Vacillating Human Correlation of SacreBLEU in Unprotected Languages

Ahrii Kim[1] and **Jinhyeon Kim**[2]
Kakao Enterprise
Gyeonggi-do, Republic of Korea
ria.i, rob.k@kakaoenterprise.com

Abstract

SacreBLEU, by incorporating a text normalizing step in the pipeline, has become a rising automatic evaluation metric in recent MT studies. With agglutinative languages such as Korean, however, the lexical-level metric cannot provide a conceivable result without a customized pre-tokenization. This paper endeavors to examine the influence of diversified tokenization schemes –word, morpheme, subword, character, and consonants & vowels (CV)– on the metric after its protective layer is peeled off.

By performing meta-evaluation with manually-constructed into-Korean resources, our empirical study demonstrates that the human correlation of the surface-based metric and other homogeneous ones (as an extension) vacillates greatly by the token type. Moreover, the human correlation of the metric often deteriorates due to some tokenization, with CV one of its culprits. Guiding through the proper usage of tokenizers for the given metric, we discover i) the feasibility of the character tokens and ii) the deficit of CV in the Korean MT evaluation.[1]

1 Introduction

For almost two decades, BLEU (Papineni et al., 2002) has been a key driver of the development of Machine Translation (MT) and MT Evaluation despite its blind spots. Marie et al. (2021) statistically support such trend, reporting that in the past decade, about 98.8% of research papers of ACL under the title of "MT" regarded it as their prime evaluation metric. However much stern warnings we have got against its use (Tan et al. 2015; Callison-Burch et al. 2006), the fact that one of the most popular metrics besides it since 2018 is its stabilized implementation SacreBLEU (Post, 2018) (Marie et al., 2021) lets us ask ourselves if this rising metric is safe for all .

The biggest strength of SacreBLEU is that it reduces the influence of pre-processing scheme on the score computation that could have fluctuated otherwise 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, this automatic metric can provide more trustworthy evaluation scores.

While it is gaining weight in the literature, its trust issue remains prominent in terms of agglutinate languages such as Korean. Languages of such typology by design require language-dependant tokenization to convey the morphological implications hardly expressible by whitespaces. Presumably for that reason, SacreBLEU specifies a customized tokenizer for some languages such as Japanese. When assessing Korean texts, therefore, the Workshop on Asian Translation directs that the texts be tokenized by MeCab-ko[2] before running any automatic metrics (Nakazawa et al., 2017), but their correlation to human judgment has not been 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 varied pre-tokenization types on the human correlation of the given metric that features three surface-based metrics: BLEU, TER (Snover et al., 2006), and ChrF (Popović, 2015). With that information, we share empirical lessons for SacreBLEU when applying it in the Korean language in MT evaluation, some of which are summarized as such:

On the segment level:

1. Almost any pre-tokenization enhances the human correlation of BLEU or TER, but not ChrF.

[1]Link to our code is available at https://github.com/kakaoenterprise/korean-sacrebleu

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

Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), pages 1 - 15

2. The character-level decomposition guarantees a feasible human correlation and fast deployment.

3. The influence of the CV level is detrimental. It degrades the human correlation of ChrF.

On the corpus level:

1. The morpheme level, in general, achieves a higher correlation, among which Kiwi and Khaiii are noteworthy.

2. Contrary to the segment level, the character-level tokens harm the human correlation of the metrics.

3. The raw score of the metrics can be inflated up to twice when different tokenizers are involved. Thus, comparing scores by simply copying from other studies is invalid.

Cost-Efficiency:

1. TER can be slower than the other two metrics by up to seven times. In the worst scenario, the metric was combined with CV and it took 360 times more than BLEU for computation.

2. No matter how beneficial the CV can be, cost-ineffectiveness is its blind spot.

2 Related Works

Recently, the research topic of word segmentation has got the limelight in many NLP tasks (Zhang et al. 2015; Park et al. 2018; Kim et al. 2020; Yongseok and Lee 2020; Park et al. 2020), especially with the outstanding achievement of subword-level pipelines such as SentencePiece (SPM) (Kudo and Richardson, 2018) or Byte-Pair Encoding (BPE) (Sennrich et al., 2016). In MT in specific, interest is growing in handling unseen vocabularies (OOVs) through an optimal token type, whereas the influence of tokenization in MT evaluation is rarely explored. Thus, this section is devoted to the studies identifying the relation between tokenization and translation quality, but with a particular focus on its language dependency.

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. (2017) suggested that their fully character-level NMT model outperformed BPE models, especially in the Finnish-English pair. Domingo et al. (2018) demonstrated that no single best tokenizer could lead to a more refined translation quality for all languages when five languages were under study. Furthermore, they remarked that such phenomenon was striking in morphologically rich languages such as Japanese.

Similarly, concerning Korean, Park et al. (2019) found that SPM Unigram allowed their NMT model to attain a higher BLEU score than simple BPE. While they mentioned that a smaller token unit was not always an answer in the case of Korean, recent studies paid more and more attention to the sub-subword token unit called *Jamo*, referring to consonants and vowels.[3] Moon and Okazaki (2020) introduced Jamo-Pair Encoding, combining Jamo with BPE. Eo et al. (2021) suggested a new division of Jamo by sub-grouping it position-wise. They demonstrated that the model with such a word decomposition outperformed Park et al. (2019).

We differ from the studies above in exploring the impact of tokenization on the MT evaluation. Our keen interest is i) to observe how vulnerable this metric is to the agglutinative languages and ii) to find a way to ensure that the metric is in line with human perception in this regard.

3 Background

This section describes the linguistic characteristics of Korean as an agglutinative language. Unlike most European languages, it features deeper layers and diversified decomposition.

