統計的機械学習による歴史的資料への濁点の自動付与

岡 照晃 （奈良先端科学技術大学院大学）[†1]

A Machine Learning Approach to Automatic Labeling of Voiced Consonant Mark for Historical Text

Teruaki Oka （Nara Institute of Science and Technology）

1 はじめに

近年，コーパスを利用した日本語研究が増えつつある．

しかし，日本語学や国語学の分野では，古い時代の資料を扱う歴史的研究が現在も大きな位置を占めている．だが，それらの分野で扱われるような歴史的資料は，コーパスとしての整備が現代語のコーパスと比べて進んでいないのが現状である．

歴史的コーパスの整備が進まない原因の一つとして，コーパス整備の際の**校訂**の，作業コストが高いことが挙げられる．校訂作業は専門家にしか行えず，作業人員を大量に集めることが難しい．またその反面，作業対象は膨大であるため，作業を完了するまでに非常に時間がかかる．

そこで本研究では，統計的機械学習手法を用い，歴史的資料の校訂作業を自動化することを最終的な目的とする．これにより，誰でも簡単に低コストかつ大規模に校訂作業を実施することが可能になると考えられる．そしてその第 1 段階として，校訂作業の中から**濁点付与**を取り上げ，自動化に取り組んだ．

2 校訂作業における濁点付与

歴史的資料の記述中にはよく，図 1 のような「濁点を付けて書いてあることが期待されるのに，濁点の付いていない文字」が含まれている．本論文ではこういった文字のことを**濁点無表記文字**と呼ぶ．濁点無表記文字は，コーパスユーザの可読性・検索性を損なわせる原因の一つである．そこで，濁点無表記文字を濁点付きの文字（**濁点文字**）に置き換える校訂作業が行われる．これが濁点付与である．

表 1 に，未校訂の資料中に存在する濁点文字（e.g. が，だ，ば，･･･ ）と濁点を付けることが可能な文字（e.g. か，た，は，･･･ ）の統計データを示した．この表を見ると，総文字数 8,423 文字に対して，その内約 19%の 1,641 文字は濁点を付けることが可能な文字（濁点無表記になっているかもしれない文字）である．人手で作業する場合は，これらすべてに対して濁点無表記か否かを網羅的に確認しなくてはならない．また，濁点が付く可能性のある文字の内，約 22%（368/1,641）が濁点無表記の濁音文字であり，例えば,「ガ」と発音する文字も「か」と表記されている．濁点付与では，見つけた濁点無表記文字の一つ一つに濁点を施していかないといけない．しかし，濁点無表記文字の数が多いだけに，非常に手間のかかる作業である．

[†1] teruaki-o@is.naist.jp

今や廣島は其名大に内外國に顯はれ苟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

（『太陽』 1925 年 2 号 p.64 より抜粋）

図 1: 濁点無表記の例．このテキストは太陽コーパスの原文から抜きだしたものである．下線を引いた文字には，通常なら濁点が期待されるが，ここでは付けられていない（濁点無表記になっている）．

表 1: 濁点と濁音の統計．近代の雑誌**明六雑誌**第 1 号（2011 年 12 月段階のデータ，総文字数:8,423）中に含まれる濁点文字と濁点をつけることが可能な文字の中から，実際の発音が清音の文字数と濁音の文字数の内訳を調査した．表中の濁点の付き得る文字の内，実際の発音が濁音である文字が濁点無表記文字である．

表記＼発音	濁音	清音	計
濁点文字	78	0	78
濁点をつけることが可能な文字	368（濁点無表記文字）	1,273	1,641
計	446	1,273	

実際，現在コーパス化が進められている近代の雑誌**国民之友**の校訂作業では，1 人の作業者が 23 ページ分 (約 1 万 6,000 文字) に濁点付与するのに，1 日を要したと報告されている[1]．これに対し，例えば，既存の校訂済みコーパスである**太陽コーパス**（国立国語研究所, 2005）は，総文字数約 1,450 万文字の規模である．

