1ZC-03

古文書文字認識における N-gramの出現頻度を利用した認識候補削減法

渡邊 一義 † 渡辺 悟 ‡ 鈴木 徹也 ‡

† 芝浦工業大学 システム理工学部 電子情報システム学科 ‡ 芝浦工業大学大学院 理工学研究科 システム理工学専攻

1 はじめに

古文書には, 未だに翻刻されていない文献が多数存在する. 古文書で使用される変体仮名は, 形は異なるが同じ読みをしたり, 形は酷似しているが異なる読みをしたりする. そのため, 変体仮名の翻刻は難しく, 人手での翻刻は非常に時間が掛かる. そこで, 我々の研究グループは, 翻刻支援システムの研究を行っている.

変体仮名の翻刻において, 画像認識のみでは, 文字に対する読みの決定は難しい. そこで我々は翻刻支援システムの構想を提案し, その一部を実装した [1, 2, 3]. システムへの入力は翻刻対象のスキャン画像であり, 出力は各変体仮名の読みである. 画像認識器と制約解消器とが連携しそれを行う. まず画像認識器が各文字を切り出し, 切り出した文字の読みの候補を列挙し制約充足問題を作成する. そして制約解消器がその制約充足問題を解く. このようにシステムは翻刻作業の一部を組み合わせ最適化問題として解く. 我々が実装したのはその制約解消器である. 単語の連接関係を利用し, 現実的な時間内で解を出力する. しかし, 一文字当たりの読みの候補数が多い場合, 解の正解率が低下する問題が発見された.

本研究の目的は, 制約解消器で求める解の正解率を高く保つために, 各文字に対する不正解となる読みの候補を削除する方法を提案することである.

以降, 第 2 節と第 3 節で先行研究を説明し, 第 4 節で提案手法を説明する. 第 5 節で実験による比較を行い, 第 6 節で結論を述べる.

2 翻刻制約充足問題

翻刻支援システムで取り扱う制約充足問題は**翻刻制約充足問題**と呼ばれ, 変数, その領域, そして制約で構成される. 変数は入力画像の変体仮名一文字分の矩形領域に対応する. 変数の領域は, その変数に割り当てる読みの候補である. 制約には充足優先度があり, 必須制約と選好制約とに分類される. 必須制約は, 必ず充足されるべき制約である. 選好制約は, なるべく充足されることが望まれる制約である. 翻刻制約充足問題は, 定義された変数間の読み順にしたがって読みを並べた結果が単語列となるという必須制約を必ず含む. その他の制約には, 等号制約 (変数の読みが等しいという制約) や単語開始制約 (その変数が単語の一文字目という制約) などがある. 変数への読みの割り当てを**解**とよぶ. 解の中で, より優先度の高い制約を満たすものが最適解となる.

Reduction of Historical Japanese Character Recognition Candidates by the Occurrence Frequencies of N-grams

Kazuyoshi WATANABE†, Satoru WATANABE‡, Tetsuya SUZUKI‡

†Department of Electronic Information Systems
College of Systems Engineering and Science
Shibaura Institute of Technology

‡Systems Engineering and Science
Graduate School of Engineering and Science
Shibaura Institute of Technology

{bp11094, mf14061, tetsuya}@sic.shibaura-it.ac.jp

3 単語の連接関係を利用した解の探索法

3.1 概要

単語の連接関係を利用した解の探索法 [2, 3] は, 以下の手順を踏む.

1. 翻刻制約充足問題から辞書を引いて読みの割り当てグラフを作成する.
2. コスト最小法と N-best 探索を併用して, グラフから N 個の解を求める.
3. 求めた N 個の解の中から, より優先度の高い制約を満たす解を残す.

