Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

機械学習による近代文語文への濁点の自動付与

岡 照晃[†] 小町 守[†] 小木曽 智信[††] 松本 裕治[†]

現代日本語のように，濁音を仮名で表記する際に必ず濁点を用いる習慣が定着したのは明治時代以降のことで，明治期の文献の中では濁音が期待される文字に濁点のない濁点無表記の場合が多い．本論文では，濁点無表記の濁音仮名文字を識別し，自動で濁点を補う手法について述べる．我々は，判定点の文字が濁点無表記文字か否かを決定する 2 値分類問題として定式化を行った．提案手法では，周辺文字列の情報のみを用いて点推定を行う．オンライン学習を採用し，大規模な『太陽コーパス』から学習を行なった．これにより提案手法は，『国民之友』において 96.016%の精度と 98.283%の再現率を達成した．

A Machine Learning Approach to Automatic Labeling of Voiced Consonants for Modern Japanese Literary Text

Teruaki Oka,[†] Mamoru Komachi,[†] Toshinobu Ogiso[††] and Yuji Matsumoto[†]

The present-day Japanese use of voiced consonant mark had established in Meiji Era. Thus, modern Japanese literary text written in Meiji Era often lacks compulsory voiced consonant marks. In this paper, we propose an approach to automatic labeling of voiced consonants for modern Japanese literary language. We formulate the task of labeling voiced consonants into binary classification problem. Our method uses as its feature set only surface information about the surrounding character strings with pointwise prediction. We use an online learning method for exploiting large datasets from Taiyo Corpus. We achieve 96.016% precision and 98.283% recall on the Kokumin_no_tomo Corpus.

1. はじめに

近年，日本語学，国語学の分野においてもコーパスを利用した研究が増えつつある．しかしながら，これらの分野で大きな位置を占めているのは古い時代の資料を扱う歴史的研究であり，そういった歴史的資料はコーパスとして整備がそれほど進んでいないのが現状である．その原因の 1 つとして，歴史的資料の中の日本語が現代のそれとは異なっていて，ある程度の知識を有した人でなければ読解が難しいことが挙げられる．

例えば現代日本語のように，濁音を仮名で表記する際に必ず濁点を用いる習慣が定着したのは明治時代以降のことで，その普及過程に当たる明治期，もしくはそれ以前に書かれた文献の中では，濁音が期待される文字に濁点のない**濁点無表記**の場合が多い．表記において濁点が使われるのは必要と判断された場合のみで，前後の文脈から類推できる場合には使われないことが多かった．図 1 に示した例文中において，下線を引いた箇所はいずれも濁音表記が期待されるにもかかわらず，濁点の付いていない文字である．図 1 のような表記をそのままテキストデータ化しただけではコーパスを利用するユーザの検索性が損なわれる可能性がある．そのためコーパスを整備する際には，濁点無表記の濁音仮名であることをコーパス中に明示しておく必要があるが，これを慣れない者が確実に行うことは難しく，さらに濁点が付く可能性のある仮名文字は 1 つの資料中に大量に含まれているため，その全てを網羅的に確認するだけでも大変な労力を必要とする．

また，こういった表記のあり方は辞書ベースでの形態素解析の適用も難しくしている．すなわち，濁音の正書法が未確立で，濁点があまりに無規則な使われ方をしていたため，それら全てをカバーしようとすると，形態素解析辞書にあらゆる濁点の脱落パターンを網羅させておく必要がある．実際，近代文語文を対象とした形態素解析辞書である近代文語 UniDic[13)]では，助動詞「ず」のような活用語も含めて，無濁点の場合の書字形も登録が行なわれている．しかし，この方法では辞書のサイズが大きくなり，解析結果の候補も増えるため，解析の精度を悪化させる原因になり得る．そのため濁点無表記の文字には，解析の前処理の段階で予め濁点が付与されることが望ましい．

本論文では，コーパス整備時における人手の負担を最小限に抑えるため，2 値分

† 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
†† 国立国語研究所
National Institute for Japanese Language and Linguistics

Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

今や廣島は其名大に内外國に顯はれ苟も時事を談す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是れ征清の大師一たひ海に航せしより　大元帥陛下大纛を此に駐め大本營となし軍務を親裁し玉ふに因てなり先つ其大勢より叙述して次第に細事に及はんとす