よって，濁点無表記のアノテーションを自動化するだけでも，校訂作業の効率向上が大いに期待できる．

3　関連研究: 形態素解析辞書を用いた手法

近代文語論説文（明治普通文）には濁点無表記が多くみられる．そのため，近代文語論説文を対象とした形態素解析辞書である**近代文語 UniDic**[2]（小木曽ら, 2008）には，無濁点の見出し語も少数だが登録されている（e.g.,「す（文語助動詞-ズ，終止形）」）．

形態素解析辞書に無濁点の見出し語を追加することで，濁点無表記を含んだ文の解析を行うことが可能になる．またそれだけでなく，解析結果から濁点付きの表記を得ることもできる（**辞書ベースの手法**）．

しかしながら，辞書ベースの手法を行うためには，校訂対象となる歴史的資料用の形態素解析辞書が必要となる．しかし，現在使用可能なものの中で実用に足るものは，近代文語 UniDic の他に中古和文を対象にした**中古和文 UniDic**（小木曽ら, 2010）のみである．そのため，中世や近世の資料にこの手法を適用することはできない．また，辞書ベースの手法では，濁点付与の性能が形態素解析の精度に依存する．そのため，濁点付与の性能を上げるた

[1] 1 日の作業時間を 5〜6 時間とした場合．
[2] バージョン 1.1

か，き，く，け，こ，　さ，し，す，せ，そ，　た，ち，つ，て，と， は，ひ，ふ，へ，ほ，　ゝ，く（くの字点）

図 2: 校訂対象文字．ここで示した 22 文字が濁点付与の対象となる文字である．

めには，学習用の形態素解析済みコーパスを新しく整備するか[3]，辞書の見出し語を増やす必要がある．しかし，これは一般にコストが高い．

加えて，近代文語 UniDic と中古和文 UniDic はいずれも，基本的に校訂済みの文を解析するために整備されている．そのため，無濁点の見出し語を追加したとしても，未校訂資料の形態素解析に利用した場合，解析結果の精度は決して高くないと考えられる．

そこで本論文では，濁点の自動付与のタスクを文字単位のクラス分類問題として定式化した．具体的には，未校訂の資料中に存在する「濁点の付く可能性のある文字（**校訂対象文字**）」を濁点文字に置き換えるべきか否か分類する問題を扱う．提案手法では，分類は点予測によって行う．点予測とは，周囲の文字に対する分類結果を参照せずに，当該分類を行う手法である．周囲の分類結果を見ないため，分類誤りが伝播することがない．そのため，表記の整っていない未校訂の資料に対しても頑健に動作することができる．また，分類の**素性**[4]にも，分類対象文字の周辺文字列の表層的な情報のみを使用し，周囲の単語境界の情報や，品詞の情報は使用しない．そのため，提案手法では，学習用コーパスとして形態素解析済みコーパスを必要とせず，形態素解析辞書も必要としない．また，形態素解析の精度に濁点付与の性能が左右されることもない．

4　提案手法: 点予測による濁点の自動付与

提案手法では，未校訂資料の中に存在する校訂対象文字に対して，それぞれ独立に「濁点を付けるべきか否か」の分類を実施する．ただし，本論文では，校訂対象文字として平仮名と踊字である**くの字点**（く，ぐ）だけを扱い，片仮名は扱わないこととした[5]．そのため，濁点付与の対象となる文字は図 2 に挙げた 22 種類である．

