E-19

文章の類似性の定量化
Evaluating Similarity between Texts

深谷 亮† 山村 毅‡ 竹内 義則† 松本 哲也† 工藤 博章† 大西 昇†
Ryo Fukaya Tsuyoshi Yamamura Yoshinori Takeuchi Tetsuya Matsumoto Hiroaki Kudo Noboru Ohnishi

1. はじめに

近年のコンピュータやインターネットの普及により，文章情報を公に発信することやその情報を手に入れることが容易になった．しかしそのような利便性の反面，コピー&ペーストによる文章の複写や加工が容易になったため，著作者が気づかないところで文章が流用され，盗作が生まれやすい環境であると言える．そこである文章が他の文章の盗作であるかどうかを判断できるシステムがあれば，盗作の発生の抑制につながるものと考えられる．本研究では，文章の類似性を定量的に評価する方法を提案する．

以下ではまず，他人の文章を真似して書かれたと考えられる文章を調査し，使われた真似の手法を挙げる．次に真似した文章の定義を明確化し，その手法に対応できる文章間の距離計算法を提案する．そして真似して書かれた文章にその距離計算法を適用し，有効性を示す．

2. 真似した文章と文章間距離の定義

他人の文章を真似して文章を書くときにどのような言い換えの手法が使われるかを調べ，真似した文章の定義を明確化するため，真似した文章が含まれると思われる学生レポートを約 200 部調査した．その中で一方が他方を真似したと考えられる文章同士は，内容についてほぼ全てをもしくは一部を写したものであった．その上で次のような文章の変形が見られた．

(1) 単語，文節の省略および追加
(2) 同義語，類義語への変換
(3) 漢字からひらがなへの変換およびその逆
(4) サ変動詞「する」の省略および追加
(5) 括弧を省略および追加した言い換え
(6) 読点の省略および追加
(7) 「AはBである」を「BはAである」に変換
(8) 助詞の省略および追加
(9) 引用符の省略および追加
(10) 態の変換
(11) 文節の順序の入れ替え
(12) 2文を1文に結合およびその逆

よって，もとの文章の全てもしくは一部に対しこれらような表層的な変化をさせた文章を，真似した文章と考えることができる．本研究における文章間距離は 2 つの文章の一方が他方を真似したものである場合に，小さな値となるものである．

3. 文章間の距離計算法

†名古屋大学大学院工学研究科

‡愛知県立大学情報科学部

文章間の距離計算法としては，全く同じ 2 つの文章について最小の値を与えるものとする．また，もとの文章に対し上で列挙した手法を用いて表層的な表現を変化させた文章についても，文章間の距離に大きな影響を及ぼさない計算法である必要がある．そのような計算法を実現すれば，2 つの文章の一方が他方を真似したものかどうかについて，人間の判断に準ずる文章間距離を計算することができると考えられる．

言い換えの手法の(4)～(12)に注目すると，これらの手法を適用する前後では文章中の単語の数，とりわけ名詞や動詞の数についてはほとんど変化がないことがわかった．そこで参考文献[1, 2]に見られるように，文章の類似性を判断する上で文章の特徴として従来からよく用いられてきた単語の頻度を本研究においても用いることにした．

また，言い換えの手法(2)が多用された場合，表現上まったく同じ単語同士をマッチングするだけでは，文章間距離は大きくなってしまうと予想できる．よってある単語の類義語が存在した場合，それらはマッチングされることが妥当であると考えられる．この問題に対しては辞書を用いて類義語や同義語の判定を行うことにした．

以上の事を踏まえ，次のような文章間距離の計算法を提案する．まず 2 つの文章 A，B に対し JUMAN Ver.3.61[3] により形態素解析を行い，単語に分割する．その結果から文章 A，文章 B に含まれる名詞と動詞の集合をそれぞれ W_A，W_B とする．次に W_A，W_B の和集合を求めることにより，文章 A，文章 B の少なくともどちらかに現れる単語リスト $W_A \cup W_B$ を作成する．そしてこの単語リストに対し，類義語の関係となる単語同士をまとめてゆくことにより，類義語のクラスタ c_i を求める．単語リストに含まれる単語を w_i，類義語のクラスタの集合を $W_A \cup^* W_B$，とし，以下に $W_A \cup^* W_B$ を求めるアルゴリズムを示す．

[1] $c_i = \{w_i\}$，$w_i \in W_A \cup W_B$ $(i = 1,2,\cdots,n)$
$W_A \cup^* W_B = \{c_1, c_2, \cdots, c_n\}$
[2] c_i の要素のいずれかと c_j の要素のいずれかが類義語である c_i，c_j の組を見つける．存在しなければ終了
[3] $c_i = c_i \cup c_j$ とする
[4] c_j を $W_A \cup^* W_B$ から取り除き，[2]へ

ここで 2 つの単語が類義語かどうかを判定する処理について説明する．2 つの単語の意味的な類似性を定量化し，閾値処理により類義語の判定を行う．この処理では，EDR 概念辞書[4]を利用する．この EDR 概念辞書は一種のシソーラスと見なすことができる．単語間の距離を計算する手法はいくつか提案されているが，ここでは最も一般的な手法を参考にした[5]．単語 i，j の語義 x，y のルートか

らの距離（深さ）をd_{ix}，d_{jy}，それらの共通の上位概念の深さをd_{xy}としたとき，2つの単語の類似度S_{ij}を，

$$S_{ij} = \max_{x,y}\left(\frac{d_{xy}\times 2}{d_{ix}+d_{jy}}\right) \quad (1)$$