3.1 Token Level

We define five meta-levels of segmentation for our experiment: word, morpheme, subword, character, and CV. The fork of a road to the classification is in the dependence of three elements: particles (or *Josa*), endings, and affixes.

- **Word**: A whitespace is a separator between this level of tokens. A token does not consider any of the three components independent.

- **Morpheme**: This token level considers particles, endings, and affixes as dependent elements. The degree of segmentation, however, varies from tokenizer to tokenizer by their tag set or algorithm.

[3]For those who are not familiar with Korean, the in-depth information about its word decomposition is provided in Appendix A.

Source Word		model 모델		Leon 레옹		Dame 데임은			before 아직		no one has strutted 그 누구도 시도한 적 없는									like 방식으로			the catwalk 캣워크를				strutted down 활보했다			
Morpheme				레	옹	데	임	은				누구	도	시도	한			없	는	방식	으로		캣워크		를	활보		했다		
						데	이	ㅁ				누	구도		하	ㄴ							캣	워크			보	했			
																							캐	엇				하	었	다
																												하	았	다
Subword							임		아	직		누구	도		한						으	로	캣					했		다
Character		모	델				임					누	구	도	시	도	한			방	식				워	크				
CV	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	ㅖ	ㅗ	ㅔ	ㅣ	ㅡ	ㅏ	ㅣ	ㅡ	ㅜ	ㅜ	ㅗ	ㅣ	ㅗ	ㅏ	ㅓ	ㅓ	ㅡ	ㅏ	ㅣ	ㅡ	ㅗ	ㅐ	ㅟ	ㅡ	ㅡ	ㅘ	ㅗ	ㅐ	ㅏ
	Jongseong		ㄹ		ㅐ		ㅁ	ㄴ		ㄱ							ㄴ	ㄱ	ㅂㅅ	ㄴ	ㅇ	ㄱ			ㅅ			ㄹ	ㄹ		ㅆ	

Table 1: All possible tokenization schemes with the tokenizers applied in this study. The English source sentence is "*Model Leon Dame strutted down the catwalk like no one has strutted before.*", and their corresponding Korean words are given by the token space.

- **Subword**: It is an arbitrary sequence of strings. It is to note that the surface form of this token resembles morphemes unless the dictionary is intentionally built at the sub-subword level. We, nevertheless, categorize it in isolation, given the absence of morphological meaning in its token.

- **Character**: This token level denotes a string. No tokenizer is needed for the decomposition.

- **CV**: It refers to the smallest token unit, Jamo, meaning consonants and vowels (CV). A certain tokenizer is required to segment a string (equal to a character) into the CV.

3.2 Tokenizer

The meta-level tokens come into shape with the help of tokenizers in most 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], data-driven Khaiii (Kakao Hangul Analyzer III)[5], a subword tokenizer SPM (Kudo and Richardson, 2018), and a CV-level tokenizer, Jamo[6]. Their systematic details are given in Appendix B.

3.2.1 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 Table 7 in Appendix C.

The prototypical tag set is preserved in Komoran or similarly in MeCab and Khaiii. The tokenizer with the most fine-grained tag set is Kkma (56 tags). It provides a detailed analysis of endings. The most coarse form is observed in Okt (19 tags), a tokenizer for Twitter. Woo and Jung (2019) report its outstanding performance in terms of typos, emojis, and punctuation. Hannanum also features a small-sized tag set (22 tags). The particle-related tags are exceptionally reduced in this tokenizer. As mentioned previously, the central divergence of the tag sets is in particles, endings, and affixes.

3.2.2 Tokenization Scenario

The exemplary sentence depicted in Table 1 gives a glimpse of all possible cases of tokens in our experiment. It illustrates that the the most diversified segmentation occurs with verbs (*strutted down*). Intriguingly, some morphological tokenizers partially employ CV, such as shown in 한 versus 하, -ㄴ (the part of *no one has strutted*). Such are the cases of Hannanum, Kkma, Komoran, Khaiii, and Kiwi.

4 Experiment

4.1 Experiment Setup

As Korean evaluation data is scarce, we have organized human evaluation of four commercial NMT systems for the English-to-Korean translation with Direct Assessment (DA), the conventional human evaluation metric employed in Conference on Machine Translation (Barrault et al., 2020). Subsequently, automatic evaluation is performed with BLEU, TER, and ChrF built in `SacreBLEU`. With the resources at hand, the correlation between the two evaluation results is computed on the segment and corpus level.

4.1.1 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 a segment split maintained.

[4] https://github.com/bab2min/Kiwi
[5] https://github.com/kakao/khaiii
[6] https://github.com/JDongian/python-jamo

		Word	Morpheme						Subword	Char	CV
			Hannanum	Kkma	Kiwi	Khaiii	Komoran	MeCab	Okt			
Ratio	Ref	1	2.04	2.27	2.24	2.24	2.22	2.06	1.78⋆	2.30	3.22	7.51‡
	Hyp	1	2.02	2.15	2.14	2.14	2.12	1.97	1.70⋆	2.20	3.16	7.23‡
Time (ms)		-	4,326.91	27,112.96‡	1,959.96	1,494.13	1,084.10	152.59	3,029.68	51.57	5.00⋆	89.07

Table 2: Given our reference and hypothesis translations, a token ratio per word is measured by category. ‡ and ⋆ denote the biggest and smallest values, respectively. In addition, the time to decompose 1,000 sample sentences is calculated in milliseconds.

- **Reference Translation**: We hire a group of professional translators to create Korean reference translations. They are advised not to post-edit MT. To guarantee the highest translation quality, one of our in-house translator double-checks the final version. The revision, nevertheless, 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 anonymized as Sys_A, Sys_B, Sys_P and Sys_Q for legal reason. The system translations are obtained on July 21, 2021.

