User-Generated Text Corpus for Evaluating Japanese Morphological Analysis and Lexical Normalization

Shohei Higashiyama[1,2], **Masao Utiyama**[1], **Taro Watanabe**[2], **Eiichiro Sumita**[1]
[1]National Institute of Information and Communications Technology, Kyoto, Japan
[2]Nara Institute of Science and Technology, Nara, Japan
{shohei.higashiyama, mutiyama, eiichiro.sumita}@nict.go.jp,
taro@is.naist.jp

Abstract

Morphological analysis (MA) and lexical normalization (LN) are both important tasks for Japanese user-generated text (UGT). To evaluate and compare different MA/LN systems, we have constructed a publicly available Japanese UGT corpus. Our corpus comprises 929 sentences annotated with morphological and normalization information, along with category information we classified for frequent UGT-specific phenomena. Experiments on the corpus demonstrated the low performance of existing MA/LN methods for non-general words and non-standard forms, indicating that the corpus would be a challenging benchmark for further research on UGT.

1 Introduction

Japanese morphological analysis (MA) is a fundamental and important task that involves word segmentation, part-of-speech (POS) tagging and lemmatization because the Japanese language has no explicit word delimiters. Although MA methods for well-formed text (Kudo et al., 2004; Neubig et al., 2011) have been actively developed taking advantage of the existing annotated corpora of newswire domains, they perform poorly on user-generated text (UGT), such as social media posts and blogs. Additionally, because of the frequent occurrence of informal words, lexical normalization (LN), which identifies standard word forms, is another important task in UGT. Several studies have been devoted to both tasks in Japanese UGT (Sasano et al., 2013; Kaji and Kitsuregawa, 2014; Saito et al., 2014, 2017) to achieve the robust performance for noisy text. Previous researchers have evaluated their own systems using in-house data created by individual researchers, and thus it is difficult to compare the performance of different systems and discuss what issues remain in these two tasks. Therefore, publicly available data is necessary for a fair evaluation of MA and LN performance on Japanese UGT.

In this paper, we present the blog and Q&A site normalization corpus (BQNC),[1] which is a public Japanese UGT corpus annotated with morphological and normalization information. We have constructed the corpus under the following policies: (1) available and restorable; (2) compatible with the segmentation standard and POS tags used in the existing representative corpora; and (3) enabling a detailed evaluation of UGT-specific problems.

For the first requirement, we extracted and used the raw sentences in the blog and Q&A site registers compiled by (the non-core data of) the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014), in which the original sentences are preserved.[2] For the second requirement, we followed the short unit word (SUW) criterion of the National Institute for Japanese Language and Linguistics (NINJAL), which is used in various NINJAL's corpora, including manually annotated sentences in the BCCWJ. For the third requirement, we organized linguistic phenomena frequently observed in the two registers as word categories, and annotated each word with a category. We expect that this will contribute to future research to develop systems that manage UGT-specific problems.

The BQNC comprises sentence IDs and annotation information, including word boundaries, POS, lemmas, standard forms of non-standard word tokens, and word categories. We will release the annotation information that enables BCCWJ applicants to replicate the full BQNC data from the original BCCWJ data.[3]

Using the BQNC, we evaluated two existing

[1] Our corpus will be available at https://github.com/shigashiyama/jlexnorm.

[2] Twitter could be a candidate for a data source. However, redistributing original tweets collected via the Twitter Streaming APIs is not permitted by Twitter, Inc., and an alternative approach to distributing tweet URLs has the disadvantage that the original tweets can be removed in the future.

[3] https://pj.ninjal.ac.jp/corpus_center/bccwj/en/subscription.html

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5532–5541
June 6–11, 2021. ©2021 Association for Computational Linguistics

	Category	Example	Reading	Translation	Standard forms
Type of vocabulary	(General words)				
	Neologisms/Slang	コピペ	*copipe*	copy and paste	
	Proper names	ドラクエ	*dorakue*	Dragon Quest	
	Onomatopoeia	キラキラ	*kirakira*	glitter	
	Interjections	おお	*ō*	oops	
	Dialect words	ほんま	*homma*	truly	
	Foreign words	ＥＡＳＹ		easy	
	Emoticons/AA	（＾－＾）			
Type of variant form	(Standard forms)				
	Character type variants	カワイイ	*kawaī*	cute	かわいい,可愛い
	Alternative representations	大きぃ	*ōkī*	big	大きい
	Sound change variants	おいしーい	*oishīi*	tasty	おいしい,美味しい
	Typographical errors	つたい	*tsutai*	tough	つらい,辛い

Table 1: Word categories in the BQNC

methods: a popular Japanese MA toolkit called MeCab (Kudo et al., 2004) and a joint MA and LN method (Sasano et al., 2013). Our experiments and error analysis showed that these systems did not achieve satisfactory performance for non-general words. This indicates that our corpus would be a challenging benchmark for further research on UGT.