以下に提案手法の概略を述べる．提案手法の詳しい説明については,（岡ら, 2011）もしくは（Oka et al., 2011）を参照．

4.1　分類に使用する素性

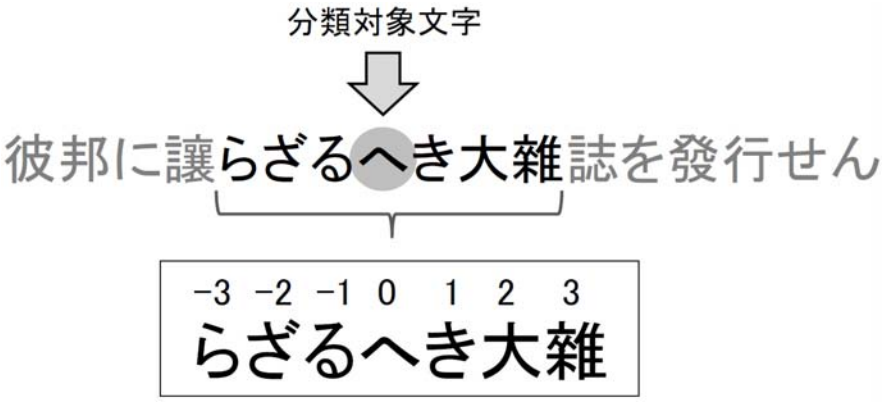
分類時の素性には，当該の分類対象文字とその周囲の文字列の表層的な情報のみを使用する．具体的には図 3 のように，分類対象文字とその左右の 3 文字の範囲内にある文字 n-gram を素性とする．文字 n-gram は，1〜3-gram までを使用する．また，各 n-gram には出現位置（分類対象文字からの相対位置）を添え字として設けている．提案手法で使用する素性は，「その n-gram がその位置に現れたか（1）否か（0）」を表す 2 値素性である．

歴史的資料は，表 1 で示しているように，完全に無濁点になっているとは限らない．所々には濁点が付いた状態の資料もある．濁点の使い方は書き手によって一定でない．そのため，あらゆる濁点の表記状態に対応するためには，こういった濁点を分類時には外しておくべき

[3]各単語のコスト（出現しやすさ）を計算するために使用される．

[4]分類の手がかりとなる情報．

[5]漢字片仮名交じり文を除いて，基本的に片仮名は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．そのため，本論文では片仮名は対象外とした．



位置-3 の文字 1-gram = ら	位置-3 の文字 2-gram = らざ	位置-3 の文字 3-gram = らざる
位置-2 の文字 1-gram = ざ	(位置-3 の文字 2-gram = らさ)	(位置-3 の文字 3-gram = らさる)
(位置-2 の文字 1-gram = さ)	位置-2 の文字 2-gram = ざる	位置-2 の文字 3-gram = ざるへ
位置-1 の文字 1-gram = る	(位置-2 の文字 2-gram = さる)	(位置-2 の文字 3-gram = さるへ)
位置 0 の文字 1-gram = へ	位置-1 の文字 2-gram = るへ	位置-1 の文字 3-gram = るへき
位置 1 の文字 1-gram = き	位置 0 の文字 2-gram = へき	位置 0 の文字 3-gram = へき大
位置 2 の文字 1-gram = 大	位置 1 の文字 2-gram = き大	位置 1 の文字 3-gram = き大雑
位置 3 の文字 1-gram = 雑	位置 2 の文字 2-gram = 大雑	

図 3: 提案手法で使用する素性．ただし，ここでは値が 1 となる素性のみを示している．

である．しかしながら，予め施されていた濁点は分類の際の証拠として有効な場合もあると考えられる．そこで，分類時には，文字 n-gram 内に含まれる濁点文字の一部〜全てより濁点を外したものも素性として参照することにした．この素性は，図 3 において，括弧付きで示してある．

また，歴史的資料の中では句読点の使い方が一貫していない．句点を読点のように使用する場合もあれば，読点を使って文末を表現する場合もある．そこで提案手法では，資料中の句読点「、。」を全て特殊記号 ⟨PUNC⟩ に置き換えることにした．

4.2 学習用事例の作成手順

学習用の事例は，学習用コーパス中にある学習用文字（図 4 参照）から作成する．学習用コーパスには校訂済みの歴史的資料を使用する．学習用事例作成の手順は以下の通りである．

