書き手の特徴を利用した学生レポートチェックシステムの構築

太田 貴久　　増山 繁
豊橋技術科学大学 知識情報工学系
{kikyu, masuyama}@smlab.tutkie.tut.ac.jp

1　はじめに

現在，教育現場において，学生レポートの剽窃（コピー&ペースト）や代筆は教師にとって深刻な問題となってきた．このような学生レポートの不正（剽窃）に対して，日本においても，[1] をはじめ，様々なソフトウェアが発表され話題となっている．これらのソフトウェアは，基本的に剽窃を含む学生レポートとその剽窃元となる文章が共に機械処理可能な状態で存在することを前提としている．しかしながら，書籍からの不正な引用や代筆が行われた場合，これらのソフトウェアでは不正を検出することは不可能である．そこで，本研究では，従来から用いられてきた文書間の類似部分を検出する機能に加え，電子化されていない剽窃元に対応するために，レポートが提出した学生自身によって記述されたか否かを判断するような機能を有する，教師の採点を補助するレポートチェックシステムを提案する．

2　関連研究

今日までに様々な学生レポートの不正を検出する研究がなされてきたが，レポートから書き手を推定することで不正を発見しようとする研究は存在しない．従来用いられてきた方法は，日本語を対象とした文書間の類似性を測る方法であり，本システムでも確実なレポートの不正検出法として利用している．このような研究としては，[2, 3, 4, 5] がある．これらの研究のうち，tf·idf を用いたベクトル空間法 [2] や，文字 n-gram による方法 [3, 4] は文章全体の類似度を求めるに留まっているため，部分的に剽窃が行われた場合に対応できないという問題がある．また，文章を文単位に分割した上で単語の頻度ベクトルを用いて類似部分の重心と分散を測る方法 [5] でも，単一の部分的な剽窃には対応できるが，複数の部分的な剽窃（最初の段落と最後の段落のコピーなど）に対応できないという問題がある．

これに対して，本システムでは，Bioinformatics で用いられる Smith-Waterman アルゴリズム [6] をベースにした手法を用いる．Smith-Waterman アルゴリズムは分子の部分マッチングを行う古典的なアルゴリズムである．文を単位として Smith-Waterman アルゴリズムを適用することで，語順を柔軟に入れ替えることができる日本語に対応しつつ，具体的な剽窃箇所を同定することが可能となり，複数の部分的な剽窃にも柔軟に対応が出来る．

3　レポートチェックシステム概要

先にも述べたように，本システムは大きく分けて 2 つの機能を有する．1 つは，従来から用いられている 2 文書の比較による類似部分検出である．もう 1 つは，提出されたレポートが本人によって記述されたか否かを分析する機能である（以下，本機能を「文体検査」と呼ぶ）．それぞれの機能の概要を以下に説明する．

3.1　類似部分検出

1 つ目の機能は，従来から存在する 2 文書間（レポートとその剽窃元）に含まれる類似部分を発見する機能である．本システムでは，太田らの手法 [7] を改良した手法を用いて類似部分を検出する．本システムでは，ある 2 文書 X, Y 中の文 x_i, y_j（$x_i \in X, y_j \in Y$）のペアがどの程度類似部分の末尾に成り得るかを求める計算式を，次のように変更した．

$$S_{i,j} = \max \begin{cases} 0 \\ s(x_i, y_j) - \theta + S_{i-1,j} \\ s(x_i, y_j) - \theta + S_{i,j-1} \\ s(x_i, y_j) - \theta + S_{i-1,j-1} \end{cases} \quad (1)$$

ここで，$s(x, y)$ は文 x_i と y_j の類似度であり，θ は文が似ているか否かを決定するしきい値を表す．このように変更することで，文を分割して剽窃した場合でも，太田らの手法 [7] より適切に類似部分の末尾を検出することが可能となる．

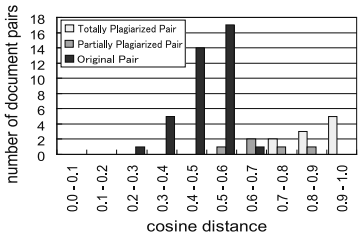
ここで，本手法の有効性を確認するために，簡単な実験を行った．実験では，あるレポートに対する，1. 似ていないレポート (Original Pair) と，2. 全体を剽窃したレポート (Totally Plagiarized Pair)，さらに，3. 部分的に剽窃したレポート (Partially Plagiarized Pair)，の類似度について，本手法と tf·idf に基づく従来手法 [2] により計測した．その結果を図 1 に示す．

図 1-(a) のように，本システムの手法は剽窃を行っていないレポートと剽窃を行っているレポートが完全に分離され，剽窃を正しく検出することに成功している．これに対して従来手法は，図 1-(b) のように，剽窃を行っていないレポートと剽窃を行っているレポートが [0.5, 0.7] の区間で混在するため判別不能となっている．

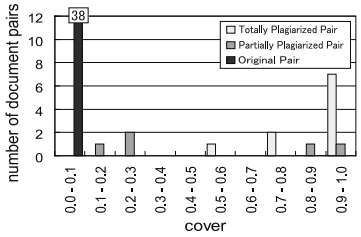
本システムでは，提出されたレポートと Web 上から取得した参考／引用文献と予測される文書のすべての可能な 2 つ組に対して上記の類似部分検出手法を用いることで剽窃を検出する．