2 Overview of Word Categories

Based on our observations and the existing studies (Ikeda et al., 2010; Kaji et al., 2015), we organized word tokens that may often cause segmentation errors into two major types with several categories, as shown in Table 1. We classified each word token from two perspectives: the type of vocabulary to which it belongs and the type of variant form to which it corresponds. For example, ニホン *nihon* 'Japan' written in *katakana* corresponds to a *proper name* and a *character type variant* of its standard form 日本 written in *kanji*.

Specifically, we classified vocabulary types into *neologisms/slang*, *proper names*, *onomatopoeia*,[4] *interjections*, *(Japanese) dialect words*, *foreign words*, and *emoticons/ASCII art (AA)*, in addition to general words.[5] A common characteristic of these vocabularies, except for general words, is that a new word can be indefinitely invented or imported. We annotated word tokens with vocabulary type information, except for general words.

From another perspective, any word can have multiple variant forms. Because the Japanese writing system comprises multiple script types including *kanji* and two types of *kana*, that is, *hiragana* and *katakana*,[6] words have orthographic variants written in different scripts. Among them, non-standard *character type variants* that rarely occur in well-formed text but occur in UGT can be problematic, for example, a non-standard form カワイイ for a standard form かわいい *kawaī* 'cute'. Additionally, ill-spelled words are frequently produced in UGT. We further divided them into two categories. The first is *sound change variants* that have a phonetic difference from the original form and are typically derived by deletions, insertions, or substitutions of vowels, long sound symbols (*chōon* "ー"), long consonants (*sokuon* "っ"), and moraic nasals (*hatsuon* "ん"), for example, おいしーい *oishīi* for おいしい *oishī* 'tasty'. The second category is *alternative representations* that do not have a phonetic difference and are typically achieved by substitution among uppercase or lowercase kana characters, or among vowel characters and long sound symbols, for example, 大きぃ for 大きい *ōkī* 'big'. Moreover, *typographical errors* can be seen as another type of variant form. We targeted these four types of non-standard forms for normalization to standard forms.

3 Corpus Construction Process

The BQNC was constructed using the following steps. The annotation process was performed by the first author.

[4] "Onomatopoeia" typically refers to both the phonomime and phenomime in Japanese linguistics literature, similar to ideophones. We follow this convention in this paper.

[5] We observed a few examples of other vocabulary types, such as Japanese archaic words and special sentence-final particles in our corpus, but we treated them as general words.

[6] Morphographic *kanji* and syllabographic *hiragana* are primarily used for Japanese native words (*wago*) and Japanese words of Chinese origin (Sino-Japanese words or *kango*), whereas syllabographic *katakana* is primarily used, for example, for loanwords, onomatopoeia, and scientific names. Additionally, Arabic numerals, Latin letters (*rōmaji*), and other auxiliary symbols are used in Japanese sentences.

Token	Translation	Standard form ID
イイ	good	38988:良い
歌	song	
です	(polite copula)	
ねェ	(emphasis marker)	28754:ね
ヨカッ	good	38988:良い_ヨカッ
タ	(past tense marker)	21642:た

Table 2: Examples of annotated text "イイ歌ですねェ" (It's a good song, isn't it?) and "ヨカッタ" (It was good.). Attributes except for token and standard form ID are abbreviated.

Standard form ID	Standard forms
21642:た	た
28754:ね	ね
38988:良い	良い,よい,いい
38988:良い_ヨカッ	良かっ,よかっ

Table 3: Examples of standard form IDs

(1) Sentence Selection We manually selected sentences to include in our corpus from the blog and Q&A site registers in the BCCWJ non-core data. We preferentially extracted sentences that contained candidates of UGT-specific words, that is, word tokens that may belong to non-general vocabularies or correspond to non-standard forms. As a result, we collected more than 900 sentences.

(2) First Annotation Sentences in the non-core data have been automatically annotated with word boundaries and word attributes, such as POS and lemma. Following the BCCWJ annotation guidelines (Ogura et al., 2011a,b) and UniDic (Den et al., 2007), which is an electronic dictionary database designed for the construction of NINJAL's corpora, we refined the original annotations of the selected sentences by manually checking them. The refined attributes were token, POS, conjugation type, conjugation form, pronunciation, lemma, and lemma ID. Additionally, we annotated each token with a word category shown in Table 1 and a standard form ID if the token corresponded to a non-standard form.

Table 2 shows two examples of annotated sentences. We annotated each non-standard token with a standard form ID denoted as "[lemma ID]:[lemma](_[pronunciation])", which is associated with the set of acceptable standard forms shown in Table 3.

(3) Second Annotation We rechecked all tokens in the sentences that we finished the first annotation and fixed the annotation criteria, that is, the definitions of vocabulary types and variant form types, and standard forms for each word. Through these steps, we obtained 929 annotated sentences.

4 Detailed Definition of Word Categories

4.1 Type of Vocabulary

Through the annotation process, we defined the criteria for vocabulary types as follows.