1. 学習用コーパスから学習用文字を 1 つ取り出す（e.g.,「が」）．
2. 取り出した文字とその左右 3 文字を合わせて 1 つの事例とみなす．この際，取り出した学習用文字が濁点文字であれば，濁点を外しておく（「が」→「か」）．
3. 取り出した学習用文字（「が」）を正解のクラスとする．

4.3 分類器

提案手法では，太陽コーパスのような大規模なコーパスからでも高速かつ高精度に分類器の学習を行うため，分類器として線形パーセプトロンを使用する．実際には，多クラスの

か，き，く，け，こ，	さ，し，す，せ，そ，	た，ち，つ，て，と，
は，ひ，ふ，へ，ほ，	ゝ，く（くの字点）	
が，ぎ，ぐ，げ，ご，	ざ，じ，ず，ぜ，ぞ，	だ，ぢ，づ，で，ど，
ば，び，ぶ，べ，ぼ，	ゞ，ぐ（くの字点）	

図 4: 学習用文字一覧.

PassiveAggressive-I（Crammer et al., 2006）を採用した.

提案手法では，各校訂対象文字ごとに「濁点をつけた文字」か「濁点を付けないまま文字」のいずれかのクラスに分類する規則（**モデル**）を作成する．例えば，校訂対象文字「か」に対して,「か」と「が」のいずれかのクラスに分類するモデルを作成する．そしてそれとは別に,「き」に対して,「き」か「ぎ」か分類するモデルを作成する.

5 濁点付与の性能評価実験

提案手法の有効性を検証するために，濁点付与の性能評価実験を行なった．今回は未校訂の近代文語論説文を対象とし，濁点付与の適合率と再現率を調べた.

5.1 実験に使用したコーパス

本実験では，学習用コーパスとして，以下の校訂済みのコーパスを使用する.

- UniDicMLJ-TRAIN:
  近代文語 UniDic のコスト算出に用いられたコーパス．形態素解析済みコーパスであり，校訂も実施済みである．ただし，量が少なく，原文の情報もコーパス中には保持されていない.

- SUN-TRAIN:
  近代語の大規模コーパスである太陽コーパス[6]から 9 割を学習用コーパスとして利用する．実際には，太陽コーパスの 1895 年 5 号，1901 年 5 号，1909 年 5 号，1917 年 5 号，1925 年 5 号を評価用コーパスとして別に分け (SUN-TEST)，残りを学習に利用することにした (SUN-TRAIN)．太陽コーパスは構造化テキストタグ付きコーパスであるため，校訂は行われているが，形態素解析までは行われていない.

また，評価には以下のコーパスを利用する.

- SUN-TEST:
  太陽コーパスの 1895 年 5 号，1901 年 5 号，1909 年 5 号，1917 年 5 号，1925 年 5 号を評価用コーパスとして利用する.

- NF-TEST:
  現在コーパス化の作業が進められている明治期の雑誌，国民之友も評価用コーパスとして利用する．ここでは，2011 年 3 月の段階で濁点付与が実施されていた 1887 年 10 号，1888 年 20 号，1888 年 30 号，1888 年 36 号を使用した.

[6]2011 年 1 月段階のデータ

表 2: コーパス内の文数と段落数の内訳.

	文数	段落数	文字総数
UniDicMLJ-TRAIN	20,330	-	604,966
SUN-TRAIN	-	70,084	6,380,398
SUN-TEST	-	6,316	619,357
NF-TEST	-	868	172,780
M6-TEST	-	1,450	252,232

表 3: 学習用事例の内訳.