（1895 年 2 号「広島の形勢」野口勝一）

図 1　濁点無表記の濁音仮名の例

類器を用いて，濁点無表記の濁音仮名[a]文字（以下，**濁点無表記文字**）を識別し，自動で濁点を補う手法について述べる．提案手法は，明治期以前のように単語のアノテーションされた資料が十分に揃っていないような時代の文に対する処理としても利用可能にすること，そして形態素解析の前処理として導入するため単語の情報は使えないことをそれぞれ考慮して，点予測による文字単位での処理を実施する．具体的には周辺文字列の情報を用いて，判定点の文字が濁点無表記文字か否かを決定する 2 値分類問題として定式化を行った．分類器の学習には，明治期の書き言葉の多様性を考慮し，できるだけ多くの訓練事例を用いて，かつそれを高速に精度よく学習させるため，オンライン学習を採用した．また，ゼロ頻度問題による学習の不足を補うべく，訓練テキスト中の文字の大半を占める漢字をクラスタリングし，クラス n-gram を素性に利用することで，分類性能の向上を図った．太陽コーパス[8]から得られた約 180 万個の事例で学習を行い，現在整備中の近代語資料の約 3 万個の事例について濁点無表記文字の検出実験を行なった結果，精度 96.016%，再現率 98.283%を達成した．また比較のために，濁点の付き得る全ての仮名文字を濁点無表記文字として検出する手法をベースラインとして設け，結果，提案手法はベースライン手法に比べて F 値において大いに性能が向上しており，本手法の有効性を確認した．

2. 関連研究

濁点無表記文字を検出して濁点を補うというタスクは，濁点無表記文字＝濁音仮名文字が清音仮名文字に置換している，と考えると，誤り検出，誤り訂正の一種だと見なすことができる．

文字レベルでの誤り検出･訂正手法では，英語の単語つづり誤りの検出と訂正についての試みが古くから行われており，代表的なものに雑音のある通信路モデルの考え方に基づいて誤りを訂正する手法 [10]がある．この手法では，1 単語を対象に，雑音（挿入，欠落，置換，転置）によって元々の文字列 X が x として観測されてしまったとき，観測文字列 x から本来の文字列 X を推測する問題として定式化が行われている．これは，事後確率 $P(X \mid x)$ を最大にする $\hat{X}$ を選択する問題として考えることができ，ベイズの定理により，

$$P(X \mid x) = \frac{P(x \mid X)P(X)}{P(x)} \tag{1}$$

で，分母は X の最大値の導出に無関係なので，

$$\hat{X} = \arg\max_X P(x \mid X)P(X) \tag{2}$$

となる．ここで，$P(X)$ は言語モデル，$P(x \mid X)$ は混同モデルと呼ばれる．言語モデルは単語 X の出現確率等から求められるが，混同モデルは，各単語について，大量のつづり誤り例を観測して求めなくてはならないため，非常にスパースになりやすい．そこで，文献 10)では，個々の誤りを独立と見なすことで近似的な値を得ている．x を X に対する 1 文字の挿入，欠落，置換，転置によって生じる文字列に限定することで，例えば，本来の文字列 access が，acress として観測される（1 文字の置換）確率は，

$$P(\text{"}acress\text{"} \mid \text{"}access\text{"}) = \frac{\text{文字“c”の文字“r”への置換誤り頻度}}{\text{文字“c”の出現頻度}}$$

として与えられる．

この手法は,日本語に対しても適用可能であるが,次のような 2 つの問題がある．まず，日本語は分かち書きがされておらず，単語の同定が容易ではない．また，異なり文字総数が英語と比べてはるかに多く，1 つの単語の長さも短い．そのため，1 文字違いの単語が多く存在し，簡単には訂正単語の候補を絞り込むことができない．

これに対し永田は，1 文を対象に，観測文字列より推測される本来の単語列の候補の中から，動的計画法を使って最尤単語列を選択する方法を提案している．[11]この手法では，言語モデルとして単語列の同時確率が用いられ，これは単語の連接確率の積で近似される．また，訂正単語候補のしぼり込みでは，単語の出現頻度と文字間での混同確率を用いた候補のランク付けを行なって，上位に現れたものを取得する．ただし日本語の場合は，任意の文字間の混同確率を求める場合にも非常にスパースになりやすい．そこで，文字を文字特徴ベクトルでクラスタリングし，図形的に似た文字のクラスを作ることによって，文字間の混同確率の代わりに文字クラス間の混同確率を導入して，スムージングを行なっている．

永田の手法は，単語あるいはそれより上位の情報までを参照して行われる．しかしながら，基本的に歴史的資料の多くは生のテキストのままで，アノテーションされたデータはないか，非常に少ないのが普通である．そのため，処理は文字単位で

a “仮名”と述べたが，実際に対象としているのは平仮名（踊字を含む）＋くの字点である．外来語や固有名詞等の限られた語の表記にしか用いられていない片仮名は対象外とした．

Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

行なえることが望ましい．また，以上で挙げた手法はいずれも，誤りの位置や種類を限定しないものであったが，濁点無表記文字の検出においては，誤りは 1 文字の置換に限られ，かつ濁点の付き得る文字に限定して生じている．