Neologisms/Slang: a newly invented or imported word that has come to be used collectively. Specifically, we used a corpus reference application called Chunagon[7] and regarded a word as a *neologism/slang* if its frequency in the BCCWJ was less than five before the year 2000 and increased to more than ten in 2000 or later.[8]

Proper names: following the BCCWJ guidelines, we regarded a single word that corresponded to a proper name, such as person name, organization name, location name, and product name, as a *proper name*. In contrast to the BCCWJ guidelines, we also regarded an abbreviation of a proper name as a *proper name*, for example, "ドラクエ" in Table 1.

Onomatopoeia: a word corresponds to onomatopoeia. We referred to a Japanese onomatopoeia dictionary (Yamaguchi, 2002) to assess whether a word is onomatopoeic. We followed the criteria in the BCCWJ guidelines on what forms of words are onomatopoeic and what words are associated with the same or different lemmas.

Interjections: a word whose POS corresponds to an interjection. Although we defined standard forms for idiomatic greeting expressions registered as single words in UniDic,[9] we did not define standard and non-standard forms for other interjections that express feelings or reactions, for example, ええ *ē* 'uh-huh' and うわあ *uwā* 'wow'.

Foreign words: a word from non-Japanese languages. We regarded a word written in scripts in the original language as a *foreign word*, for example, English words written in the Latin alphabet such as "plastic". Conversely, we regarded loanwords written in Japanese scripts (hiragana, katakana, or kanji) as general words, for example, プラスチッ

[7] https://chunagon.ninjal.ac.jp

[8] The original sentences were from posts published between 2004 and 2009.

[9] Eight greeting words exist, for example, ありがとう *arigatō* 'thank you' and さようなら *sayōnara* 'see you'.

ク ‘plastic’. Moreover, we did not regard English acronyms and abbreviations written in uppercase letters as foreign words because such words are typically also written in the Latin alphabet in Japanese sentences, for example, ＳＮＳ.

Dialect words: a word from a Japanese dialect. We referred to a Japanese dialect dictionary (Sato, 2009) and regarded a word as a *dialect word* if it corresponded to an entry or occurred in an example sentence. We did not consider normalization from a dialect word to a corresponding word in the standard Japanese dialect.

Emoticons/AA: nonverbal expressions that comprise characters to express feelings or attitudes. Because the BCCWJ guidelines does not explicitly describe criteria on how to segment emoticon/AA expressions as words, we defined criteria to follow emoticon/AA entries in UniDic.[10]

4.2 Type of Variant Form

There are no trivial criteria to determine which variant forms of a word are standard forms because most Japanese words can be written in multiple ways. Therefore, we defined standard forms of a word as all forms whose occurrence rates were approximately equal to 10% or more in the BCCWJ among forms that were associated with the same lemma. For example, among variant forms of the lemma 面白い *omoshiroi* ‘interesting’ or ‘funny’ that occurred 7.9K times, major forms 面白い and おもしろい accounted for 72% and 27%, respectively, and other forms, such as オモシロイ and オモシロい, were very rare. In this case, the standard forms of this word are the two former variants. We annotated tokens corresponding to the two latter non-standard forms with the standard form IDs and the types of variant forms. We defined criteria for types of variant forms as follows.

Character type variants: among the variants written in different scripts, we regarded variants whose occurrence rates were approximately equal to 5% or less in the BCCWJ as non-standard forms of *character type variants*. Specifically, variants written in kanji, hiragana, or katakana for native words and Sino-Japanese words, variants written in katakana or hiragana for loanwords, variants written in uppercase or lowercase Latin letters for English abbreviations are candidates for character type variants. We assessed whether these candidates were non-standard forms based on the occurrence rates.

Alternative representations: a form whose internal characters are (partially) replaced by special characters without phonetic differences. Specifically, non-standard forms of *alternative representations* include native words and Sino-Japanese words written in historical kana orthography (e.g., 思ふ for 思う *omō*/*omou* ‘think’), and loanwords written as an unusual[11] katakana sequence (e.g., オオケストラ for オーケストラ ‘orchestra’). Additionally, *alternative representations* include substitution with respect to kana: substitution of the long vowel kana by the long sound symbol (e.g., おいし〜 for おいしい *oishī* ‘tasty’), substitution of upper/lowercase kana by the other case (e.g., ゎたし for わたし *watashi* ‘me’), and phonetic or visual substitution of kana characters by Latin letters and symbols (e.g., かわＥ for かわいい *kawaī* ‘cute’ and こωにちは for こんにちは *konnichiwa* ‘hello’).

Sound change variants: a form whose pronunciation is changed from the original form. Specifically, *sound change variants* include the insertion of special moras (e.g., 強ーい *tsuyōi* for 強い *tusyoi* ‘strong’), deletion of moras (e.g., くさ *kusa* for くさい *kusai* ‘stinking’), and substitution of characters/moras (e.g., っす *ssu* for です *desu* polite copula and すげえ *sugē* for すごい *sugoi* ‘awesome’).