事例のクラス / 学習用コーパス	濁音文字	清音文字	合計
UniDicMLJ-TRAIN	26,123	110,974	137,097
SUN-TRAIN	208,099	962,580	1,170,679

・ M6-TEST:

Oka et al.(2011）では，評価用コーパスとして上記 2 つのコーパスのみを利用している．本論文ではさらに，太陽や国民之友と同じ，明治期の雑誌である**明六雑誌**（全 43 号）[7]も評価用コーパスとして利用する．ただし，明六雑誌はほとんどの記事が漢字片仮名交じり文で記述されている．そのため，ここでは，全ての片仮名文字を平仮名文字に直して用いることにした．

評価に使用する太陽コーパス・国民之友・明六雑誌（3 雑誌コーパス）はいずれも校訂済みのコーパスである．しかし，タグを使って原文が保持されている．そこで，実験を実際のタスクに近づけるため，評価にはタグから再現した原文を使用する．

また，明治期において，句読点の使い方はまだ確定していなかった．そのため，明確な文境界を定めることは難しい．今回使用する 3 雑誌コーパスでも，文境界の明確なアノテーションは行われていない．そこで今回，3 雑誌コーパスは学習・評価の両方において，段落単位で使用することにした．実際のタスクでも文境界が定められていないことが多いため，これは，より実際のタスクに近い設定といえる．ただし，UniDicMLJ-TRAIN は近代文語 UniDic の学習に使用されたコーパスであるため，文境界は明確にされている．そこで，UniDicMLJ-TRAIN のみ，文単位で使用することにした．事例抽出の際には，文（or 段落）の頭と末尾に，それぞれ文頭，文末を表す特殊記号 ⟨BOS⟩ と ⟨EOS⟩ を設ける．

太陽コーパス・国民之友・明六雑誌には口語で書かれた記事も含まれている．この実験では，文語を扱う．そのため，口語の記事や引用は学習用コーパス・評価用コーパスのいずれからも全て除外した．

口語文を除いた各コーパスの文数と段落数，総文字数の内訳を表 2 に示す．また，学習用事例と評価用事例の内訳を表 3 と表 4 に示す.

[7]2011 年 12 月段階のデータ

表 4: 評価用事例の内訳.

評価用コーパス ＼ 事例のクラス	濁音文字	清音文字	合計
SUN-TEST	899	92,803	93,702
NF-TEST	3,842	25,418	29,260
M6-TEST	6,219	39,314	45,533

5.2 比較手法: 辞書ベースの手法

辞書ベースの手法を比較手法として設定した．この手法では，以下の手順で近代文語 UniDic を拡張し，拡張した辞書を用いた形態素解析の結果から濁点付与を行う．

1. 近代文語 UniDic の全ての見出し語から濁点を全て外す．ただし，各語のフィールド中には濁点を外す前の表記を保存しておく．

2. 無濁点にした UniDicMLJ-TRAIN を用いて，1 で作成した辞書の各語のコストを計算する．

3. 各語のフィールド中に残しておいた濁点付きの表記から，濁点の一部〜すべてが補われた見出し語を復元し，辞書に追加する．ただし，この時追加する語のコストは，無濁点の場合のコストと同一とする．

この辞書を使えば，濁点が所々抜け落ちたような文でも形態素解析を行うことができる．また，各語のフィールド中には濁点付きの表記が保存されているため，形態素解析の結果から濁点付きの表記を復元できる．

ただし，辞書ベースの手法では，形態素解析済みの近代語のコーパス（現時点で UniDicMLJ-TRAIN のみ）からでしか学習が行えないという欠点がある．また，評価用コーパスはいずれも文のアノテーションが明確に行われていないため，形態素解析は段落単位で行う．

形態素解析器には MeCab[8] を使用する．

5.3 濁点付与性能評価実験

各手法を用いてそれぞれの評価用コーパスに濁点付与を行い，評価を行なった．ここでは濁点付与の適合率，再現率と，その 2 つの調和平均である F 値で評価した．各値の計算方法は以下の通り．

$$\text{適合率} = \frac{\text{正しく濁点を付けた文字数}}{\text{濁点を自動付与した文字数}} \times 100[\%] \tag{1}$$

$$\text{再現率} = \frac{\text{正しく濁点を付けた文字数}}{\text{評価用コーパス中の濁点無表記文字数}} \times 100[\%] \tag{2}$$

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{3}$$

結果を表 5 に示す．

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

表 5: 濁点付与の性能評価.