と定義する．この類似度は，2つの単語が同じ単語である場合に1，全く無関係である場合に0になる．本研究では経験的に$S_{ij} > 0.85$となる単語同士を類義語と判定することにする．

2つの文章A，Bに対し，先に求めた類義語のクラスタc_iに含まれる単語について，頻度を求める．そして文章A，Bの特徴ベクトルf_A，f_Bを次のように定義する．

$$f_X = (h_X(c_1), h_X(c_2), \cdots) \quad (2)$$

ここで$h_X(c)$は文章X（=A, B）における類義語のクラスタcに含まれる単語の頻度の和を表し，$c_i \in W_A \cup^* W_B (i=1,2,\cdots)$である．そして文章Aと文章Bの間の距離は$f_A$，$f_B$の差の1ノルムとする．つまり，

$$|f_A - f_B|_1 = \sum_{k=1}^{n} |a_k - b_k| \quad (3)$$

と計算する．この文章間距離は文章Aと文章Bで同じ名詞と動詞が同じ回数使われている場合や，一方が他方の単語を単純な類義語に変換しただけである場合に対しては最小値0となり，使われている名詞と動詞が全く違う場合に対しては最大値2となる．

4. 実験と考察

実験では，ある文章とそれを真似した文章の距離と，ある文章とそれとは無関係の文章の距離とを比較することで，本手法の評価を行う．2つの文章（文章1，文章2）を用意し，それぞれを15名の人に真似して書いてもらって真似た文章のデータを作成した．また，小説，ニュース等から文章1，文章2とほぼ同じ長さの文章を10例ずつ選び出して，無関係な文章のデータを作成した．

このようにして集めた文章から次のような4つのデータセットA，B，C，Dを用意した．

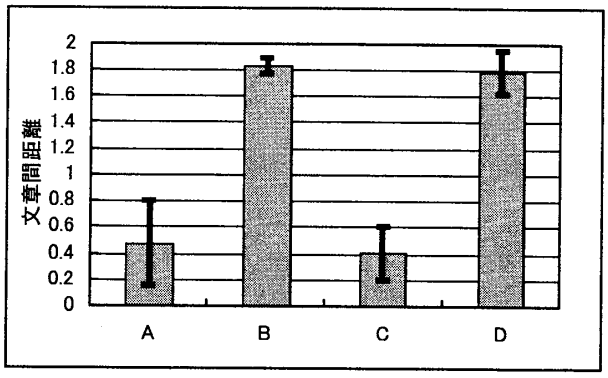

図1　文章間距離の平均と標準偏差

A：文章1とそれを真似した文章の組（15組）
B：文章1とそれとは無関係な文章の組（10組）
C：文章2とそれを真似した文章の組（15組）
D：文章2とそれとは無関係な文章の組（10組）

これらのデータセットに対し本研究の文章間距離計算をした結果を平均と標準偏差により図1に示す．

図から明らかなように，オリジナルの文章とそれを真似した文章の組の文章間距離は小さな値となったが（A，C），オリジナルの文章とそれとは無関係な文章の組の文章間距離は大きな値となり（B，D），明確な差が生じた．

真似た文章のデータの中にはオリジナルの文章の内容をほぼ丸写ししたものもいくつかあった．しかし一方で，オリジナルの文章とは文の出現順序等が大幅に変えられていたものもあった．このような文章に対しても，本研究における距離計算法は良好な結果を算出していた．これは文章の特徴として単語及びその類似語の頻度を用いているからである．このことから，本研究の距離計算法が有効性であることがわかる．

5. おわりに

本研究では学生レポートを調査し，他人の文章を真似するときに使われる手法を列挙した．真似した文章を明確化し，その手法に対応できるように単語の頻度，類義語判定を利用した文章間の距離計算法を提案した．そして真似した文章に対し，本手法の有効性を示すことができた．

ただ，現在の距離計算法は文章全体における単語の頻度を比較しているため，一部分だけを真似した文章については良好な結果は得られない．これについては何らかの形で部分マッチングを行う必要があると考える．また本研究のアプリケーションとしては，我々学生にとって身近な問題として，多数の学生レポートの中から他人のレポートを真似して作られたものを発見するシステムが考えられる．学生レポートでは真似していない文章同士でもある程度文章間距離は近くなることや，与えられた課題によって文章間距離が変化することが予想される．よってその中から疑わしいレポートを見つけるためには統計的な知識を利用して，学生レポートの組み合わせの中から文章間距離が外れ値となるもの判断できなければならないと考える．

参考文献

[1] 谷村正剛，田中（石井）久美子，中村裕志：異なる発信元からのWWWニュース記事の内容に基づく対応付け，情処学NL研報，146-14，pp.89-94，2001
[2] S. Singh, F. J. Tweedie: Neural Networks and Disputed Authorship: New Challenges, 'Artificial Neural Networks' 26-28 June 1995 Conference Publication No. 409, IEE, pp.24-28, 1995
[3] 黒橋禎夫，長尾真：日本語形態素解析システムJUMAN version 3.61，京都大学大学院情報学研究科，1998
[4] 日本電子化辞書研究所：EDR電子化辞書仕様説明書，日本電子化辞書研究所，1995
[5] 長尾真：自然言語処理，岩波書店，1996