Typographical errors: a form with typographical errors derived from character input errors, kana-kanji conversion errors, or the user’s incorrect understanding. For example, つたい *tsutai* for つらい *turai* ‘tough’ and そｒ for それ *sore* ‘it’.

5 Evaluation

We present the statistics of the BQNC in Table 4. It comprises 929 sentences, 12.6K word tokens, and 767 non-standard word tokens. As shown in Table 6, the corpus contains tokens of seven types of vocabulary and four types of variant form. Whereas there exist fewer than 40 instances of neologisms/slang, dialect words, foreign words, and

[10] For example, if characters expressing body parts were outside of punctuation expressing the outline of a face, the face and body parts were segmented, but both were annotated with *emoticons/AA*, for example, “ｍ（．＿．）ｍ” → “ｍ|（．＿．）|ｍ”.

[11] We assessed whether a form is unusual if its occurrence rate was approximately equal to 5% or less in the BCCWJ similar to the case of character type variants.

Regi- ster	# sent	# word token	# word type	# NSW token	# NSW type
Q&A	379	5,649	1,699	320	221
Blog	550	6,951	2,231	447	257
Total	929	12,600	3,419	767	420

Table 4: Statistics of the BQNC. NSW represents non-standard word.

Task	MeCab			MeCab+ER		
	P	R	F	P	R	F
SEG	89.2	95.1	92.1	**93.5**	**96.5**	**95.0**
POS	87.5	93.3	90.3	**91.4**	**94.3**	**92.8**
NOR	–	–	–	55.9	25.8	35.3

Table 5: Overall performance

typographical errors, each of the other category has more than 100 instances. Our corpus contains a similar number of non-standard tokens to Kaji and Kitsuregawa (2014)'s Twitter corpus (1,831 sentences, 14.3K tokens, and 793 non-standard tokens) and Osaki et al. (2017)'s Twitter corpus (1,405 sentences, 19.2K tokens, and 768 non-standard tokens). The former follows the POS tags for the Japanese MA toolkit JUMAN and the latter follows the authors own POS tags that extend NINJAL's SUW.

In the following subsections, we evaluate the existing methods for MA and LN on the BQNC and discuss correctly or incorrectly analyzed results.

5.1 Systems

We evaluated two existing methods. First, we used MeCab 0.996 (Kudo et al., 2004),[12] which is a popular Japanese MA toolkit based on conditional random fields. We used UniDicMA (unidic-cwj-2.3.0)[13] as the analysis dictionary, which contains attribute information of 873K words and MeCab's parameters (word occurrence costs and transition costs) learned from annotated corpora, including the BCCWJ (Den, 2009).

Second, we used our implementation of Sasano et al. (2013)'s joint MA and LN method. They defined derivation rules to add new nodes in the word lattice of an input sentence built by their baseline system, JUMAN. Specifically, they used the following rules: (i) sequential voicing (*rendaku*), (ii) substitution with long sound symbols and lowercase kana, (iii) insertion of long sound symbols and lowercase kana, (iv) repetitive onomatopoeia (XYXY-form[14]) and (v) non-repetitive onomatopoeia (XQY*ri*-form and XXQ*to*-form). For example, rule (iii) adds a node of 冷たぁぁい *tsumetāi* as a variant form of 冷たい *tsumetai* 'cold' and rule (iv) adds a node of うはうは *uhauha* 'exhilarated' as an onomatopoeic adverb. if the input sentences contain such character sequences.

The original implementation by Sasano et al. (2013) was an extension of JUMAN and followed JUMAN's POS tags. To adapt their approach to the SUW, we implemented their rules and used them to extend the first method of MeCab using UniDicMA. We set the costs of the new nodes by copying the costs of their standard forms or the most frequent costs of the same-form onomatopoeia, whereas Sasano et al. (2013) manually defined the costs of each type of new word. We denote this method by MeCab+ER (Extension Rules). Notably, we did not conduct any additional training to update the models' parameters for either methods.

5.2 Overall Results

Table 5 shows the overall performance, that is, **P**recision, **R**ecall, and **F**$_1$ score, of both methods for **SEG**mentation, **POS** tagging[15] and **NOR**malization.[16] Compared with well-formed text domains,[17] the relatively lower performance (F$_1$ of 90–95%) of both methods for segmentation and POS tagging indicates the difficulty of accurate segmentation and tagging in UGT. However, MeCab+ER outperformed MeCab by 2.5–2.9 F$_1$ points because of the derivation rules. Regarding the normalization performance of MeCab+ER, the method achieved moderate precision but low recall, which indicates its limited coverage for various variant forms in the dataset.

