様式 C-7-2

自己評価報告書

平成２３年　５月２５日現在

機関番号：32602
研究種目：若手研究（Ｂ）
研究期間：2008～2011
課題番号：20700224
研究課題名（和文）様々な種類の文書に対応した汎用性の高い著者推定手法

研究課題名（英文）　Robust method for authorship attribution applicable to various document types

研究代表者
安形　輝（AGATA TERU）
亜細亜大学国際関係学部准教授
研究者番号：80306505

研究分野：図書館・情報学

科研費の分科・細目：情報学・図書館情報学・人文社会情報学

キーワード：著者推定、圧縮プログラム

１．研究計画の概要

目的は、異なるタイプの文書、長さの異なる文書（特に長さの短い文書）、異なる言語の文書を対象として著者推定実験を行い、実験結果に基づく改善をこの手法に加えることで、より汎用性が高い著者推定手法を開発することである。さらに、今回、構築する著者推定実験用テスト集合を公開することで、著者推定実験の標準化に貢献する。

２．研究の進捗状況

著者推定実験対象とする文書は具体的には(1)真贋が問題となっている福沢諭吉の新聞論説記事、(2)電子メール、(3)コミック資料、(4)ソフトウェアのプログラムである。このうち、(1)福沢諭吉関係の新聞論説記事については調査対象の選定、原資料の入手、画像データの作成までは完了している。現在は画像データから OCR データに基づき、テキストデータを入力している。すでにデータの入力が済んだ部分については試行的な実験を行った。

３．現在までの達成度

③やや遅れている。

理由は、当初の実験対象文書である福沢諭吉関連の新聞論説記事のテキストデータの作成が容易でなかったことにある。計画では「福沢諭吉全集」の画像データから OCR によりテキストデータを作成する予定であった。しかし、全集に収録される時点で原資料である「時事新報」記事に編集が加わっていることが明らかとなり、全集の元資料である「時事新報」の複写からテキストデータを作成するよう計画を変更した。この資料は非常に古い印刷物の複写しか入手できないため、画質が荒く、旧字体の活字が用いられている。そのため、OCR によるテキスト抽出はほぼ不可能に近い。結果として、一部の OCR 結果に基づき、ほとんどをアルバイトが入力する作業が発生している。

４．今後の研究の推進方策

研究期間中に作業が完了するかの判断がつかない福沢諭吉関係の新聞論説記事だけでなく、並行して、電子メール、コミック資料、プログラムなどの異なるタイプの文書に関するデータ集合を整備することで実験作業を進めていく。

５．　代表的な研究成果

（研究代表者、研究分担者及び連携研究者には下線）