評価用コーパス	手法	学習用コーパス					
		UniDicMLJ-TRAIN			UniDicMLJ-TRAIN + SUN-TRAIN		
		適合率.[%]	再現率.[%]	F 値	適合率.[%]	再現率.[%]	F 値
SUN-TEST	辞書ベース	50.9	91.8	65.5	-	-	-
	提案手法	54.7	85.2	66.6	71.2	97.0	82.1
NF-TEST	辞書ベース	93.3	96.5	94.9	-	-	-
	提案手法	95.1	94.5	94.8	96.0	98.3	97.1
M6-TEST	辞書ベース	90.1	95.9	92.9	-	-	-
	提案手法	93.4	92.4	92.9	94.7	98.1	96.4

学習用コーパスを UniDicMLJ-TRAIN でそろえた場合，提案手法は，辞書ベースの手法に比べて低い再現率を示している．しかし，適合率は辞書ベースの手法よりも高いため，F 値においてはほとんど同じ性能が得られた．

また，提案手法は辞書ベースの手法に比べて低コストで学習用コーパスを追加できるという利点がある．学習用コーパスとして SUN-TRAIN を追加したとき，提案手法は適合率，再現率，F 値のすべてにおいて，辞書ベースの手法よりも高い性能を示した．このように，比較的簡単に性能を上げられるという点において，提案手法の優位性が確認できた．

5.4 エラー分析

提案手法のエラー分析を行った結果，以下に挙げた語と語の間で，濁点付与に失敗する傾向がみられた．

- 格助詞「が」・接続助詞「が」と，終助詞「か」・並列助詞「か」
- 打消しの助動詞「ず」と，サ変動詞「す」
- サ変動詞と，ザ変動詞
- 接続助詞「ば」と，係助詞の「は」
- 当時，語形上揺れがあった語（e.g.,「願わくば」と「願わくは」）
- 濁点を付けても付けなくてもどちらでもよさそうな語（e.g.,「結び」と,「結ひ」,「出て」と,「出で」）

また，提案手法と辞書ベースの手法のエラーを比較したとき，濁点付与誤りの傾向に大きな差は見られなかった．ただし，辞書ベースの手法特有のエラーとして，以下の語間での間違いが多くみられた．

- 接続助詞の「て」と接続助詞の「で」
- 接続助詞の「とも」と接続助詞「ども」

提案手法では，単語境界の情報を分類に使用しない．そのため，以下のようなエラーが数例見つかった（濁点付与に失敗している文字を太字にして示している）.

猶ほ兒玉氏の力強**が**りとや思ひけん

表 6: 形態素解析性能の改善度の比較（SUN-TEST）.

	単語分割			品詞認定			語彙素認定		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
原文	99.866	99.965	99.916	99.596	99.498	99.547	99.575	99.478	99.527
辞書ベース	99.886	99.842	99.864	99.300	99.342	99.321	98.815	98.858	98.837
提案手法で前処理	**99.943**	**99.972**	**99.957**	**99.727**	**99.698**	**99.713**	**99.715**	**99.686**	**99.700**

表 7: 形態素解析性能の改善度の比較（NF-TEST）.