新納は対象を文中の平仮名列に限定し，その部分文字列である平仮名（文字）n-gram を用いて誤りの検出と訂正を行う方法を提案している．[15]この手法では，誤りの検出を平仮名 n-gram の頻度に基づいて行なっており，与えられた平仮名文字列中に存在する平仮名 n-gram それぞれの訓練コーパスでの出現頻度を求めて，その中の最小値（n-gram 最小頻度）が，設定されたしきい値以下ならば，その平仮名列には誤りが含まれていると判定する．また訂正では，欠落，挿入，置換，転置によって観測平仮名列が得られるような訂正候補を列挙し，その中から n-gram 最小頻度が最大となる平仮名列へと訂正を実施する．新納の手法は，比較的 1 文内での平仮名の割合が大きい現代日本語書き言葉を対象としているが，近代文語文は，漢文訓読文に近く，1 文を構成する文字も時代が古くなるほどに漢字の占める割合が大きくなるため，我々の問題設定とは異なる．

そこで本論文では，濁点の付き得る仮名文字だけを対象として，点予測を用いた検出を試みる．点予測とは，分類時の素性として，周囲の単語境界や品詞情報等を参照せずに，周辺の文字列の情報のみを参照する方法である．例えば，Neubig らは Sassano の Support Vector Machine（SVM）を用いた単語分割手法 [9]を発展させ，点予測に基づく日本語の形態素解析 [4]を行なっている．提案手法では，Sassano が単語分割に用いた 2 値素性を拡張し，当該文字が濁点無表記文字か否かの分類に用いることにした．また，Sassano や Neubig らが線形分類器を用いているのに対し，提案手法では，大量の事例を高速かつ精度よく分類するために，オンライン学習手法である Passive Aggressive algolithm[7]に多項式カーネルを適用した．[12]

提案手法においても永田の手法と同じように漢字をクラスタリングするが，永田が OCR 誤り訂正のタスクのため，文字形状の類似性でクラス分けを行い，混同確率のスムージングに用いていたのに対して，濁点の自動付与タスクでは連濁など，周辺文脈の影響を受けるため，提案手法では，漢字をそれに接続する文字の傾向でクラス分けすることで，分類器の学習のスパース性解消を試みた．

3. 点予測を用いた濁点無表記文字の検出

本論文では，近代文語文を対象に，濁点無表記文字を点予測によって検出し，自動的に濁点を補う方法について述べる．具体的には，濁点が付与され得る全ての仮名文字（平仮名＋くの字点）について，当該文字とその前後の文字列を入力し，濁点無表記文字か否かを出力する 2 値分類問題として定式化を行う．

判定点
↓
i-3 i-2 i-1 i i+1 i+2 i+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>

文字1-gram:	-3/ら	-2/ざ -2/さ	-1/る	0/へ	1/き	2/大 2/<B90>	3/雑 3/<B74>
文字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>	
文字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>		
文字種1-gram:	-3/H	-2/H	-1/H	0/H	1/H	2/C	3/C
文字種2-gram:	-3/HH	-2/HH	-1/HH	0/HH	1/HC	2/CC	
文字種3-gram:	-3/HHH	-2/HHH	-1/HHH	0/HHC	1/HCC		
濁音化可能性1-gram:	-3/0	-2/2 -2/1	-1/0	0/1	1/1	2/0	3/0
濁音化可能性2-gram:	-3/02 -3/01	-2/20 -2/10	-1/01	0/11	1/10	2/00	
濁音化可能性3-gram:	-3/020 -3/010	-2/201 -2/101	-1/011	0/110	1/100		

図 2　濁点付与に使用する素性

3.1 分類に用いる素性

点予測による濁点無表記文字の検出では，以下の3種類の素性（出現したか否かの2値素性）を参照する2値分類器によって分類を行なっている（図2参照）．

（1）**文字n-gram**：　1文内における当該仮名文字位置iから前後 3 文字の範囲中に存在する全ての部分文字列（最大長さ3）．iからの相対位置を併記して用いる．

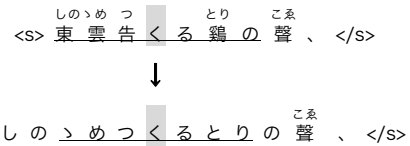
（2）**文字種n-gram**：　各文字を文字種に変換した列を対象とする以外は文字n-gramと同じである．文字種は，平仮名（H）,片仮名（K）,漢字（C），英字（L），数字（D），くの字点（V）,句点（S），読点（c），文頭<s>（s），文末</s>（/s），その他（O）の全11種である．

（3）**濁音化可能性n-gram**：　各文字を，濁点を付けることができない文字（0），濁点を付けることができる文字（1），既に濁点が付けられている文字（2）に置換した列を対象とする以外は文字n-gramと同じである．