- **Token Ratio & Time**: Given a word ($ratio = 1.0$), an average token ratio per token type is displayed in Table 2. The size of character and CV tokens are about 1.5 and 4 times larger than that of the average morpheme tokens. In addition, time taken to process 1,000 sentences is logged per token unit. The character level is about 5,000 times faster than Kkma.

In terms of normalizing data, errors in the source test sets and their subsequent impact on the system translations as discussed in Kim et al. (2021) remain undealt with. Only some minor technical issues, i.e. a single quote (') versus a backtick (`), are normalized.

4.1.2 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 set of more or less 300 translated sentences. The contextual information of the documents is maintained to help them consider when making a judgment. They are allowed to reverse their previous decisions within a document boundary.

Regarding their qualification, 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 16,116 (8,058 of Adequacy and Fluency) evaluation data. The judgment on Fluency is only utilized as supplementary information.

4.1.3 Quality Control

Out of the 8,058 Adequacy judgments, the first 10 judgments of each evaluator are removed from the calculation. The scores are then normalized with judge-wise Z-scores. Then, Inter-Quartile Range (IQR) is computed as in Equation 1, where Q_1 and Q_3 signify the first and third quartile values and x denotes outliers that fall into the two categories. Having removed 4.1% of the data, we base our observation on 7,727 judgments.

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

or

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

4.1.4 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 are measured against the human Adequacy judgment by:

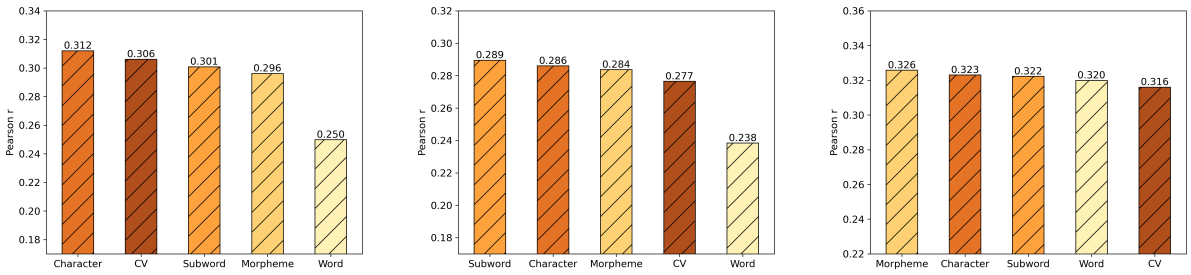
$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H}) \cdot (M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2}\sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}} \quad (2)$$

where H and M refer to the machine and human DA scores, respectively, and $\overline{H}$ and $\overline{M}$, their mean values. The Pearson's r measures the linear relationship between the two variables. During the process, some of the issues have concerned us:

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

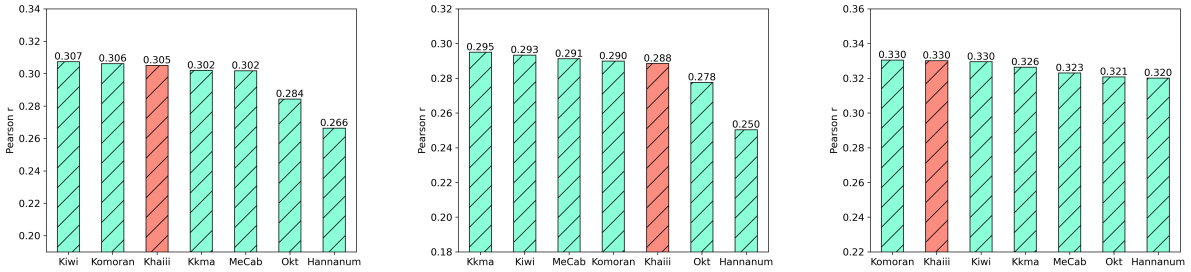
		Default	Word	Morpheme							Subword	Char	CV
				Hannanum	Kkma	Kiwi	Khaiii	Komoran	MeCab	Okt			
BLEU	*ngrams*	4	1	2	2	2	2	2	2	2	2	2	5
ChrF	*char_order*	6	3	3	3	3	3	3	3	3	3	3	5
	word_order	0	0	0	0	1	1	1	1	1	1	2	0

Table 3: The adjusted parameters of BLEU and ChrF per token type.



(a) BLEU (b) TER (c) ChrF

Figure 1: The Pearson correlation on the segment level: concerning the meta-token level. The morpheme corresponds to the average value of all morpheme tokens.



(a) BLEU (b) TER (c) ChrF

Figure 2: The Pearson correlation on the segment level: concerning the morpheme level. Khaiii is in red to inform its different basis.

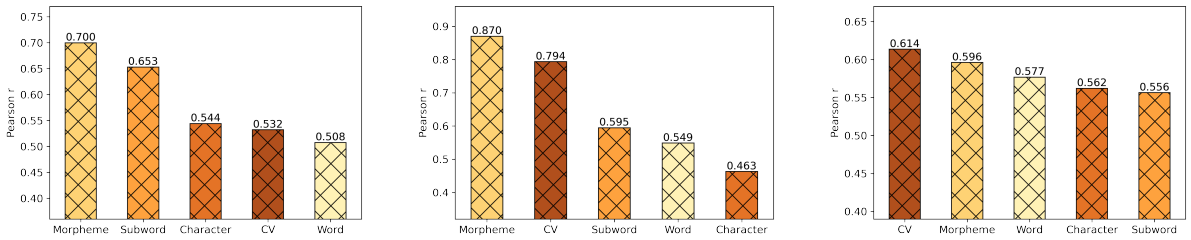
- **Do we adjust n-gram parameters?**
  The BLEU score is a geometric mean of four-grams. As the token unit is divergent, on the one hand, we attempt to avoid a circumstance where any tokenizer benefits from the n-gram parameter. On the other, the default word n-gram of ChrF is zero, which leads to the same conclusion for some tokens. To make the consequence of the token unit clear and compatible, we have organized a preliminary study to obtain the best-correlated n-gram parameters per token typology. The result is provided in Table 3 along with the default values.

