Automatic Labeling of Voiced Consonants for Morphological Analysis of Modern Japanese Literature

Teruaki Oka†
teruaki-o@is.naist.jp

Mamoru Komachi†
komachi@is.naist.jp

Toshinobu Ogiso‡
togiso@ninjal.jp

Yuji Matsumoto†
matsu@is.naist.jp

†Nara Institute of Science and Technology
‡National Institute for Japanese Language and Linguistics

Abstract

Since the present-day Japanese use of voiced consonant mark had established in the Meiji Era, modern Japanese literary text written in the Meiji Era often lacks compulsory voiced consonant marks. This deteriorates the performance of morphological analyzers using ordinary dictionary. In this paper, we propose an approach for automatic labeling of voiced consonant marks for modern literary Japanese. We formulate the task into a binary classification problem. Our pointwise prediction method uses as its feature set only surface information about the surrounding character strings. As a consequence, training corpus is easy to obtain and maintain because we can exploit a partially annotated corpus for learning. We compared our proposed method as a preprocessing step for morphological analysis with a dictionary-based approach, and confirmed that pointwise prediction outperforms dictionary-based approach by a large margin.

1 Introduction

Recently, corpus-based approaches have been successfully adopted in the field of Japanese Linguistics. However, the central part of the fields has been occupied by historical research that uses ancient material, on which fundamental annotations are often not yet available.

Despite the limited annotated corpora, researchers have developed several morphological analysis dictionaries for past-day Japanese. National Institute for Japanese Language and Linguistics creates Kindai-bungo UniDic,[1] a morphological analysis dictionary for modern Japanese literary text,[2] which achieves high performance on analysis for existing electronic text (e.g. Aozora-bunko, an online digital library of freely available books and work mainly from out-of-copyright materials).

However, the performance of morphological analyzers using the dictionary deteriorates if the text is not normalized, because these dictionaries often lack orthographic variations such as Okuri-gana,[3] accompanying characters following Kanji stems in Japanese written words. This is problematic because not all historical texts are manually corrected with orthography, and it is time-consuming to annotate by hand. It is one of the major issues in applying NLP tools to Japanese Linguistics because ancient materials often contain a wide variety of orthographic variations.

For example, there is an issue of voiced consonant marks. Any "Hiragana" character and "Katakana" character (called Kana character altogether) represent either consonant (k, s, t, n, h, m, y, r, w) onset with vowel (a, i, u, e, o) nucleus or only the vowel (except for nasal codas N). Furthermore, the characters alone can not represent syllables beginning with a voiced consonant (g, z, d, b) in current orthography. They are spelled with Kana and a voiced consonant mark (゛) to the upper right (see Figure 1). However, confusingly, it was not ungrammatical to put down the character without the mark to represent voiced syllable

[1] http://www2.ninjal.ac.jp/lrc/index.php?UniDic

[2] In historical linguistics, the phrase "modern Japanese" refers to the language from 1600 on to the present in a broad sense. However, most Japanese people regard the phrase to the Meiji and Taisho Era; we also use the phrase to intend the narrower sense.

[3] In Japanese Literature, both Kana (phonogramic characters) and Kanji (ideographic characters) are used together. Generally, conjugated form is ambiguous, given the preceding Kanji characters. However, the character's pronunciation can also be written using Kana characters. Thus, the pronunciation's tailing few syllables are hanged out (Okuri), using Kana (gana) characters for disambiguating the form. Although the number of Okuri-gana is fixed for each Kanji character now, it was not fixed in the Meiji Era.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 292–300,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

ga	→	が:か (ka) + ゛	da	→	だ:た (ta) + ゛
gi	→	ぎ:き (ki) + ゛	di	→	ぢ:ち (ti) + ゛
gu	→	ぐ:く (ku) + ゛	du	→	づ:つ (tu) + ゛
ge	→	げ:け (ke) + ゛	de	→	で:て (te) + ゛
go	→	ご:こ (ko) + ゛	do	→	ど:と (to) + ゛
za	→	ざ:さ (sa) + ゛	ba	→	ば:は (ha) + ゛
zi	→	じ:し (si) + ゛	bi	→	び:ひ (hi) + ゛
zu	→	ず:す (su) + ゛	bu	→	ぶ:ふ (hu) + ゛
ze	→	ぜ:せ (se) + ゛	be	→	べ:へ (he) + ゛
zo	→	ぞ:そ (so) + ゛	bo	→	ぼ:ほ (ho) + ゛

Figure 1: Spelling syllables beginning with the voiced consonants g, z, d and b by Hiragana characters with a voiced consonant mark.