	単語分割			品詞認定			語彙素認定		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
原文	98.607	99.406	99.005	95.878	95.112	95.494	95.696	94.932	95.312
辞書ベース	99.860	99.662	99.761	95.859	96.048	95.953	95.225	95.413	95.319
提案手法で前処理	**99.935**	**99.904**	**99.919**	**99.490**	**99.520**	**99.505**	**99.465**	**99.495**	**99.480**

ただし，同じようなエラーは辞書ベースの手法でも生じている．以下の例のように，辞書ベースの手法でも単語分割に失敗し，その結果，濁点付与を失敗することがある．

人 | の | 氣附:名詞（キヅケ） | が:助詞 | さる:連体詞 | 所ろ:名詞 | 、

5.5 形態素解析精度の改善度の比較

近代文語 UniDic は本来，校訂済みの文を解析するために整備されている．そのため，未校訂の資料の形態素解析に利用した場合，結果の精度は高くない．そこで，提案手法を形態素解析の前処理に用いることで，形態素解析の性能がどれほど改善できるか調査した．

ただし実際には，評価用コーパスの校訂済み本文を近代文語 UniDic を用いて形態素解析し，正解データとしている．そして，原文に濁点付与を行うことで，どこまでその結果に近づけるかを評価した．

提案手法は，UnidicMLJ-TRAIN + SUN-TRAIN で学習を行なったモデルを使用する．また，形態素解析器には MeCab を利用した．

評価は，単語分割，品詞認定，語彙素認定の 3 段階で行う．それぞれの段階における適合率・再現率・F 値を調査した[9]．

また，前処理に提案手法を使用せず，辞書ベースの手法で，濁点付与と形態素解析を同時に実施した場合と比較を行なった．ただし，近代文語 UniDic にはもともと少数であるが，無濁点の見出し語が含まれている．2 つの手法を公平に比較するため，提案手法で前処理された資料を解析する近代文語 UniDic からは，無濁点の見出し語を取り除いている．

結果を表 6〜8 に示す．この結果を見ると，提案手法を用いて濁点付与を行うことで，校訂済みテキストを解析するのとほぼ同等の性能を実現することが可能だと分かった．

[9]品詞認定，語彙素認定の性能は MeCab の評価用スクリプト mecab-system-eval を利用して求めた．

表 8: 形態素解析性能の改善度の比較（M6-TEST）.

	単語分割			品詞認定			語彙素認定		
	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
原文	98.442	99.523	99.980	95.679	94.648	95.161	95.392	94.365	94.876
辞書ベース	99.857	99.788	99.822	95.762	95.828	95.795	94.997	95.063	95.030
提案手法で前処理	**99.898**	**99.930**	**99.914**	**99.398**	**99.367**	**99.382**	**99.353**	**99.322**	**99.338**

6 おわりに

本論文では，点予測を用い，文字単位で濁点の自動付与を行う手法を提案した．太陽コーパスで学習を行い，近代語の資料に対して濁点付与を行なった結果，国民之友，明六雑誌に対しては適合率，再現率共に約 95%以上の性能で濁点付与が行えた．また，提案手法を前処理に用いることで，未校訂の資料でも校訂済みの資料と同程度の性能で形態素解析が行えるようになることが分かった．

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．

文献

[1] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer (2006) 「Online Passive-Aggressive Algorithms」 *Journal of Machine Learning Research*, 7 pp. 551-585.

[2] Teruaki Oka, Mamoru Komachi, Toshinobu Ogiso, Yuji Matsumoto (2011) 「Automatic Labeling of Voiced Consonants for Morphological Analysis of Modern Japanese Literature」 In *Proceedings of the 5th International Joint Conference of Natural Language Processing* (IJCNLP 2011), pp. 292-300.

[3] 小木曽智信, 小椋秀樹, 近藤明日子 (2008) 「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.

[4] 小木曽智信, 小椋秀樹, 田中牧郎, 近藤明日子, 伝康晴 (2010) 「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, 2010-CH-85:4.

[5] 岡照晃, 小町守, 小木曽智信, 松本裕治 (2011) 「機械学習による近代文語文への濁点の自動付与」 情報処理学会研究報告 自然言語処理研究会報告, 2011-NL-201:6, pp. 1-8.

[6] 国立国語研究所編 (2005) 『太陽コーパス』 国立国語研究所資料集 15, 博文館新社.