- **TER scores over 1.0**
  Theoretically speaking, a TER score of 1.0 represents a total mismatch between a hypothesis and reference. Yet, when a reference is too short for its hypothesis, the computation is programmed to exceed 1.0, which becomes an outlier to the Pearson correlation. We, thus, normalize such cases by cutting down to 1.0.

- **Is the sample size enough?**
  Koehn (2004) reported that they reached a near 100% confidence with 3,000 samples when assessing MT systems with BLEU. In light of their work, we believe that our sample size is affordable to draw a valid conclusion.

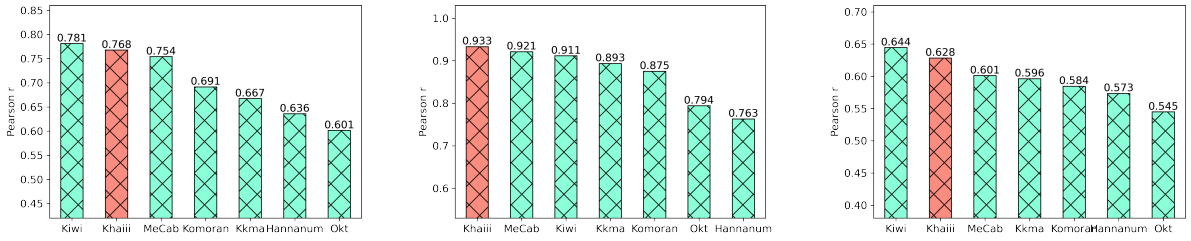
4.2 Experiment Result

The Pearson correlation of `SacreBLEU` to human DA scores when with different token types is reported on the segment and corpus level. On each level, the results are organized by the meta level, with the morpheme represented by the average score of seven types. Afterward, the morpheme tokens are compared among themselves. Khaiii is



(a) BLEU (b) TER (c) ChrF

Figure 3: The Pearson correlation on the corpus level: concerning the meta-token level.



(a) BLEU (b) TER (c) ChrF

Figure 4: The Pearson correlation on the corpus level: concerning the morpheme level.

highlighted with a different color to present its algorithmic divergence.

4.2.1 Segment Level

Figure 1 and Figure 2 reports the Pearson correlation of the meta- and morpheme level, respectively. The scores range from 0.23 to 0.33.

BLEU achieves better human correlation when the token is more fine-grained. When a sentence is not decomposed, the score is likely to lose validity. The best fit for this metric is a character ($r = 0.312$). Among the morphemes, we witness an insignificant correlation of MeCab.

The result of **TER** coincides with BLEU in that any tokenizer can enhance the correlation of the metric. The result shows that SPM goes best with this metric. It is also noticeable that CV results in a poor correlation. Moreover, Khaiii is insignificant to this metric.

ChrF has obtained relatively consistent correlations in all token types despite its re-adjusted parameters. The morpheme level is best suited for this metric, among which Khaiii stands out for a good reason and CV for a wrong reason. CV often deteriorates the correlation of ChrF.

We conclude that any pre-tokenization is essential for BLEU and TER, while ChrF should be approached with caution on the segment level. On the bright side, the performance of Kiwi is noteworthy among the morpheme tokenizers. Furthermore, as a whole, we stress the effectiveness of the character-level segmentation, which guarantees a fast deployment and the human correlation that is often better than MeCab. On the other side, the CV level is undependable in the Korean MT evaluation, unlike in other NLP tasks. Furthermore, Hannanum and Okt are not an option for this task.

4.2.2 Corpus Level

Figure 3 to Figure 4 depict the result of the meta- and morpheme levels, respectively. The score ranges from 0.46 to 0.93, which is much higher and broader than the segment level.

On the meta level, the morpheme tokens are likely to attain a higher correlation to human judgment in all cases. Moreover, the performance of Kiwi and Khaiii is striking. However, the correlation of TER and ChrF degrades with character tokens or SPM in the case of ChrF. Such a tendency is in clear contrast to the finding observed at the segment level.

Additionally, the raw scores of each metric are compared to human DA scores, as shown in Table 4. As expected from the characteristics of the lexical matching system, the smaller units result in higher raw scores, which, however, can soar up to twice in the case of BLEU (from 28.1 to 48.5 in Sys_A). Likewise, the most severe version of TER

	Ave. DA ↑	Ave. z	Word	Okt	MeCab	Komoran	Kkma	Kiwi	Khaiii	Hannanum	SPM	Character	CV
Sys_A	68.783	0.203	28.099	33.398	38.341	40.275	40.986	41.022	40.005	36.939	41.015	48.712	48.467
Sys_B	67.160	0.112	28.932	34.351	39.185	41.007	41.920	41.997	40.881	37.793	41.948	49.553	49.188
Sys_P	64.688	0.027	23.941	30.415	35.605	36.621	37.236	38.458	37.034	32.902	37.213	45.924	45.098
Sys_Q	57.734	-0.220	25.941	31.382	35.602	37.304	38.063	38.138	36.939	34.058	38.155	47.096	46.602