	Voiced	Voiceless
Marked	3,124	0
Ambiguous	4,032	29,332

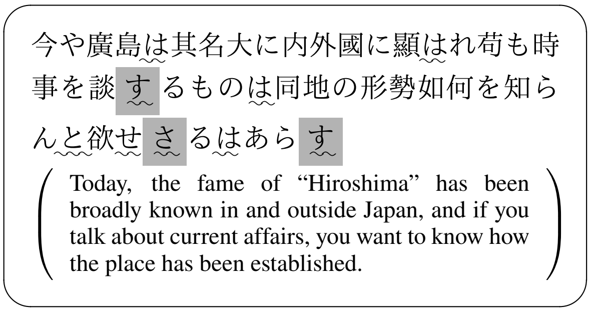
Table 1: The contingency table of observed frequencies of characters and voiceness.

until the Meiji Era, because Japanese orthography dates back to the Meiji Era. Consequently, modern Japanese literary text written in the Meiji Era often lacks compulsory voiced consonant marks. The mark was used only when the author deems it necessary to disambiguate; and it was not often used if one can infer from the context that the pronunciation is voiced.

Figure 2 shows characters which lack the voiced consonant mark even though we expect it to be marked in the text. Hereafter, we call such characters as “unmarked characters.” Also, we call the characters to which the voiced consonant mark can be attached as “ambiguous characters.” In Table 1, we present the statistics of the voiced consonants in “Kokumin-no-tomo” corpus which we will use for our evaluation. As you can see, 12% of the ambiguous characters are actually voiced but not marked. In addition, 44% of the voiced characters have the voiced consonant mark, showing the variation of using the voiced consonant mark in the corpus.

In the modern Japanese literary text, orthographic variations are not only the unmarked. However, unmarked characters appear a lot in the text and can be annotated easily by hand. Thus, we can get natural texts for evaluation of our method at low cost (in fact, it cost only a few weeks to annotate our above-mentioned test corpus). Therefore, we decided to begin with attaching voiced consonant mark for unmarked characters as a starting point for normalizing orthographic variations.

Basically, Kindai-bungo UniDic is created for a fully annotated sentence that does not include unmarked characters, and thus if the target sentence includes unmarked character(s), the performance can degrade considerably.



今や廣島は其名大に内外國に顯はれ荷も時
事を談するものは同地の形勢如何を知ら
んと欲せさるはあらす

(Today, the fame of “Hiroshima” has been broadly known in and outside Japan, and if you talk about current affairs, you want to know how the place has been established.)

Figure 2: Example of sentences that includes unmarked characters. This text is an excerpt from “The tide of Hiroshima”: Katsuichi Noguchi, Taiyo, No.2, p.64 (1925). Wavy-underlined characters are ambiguous character, and gray-boxed characters are unmarked character.

There are two major approaches to handle this problem: a dictionary-based approach and a classification-based approach.

First, the dictionary-based approach creates a dictionary that has both original spellings and modified variants without the mark. For example, Kindai-bungo UniDic includes both entries “ず (zu)” and “す (zu)” for frequent words such as “ず (zu)” in auxiliary verb. This allows morphological analysis algorithms to learn the weights of both entries all together from a corpus annotated with part-of-speech tags in order to select appropriate entries during decoding.

Second, the classification-based approach employs a corpus annotated with unmarked characters to learn a classifier that labels the voiced consonant mark for unmarked characters. Unlike the dictionary-based approach, the classification-based approach does not require part-of-speech tagged nor tokenized corpora. Since it is easier for human annotators to annotate unmarked characters than word boundaries and part-of-speech tags, we can obtain a large scale annotated corpus at low cost.