1つの事例を作る際には，訓練テキストの地の文だけでなく，参照する前後3文字の間に濁点の付いた仮名文字がある場合には，そこから再現できる濁点の一部～全全てが脱落した文字列のパターン全てと（ex. 図2では例えば位置-3の文字3-gram

Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

(a) ルビの展開：



(b) 踊字の展開：

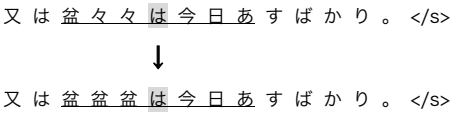


図 3 ルビと踊り字の展開

として「らざる」と「らさる」の2つを参照している)，当該仮名文字の前後の漢字のルビを開いた文字列[b]（図3a）と前後の踊字を開いた文字列（図3b）を同時に参照する．ただし簡単のため，ルビが振られた地の文の文字列に，濁点の付き得る平仮名もしくは，くの字点が含まれる場合はルビを開くことはせず，また展開できる踊字は唯一濁点の付き得ない「々」のみとした．また，踊字にルビが振られていることもあるため，踊字の展開とルビの展開は別々に行うこととした．

3.1 漢字のクラスタリング

近代の総合雑誌を基に作られた太陽コーパス（文語文記事）を構成する文字の内訳を調査した結果，表1が得られた．近代文語文は文体が訓読文に近いため，その大半を占めているのが漢字である．しかし，各漢字ごとの出現頻度は様々であり，数回程度しか出現しなかったような漢字については，その使われ方が網羅的に現れず，その漢字の周りで学習がスパースになってしまう．例えば，「深」という漢字が訓練テキスト中に１度しか現れず，それが「深い」というフレーズであったとすると，「深けれは」というフレーズの「は」が濁点無表記文字か否かの判定には「深」を有効に活用することができない．そこで，漢字をクラスタリングすることで，文字n-gramからクラスn-gram[6]を作成し，素性に追加した．これにより，例えば「深」と「寒」が同じ漢字クラスXに属して，かつ「寒」が「寒ければ」というフレーズで訓練テキスト中に表れていたとすると，「Xければ」というフレーズは「寒ければ」で学習されているため，「深ければ」という事例がコーパス中に存在しなくても，間接的に「深」を「濁点無表記文字である」ことを優位にする材料として活用できる．

クラスタリングは文献2)の手法を参考に，ある１つの漢字（$char_{kanji}$）に接続する2文字（$char_1 char_2$）の分布 $P(char_1 char_2 \mid char_{kanji})$ を用いて行なった．漢字に後続する2文字の分布と，前方に現れる2文字の分布についてそれぞれクラスタリングを行い，判定点からの相対位置が負の漢字に前者，正の漢字に後者をそれぞれ使用した．

b 近代文語文は，所謂漢文訓読文に近く，ルビが多用され，時には地の文がほぼ総ルビになっているような場合もある．

表 1 太陽コーパス文語文記事内文字内訳

平仮名	76
片仮名	93
漢字	6,807
英数字	89
その他	83
計	7,148

4. 評価実験

提案手法の有効性を評価するため，現在整備中の近代語資料を用いて濁点無表記文字の検出実験を行い，濁点の自動付与を実施した．

4.1 太陽コーパス

太陽コーパス[8]は，近代の総合雑誌『太陽』(1895～1928）を対象として作成された大規模な構造化テキストタグ付きコーパスである．記事の分量，ジャンル，カバーする年代の広さにおいて十分な量を備えており，また，できるだけ原文の情報を残すように整備されているため，タグを使って漢字のルビや踊り字を開いた表記の情報なども併記されている．濁点無表記文字も濁点付きに人手で修正されており，これにも原文では濁点が付いていなかったことを示すタグが併記されている．一定しない仮名遣いなど近代書き言葉の多様性を考慮すると，分類器の学習にはなるべくルビ等の原文上の情報を多く使用できることが望ましく，本論文では太陽コーパスを分類器の学習に利用する．

近代語の資料は現在，少しずつ整備が進められてはいるが，冒頭で述べたような理由から，濁点無表記文字のアノテーションでさえそれほど進んでいない．そのため，既に整備を終えた太陽コーパスを元に，機械学習等の統計的な手法を用いて自動でさらに大量のコーパスを整備できることが望まれている．以下の実験では，現在整備中の近代語資料を評価に用意して，太陽コーパスでの学習の有効性を確認する．

Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

表 2　事例数内訳

	正例数	負例数	計
訓練事例（太陽コーパス）	355,195	1,519,323	1,874,518
評価事例（国民之友）	3,843	25,451	29,294