(a) BLEU

	Ave. DA ↑	Ave. z	Word	Okt	MeCab	Komoran	Kkma	Kiwi	Khaiii	Hannanum	SPM	Character	CV
Sys_A	68.783	0.203	82.811	68.223	64.142	63.041	62.253	62.352	63.412	67.833	62.391	57.718	52.932
Sys_B	67.160	0.112	82.334	67.332	63.519	62.585	61.545	61.649	62.867	67.249	61.083	56.364	51.962
Sys_P	64.688	0.027	89.652	69.882	64.898	64.859	63.479	62.983	64.346	71.199	65.914	62.163	54.063
Sys_Q	57.734	-0.220	86.699	70.356	66.611	65.641	64.751	64.758	66.126	71.199	64.767	59.771	54.697

(b) TER

	Ave. DA ↑	Ave. z	Word	Okt	MeCab	Komoran	Kkma	Kiwi	Khaiii	Hannanum	SPM	Character	CV
Sys_A	68.783	0.203	44.897	46.508	47.544	48.904	46.326	49.299	48.763	46.019	47.932	47.887	53.140
Sys_B	67.160	0.112	45.725	47.345	48.370	49.635	47.131	50.096	49.560	46.826	48.807	48.707	53.807
Sys_P	64.688	0.027	42.742	44.171	45.342	46.182	43.796	47.017	46.354	43.401	45.357	45.699	51.198
Sys_Q	57.734	-0.220	43.505	45.134	46.031	47.166	44.639	47.557	47.011	44.378	44.378	46.533	51.775

(c) ChrF

Table 4: The raw scores of the metrics of the four MT systems by token type along with the human DA scores and their z-scores. The highest scores are in blue & red.

scores is before the tokenization (82.33 - 89.69). The ChrF scores, on the other hand, fluctuate moderately from 44.9 to 53.1 (in Sys_A). We, therefore, advise not to copy raw SacreBLEU scores from any studies when this language is concerned.

While so, we discover a substantial problem that the system rankings calculated by the automatic metrics do not comply with the human judgment at all. As the highest scores in blue and red demonstrate such a trend, 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 $[Sys_A = 2, Sys_B = 1, Sys_P = 3, Sys_Q = 4]$. In the worst case, the third and fourth ranks are swapped according to BLEU when tokenized by MeCab, Kiwi, or Khaiii. Such an erroneous conclusion by the metrics can be drawn due to either the small number of systems or possible outlier systems in the experiment setup (Mathur et al., 2020). We leave the verification of this issue to our future work.

5 Extra Meta-Evaluation

As an extended work, we investigate the influence of pre-tokenization on other homogeneous automatic metrics: NLTK-BLEU[8], GLEU[9] (Wu et al., 2016), NIST[10], RIBES (Isozaki et al., 2010), CharacTER (Wang et al., 2016), and EED (Stanchev et al., 2019). We compute the Person correlation r of a total of nine metrics per tokenization on the segment and corpus level under the same environment. The results are provided in Figure 5 through Figure 8 in Appendix D.

5.1 Segment Level

Albeit minor differences from SacreBLEU, NLTK-BLEU is most benefited from the CV level, not the character level. GLEU features a more robust correlation to any given token type than BLEU. Consistent with such a tendency, the CV level increases the correlation of RIBES. Interestingly enough, however, NIST turns out to be vulnerable to any token types except SPM, and the scope of the scores is markedly low (0.1 - 0.19).

In terms of edit-distance-based metrics, the result does not vacillate much and, at the same time, presents high human correlations. CharacTER favors the morpheme level, such as Komoran. EED, on the other hand, does not favor any token types. The more decomposed a token is, the lower the human correlation becomes in this metric.

To summarize, there is a good chance that the CV level enhances the correlation of many n-gram-based metrics such as BLEU. The metrics that a word should be left as it is are NIST and EED.

5.2 Corpus Level

On the corpus level, the morphological tokens are predominantly helpful in obtaining a higher human correlation, as in the case of BLEU, GLEU, and NIST. Among the morphemes, the role of Kiwi is

[8] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[9] https://www.nltk.org/_modules/nltk/translate/gleu_score.html

[10] https://www.nist.gov/itl/iad/mig/metrics-machine-translation-evaluation/

	Word	Kkma	Hannanum	Okt	Komoran	MeCab	Khaiii	**Kiwi** ↑	Subword	Character	CV
EED	0.095	0.094	0.093	0.089⋆	0.092	0.093	0.098	0.096	0.094	0.096	0.201
BLEU	0.110	0.110	0.108	0.107	0.111	0.109	0.134	0.106⋆	0.106⋆	0.108	0.128
ChrF	0.111	0.113	0.108⋆	0.115	0.121	0.115	0.121	0.115	0.115	0.129	0.147
CharacTER	0.284⋆	0.827	0.633	0.434	0.77	0.679	0.763	0.816	0.792	2.391	366.65
GLEU	1.018	1.059	1.075	1.036	1.029	1.060	1.038	1.002	0.961⋆	0.979	1.068
NIST	1.016	1.061	1.044	1.042	1.082	1.033	1.011⋆	1.032	1.032	1.085	1.119
NLTK-BLEU	1.072	1.016	0.982	1.011	0.994	1.036	1.140	1.037	0.981⋆	1.020	1.028
RIBES	1.011⋆	3.888	2.867	1.791	3.360	2.735	3.441	3.578	3.476	13.094	628.96
TER	0.332⋆	9.849	5.236	2.413	8.232	5.061	7.768	7.653	8.106	24.933	362.18

Table 5: The time of each metric to compute a score for 100 sentences when combined with different token units. The value is sorted by Kiwi (unit: seconds). The best scores are with a star(⋆) and the abnormal cases are stressed in blue.

significant. This token type is, however, detrimental to RIBES, which scores the highest correlation in this experiment. The character level, on the other hand, is beneficial to this metric. In the case of CharacTER and NIST, the correlation is degraded with word decomposition by the CV or character level.