5.3 Results for Each Category

Table 6 shows the segmentation and POS tagging recall for both methods for each category. In contrast to the sufficiently high performance for general words, both methods performed worse for words of characteristic categories in UGT; micro average recall was at most 79.6% for segmentation

[12] https://taku910.github.io/mecab/

[13] https://unidic.ninjal.ac.jp/

[14] "X" and "Y" represent the same kana character(s) corresponding to one mora, "Q" represents a long consonant character "っ/ッ", "*ri*" represents a character "り/リ", and "*to*" represents a character "と/ト".

[15] We only evaluated top-level POS.

[16] We regarded a predicted standard form as correct if the prediction was equal to one of the gold standard forms.

[17] For example, Kudo et al. (2004) achieved F$_1$ of 98–99% for segmentation and POS tagging in news domains.

Category	#	MeCab		MeCab+ER	
		SEG	POS	SEG	POS
Dialect words	23	91.3	78.3	**95.7**	**82.6**
Proper names	103	87.4	84.5	**88.4**	**85.4**
Onomatopoeia	218	79.8	73.4	**87.2**	**77.1**
Foreign words	14	78.6	78.6	78.6	78.6
Emoticons/AA	270	73.7	64.1	73.3	63.3
Interjections	174	64.9	**53.5**	**72.4**	48.9
Neologisms/Slang	37	67.6	67.6	67.6	67.6
Sound change var.	419	50.6	47.5	**82.6**	**76.4**
Char type var.	248	71.0	62.9	**78.2**	**69.4**
Alternative rep.	132	65.2	54.6	**76.5**	**69.0**
Typos	23	47.8	30.4	47.8	30.4
Non-gen/std total	1565	68.9	61.9	79.6	70.4
Standard forms of general words	11K	98.9	97.7	98.9	97.7

Table 6: Recall for each category (SEG and POS)

Category	#	MeCab+ER
Sound change variants	419	37.0
Character type variants	248	0.0
Alternative representations	132	32.6
Typographical errors	23	0.0

Table 7: Recall for each category (normalization)

MeCab\MeCab+ER	T	F
T	11955	32
F	200	413

Table 8: Number of correct (T) or incorrect (F) segmentation for two methods

and 70.4% for POS tagging ("non-gen/std total" column). MeCab+ER outperformed MeCab particularly for onomatopoeia, character type variants, alternative representations, and sound change variants. The high scores for dialect words were probably because UniDicMA contains a large portion of (19 out of 23) dialect word tokens. Interjection was a particularly difficult vocabulary type, for which both methods recognized only approximately 50% of the gold POS tags. We guess that this is because the lexical variations of interjections are diverse; for example, there are many user-generated expressions that imitate various human voices, such as laughing, crying, and screaming.

Table 7 shows the recall of MeCab+ER's normalization for each category. The method correctly normalized tokens of alternative representations and sound change variants with 30–40% recall. However, it completely failed to normalize character type variants not covered by the derivation rules and more irregular typographical errors.

5.4 Analysis of the Segmentation Results

We performed error analysis of the segmentation results for the two methods. Table 8 shows a matrix of the number of correct or incorrect segmentations for the methods for gold words. There existed 32 tokens that only MeCab correctly segmented (T-F), 200 tokens that only MeCab+ER correctly segmented (F-T), and 413 tokens that both methods incorrectly segmented (F-F).

In Table 9, we show the actual segmentation/normalization examples using the methods for the three cases; the first, second, and third blocks show examples of T-F, F-T, and F-F cases, respectively. First, out of 32 T-F cases, MeCab+ER incorrectly segmented tokens as onomatopoeia in 18 cases. For example, (a) and (b) correspond to new nodes added by the rules for the XQY*ri*-form and XYXY-form onomatopoeia, respectively, even though (a) is a verb phrase and (b) is a repetition of interjections.

Second, out of 200 F-T cases that only MeCab+ER correctly segmented, the method correctly normalized 119 cases, such as (c), (d), and the first word in (g), and incorrectly normalized 42 cases, such as (e) and the second word in (f). The remaining 39 cases were tokens that required no normalization, such as the first word in (f), the second word in (g), and (h). The method correctly normalized simple examples of sound change variants (c: しーかーも for しかも) and alternative representations (d: ぉぃら for おいら) because of the substitution and insertion rules, but failed to normalize character type variants (f: やきゅー for 野球) and complicated sound change variants (e: んまぃ for うまい).

Third, out of 413 F-F cases, 148 tokens were complicated variant forms, including a combination of historical kana orthography and the insertion of the long sound symbol (i), a combination of the character type variant and sound change variant (j), a variant written in *romaji* (k). The remaining 265 tokens were other unknown words, including emoticons (l), neologisms/slang (m), and proper names (n).[18]

5.5 Analysis of the Normalization Results

Table 10 shows the detailed normalization results for MeCab+ER. Among 767 non-standard words (Gold), the method correctly normalized 198 true positives (TP) and missed 569 (58+511) false nega-

[18] 社割 *shawari* is an abbreviation of 社員割引 *shain waribiki* 'employee discount'. ガルバディア 'Galbadia' is an imaginary location name in the video game Final Fantasy.