4.2 実験に使用したデータ

実験には，分類器の学習に，予め濁点無表記文字に濁点が補われている太陽コーパス文語文記事（全538,944文）を用いた．訓練事例の獲得は，テキスト中の濁点のついた濁音仮名文字と濁点の付き得る清音仮名文字全てから行い，濁点の付いた仮名文字を判定点として正例（ただし，事例作成直前に判定点の仮名文字からのみ濁点を外しておく），濁点の付き得る清音仮名文字を判定点として負例をそれぞれ取得した．また評価には，現在整備が進められている途中で濁点無表記文字がそのまま残されている明治期の総合雑誌『国民之友』（1887年10号，1888年20号，1888年30号，1888年36号）の文語文記事（全9,818文）を用いた．評価事例の獲得は，テキスト中の濁点の付き得る仮名文字全てから行い，濁点無表記文字を判定点として正例，濁点を付けることが可能な清音仮名文字を判定点として負例をそれぞれ取得した．

取得した訓練事例と評価事例の内訳を表2に示す．

4.3 濁点無表記文字の検出性能評価

4.3.1 実験概要

自然言語処理の分野でよく用いられる 2 値分類器に Support Vector Machine (SVM)があるが，今回のように訓練事例数が多い場合には，計算時間が非常に大きくなってしまう．[c]また，近代書き言葉の多様性を考慮すると，使用可能な事例は可能な限り全て使用できることが望ましい．そこで，本論文では 2 値分類器に，オンライン学習アルゴリズムである Passive Aggressive algorithm（PA）[7]を採用し，実際には PA-I に 2 次の多項式カーネルを使用することで，組み合わせ素性を考慮した．学習時の反復回数は 20 回に設定した．実験のための PA-I を実装したツールしては，opal[d]を利用した．

漢字のクラスタリングは，太陽コーパス文語文記事の地の文を用いて行なった．クラスタリングにはRepeated Bisection[3]を実装したツールbayon[e]を使用した．また，

1	又
18	劈 焚 撒 繙 叩 描 蒔 捌 措 湧 吐 惹 溶 牽 欺 掻 飽 閃 咲 届
	赴 曳 呟 儲 欠 轟 傾
20	具 換 答 備 携 稱 傳 衰 考 加 對 會
58	噴 淺 稠 永 濃 鋭 短 幼 青 赤 瞬 暗 辛 低 乾 完 脆 勘 臭 厚
	宜 寒 遅 細 遠 淡 輕 多 凄 深 若 高 暫 強 斯 早 無 訊 薄
68	萠 肥 殖 越 消 燃 癒 潰 沸 絶 禁

図 4　漢字のクラスタリング結果（クラス数 100, 漢字に後続する文字の分布でクラスタリング）

表 3　カーネルと漢字クラスの比較実験結果

	カーネル	漢字クラス数[f]	正則化パラメータ	精度[%]	再現率[%]	F 値
Baseline	-	-	-	13.119	100.00	0.23195
提案手法 (PA-I)	線形	-	0.0016	95.400	98.230	0.96794
		50	0.00155	95.593	98.205	0.96881
		100	0.003	95.718	98.309	0.96996
		500	0.0015	95.762	98.205	0.96968
		1000	0.0019	95.547	98.257	0.96883
	2 次の多項式	-	0.00003	96.084	98.335	0.96682
		50	0.000012	95.903	98.074	0.96976
		100	0.00003	96.016	**98.465**	**0.97225**
		500	0.00003	96.157	98.309	0.97221
		1000	0.00002	**96.179**	98.257	0.97207

分布 $P(char_1 char_2 \mid char_{kanji})$は，統計的言語モデル作成ツールPalmkit[g]を使用し，文字単位のbigram確率 $P(char_1 \mid char_{kanji})$とtrigram確率 $P(char_2 \mid char_{kanji} char_1)$をそれぞれ求め，式(3)のように計算した．スムージングにはWitten Bellを使用した．

$$P(char_1 char_2 \mid char_{kanji}) = P(char_1 \mid char_{kanji}) P(char_2 \mid char_{kanji} char_1) \quad (3)$$

ここで，$char_1 char_2$の取り得る値は（$char_i (i = 1,2)$の取り得る値が太陽コーパス文語文記事中に現れる異なり文字総数通りあることから，）7148^2通り存在する．そのため，Palmkitでn-gram モデルを作る際には予め，平仮名，くの字点以外の文字を素性の文字種n-gramで用いた文字種へと置き換えておくことをし，クラスタリン

c 線形カーネルに関しては様々な高速化手法が提案されている．[1]しかし，自然言語処理で一般に使用される素性間の関係を考慮した組み合わせ素性を使うには，それらを明示的に展開せねばならない．

d http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal

e http://code.google.com/p/bayon

f 漢字クラス数は，例えば 50 のときは前方に表れる文字の傾向で 50 個のクラスに分割し，後続する文字の傾向でも 50 個のクラスに分割したことを表している．