5.3 Computation Time

Table 5 describes the time to compute metric scores of 100 sentences per token type. From the perspective of token type, the more fine-grained token type takes more time. For instance, treating CV takes 100 times more than words in TER. No matter how good the CV level can be, inefficiency is its blind spot.

From the viewpoint of automatic metrics, RIBES, TER, and CharacTER are one of the most time-consuming ones. The pairing with CV and RIBES, for instance, would end in taking up about 630 seconds (10 minutes) to deal with 100 sentences. On the contrary, EED boasts the utmost efficiency.

6 Limitations & Future Works

We acknowledge some limitations this work has to embrace. First of all, the number of systems in question is small, which, in part, has led to an arguable conclusion on the corpus level. Furthermore, all of the systems are online APIs. Second, while questioning the influence of token type on the agglutinative languages, we base our study solely on Korean.

It is of our future interest to probe into the consequence of token types in other comparable languages other than Korean. We also intend to scale up the experiment by employing state-of-the-art NMT models.

7 Conclusion

This paper analyzes the influence of diversified token units on the human correlation of `SacreBLEU` on both segment and corpus levels when it comes to agglutinative languages such as Korean by performing meta-evaluation with Pearson correlation. We demonstrate that the pre-tokenization with a fit-for-all token type is not always an optimal choice in Korean MT evaluation. We summarize some of the valuable lessons:

- BLEU and TER should always be accompanied by a segmentation process beforehand.
- Tokenizer should be carefully selected in ChrF.
- The human correlation of some metrics, which are mostly related to edit distance, is easily degraded by token type.
- The CV level is beneficial to some metrics. However, its exponential computation time makes it unprofitable in the MT evaluation.
- Instead, we discover the possibility of a character-level segmentation as a quick and easy substitute on the segment level.
- However, the morpheme level is recommended on the corpus level such as Kiwi or Khaiii, among others.
- The raw score on the corpus level can be inflated up to twice. We strongly advise against copying scores from other studies.

Acknowledgements

Special thanks to the members of Business Automation for their thoughtful comments and sound discussions.

References

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

Alfred V. Aho and Margaret J. Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

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.

Edward Fredkin and Bolt Beranek. 1960. Trie memory. *Communications of the ACM*, pages 490–499.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, page 944–952, USA. Association for Computational Linguistics.

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. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. Overview of the 4th workshop on Asian translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1–54, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

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.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

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.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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 Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Word Decomposition

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

The general definition of a word, as shown above, conjectures that it is segmented with whitespaces. While such is the case of most European languages, it is arguable in Korean whose words do not always accompany spaces between themselves, depending on schools. Here we illustrate three approaches in defining a word: *comprehensive*, *compromising*, and *analytic*. Their views on the independence of post-positional particle, ending, or affix as a word diverge (Nam et al., 2019), as displayed respectively in Table 6 of Level *Word*.

Following the comprehensive standpoint, what is typically understood as a word in Western languages is equivalent to *Eojeol* in Korean. Those with the compromising perspective perceives that endings and affixes are not a word while the analytic school recognizes the independence of endings. That much active discussion is possible with the morpheme boundary as well, due to the fact that a character is divisible.

In other words, **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 6 is also a character, but at the same time it is a combination of two consonants (ㅇ , ㄹㄱ) and one vowel (ㅣ). We call this sub-layer *Jamo* (ㅇ - ㅣ - ㄹㄱ) in Korean or CV in this paper, the abbreviated form from the initial letters of consonant (자음/ja-eum/) and vowel (모음/mo-eum/).

CV is position-wise; it is situated in a fixed position of *Choseong* (initial, ㅇ), *Jungseong* (middle, ㅣ), and *Jongseong* (final, ㄹㄱ), respectively. Some affixes or morphemes take the form of Jongseong, making a diversified token scenario between the morpheme and CV level.

B Architecture of the Morpheme Analyzers

This section delves into the detailed architecture of the morpheme analyzers mentioned in this paper. The aforementioned analyzers are grouped into dictionary-based and data-based by their core algorithm.

B.1 Dictionary-based

Most of the tokenizers applied in this paper belongs to this category. The first step of the tokenization is that when encountered a word, all possible morphological scenarios are represented with some probabilities by referring to a dictionary that contains vocabularies and their morphological information. The next step is to find the optimal morpheme combination that maximizes the observed probability, with the assumption being that the output morpheme m_k of position k is determined by its previous output m_{k-1} and its k^{th} character c_k. Then, as a final procedure m_k is tagged.

For the agglutinative languages whose characters are always divisible, the decomposition depth should be determined whether to separate the character into the CV level. In that sense, we will denominate each case as *non-CV* and *CV level* for convenience's sake.

The non-CV-level decomposition is performed in Kkma, Okt, and Hannanum in our case. Candidate tokens are generated by restoring from the dictionary, and their probabilities are calculated by Dynamic Programming. The CV level segmentation, on the other hand, is the case of Komoran and Kiwi. The probability is calculated by Aho-Corasick string-matching algorithm (Aho and Corasick, 1975) applied on the dictionary which is structured as a look-up table called Tries (Fredkin and Beranek, 1960) of CV.