	VT	Gold SEG&SForms	Reading	Translation	MeCab result	MeCab+ER result
(a)		はっ|たり	*haQ|tari*	paste and	はっ|たり	はったり
(b)		こら|こら	*kora|kora*	hey hey	こら|こら	こらこら
(c)	S	しーかーも [しかも]	*shīkāmo*	besides	しー|かー|も	しーかーも [しかも]
(d)	A	ぉぃら [おいら,オイラ]	*oira*	I	ぉ|ぃ|ら	ぉぃら [おいら]
(e)	S	んまぃ [美味い,旨い,うまい]	*mmai*	yummy	ん|ま|ぃ	んまぃ [んまい]
(f)	C,A	も|やきゅー [野球]	*mo|yakyū*	also, baseball	もや|きゅー	も|やきゅー [やきゅう]
(g)	S	たしーか [確か,たしか] |に	*tashīka|ni*	surely	た|し|ー|かに	たしーか [たしか] |に
(h)		ふぅ〜〜ん	*fūn*	hmm	ふぅ〜|〜|ん	ふぅ〜〜ん [ふん]
(i)	S	ませう〜 [ましょう]	*mashō*	let's	ませ|う|〜	ませ|う〜 [う]
(j)	C,S	けこーん [結婚]	*kekōn*	marriage	け|こーん	け|こー [こう] |ん
(k)	A	ください|ｎｅ [ね]	*kudasai|ne*	Won't you. . .?	ください|ｎ|ｅ	ください|ｎ|ｅ
(l)		（＾へ＾）			（|＾|へ|＾|）	（|＾|へ|＾|）
(m)		社割	*shawari*	employee discount	社|割	社|割
(n)		ガルバディア	*garubadhia*	Galbadia	ガルバ|ディア	ガルバ|ディア

Table 9: Segmentation and normalization results (shown in "[]") by MeCab and MeCab+ER. Incorrect results are written in gray. VT represents variant type. C, A, and S represent character type variant, alternative representation, and sound change variants, respectively. Gold SEG&SForms represent the gold segmentation and gold standard forms (shown in "[]").

Total		T-SEG				F-SEG
Gold	767	TP	198	FN	58	511
Pred	354	TP	198	FP	99	57

Table 10: Detailed normalization results for MeCab+ER

tives (FN). Similarly, among 354 predictions (Pred), the methods incorrectly normalized 156 (99+57) false positives (FP). We further divided FN and FP according to whether they were correctly segmented (T-SEG) or not (F-SEG).

We do not show TP and FN examples here because we already introduced some examples in §5.4. Among the FP examples, some of them were not necessarily inappropriate results; normalization between similar interjections and onomatopoeia was intuitively acceptable (e.g., おお〜 was normalized to おお *ō* 'oh' and サラサラ〜 was normalized to サラサラ *sarasara* 'smoothly'). However, we assessed these as errors based on our criterion that interjections have no (non-)standard forms and the BCCWJ guidelines that regards onomatopoeia with and without long sound insertion as different lemmas.

5.6 Discussion

The derivation rules used in MeCab+ER improved segmentation and POS tagging performance and contributed to the correct normalization of parts of variant forms, but the overall normalization performance was limited to F_1 of 35.3%.

We classified the main segmentation and normalization errors into two types: complicated variant forms and unknown words of specific vocabulary types such as emoticons and neologisms/slang. The effective use of linguistic resources may be required to build more accurate systems, for example, discovering variant form candidates from large raw text similar to (Saito et al., 2017), and constructing/using term dictionaries of specific vocabulary types.

6 Related Work

UGT Corpus for MA and LN Hashimoto et al. (2011) developed a Japanese blog corpus with morphological, grammatical, and sentiment information, but it contains only 38 non-standard forms and 102 misspellings as UGT-specific examples. Osaki et al. (2017) constructed a Japanese Twitter corpus annotated with morphological information and standard word forms. Although they published tweet URLs along with annotation information, we could only restore parts of sentences because of the deletion of the original tweets. Sasano et al. (2013); Kaji and Kitsuregawa (2014); Saito et al. (2014, 2017) developed Japanese MA and LN methods for UGT, but most of their in-house data are not publicly available.

For English LN, Han and Baldwin (2011) constructed an English Twitter corpus and Yang and Eisenstein (2013) revised it as LexNorm 1.2. Baldwin et al. (2015) constructed an English Twitter corpus (LexNorm2015) for the W-NUT 2015 text normalization shared task. Both LexNorm 1.2 and LexNorm2015 have been used as benchmark datasets for LN systems (Jin, 2015; van der Goot, 2019; Dekker and van der Goot, 2020).

For Chinese, Li and Yarowsky (2008) published a dataset of formal-informal word pairs collected from Chinese webpages. Wang et al. (2013) released a crowdsourced corpus constructed from microblog posts on Sina Weibo.