g http://palmkit.sourceforge.net

Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

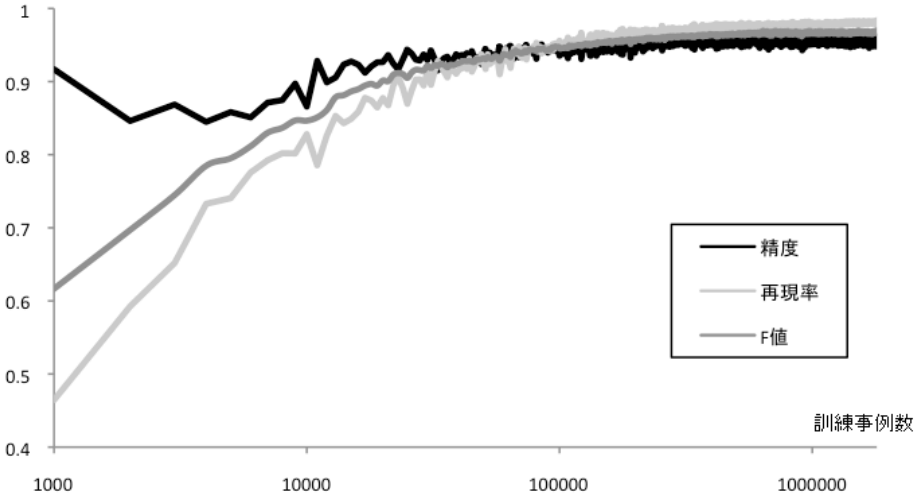


図 5　訓練事例数に対する検出性能

グに使用する分布ベクトルの次元数を制限した．[h]漢字のクラスタリングを行なった結果の一例を図4に示す．図4は後続する文字列の傾向でクラス分けを行なった結果のため，活用が似通った漢字が１まとめにされている一方で，漢字クラス１の「又」のように，直後に平仮名よりも「、(読点)」の続くような漢字も考慮してクラスタリングが行えている．

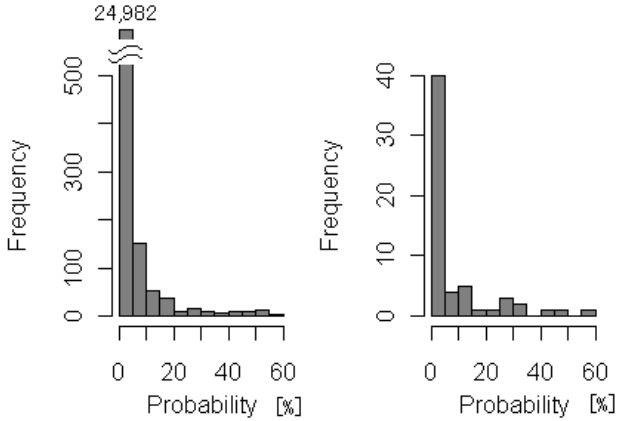
比較のために，濁点の付き得る全ての仮名文字を濁点無表記文字として検出する手法をベースラインに設定した．

評価は，濁点無表記文字の検出に対する精度と再現率，及び，それらの調和平均である F 値により行なった．精度，再現率はそれぞれ以下を測定した．

$$\text{精度} = \frac{\text{正しく検出された濁点無表記文字数}}{\text{検出された濁点無表記文字数}}$$

$$\text{再現率} = \frac{\text{正しく検出された濁点無表記文字数}}{\text{正解の濁点無表記文字数}}$$

h ある漢字の後に続く漢字の傾向や前方に現れる英数字の傾向などでクラスタリングをしたとしても，その傾向は素性を作る際の文字列中には現れない．そのため，そのようにして作られたクラスは分類の際の判定点に当たる仮名文字が濁点無表記か否かを決めるのにはさほど役立たない．そこで，ここでは判定点にならないような，文字自体の詳細は省き，大雑把な文字種に束ねることで，次元を削減している．



(a)清音仮名と正しく分類された事例　(b)誤って清音仮名に分類された事例

図 6　分類結果別確率頻度

4.3.1 実験結果

提案手法ならびにベースラインの精度，適合率及び F 値を表 3 に示す．提案手法は，2 次の多項式カーネルを用いて，漢字を 100 個のクラスに分けた場合において F 値が最大となり，精度で 96.016%，再現率で 98.465%を達成した．これによって，太陽コーパスでの学習が別の雑誌資料である国民之友の濁点無表記の検出にも有効であることが確認できた．

検出の性能は 2 次の多項式カーネルを用いた方が線形よりも若干高くなっている．これは，組み合わせ素性が考慮されることで，例えば，漢字クラス導入によるスムージングの結果，精度が低下してしまうのを“元の漢字+漢字クラス”という素性で防ぐことができたためだと考えられる．また，F 値の比較において，提案手法はベースラインを大きく上回っており，これによって提案手法の有効性も確認できた．