Therefore, in this paper, we propose a classification-based approach to automatic labeling of voiced consonant marks as a preprocessing step for morphological analysis for modern Japanese literary language.

We formulate the task of labeling voiced consonant marks into a binary classification problem. Our method uses as its feature set only surface information about the surrounding character strings with pointwise prediction, whose training data are available at low cost. We use an online learning method for learning large spelling variation from massive datasets rapidly and accurately. Thus, we can improve its performance easily by increasing amount of training data. In addition, we perform clustering of Kanji, which is abundant in the training data, and employ class n-grams for addressing the data sparseness problem. We compared our classification-based approach with the dictionary-based approach and showed that the classification-based method outperforms the dictionary-based method, especially in an out-of-domain setting. We also conducted an experiment to demonstrate that automatic labeling of unmarked characters as a pre-processing step improves the performance of morphological analysis of historical texts without normalization by a large margin, taking advantage of large scale annotated corpus of unmarked characters.

The rest of this paper is organized as follows: In section 2 we describe related work of automatic labeling of Japanese voiced consonant marks. Section 3 details our proposed classification-based method using pointwise prediction. We then explain experimental settings and results in section 4. Section 5 concludes our work and presents future work.

2 Related Work

If we assume an unmarked character as substitution error of one voiced consonant to one voiceless consonant, the task of detecting an unmarked character can be considered as a kind of error correction. In English, we can perform error correction for the one character's error by word-based approach. However, in Japanese, we cannot simply apply word-based approach because sentences are not segmented into words.

Nagata (1998) proposed a statistical method using dynamic programming for selecting the most likely word sequences from candidate word lattice estimated from observed characters in Japanese sentence. In this method, the product of the transition probability of words is used as a word segmentation model. However, most of the historical materials that we deal with are raw text, and there exist little, if any, annotated texts with words and part-of-speech tags. Thus, a word segmentation model learned from such a limited amount of data is unreliable. Unlike Nagata's method, our classification-based method does not rely on word segmentation and can exploit low-level annotation such as voiced consonant mark, which is available quite easily.

In addition, Nagata performed clustering of characters for smoothing confusion probability among characters to narrow down correction candidates. We also perform clustering on Kanji for addressing the data sparseness problem. Though Nagata uses character's shape for clustering, we instead use neighboring characters of the Kanji character. The intuition behind this is that whether to attach voiced consonant mark is affected by surrounding contexts, like sequential voicing.

On contrary, Shinnou (1999) proposed an error detection and correction method that does not perform word segmentation. He restricts the target to Hiragana characters and uses Hiragana n-gram that is a substring of the characters. In his method, error detection is determined by the Hiragana n-gram frequency. One counts each Hiragana n-gram frequency in training corpus and judges whether the string includes error by checking if the smallest frequency among them (minimum frequency of n-gram) is larger than a threshold value. After error detection, one enumerates candidate strings and corrects the input string to the string that has the largest minimum frequency of n-gram compared to other candidates.

The reason why Shinnou restricts targets to Hiragana characters is that it narrows down candidates of error correction. He used the fact that the number of Hiragana characters is 50 at most while the total number of distinct characters is more than 6,000 in Japanese. This method works well for present-day Japanese literary texts that contain relatively long Hiragana character strings. However, modern Japanese texts contain many Kanji characters and relatively short Hiragana character strings because modern Japanese texts are similar to Kanbun-kundokubun, or the Japanese reading of a Chinese text. Therefore, Hiragana n-grams fail to model error detection well for modern Japanese texts. Moreover, error correction of unmarked characters is much simpler than error correction of all the Hiragana. Our method differs from Shinnou's method in that we focus on automatic labeling of voiced consonant marks and employ a discriminative character n-gram model using a classification-based method. Although Shinnou's generative model is not capable of using overlapping features, our classification-based approach allows flexible feature design such as including character types that may help classification on unmarked characters. In addition, Shinnou's method requires a fully annotated corpus with unmarked characters even though there is a large amount of raw text in modern literary Japanese.