B.2 Data-driven

Khaiii is the sole analyzer that fits in to this category in this paper. While the previous dictionary-based tokenizers consider the word decomposition as an analysis problem, Khaiii approaches it as a classification problem of determining a morpheme tag for a given input character. One of the main challenges is the disharmonious token length of input and output observed in some cases such as shortened words whose restoration involves the CV-level segmentation. As an instance, the verb 했다 (did) can be segmented into 하/VX + 였/EP + 다/VV. It is clear that just by combining 하 and 였 the original morpheme 했 is not able to be achieved at a character level (하였 vs. 했).

While Recurrent Neural Networks (RNN) is a popular baseline in this regard, Khaiii adopts Convolutional Neural Networks (CNN) to maintain the information of input character and its corresponding output tag. In addition, CNN can speed up the

Level	Denomination	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
Character	Eumjeol	-	-	-	혜, -미, -가, 동, -화, -를, 읽, -었, -다
CV	Jamo	-	-	-	ㅎ, -ㅖ, ㅁ, -ㅣ, ㄱ, -ㅏ, ㄷ, -ㅗ, -ㅇ, ㅎ, -ㅘ, ㄹ, -ㅡ, -ㄹ, ㅇ, -ㅣ, -ㄹㄱ, ㅇ, -ㅓ, -ㅆ, ㄷ, -ㅏ

Table 6: Level of word decomposition in Korean, indicating an open discussion about defining a word (Nam et al., 2019).

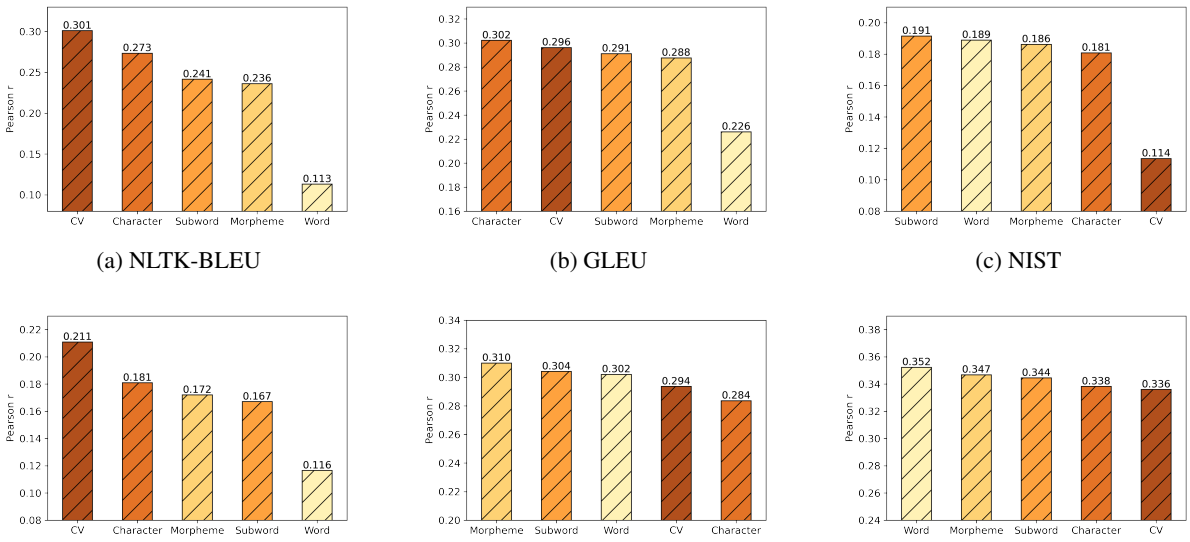
process. More in-depth architecture is provided in their git page. The model is trained with Sejong Corpus provided by Sejong Project, together with a manually created 6k words. After rooting erroneous sentences out, the size of the corpus reaches about 10.3 million words/Eojeol).

C Tag Sets of Korean Tokenizers

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			
	connective ending	**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	SSO	SS		SS	SS
						SSC				
	~ - _		SP		SP	SC	SP		SP	SP
	others		SO		SO	SY	SO		SO	SO
Etc.	**Chinese character**		SW	Foreign	SW		SW		SW	SW
	foreign word		SH		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: Tag sets of Sejong Project and seven Korean tokenizers.