3.2 読みの割り当てグラフ

読みの割り当てグラフとは変数への読みの割り当てと読み順とを表した有向グラフである. ノードには変数, 読み, 単語の種類, 生起コストが関連づけられており, リンクには単語間の連接コストが付与されている. 生起コストは単語の現れにくさ, 連接コストは単語の繋がりにくさをそれぞれ示す. 特別なノードとして文頭を示すノードと文末を示すノードとがある. 図 1 はある読みの割り当てグラフの一部である.

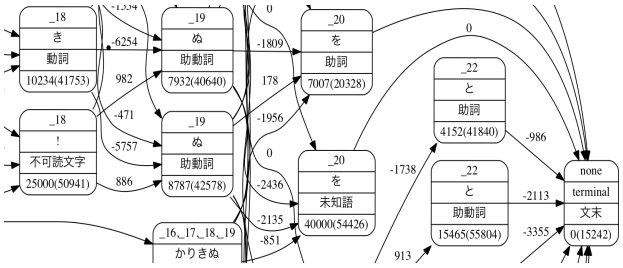


図 1: 読みの割り当てグラフの一部

3.3 コスト最小法と N-best 探索

コスト最小法 [4] は, 単語の生起コストと連接コストとの総和が最小となるパスを最適解として選択する方法である. N-best 探索 [5] は上位 N 個の解を求める探索法である. 翻刻支援システムでは, コスト最小法と N-best 探索とを合わせて, 文頭から文末までのコストの総和が少ない順に N 個の解を求める.

4 N-gramの出現回数に基づく領域簡約

翻刻済み文献中での N-gram の出現回数を利用して, 一文字あたりの読みの候補数を削減する方法を提案する. N-gram とは文章中に連続して出現する, N 文字からなる文字列である [4]. 読みの候補数を削減する処理は, 3.1 節の手順 1 を行う前に適用し, 翻刻制約充足問題を作り直す.

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.

4.1 スコアの計算方法

r を変数 x_j の領域 D_j のある要素とし, r の変数 x_j の読みの候補としてのスコアを次のように計算する.

1. 変数の読み順上で, x_j を含み, 長さ N の変数列の集合を V_s とする.
2. スコア計算に使用する N-gram の集合 G を次式で求める.

$$G = \bigcup_{(x_{i_1}, x_{i_2}, \ldots x_{i_N}) \in V_s} D_{i_1} \times D_{i_2} \times \ldots D_{i_N}$$

3. 変数 x_j の読みの候補 r に対するスコア $score(r)$ を N-gram の出現回数から, 次式で求める.

$$score(r) = \sum_{g \in G} F(g)$$

関数 F は文字列 g から自然数への写像であり, 翻刻済み文献中での文字列 g の出現回数を与える. このスコアは, 読みの候補 r とその前後に位置する読みの候補との繋がりやすさを示す. スコアが大きいほど前後に位置する読みの候補との関連が強いことを示す.

4.2 読みの候補の削除方法

各読みの候補に対して計算したスコアを比較し, 読みの候補を削除する方法を説明する.

4.2.1 読みの候補削除法 1

読みの候補削除について, 以下の条件を設定する.

- 各変数について, スコアの得点が上位 2 位以内でない読みの候補を削除対象として選択する.
- 削除対象として選ばれた読みの候補を削除する.

ただし, 削除対象となった読みの候補について, 前後に位置する読みの候補との組み合わせで, 助詞が形成される場合は削除を行わない. 助詞は古典文法の文献 [6] に記載されているものとする.

4.2.2 読みの候補削除法 2

決定木を用いて読みの候補を削除するかどうかを判定する. 決定木の構築には機械学習アルゴリズム C4.5[7] を用いる. 決定木の入力は次の 5 つの属性である. それは, 読み, 読みの候補数, スコア, スコア計算に使用した N-gram の集合 G の要素数, 集合 G の要素で最も出現回数が多い文字列の出現回数である. 正解の読みを含む翻刻制約充足問題をランダムで作成し, 学習データを構築した.