3 Detecting Unmarked Character with Pointwise Prediction

We formulate the task of automatic labeling of unmarked character into a binary-classification problem. More precisely, we build a binary classifier for detecting whether the target character is unmarked or not.

In our classifier, we use only surface information about one target character and its surrounding characters, and the classifier output is either unmarked (+1) or not (-1). Since proposed method does not require a corpus annotated with word boundaries or part-of-speech tags for learning, we take advantage of a large modern a Japanese corpus, Taiyo-Corpus,[4] which is based on Japanese magazines from the Meiji Era. This corpus is not annotated with neither word boundaries nor part-of-speech tags but is manually annotated with unmarked characters.

We employed pointwise prediction which makes a single independent decision at each point: ambiguous Hiragana character or Kunoji-ten[5].[6] Therefore, our method can learn from partially annotated corpora (Neubig and Mori, 2010) including raw corpora of modern Japanese literary text, and thus it is easy to obtain training data.

Neubig et al. (2011) extend the word segmentation method proposed by Sassano (2002) to Japanese morphological analysis using pointwise prediction. In our method, we adopt the binary features from (Sassano, 2002) to this task. Unlike Sassano and Neubig et al. who use an SVM, we use an online Passive-Aggressive algorithm for exploiting large datasets while achieving high accuracy.

3.1 Features for Classification

Our approach builds a binary classifier that uses binary features indicating whether the following n-grams exist or not (shown in Figure 3).

3.1.1 Character n-grams

These features correspond to character n-grams that surround the target character. Only characters within a window of three characters are used in classification ($n \leq 3$). These n-grams are referred with relative position from the target character.

If given sentence is $c_1 c_2 \cdots c_m$ and target character is c_i, character n-grams are $(-3/c_{i-3}c_{i-2}c_{i-1}$, $-2/c_{i-2}c_{i-1}c_i$, $-1/c_{i-1}c_i c_{i+1}$, $0/c_i c_{i+1}c_{i+2}$, $1/c_{i+1}c_{i+2}c_{i+3}$, $-3/c_{i-3}c_{i-2}$, $-2/c_{i-2}c_{i-1}$, $-1/c_{i-1}c_i$, $0/c_i c_{i+1}$, $1/c_{i+1}c_{i+2}$, $2/c_{i+2}c_{i+3}$, $-3/c_{i-3}$, $-2/c_{i-2}$, $-1/c_{i-1}$, $0/c_i$, $1/c_{i+1}$, $2/c_{i+2}$, $3/c_{i+3})$.

3.1.2 Character type n-grams

These features are similar to previously mentioned character n-grams with only the modification of replacing the character itself with the character type. We deal with eleven character types, Hiragana/H, Katakana/K, Kanji/C, Odoriji/O, Latin/L, Digit/D, dash/d, stop and comma/S, BOS ($\langle s \rangle$)/B, EOS ($\langle /s \rangle$)/E and others/o as the character types.

3.1.3 Markedness n-grams

These features are also similar to character n-grams with only the modification of replacing the character itself with 0 (voiced consonant mark cannot be attached), 1 (the mark can be attached) and 2 (it already has the mark).

3.2 Clustering on Kanji

In modern Japanese literary text, various Kanji characters were found commonly even in a sentence compared to nowadays. However, the frequency of each Kanji character varies. Learning tends to be sparse around a Kanji character that appears only several times in training corpus. For example, if "深" (deep) appeared only once in training corpus as in a word "深い" (is deep), then we will not be able to use the information "深" in a phrase "深ければ" (if it is deep) when we classify a character "は" in "深ければ."

[4] http://www2.ninjal.ac.jp/lrc/index.php?%C2%C0%CD%DB%A5%B3%A1%BC%A5%D1%A5%B9

[5] Kunoji-ten is a iteration mark, either "く" or "ぐ".

