JKO	JKO	JKO		JKO	JKO
		adverbial	JKB		JKB	JKB	JKM		JKM	JKM
		vocative	JKV		JKV	JKV	JKI		JKI	JKI
		quotation	JKQ		JKQ	JKQ	JKQ		JKQ	JKQ
	auxiliary		JX		JX	JX	JX	JX	JX	JX
	conjunctive		JC		JC	JC	JC		JC	JC
	predicative		-		-	-	-	JP	-	-
Dependent	**pre-final ending**	**honorific**	EP	PreEomi	EP	EP	EPH	EP	EP	EP
		tense					EPT			
		politeness					EPP			
	sentence-closing ending	**declarative**	EF	Eomi	EF	EF	EFN	EF	EF	EF
		interrogative					EFQ			
		imperative					EFO			
		requesting					EFA			
		interjective					EFI			
		honorific					EFR			
	connectiveending	**equal**	EC	EC		EC	ECE	EC	EC	EC
		auxiliary					ECS			
		dependent					ECD			
	transformative ending	**nominal**	ETN		ETN	ETN	ETN	ET	ETN	ETN
		adnominal	ETM		ETM	ETM	ETD		ETD	ETD
	prefix	**substantive**	XPN	-	XPN	XPN	XPN	XP	XPN	XPN
		predicative	-	-	-	-	XPV		-	-
	suffix	**derived noun**	XSN	Suffix	XSN	XSN	XSN	XS	XSN	XSN
		derived verb	XSV		XSV	XSV	XSV		XSV	XSV
		derived adverb	XSA		XSA	XSA	XSA		XSA	XSA
	root	**root**	XR	-	XR	XR	XR	-	XR	XR
Punctuation	**. ? !**		SF	Punctuation	SF	SF	SF	S	SF	SF
	…		SE		SE	SE	SE		SE	SE
	“ ” ‘ ’ ( )		SS		SS	SSO여는 괄호 (, [SSC닫는 괄호),]	SS		SS	SS
	~- _		SP		SP	SC	SP		SP	SP
	others		SO		SO	SY	SO		SO	SO
	Chinese character		SW		SW		SW		SW	SW
Etc.	**foreign word**		SH	Foreign	SH	SH	OH	F	SH	SH
	number		SL	Alpha	SL	SL	OL		SL	SL
	unknown noun		SN	Number	SN	SN	ON	-	SN	SN
	unknown verb		NF	Unknown	NF	-	UN	-	ZN	UN
	unknown		NV		NV	-		-	ZV	
	unknown		NA		NA	-		-	ZZ	
	consonant/vowel		-	KoreanParticle	-	-	-	-	SWK	-
	hashtag		-	Hashtag	-	-	-	-	-	W_HASHTAG
	user name		-	ScreenName	-	-	-	-	-	W_MENTION
	email		-	Email	-	-	-	-	-	W_EMAIL
	url		-	URL	-	-	-	-	-	W_URL

Table 7: Summary of the tag sets of Sejong and seven tokenizers. The tag sets of KoNLPy tokenizers are referred to KoNLPy.[9]