5 実験と評価

従来手法と比べて, 提案手法を適用した場合に解の正解率が改善されるかを調べる. 実験は次のように行った. 翻刻制約充足問題は, 伊勢物語の一部を利用して人手で作成されたものを使用する. N-gram の出現回数を数える対象は, 翻刻済みである源氏物語 [8](全 54 話, 約 60 万字) とした. 翻刻済み文献の文量を考慮して N-gram の出現回数は N=3 の場合を適用する. 解の探索について, 3.1 節の手順 2 で求める解の個数は 10 個とする. 正解率は, 3.1 節の手順 3 によって選別された解の正解率の平均を取り, 平均正解率とする.

従来手法のみによる実験結果を表 1, 読みの候補削除法 1 を用いた場合の実験結果を表 2 に示す. 表 1, 表 2 より, 提案手法によって読みの総候補数が減少し, 平均正解率が向上していることを確認できる. しかし, 正解の読みの候補を削除してしまい, 正解含有率が低下する場合がある. 表 2 で平均正解率が約 25%低下した ise_3 の読みの候補数は, 一文字当たり高々4 つである. よって, 画像認識で挙げられる読みの候補は一文字につき 4 つ以下で, 且つ正解を含む精度が必要である.

表 1: 従来手法のみで解を求めた場合の結果

問題	読みの総候補数	問題の正解含有率	平均正解率
ise_1	121	1.000	0.932
ise_2	165	1.000	0.743
ise_3	212	1.000	0.617

表 2: 読みの候補削除法 1 を用いた場合の実験結果

問題	読みの総候補数	問題の正解含有率	平均正解率
ise_1	119	1.000	0.951
ise_2	123	1.000	0.902
ise_3	134	0.957	0.749

表 3: 読みの候補削除法 2 を用いた場合の実験結果

問題	読みの総候補数	問題の正解含有率	平均正解率
ise_1	80	0.915	0.891
ise_2	86	0.702	0.679
ise_3	111	0.787	0.598

読みの候補削除法 2 を用いた場合の実験結果を表 3 に示す. 表 3 から, 問題の正解含有率が 8.5%から約 30%低下したことを確認できる. 加えて, 平均正解率も約 11%から約 40%低下した. 今回の機械学習では, 読みの候補について正解, 不正解の分類が上手く出来なかった.

6 おわりに

翻刻制約充足問題において, N-gram の出現回数を利用した読みの候補削減法を提案した. 提案手法によって, 正解の読みを残しつつ, 不正解の読みが削除でき, 解の平均正解率が向上することを実験より確認できた.

今後は, 画像認識器の出力結果を用いた場合で, 提案手法の効果を確認する必要がある. また, 文法を考慮した解の探索を行うことで, より良い解が得られる可能性があると考えられる.

参考文献

[1] 新井侑太. 制約充足による手書き変体仮名認識. 芝浦工業大学修士論文, 2013.

[2] 渡辺悟. 手書き変体仮名認識における単語の連接関係を利用した解の削減. 芝浦工業大学総合研究論文, 2014.

[3] 渡辺悟. 手書き変体仮名認識における単語の連接関係を利用した解の絞り込み. 情報処理学会第 76 回全国大会講演論文集, pp. 2_199-2_200, 2014.

[4] 奥村学. 自然言語処理の基礎. コロナ社, 2010.

[5] 永田昌明. 統計的言語モデルと N-best 探索を用いた日本語形態素解析法. 情報処理学会論文誌, Vol. 40, No. 9, pp. 3420-3431, 1999.

[6] 浜本純逸 監修, 黒川行信 編著. 読解をたいせつにする体系古典文法. 数研出版, 2008.

[7] J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1992.

[8] 米国議会図書館蔵『源氏物語』翻字本文, http://textdb01.ninjal.ac.jp/LCgenji/ (2014 年 5 月 3 日アクセス)

Copyright ©2015 Information Processing Society of Japan.
All Rights Reserved.