[6] Katakana characters had been used for specific words like adopted words and proper nouns. Thus, we excluded Katakana characters in this paper.

target character position
↓
-3 -2 -1 0 1 2 3
⟨s⟩彼邦に譲らざるへき大雑誌を發行せんと計畫したるも、⟨/s⟩
（ Though we planned to publish a big magazine that compares favorably with the one in that country, ）

Character 1-gram:	-3/ら	-2/ざ -2/さ	-1/る	0/へ	1/き	2/大 2/⟨B90⟩	3/雑 3/⟨B74⟩
Character 2-gram:	-3/らざ -3/らさ	-2/ざる -2/さる	-1/るへ	0/へき	1/き大 1/き ⟨B90⟩	2/大雑 2/⟨B90⟩⟨B74⟩	
Character 3-gram:	-3/らざる -3/らさる	-2/ざるへ -2/さるへ	-1/るへき	0/へき大 0/へき ⟨B90⟩	1/き大雑 1/き ⟨B90⟩⟨B74⟩		
Character type 1-gram:	-3/H	-2/H	-1/H	0/H	1/H	2/C	3/C
Character type 2-gram:	-3/HH	-2/HH	-1/HH	0/HH	1/HC	2/CC	
Character type 3-gram:	-3/HHH	-2/HHH	-1/HHH	0/HHC	1/HCC		
Markedness 1-gram:	-3/0	-2/2 -2/1	-1/0	0/1	1/1	2/0	3/0
Markedness 2-gram:	-3/02 -3/01	-2/20 -2/10	-1/01	0/11	1/10	2/00	
Markedness 3-gram:	-3/020 -3/010	-2/201 -2/101	-1/011	0/110	1/100		

Figure 3: Feature for classification of unmarked characters.

Therefore, we carry out clustering on Kanji characters and add character class n-gramin feature sets. For example, if “深” and “寒” (cold) belong to the same class X, and “寒” appears in training corpus as in a phrase “寒ければ” (if it is cold), then features corresponding to a phrase “X ければ” (if it is X) will be learned from “寒ければ.” As a result, we will be able to exploit “深” as evidence of detecting “は” in “深ければ” as unmarked character.

Clustering was performed on Kanji characters with the subsequent and the previous two characters individually based on (Pereira et al. 1993).

A Kanji character that appears left of the target character is replaced with the class of the former-clusters and that appears right is replaced with the class of the latter-clusters.

4 Experiments

We conducted two experiments for evaluating our method as follows.

4.1 Experimental Settings

We compare three approaches for automatic labeling of unmarked character as a pre-processing to morphological analysis on historical texts.

First, we built a naive generative model as baseline for labeling voiced consonant mark. This method labels voiced consonant marks that maximize the likelihood of a sentence by using a character 3-gram model. One deficiency of the baseline method is that it requires a fully annotated corpus with the marks.

Second, for the dictionary-based approach, we created a dictionary and corpus from the same training corpus used by the Kindai-bungo UniDic (U-Train) with all the marks removed. We preserved the original orthography in the field of each entry. We then trained a morphological analyzer[7] using the dictionary and corpus. Finally, we added to the dictionary entries with which we partially (or completely) replaced voiced consonant marks. This method assigns voiced consonant marks and performs morphological analysis jointly. However, it requires an annotated corpus with both the marks, word segmentation and part-of-speech tags, which are scarce to obtain.

Third, we constructed a proposed classifier from an annotated corpus with the voiced consonant marks. Our method does not need the information of word segmentation and part-of-speech. There-

[7] http://mecab.sourceforge.net/

Training corpus	positive	negative	all
U-Train	25,910	111,511	137,421
T-Train	208,097	966,308	1,174,405
K-Train	24,185	-	24,185

Table 2: Number of instances in each training corpus.

Test corpus	positive	negative	all
T-Eval	899	93,022	93,921
K-Eval	3,843	25,461	29,304

Table 3: Number of instances in each test corpus.

fore we can take advantage of Taiyo-Corpus. We use only articles written in a literary style in the corpus (398,077 sentences). We use 10% of this corpus for evaluation (T-Eval, including 33,847 sentences), and the rest for training (T-Train, including 364,230 sentences).

