様　式　Ｃ－７－１

平成２６年度科学研究費助成事業（科学研究費補助金）実績報告書（研究実績報告書）

1．機関番号　1 4 6 0 3　　2．研究機関名　奈良先端科学技術大学院大学

3．研究種目名　特別研究員奨励費　　4．研究期間　平成２５年度～平成２６年度

5．課題番号　２５・９９３５

6．研究課題名　統計的機械学習を用いた歴史的資料の校訂の自動化に関する研究と自動校訂ツールの開発

7．研究代表者

研究者番号	研究代表者名	所属部局名	職名
	オカ　テルアキ 岡　照晃	情報科学研究科	特別研究員(DC2)

8．研究分担者

研究者番号	研究分担者名	所属研究機関名・部局名	職名

9．研究実績の概要

平安時代や明治時代といった古い時代の文献資料（歴史的資料）のコーパス化作業は，人手の校訂作業がコスト高であるため，現代語に比べて遅れている．そこで本研究では，統計的機械学習の手法を用い，コンピュータによる校訂作業の自動化を目的とする．校訂とは，コーパスユーザの可読性・検索性を向上させるために表記を整える作業であり，本研究では特に表記の標準化を自動化の対象としている．
例えば，歴史的資料の中には，「及ひ（オヨビ）」のように濁音が期待されるのに濁点の付いていない文字（濁点無表記文字）や，歴史的仮名遣と一致しない仮名遣など，表記のバリエーションが多く含まれる．表記のバリエーションはコーパスを検索する際の障害となるため，コーパス整備時には表記を標準化する作業が必要となる．
本研究が扱った表記のバリエーションは以下の5種類である．
①濁点無表記　e.g.，及ひ（オヨビ）②仮名遣の不統一　e.g.，用い（モチイ），用ひ（モチイ），用ゐ（モチイ）③送り仮名の不統一　e.g.，限り，限ぎり，限（カギリ）④踊字による省略　e.g.，及ば/ゝ（オヨバ/バ），恐る々々（オソルオソル）
⑤　漢字片仮名交じり文　　e.g.，裁判官ハ刑法ノ宣告又ハ懲戒ノ処分ニ由ルノ外其ノ職ヲ免セラルヽコトナシ
本研究では，統計的機械学習を用いた日本語自動形態素解析と表記の標準化を同時に実施することで，高精度な表記の標準化の実現を目指す．本年度は，前年度に開発した辞書引き手法に加え，Augmented-Loss Trainingと呼ばれる手法を採用し，形態素解析と表記の標準化を同時に学習できるツールを開発した．Augmented-Loss Trainingを採用したことで，これまでは形態素解析の学習に使用できなかった，単語分割や品詞タグ付けの行われていない太陽コーパスのような表記整理済みコーパスを学習に使用可能となった．

10. キーワード

(1) 校訂	(2) 歴史的資料	(3) 表記の標準化	(4) 形態素解析
(5) 歴史コーパス	(6)	(7)	(8)

11. 現在までの達成度

（区分）
（理由） 26年度が最終年度であるため、記入しない。

12. 今後の研究の推進方策

（今後の推進方策）
26年度が最終年度であるため、記入しない。

13.研究発表（平成26年度の研究成果）

〔雑誌論文〕 計(0)件　うち査読付論文 計(0)件

著者名	論文標題

雑誌名	査読の有無	巻	発行年	最初と最後の頁

掲載論文のDOI（デジタルオブジェクト識別子）

〔学会発表〕 計(1)件　うち招待講演 計(0)件

発表者名	発表標題
岡照晃, 松本裕治	形態素解析との同時最適化による歴史的資料の自動表記整理

学会等名	発表年月日	発表場所
情報処理学会研究報告 第216回自然言語処理研究会 第101回音声言語情報処理研究会 合同研究会	2014年05月22日～2014年05月22日	東京工業大学

〔図書〕 計(0)件

著者名	出版社

書名	発行年	総ページ数

14.研究成果による産業財産権の出願・取得状況

〔出願〕 計(0)件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	出願年月日	国内・外国の別

〔取得〕 計（0）件

産業財産権の名称	発明者	権利者	産業財産権の種類、番号	取得年月日 出願年月日	国内･外国の別

15.備考

（注）･印刷に当たっては、A4判（縦長）･両面印刷すること。