本論文では比較的整備が進み，学習に利用できるデータの多い近代語を対象とした．しかし，今後は提案手法を近世等の整備がさほど進んでいない文書に対しても利用することを考えており，その場合には，今回のように大量の訓練データがあるとは限らない．そこで，訓練データ数を変化させた場合の提案手法の性能の違いを見るために，精度と再現率，F 値をそれぞれ図 5 にプロットした．ただし，漢字クラスは使用せず，カーネルも線形を使用している．精度は，比較的訓練事例数が少なくても高く取れているが，再現率が 90%を越えるには大体 10 万個の事例が必要

Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

<s>若し此の大勢の動く<d prob="98.006%">が</d>まゝに乗り行か<d prob="96.664%">ば</d>、</s>
<s>結局如何なる塲所に到る可き乎、</s>
<s>實に今日は油斷のならぬ時節ぞかし、</s>
<s>保守的の反動豈に偶然ならんや、</s>
<s>必ら<d prob="99.955%">ず</d>他に此れを激成するものな<P prob="16.524%">か</P>ら<d prob="99.937%">ず</d>、</s>
<s>激成するものとは何<d prob="99.998%">ぞ</d>や、</s>
<s>貴族的急進派の運動是れなり、</s>
<s>其の運動は吾人<d prob="99.936%">が</d>詳説する迄もなし、</s>
<s>人若し頭を轉<d prob="99.996%">じ</d>て昨日の世界を看は、</s>
<s>以て其の如何なるものなりやを知らん、</s>

図 7 『国民之友』に対する濁点自動付与例

#1 <s>半は反對するか如きことあらは、</s>
#2 <s>如何に煽動者出て來るも、</s>
#3 <s>今日よりして戰を挑む積りなるか、</s>
#4 <s>谷將軍は余腐儒生と縣を同ふするが故に、</s>
#5 <s> 魯西亞は斯く土地の廣大に釣合はす人口甚僅少なりと雖</s>

図 8 清音を誤って濁点無表記文字と検出してしまった例

#6 <s>亦た茲に存すと謂はざる可からず、</s>
#7 <s>然らされば止め～～の叱聲四に起る</s>
#8 <s>は孔子の成語なれと其の語中に支那に限きる特種の意趣もなけれは</s>
#9 <s>願くは我大臣諸公をして其の長所に於て民間論者と爭はしめんことを</s>
#10 <s>殘るーツか。</s>

図 9 濁音を濁点無表記文字と検出できなかった例

であることがわかる（およそ 3 万文）．[i]これは，訓練事例中の負例数が正例数に比べて多いためだと思われる．また，訓練事例数が多いほど検出性能は向上しているため，事例数を更に増やすことで今以上の性能が得られると予想される．

i 同じ日本語である，という理由から，学習データが足りないのなら十分に量のある現代語のテキストで学習するといった方法が考えられる．しかし，古い時代の日本語は，使用している文字や語彙，文法，表記法が現代のものと，極端でないにしても確実に異なるため，そういった手法はあまり上手くいかないことが文献14)において報告されている．

4.4 濁点の自動付与を用いたアノテーション支援

濁点の自動付与の応用例として，アノテーション支援に用いる方法について述べる．まず，分類の結果，濁点無表記文字と識別された文字に濁点の自動付与を実施する．ただし，コーパスを利用する側にとっては，原文には濁点が付いていなかったという情報も有用なため，残しておく必要がある．そこで，濁点の自動付与を実施したことを示すタグ（<d></d>）を設ける．また，表3で示している通り，提案手法における濁点無表記文字の検出精度は100%ではないため，濁点の付与を自動で行なったとしてもその結果を最終的には人手で確認する必要がある．その際の判断基準としてタグ内に濁点付与の確率値を記述する．これにはPA-Iの分類スコアをシグモイド関数を使って確率値に直したものを使用する.[5][j]また，検出から漏れてしまった濁点無表記文字についてもやはり，作業者が確認して回収する必要があるが．その作業はタグの付いている箇所が間違っているか否かを確認するよりも手間がかかる．そのため，可能な限り濁点無表記文字へのタグ付けは，再現率が100%に近いことが望まれる．

図6に示したように，分類スコア（2次の多項式+漢字クラス数100）から得られた確率値は，清音仮名文字であると正しく分類された事例のほとんどが，10%以下の値となっている．そこで，分類では濁点無表記文字と判定されなくとも，分類スコアから得られる確率が10%以上ならば，タグ（<P></P>）を設けて人手での確認を促すようにした．ただし，このタグで囲った文字には濁点は付けない．これにより，濁点付与の性能自体はそのままに，評価用テキスト中の濁点無表記文字の98.956%をカバーすることができる．