For evaluation, we prepared a modern Japanese magazine "Kokumin-no-Tomo" corpus (85,291 sentences). It is not annotated with word boundaries nor part-of-speech tags. From the corpus, we use four numbers for testing, No.10, 20, 30 and 36, which we had finished annotating voiced consonant mark at the time (K-Eval, including 10,587 sentences), and the rest for training (K-Train, including 74,704 sentences).

4.2 Preparing Training and Test Corpus

We extract training instances from all ambiguous characters. We regard instances with the mark as positive instances and instances without the mark as negative instances. Note that we detach voiced consonant mark from target character when extracting training instances. Although we extract test instances in a similar manner, we do not count characters originally with the mark at testing. In other words, we evaluate the accuracy only on unmarked characters present in real world setting. We show per instance breakdown of training and evaluation instances in Tables 2 and 3.

4.3 Tools

In this paper, we use an online Passive Aggressive algorithm, specifically PA-I for learning a binary classifier with (Yoshinaga et al. 2010).[8] We use a linear kernel and set the iteration number to 20. Also, we optimized the regularization parameter C by performing 10-fold cross-validation on the training corpus.

We performed clustering on Kanji with narrative sentences in training corpus. We used a clustering tool bayon[9] that implements the Repeated Bisection algorithm, which is a variant of the k-means algorithm. We use the product of probability of character bigram $P(char_1|char_{kanji})$ and trigram $P(char_2|char_{kanji}char_1)$ as distributions of two characters connecting to Kanji $P(char_1char_2|char_{kanji})$. Probabilities of character bigram and trigram are calculated by using the language modeling toolkit Palmkit.[10] We use Witten Bell smoothing. For computational efficiency, we replaced characters that are not Hiragana or Odori-ji with character type when creating the language model.

4.4 Experiment 1: intrinsic

In our first intrinsic experiment, we compared the precision, recall and F-measure of labeling voiced consonant mark with three approaches.

Table 4 presents the results of the intrinsic evaluation. The proposed method outperforms other methods in terms of precision and F-measure using the same training corpus. Moreover, by adding T-Train, the proposed method achieves the best performance in all evaluation metrics including recall. This is because our proposed method can benefit from a large-scale annotated corpus with voiced consonant marks, which is not possible for the dictionary-based method since it requires fully annotated corpus with words and part-of-speech tags. Although the baseline method can use corpora annotated with voiced consonant marks and achieves comparable performance to the proposed method regarding recall, its precision is inferior to the proposed method by a large margin. We suppose that this improvement comes from discriminative learning of the language model, which enables us to design flexible features. Generally, precisions are lower in T-Eval than in K-Eval over all methods. This is because T-Eval has relatively few positive instances and most of the instances are difficult to judge whether they are unmarked or not even for human.

In the baseline and the proposed method, performance is improved further by increasing amount of training data. By adding T-Train for U-Train, F-measure increases more than 10-points in T-Eval. We show in Figure 4 the change in recall

[8] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/

[9] http://code.google.com/p/bayon/

[10] http://palmkit.sourceforge.net/

	Training corpus	Number of Kanji class(k)	Test corpus					
			T-Eval			K-Eval		
			Prec.[%]	Rec.[%]	F	Prec.[%]	Rec.[%]	F
Baseline	U	-	35.085	86.203	49.872	91.276	91.344	91.310
	U+T	-	55.248	94.702	69.784	94.141	93.651	93.895
Dictionary-based	U	-	51.525	92.102	66.082	93.473	96.513	94.969
Proposed method	U	-	58.594	83.426	68.831	95.675	94.405	95.036
		50	64.061	83.871	72.640	96.235	93.781	94.992
		100	64.098	84.205	72.788	**96.401**	94.093	95.233
		500	60.430	84.427	70.441	95.982	94.483	95.227
		1000	59.745	83.537	69.666	95.718	94.223	94.965
	U+T	-	70.943	95.328	81.347	96.120	97.996	97.049
		50	71.993	95.217	81.992	96.073	98.048	97.050
		100	72.472	94.883	82.177	96.146	98.022	**97.075**
		500	71.704	94.994	81.722	96.120	97.996	97.049
		1000	**72.727**	95.216	**82.466**	96.288	97.866	97.071
	U+T+K	-	70.723	**95.661**	81.323	95.955	98.152	97.041
		50	72.236	95.216	82.149	95.953	98.100	97.015
		100	72.054	95.217	82.032	95.977	98.074	97.014
		500	71.836	95.328	81.931	95.883	**98.179**	97.017
		1000	71.956	95.328	82.009	96.001	98.074	97.026