Classification of Linguistic Phenomena in UGT

To construct an MA dictionary, Nakamoto et al. (2000) classified unknown words occurring in Japanese chat text into contraction (e.g., すげー for すごい *sugoi* 'awesome'), exceptional kana variant (e.g., こんぴゅーた for コンピュータ 'computer'), abbreviation, typographical errors, filler, phonomime and phenomime, proper nouns, and other types. Ikeda et al. (2010) classified "peculiar expressions" in Japanese blogs into visual substitution (e.g., わたU for わたし *watashi* 'me'), sound change (e.g., でっかい for でかい *dekai* 'big'), kana substitution (e.g., びたみん for ビタミン 'vitamin'), and other unknown words into similar categories to Nakamoto et al. (2000). Kaji et al. (2015) performed error analysis of Japanese MA methods on Twitter text. They classified missegmented words into a dozen categories, including spoken or dialect words, onomatopoeia, interjections, emoticons/AA, proper nouns, foreign words, misspelled words, and other non-standard word variants. Ikeda et al. (2010)'s classification of peculiar expressions is most similar to our types of variant forms and Kaji et al. (2015)'s classification is most similar to our types of vocabulary (shown in Table 2), whereas we provide more detailed definitions of categories and criteria for standard and non-standard forms. Other work on Japanese MA and LN did not consider diverse phenomena in UGT (Sasano et al., 2013; Saito et al., 2014).

For English, Han and Baldwin (2011) classified ill-formed English words on Twitter into extra/missing letters and/or number substitution (e.g., "b4" for "before"), slang (e.g., "lol" for "laugh out loud"), and "others". van der Goot et al. (2018) defined a more comprehensive taxonomy with 14 categories for a detailed evaluation of English LN systems. It includes phrasal abbreviation (e.g., "idk" for "I don't know"), repetition (e.g., "soooo" for "so"), and phonetic transformation (e.g., "hackd" for "hacked").

For Chinese, Li and Yarowsky (2008) classified informal words in Chinese webpages into four types: homophone (informal words with similar pronunciation to formal words, e.g., 稀饭 ⟨xīfàn⟩[19] "rice gruel" for 喜欢 ⟨xǐhuan⟩ "like"), abbreviation and acronym (e.g., GG for 哥哥 ⟨gēge⟩ "elder brother"), transliteration (informal words are transliteration of English translation of formal words, e.g., 3Q ⟨sānqiu⟩ for 谢谢 ⟨xièxie⟩ "thank you"), and "others". Wang et al. (2013) also classified informal words in Chinese microblog posts similar to Li and Yarowsky (2008).

Methods for MA and LN In the last two decades, previous work has explored various rules and extraction methods for formal-informal word pairs to enhance Japanese MA and LN models for UGT. Nakamoto et al. (2000) proposed an alignment method based on string similarity between original and variant forms. Ikeda et al. (2010) automatically constructed normalization rules of peculiar expressions in blogs, based on frequency, edit distance, and estimated accuracy improvements. Sasano et al. (2013) defined derivation rules to recognize unknown onomatopoeia and variant forms of known words that frequently occur in webpages. Their rules were also implemented in a recent MA toolkit Juman++ (Tolmachev et al., 2020) to handle unknown words. Saito et al. (2014) estimated character-level alignment from manually annotated pairs of formal and informal words on Twitter. Saito et al. (2017) extracted formal-informal word pairs from unlabeled Twitter data based on semantic and phonetic similarity.

For English and Chinese, various classification methods for normalization of informal words (Li and Yarowsky, 2008; Wang et al., 2013; Han and Baldwin, 2011; Jin, 2015; van der Goot, 2019) have been developed based on, for example, string, phonetic, semantic similarity, or co-occurrence frequency. Qian et al. (2015) proposed a transition-based method with append(x), separate(x), and separate_and_substitute(x,y) operations for the joint word segmentation, POS tagging, and normalization of Chinese microblog text. Dekker and van der Goot (2020) automatically generated pseudo training data from English raw tweets using noise insertion operations to achieve comparable performance without manually annotated data to an existing LN system.

[19] Pinyin pronunciation is shown in "⟨⟩".

7 Conclusion

We presented a publicly available Japanese UGT corpus annotated with morphological and normalization information. Our corpus enables the performance comparison of existing and future systems and identifies the main remaining issues of MA and LN of UGT. Experiments on our corpus demonstrated the limited performance of the existing systems for non-general words and non-standard forms mainly caused by two types of difficult examples: complicated variant forms and unknown words of non-general vocabulary types.

In the future, we plan to (1) expand the corpus by further annotating of 5–10 times more sentences for a more precise evaluation and (2) develop a joint MA and LN method with high coverage.

Acknowledgments

We would like to thank the anonymous reviewers and area chairs for their constructive comments. We used the Balanced Corpus of Contemporary Written Japanese to construct our corpus.

References

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.