図 7 に実際にタグ付けを行なった例を示す．また，濁点無表記文字の検出に失敗した例の一部を図 8 と図 9 に示す．図 8 と図 9 はいずれも濁点の自動付与を行う前の文に，検出に失敗した文字へと下線を引く形で誤りを明示している．誤検出あるいは検出漏れの大部分は，図 8#1～3 や図 9 #6, 10 に示したような「半ば」と「半は」,「出で」と「出て」,「存ず」と「存す」,「が（接続助詞 or 格助詞）」と「か（係助詞）」等，その 1 文を見ただけでは人手での判断にも迷うようなものであった．これらは前後の文脈を参照すれば分かる場合もあるが,「存す」と「存ず」などは一概には決められない例である．明治期は標準的な語形･表記が必ずしも確立していない時期であるため，濁点無表記文字を濁点付きの仮名に人手で修正を行なった太陽コーパスの中でさえ「願くば」と「願くは」が混在している．人手の区別においては，記事全体で統一するか，そのまま放置されるといった方法で処理されているため，そういったものはパターン化してしまい，提案手法とは別の処理を適用することで対処すべきだと考えられる．また，図 9#8 は送り仮名の使い方が一般的な「限る」と異なるために生じたエラーだと考えられるが，これについては，今後の

j あくまで分類後に確率を付けるのであって，この確率値に基づいて分類を行う訳ではない．

Vol.2011-NL-201 No.6
Vol.2011-SLP-86 No.6
2011/5/16

課題でもある仮名遣いの正規化を実施することで解消できると考えられる.

5.おわりに

本論文では，点予測を用いた濁点無表記文字の検出と濁点の自動付与に関する手法を提案した．太陽コーパスで学習を行い，同時期の雑誌資料である国民之友で評価を行った結果，精度，再現率共に 96%以上の性能で検出が行えた．また，分類のスコアから得られる確率を用いることで，約 99%の濁点無表記文字にタグを付けることができた.

謝辞 本研究は、国立国語研究所の共同研究プロジェクト「統計と機械学習による日本語史研究」による研究成果の一部である。

参考文献

1) Cho-Jui Hsien, Kai-Wei Chang, Chin-Jen Lin: A Dual Descent Method for Large-scale Linear SVM, ICML, pp.1-12 (2008).
2) Fernando Pereira, Naftali Tishby, and Lillian Lee: Distributional Clustering of English Words, ACL-31, pp.183-190 (1993) .
3) George Karypis: CLUTO ⋆ A Clustering Toolkit, University of Minnesota, Department of Computer Science, Technical Report #02-017 (2003).
4) Graham Neubig, Yosuke Nakata, Shinsuke Mori: Pointwise Predication for Robust, Adaptable Japanese Morphological Analysis, ACL-49 (2011) to appear.
5) Hsuan-Tien Lin, Chih-Jen Lin, Ruby C. Weng: A Note on Platt's Probabilistic Outputs for Support Vector Machines, Machine Learning, Vol.68, Issue.3, pp.267-276 (2007).
6) 北研二，辻井潤一：確率的言語モデル，言語と計算 4，東京大学出版会， pp. 72-76 (1999).
7) Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer: Online passive-aggressive algorithms, Journal of Machine Learning Research 7, pp.551-585 (2006).
8) 国立国語研究所編：太陽コーパス，国立国語研究所資料集 15，博文館新社(2005).
9) Manabu Sassano: An Empirical Study of Active Learning with Support Vector Machines for Japanese Word Segmentation, ACL-40, pp.505-512 (2002).
10) Mark. D. Kernighan, Kenneth. W. Church, and William. A. Gale: A spelling correction program based on a noisy channel model, COLING-90, pp.205-210 (1990).
11) 永田昌明: 文字類似度と統計的言語モデルを用いた日本語文字認識誤り訂正法，電子情報学会論文誌(D-II), Vol.J81-D-II, No.11, pp.2624-2634 (1998).
12) N. Yoshinaga and M. Kitsuregawa: Kernel Slicing: Scalable Online Training with Conjunctive Features, COLING, pp.1245-1253, (2010).
13) 小木曽智信，小椋秀樹，近藤明日子: 近代文語文を対象とした形態素解析辞書の開発，言語処理学会第 14 回年次大会発表論文集, pp.225-228 (2008).
14) 小木曽智信，小椋秀樹，田中牧郎，近藤明日子, 伝康晴: 中古和文を対象とした形態素解析辞書の開発，情報処理学会研究報告，Vol.2010-CH-85 No.4 (2010).
15) 新納浩幸: 平仮名 N-gram による平仮名列の誤り検出とその修正，情報処理学会論文誌，Vol.40, No. 6, pp.2690-2698 (1999).