Table 4: Performance of intrinsic evaluation: labeling voiced consonant mark.

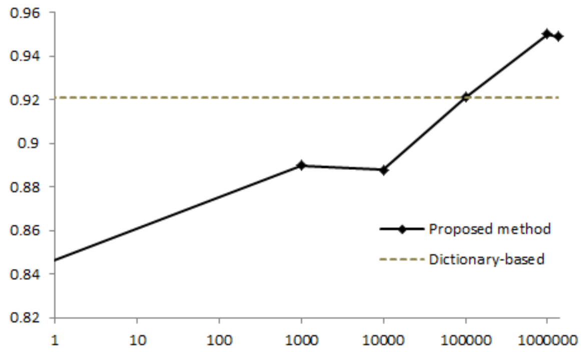


Figure 4: Improvement of recall with adding training instances.

when adding training instances from T-Train to U-Train in T-Eval (k=100). We confirmed that with just 1,000 instances added, recall increased 0.05 with the proposed method. Moreover, the proposed method's recall exceeded that of the dictionary-based approach after 100,000 instances were added. Although the F-measure was degraded by adding positive instances from K-Train, recall improved in K-Eval since positive instances add evidence for decision on voiced consonant marks. Apparently, it is effective to add instances from the same domain. However, the baseline and dictionary-based methods are not capable of using partially annotated corpora like K-Train. Our method employs pointwise prediction to make use of partially annotated corpus. Thus, we confirmed the effectiveness of using partially annotated corpora.

In addition, the proposed method shows the highest performance in k=1,000 for T-Eval and k=100 for K-Eval, respectively, when learned on T-Train and U-Train. In all settings, clustering improves precision while recall sometimes deteriorates. The performance gain is even larger when training data is scarce (See the results of U-Train). From this fact, we confirmed the effectiveness of clustering on Kanji for addressing the data sparseness problem.

Table 5 lists our features and their performance. Because the performance of detection degrades drastically when we subtract Character n-gram from All, this feature is crucial for determining unmarked characters. This is another piece of evidence that discriminative language model works quit well for this task. On the other hand, both Character type n-gram and Markedness n-gram contribute to improvement of precision. As a result, F-measure increases using those features.

We also investigated errors of the classification on our method. Although we found some errors which due to lack of training data, we found errors which are difficult to determine without discourse context, like “か”(ka) of binding particle or auxiliary verb and “が”(ga) of case-marking particle or auxiliary verb. However, these instances are difficult even for human to determine whether unmarked or not. Since the basic policy is to use the mark when there is ambiguity, the absence of the mark in an ambiguous case can be considered as evidence of non-unmarked character. Moreover,

		T-Eval/K-Eval	
Feature	Prec.[%]	Rec.[%]	F
Character n-gram only	70.041/95.882	**95.439**/98.152	80.791/97.004
All − Character n-gram	2.521/20.000	1.001/ 0.156	1.433/0.310
All − Character Type n-gram	70.651/96.028	95.328/98.126	81.115/97.066
All − Markedness n-gram	69.764/95.884	95.217/**98.205**	80.527/97.031
All	**72.472/96.146**	94.883/98.022	**82.177/97.075**

Table 5: Performance of each feature and their combination.

our method can not refer the discourse information since we only employed local context of character n-grams. Therefore, our method excessively tend to classify characters into unmarked. On the other hand, we found instances for which both unmarked and marked form are acceptable, like “結び”(tie) and “結ひ”(tie). Note that “結び” and “結ひ” are pronounced differently as “musubi” and “yui,” respectively. These instances seem to be the cause of degradation of precisions in T-Eval. For Odori-ji, it tends to fail classification because they not only depend on information of previous consonants but also on common practice such as “かえすぐ(がえす)”(again and again).