Kelly Dekker and Rob van der Goot. 2020. Synthetic data for English lexical normalization: How close can we get to manually annotated data? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6300–6309, Marseille, France. European Language Resources Association.

Yasuharu Den. 2009. A multi-purpose electronic dictionary for morphological analyzers [in Japanese]. *Journal of the Japanese Society for Artificial Inteligence*, 34(5):640–646.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics [in Japanese]. *Japanese Linguistics*, 22:101–123.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato, and Masaaki Nagata. 2011. Construction of a blog corpus with syntactic, anaphoric, and sentiment annotations [in Japanese]. *Journal of Natural Language Processing*, 18(2):175–201.

Kazushi Ikeda, Tadashi Yanagihara, Kazunori Matsumoto, and Yasuhiro Takishima. 2010. Automatic rule generation approach for morphological analysis of peculiar expressions on blog documents [in Japanese]. *IPSJ Transactions on Databases*, 3(3):68–77.

Ning Jin. 2015. NCSU-SAS-ning: Candidate generation and feature engineering for supervised lexical normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 87–92, Beijing, China. Association for Computational Linguistics.

Nobuhiro Kaji and Masaru Kitsuregawa. 2014. Accurate word segmentation and pos tagging for Japanese microblogs: Corpus annotation and joint modeling with lexical normalization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 99–109, Doha, Qatar. Association for Computational Linguistics.

Nobuhiro Kaji, Shinsuke Mori, Fumihiko Takahashi, Tetsuro Sasada, Itsumi Saito, Keigo Hattori, Yugo Murawaki, and Kei Utsumi. 2015. Kētaiso kaiseki no error bunseki (Error analysis of morphological analysis). In *Proceedings of the 21th Annual Meeting of the Association for Natural Language Processing Workshop "Error Analysis on Natural Language Processing"*, Kyoto, Japan.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Zhifei Li and David Yarowsky. 2008. Mining and modeling relations between formal and informal Chinese phrases from web corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1031–1040, Honolulu, Hawaii. Association for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Yasunari Nakamoto, Kazuya Mera, and Aizawa teruaki. 2000. Using sequence alignment to improve the morphological analysis [in Japanese]. *IPSJ SIG Technical Report*, 2000(11):87–93.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yutaka Hara. 2011a. Gendai nihongo kakikotoba kinkō corpus kētairon kitēshū dai 4 ban ge (Regulations of morphological information for balanced corpus of contemporary written Japanese 4th edition volume 2). *NINJAL Internal Reports*.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yutaka Hara. 2011b. Gendai nihongo kakikotoba kinkō corpus kētairon kitēshū dai 4 ban jō (Regulations of morphological information for balanced corpus of contemporary written Japanese 4th edition volume 1). *NINJAL Internal Reports*.

Ayaha Osaki, Yoshiaki Kitagawa, and Mamoru Komachi. 2017. Nihongo Twitter bunsho wo taishō to shita kēretsu labeling ni yoru hyōki sēkika (Text normalization by sequence labeling for Japanese Twitter documents). *IPSJ SIG Technical Report*, 2017-NL-231(12):1–6.

Tao Qian, Yue Zhang, Meishan Zhang, Yafeng Ren, and Donghong Ji. 2015. A transition-based model for joint segmentation, POS-tagging and normalization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1837–1846, Lisbon, Portugal. Association for Computational Linguistics.

Itsumi Saito, Kyosuke Nishida, Kugatsu Sadamitsu, Kuniko Saito, and Junji Tomita. 2017. Automatically extracting variant-normalization pairs for Japanese text normalization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 937–946, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Itsumi Saito, Kugatsu Sadamitsu, Hisako Asano, and Yoshihiro Matsuo. 2014. Morphological analysis for Japanese noisy text based on character-level and word-level normalization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1773–1782, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. 2013. A simple approach to unknown word processing in japanese morphological analysis. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 162–170, Nagoya, Japan. Asian Federation of Natural Language Processing.

Ryoichi Sato. 2009. *Todōfuken betsu zenkoku hōgen jiten (Dialect Dictionary of Japanese Prefectures)*. Sanseido.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2020. Design and structure of the Juman++ morphological analyzer toolkit. *Journal of Natural Language Processing*, 27(1):89–132.

Rob van der Goot. 2019. MoNoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206, Florence, Italy. Association for Computational Linguistics.

Rob van der Goot, Rik van Noord, and Gertjan van Noord. 2018. A taxonomy for in-depth evaluation of normalization for user generated content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

Aobo Wang, Min-Yen Kan, Daniel Andrade, Takashi Onishi, and Kai Ishikawa. 2013. Chinese informal word normalization: an experimental study. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 127–135, Nagoya, Japan. Asian Federation of Natural Language Processing.

Nakami Yamaguchi. 2002. *Giongo gitaigo jiten (Phonomime and Phenomime Dictionary)*. Kodansha.

Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 61–72, Seattle, Washington, USA. Association for Computational Linguistics.