3.2　文体検査

2 つ目の機能はレポートに見られる書き手の特徴を分析し，そのレポートを提出者本人が記述したか否かを判定する文体検査機能である．文体検査は太田らの手法 [8] をベースに行われる．太田らの手法 [8] は，「読点の出現頻度」，「助詞の trigram」，「自立語の品詞」を書き手の特徴とし，書き手がこれらの特徴をどの程度重要視しているかを最大エントロピー法により学習し利用している．太田らの手法 [8] は，



(a) tf·idf-based method [2]



(b) Our Method

図 1: 実験結果

ある入力 i（o の深層構造）から出力 o（実際の文）が生成される確率を以下の式で定義している．

$$P(o|i) = \frac{1}{Z(i)} \exp\left(\sum_k w_k v_k(o,i)\right) \quad (2)$$

$$Z(i) = \sum_{o \in \mathrm{Gen(i)}} \exp\left(\sum_k w_k v_k(o,i)\right) \quad (3)$$

ここで，w_k が制約の重みを表す．また，$v_k(o,i)$ は入力 i のもとで出力 o が制約 c_k に違反する数，$\mathrm{Gen}(i)$ は入力 i に対する出力候補の集合を表す．さらに，太田らの研究 [8] では，文章 S の生成確率 $P(S)$ を次のように定義している．

$$P(S) = \prod_{s \in S} P(s|s') \quad (4)$$

ここで，s は文章 S 中の文を，s' は s に対応する入力を表す．しかしながら，深層構造を実際の文から推定することは困難を極める．そこで，本手法では，s' は文 s までに出現した名詞の集合とした．

本システムでは太田らの手法 [8] からの変更点として，書き手の特徴を，金の研究 [9, 10, 11] を参考に,「読点の位置」,「助詞の trigram」,「品詞の遷移」の 3 種類に変更した．最終的には，本システムでは上記 $P(S)$ を，レポートが提出者本人によって記述された確率として扱い，この値を採点者に提示する．

本手法に対する簡単な実験として，1000 名分のブログに対して適用した．[1]この結果，生成確率の高い上位 1 人に正しいブログの書き手が来る精度は，80.1% を達成した．ただし，より文体の制限が強いレポート—ブログよりもレポートの方がフォーマルな文体が求められる—を対象にした場合，さらに精度が低下すると予測できる．

[1]本来ならば，類似部分検出手法の実験と同様にレポートを対象に行うべきだが，データ不足のため今回はブログを利用した

4 おわりに

本研究では学生レポートに含まれる不正を検出するための 2 つの方法を用いる学生レポートチェックシステムを提案した．2 つの方法の 1 つ目は従来から用いられてきた 2 文書間の類似部分を発見する方法で，もう 1 つはレポートに見られる書き手の特徴を解析する方法である．これらの方法を併用することで，従来は検出できなかった様々なレポートの不正を検出することが可能となる．ただし，現状では，学生レポートのデータ数が少ないため，文体解析において十分な精度を達成できるか否かは不明である．今後，更なるデータ収集を行い，より精度の高い文体解析を目指す．

謝辞

本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の支援により行われた．

参考文献

[1] 株式会社アンク : “コピペルナー”, http://www.ank.co.jp/works/products/copypelna/.

[2] 小河, 岩堀, 岩田 : “情報メディア教育における類似レポート判定システムの構築”, 平成 13 年度電気関係学会東海支部連合大会講演論文集, **604**, p. 304 (2001).

[3] 村田, 黒岩, 高橋, 白井, 小高, 小倉 : “学生レポートの n-gram による類似度評価の検討”, 情報科学技術フォーラム（FIT）2002 講演論文集, pp. 101–102 (2002).

[4] 高橋, 宮川, 小高, 白井, 黒岩, 小倉 : “Web サイトからの剽窃レポート発見支援システム”, 電子情報通信学会論文誌, **J90-D**, 11, pp. 2989–2999 (2007).

[5] 深谷, 山村, 工藤, 松本, 竹内, 大西 : “単語の頻度統計を用いた文章の類似性の定量化”, 電子情報通信学会論文誌, **J87-DII**, 2, pp. 661–672 (2004).

[6] T. F. Smith and M. S. Waterman: “Identification of common molecular subsequences”, Journal of Molecular Biology, **147**, pp. 195–197 (1981).

[7] 太田, 増山 : “学生レポート採点支援のためのレポート類似部分発見手法”, 電子情報通信学会技術研究報告, pp. 37–42 (2006).

[8] 太田, 増山 : “青空文庫を対象とした書き手の識別とその応用”, 言語処理学会第 15 回年次大会発表論文集, pp. 679–680 (2009).

[9] 金 : “読点の打ち方と著者の文体特徴”, 計量国語学, **19**, 7, pp. 317–330 (1994).

[10] 金 : “助詞の n–gram モデルに基づいた書き手の識別”, 計量国語学, **23**, 5, pp. 225–239 (2002).

[11] 金, 村上 : “ランダムフォレスト法による文章の書き手の同定”, 数理統計, **55**, 2, pp. 255–268 (2007).