4.5 Experiment 2: extrinsic

As a second extrinsic experiment, we investigated how effective these approaches are at improving accuracy of morphological analysis.

To create gold-standard annotation for morphological analysis, we take the result of morphological analysis for the corpus annotated with voiced consonant marks using the standard version of Kindai-bungo UniDic. Since the word and part-of-speech information are not available in Taiyo and Kokumin-no-Tomo corpus, this constitutes the upper bound of the morphological analysis performance on these data.

We evaluated the result of morphological analysis for two methods. First, we tested the dictionary-based method by performing morphological analysis using the same Kindai-bungo Unidic with additional entries that partially (or all) without voiced consonant marks as we described in section 4.1. Second, we evaluated the proposed method by pre-processing the unlabeled test corpus with the proposed method and performing morphological analysis using the standard version of Kindai-bungo Unidic. Then, we calculated the agreement rate between each method and the gold standard by counting how many sentences are identical to the gold standard. We compared each word's parts-of-speech tags and lexemes for the comparison.

	Taiyo	Kokumin-no-Tomo
Dictionary-based	91.479 [%]	88.968 [%]
Proposed method	99.016 [%]	96.647 [%]

Table 6: Performance of extrinsic evaluation: agreement rate of morphological analysis result.

Table 6 shows the results of the extrinsic evaluation. As you can see, the proposed method gives higher agreement with the gold standard in morphological analysis results than the dictionary-based approach, thanks to the large scale Taiyo corpus annotated with voiced consonant marks. In these experiments, we confirmed that pre-processing with the proposed method is effective for improving morphological analysis of unnormalized modern Japanese literary text.

5 Conclusion

In this paper, we proposed a pointwise approach to label voiced consonant marks for modern Japanese literary text. We confirmed that pointwise prediction outperforms the dictionary-based approach by a large margin. By using the proposed method as pre-processing, morphological analysis results become much closer to the gold standard than using the dictionary-based approach.

Also, we are using the method for annotating the modern Japanese literature. Thanks to the method, we are able to accelerate manual annotation with considerably small effort.

One limitation is that we only deal with unmarked characters in this work. In modern Japanese literary text, there are other orthographic variations such as Okuri-gana and Kana-usage as well. As our future work, we will work on normalizing these variations for improving accuracy of morphological analysis.

We hope this work will encourage further investigation into historical work.

References

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional Clustering of English Words. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (ACL-93)*:183-190.

George Karypis. 2003. CLUTO A Clustering Toolkit. *University of Minnesota, Department of Computer Science, Technical Report* #02-017.

Graham Neubig, Shinsuke Mori. 2010. Word-based Partial Annotation for Efficient Corpus Construction. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*:2723-2727.

Graham Neubig, Yosuke Nakata, Shinsuke Mori. 2011. Pointwise Predication for Robust, Adaptable Japanese Morphological Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*:529-533.

Hiroyuki Shinnou. 1999. Detecting and Correction for Errors in Hiragana Sequences by a Hiragana Character N-gram. *Journal of Information Processing Society of Japan*,40(6):2690-2698.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research 7*:551-585.

Manabu Sassano. 2002. An Empirical Study of Active Learning with Support Vector Machines for Japanese Word Segmentation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*:505-512.

Masaaki Nagata. 1998. A Japanese OCR Error Correction Method Using Character Shape Similarity and Statistical Language Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistic (COLING-ACL '98)*:922-928.

Naoki Yoshinaga and Masaru Kitsuregawa. 2010. Kernel Slicing: Scalable Online Training with Conjunctive Features. In *Proceedings of the 23th International Conference on Computational Linguistic (COLING 2010)*:1245-1253.