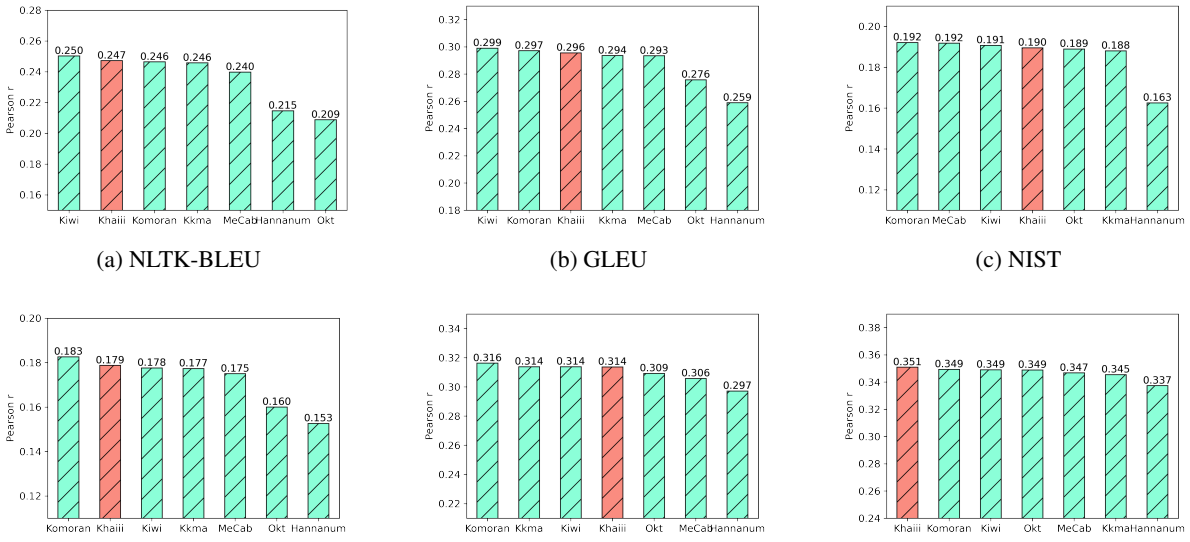
D Meta-Evaluation



(a) NLTK-BLEU (b) GLEU (c) NIST

(d) RIBES (e) CharacTER (f) EED

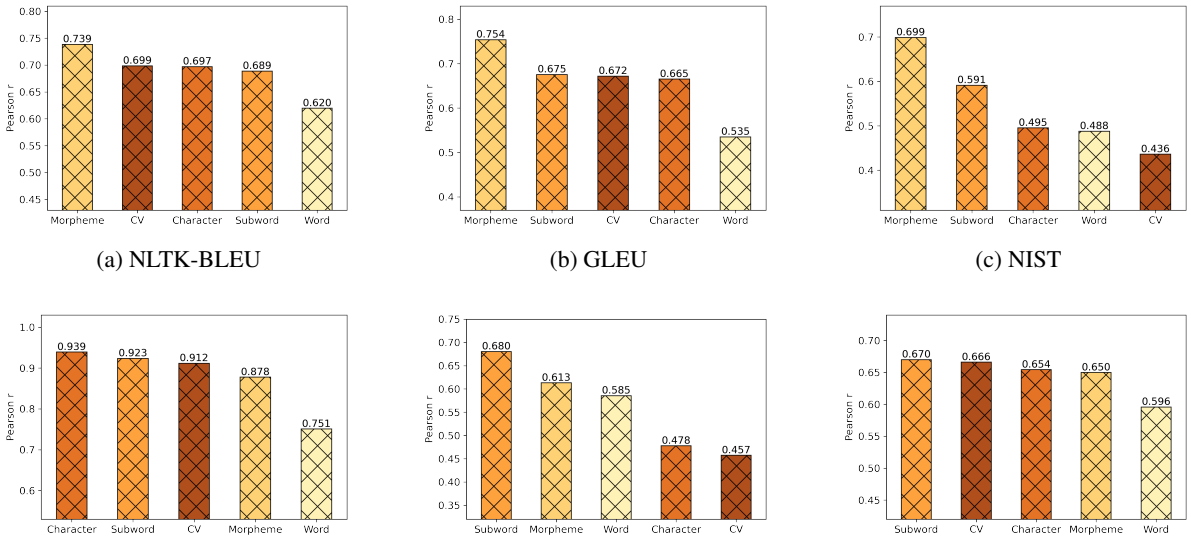
Figure 5: The Pearson correlation on the segment level: concerning the meta-level



(a) NLTK-BLEU (b) GLEU (c) NIST

(d) RIBES (e) CharacTER (f) EED

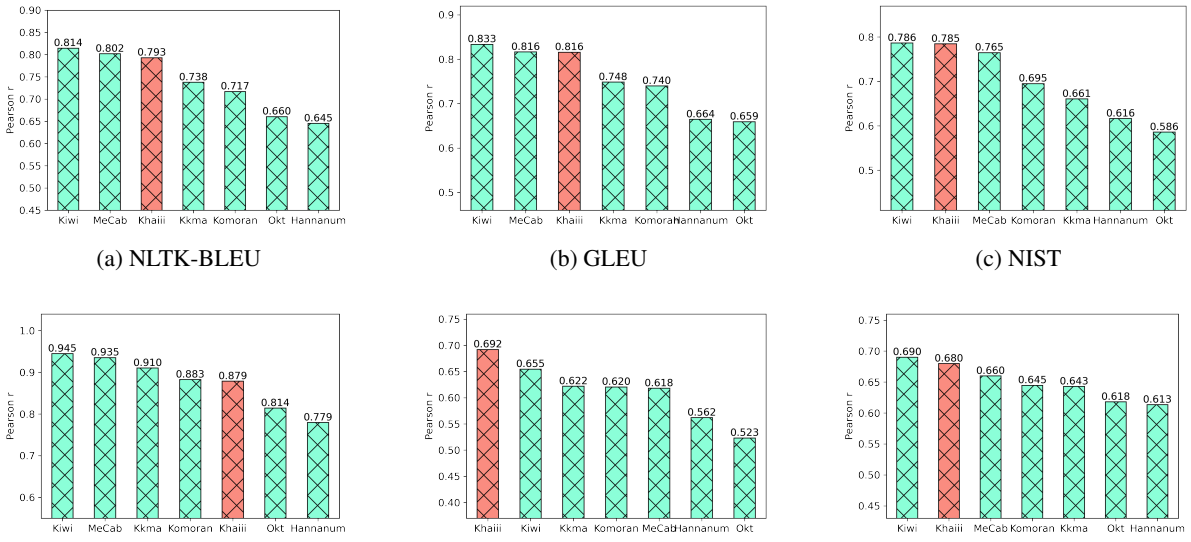
Figure 6: The Pearson correlation on the segment level: concerning the morpheme level



(a) NLTK-BLEU (b) GLEU (c) NIST

(d) RIBES (e) CharacTER (f) EED

Figure 7: The Pearson correlation on the corpus level: concerning the meta-level



(a) NLTK-BLEU (b) GLEU (c) NIST

(d) RIBES (e) CharacTER (f) EED

Figure 8: The Pearson correlation on the corpus level: concerning the morpheme level