自 然 言 語 処 理 153－10
（２００３．１．２０）

頻度統計と概念辞書を用いた文章の類似性の定量化

深谷 亮† 山村 毅‡ 工藤博章† 松本哲也† 竹内義則† 大西 昇†

† 名古屋大学大学院情報工学専攻

(〒 464-8603 名古屋市千種区不老町, TEL 052-789-3811)

‡ 愛知県立大学情報科学部

(〒480-1198 愛知県愛知郡長久手町大字熊張字茨ケ廻間 1522-3, TEL 0561-64-1111)

E-mail: † {fukaya, kudo, matumoto, takeuchi, ohnishi}@ohnishi.nuie.nagoya-u.ac.jp
‡ yamamura@ist.aichi-pu.ac.jp

概要:
本研究では，他人の文章を真似して作成された文章を発見するための文章間類似度の計算法を提案する．真似した文章の多くは，もとの文章に含まれる文と類似した文から構成され，類義語・同義語へ言い換えることなどにより表層的な表現を変化させる．そこで，本手法では各文章を構成される文単位で照合し，表層的な表現の変化に対応するため単語の頻度と概念辞書を用いる．本手法による類似度により，同一テーマで記述された文章と真似して書かれた文章とを明確に区別することができることを示す．

キーワード 文書処理 類似性 概念辞書 単語の頻度

Measuring Similarity between Documents using Term Frequency and Concept Dictionary

Ryo Fukaya† Tsuyoshi Yamamura‡ Hiroaki Kudo† Tetsuya Matsumoto† Yoshinori Takeuchi† Noboru Ohnishi†
†School of Engineering, Graduate School, Nagoya University
(Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan, TEL 052-789-3811)
‡Faculty of Information Science and Technology, Aichi Prefectural University
(1522-3 Ibaragabasama, Kumabari, Nagakute-cho, Aichi-gun, Aichi 480-1198, Japan, TEL 0561-64-1111)

E-mail: † {fukaya, kudo, matumoto, takeuchi, ohnishi}@ohnishi.nuie.nagoya-u.ac.jp
‡ yamamura@ist.aichi-pu.ac.jp

ABSTRACT:
In this research, we propose a method of calculating the similarity between documents for identifying the imitated document. Many of imitated documents consist of the sentences similar to the sentence contained in an original documents, and their surface-expression changes by transposing to a synonym etc. Then, our method compares two documents in the sentence unit that constitutes each document by using term frequency and concept dictionary. We show that we can distinguish clearly the document described by the same theme, and the imitated document by using our method.

Keywords document processing, similarity, concept dictionary, term frequency

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り，そのため，著作者が気づかないところで文章が流用され，オリジナリティの存在しない真似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真似であるかどうかを判断できるシステムがあれば，真似の発生の抑制につながるものと考えられる．そのためにはまずコンピュータが，真似かどうかという観点から文章の類似性を定量的に扱える必要がある．

過去の研究において，ある文章が他人の文章の真似であるかどうかを判断することを目的としたテキスト分析の試みは少ない．テキスト分類において主題分類や著者分類などのための尺度とは異なる，様々な文章の類似性の評価方法を確立する必要性という観点から，本研究は新たなアプローチを提案するものでもある．

これまでに真似であるかどうかを判断することを目的とした研究としては，文献[1][2]が挙げられる．それらの研究では，学生レポートを処理対象としている．文献[1]ではｎグラム解析を用いた類似度計算方法を提案し，一方，文献[2]では $tf \cdot idf$ とコサイン尺度を用いたベクトル空間法による類似度計算法を提案している．いずれの研究においても，多少手が加えてある

程度の真似においては良好な結果を残している．しかし，真似した文章においてよく見られる，もとの文章に存在する単語を類義語・同義語に変換するということが行われていた場合，それだけで類似度が小さくなってしまう問題がある．また，もとの文章の抄録という形での真似や部分的に真似が含まれる場合における検討はなされていない．

本論文でははじめに，一般的な類似文と真似の違いについてその考えを述べる．次に他人の文章を真似して書かれた文章を調査し，使われた真似の手法を細分化する．そして，その手法に対応できる文章間の類似性計算法を提案する．最後に真似して書かれた文章にその類似性計算法を適用し，有効性を示す．

2. 類似した文章

一口にある文章とある文章が似ていると言う場合においても，その似ていることについては様々な解釈の仕方がある．本研究では主に次のように，類似文章を分類することができると考える．

(1) 分野が類似している文章:
同じ分野に分類できる．その表現や論述の様式は様々である．

(2) 概念レベルで類似している文章:
描写されている事柄に違いはあっても，その根底にある概念 (ストーリー) は同じである．

(3) 記述内容が同一である文章:
同一の事柄を述べているが，その表現や論述の様式は様々である．例えば，新聞報道の記事やオリジナルの学生レポート．

(4) 学生レポートに見られるような真似の文章:
一般的にコピーと認知される．もとの文章の表現を表層的に変化させることで作られる．

従来，このように分類される類似文章のうち，文献[3][4]などにおいて行われている文書クラスタリングの研究で処理対象とされてきた類似文章は(1)，(3)に分類されるものである．一方，本研究では(4)に分類される真似の文章を取り上げる．そして，真似である文章と，その他に分類される類似文章とを明確に分離する文章間の類似性計算法を確立することを目的とする．

そのためには，真似して作成された文章ともとの文章との差異に注目し，その変化に対応する形で文章の類似性を計算するアプローチが考えられる．そこで，他人の文章を真似して書かれた文章を観察し，変化の様子を調査した．他人の文章を真似した例を図 1 に示す．

もとの文章：
日本語文の形態素解析の目的は，文章を辞書に登録されている単語の可能な組み合わせで分割し，構文上の役割を決定することである。問題は，英語やフランス語といった言語と違い，日本語は単語の間に空白をいれずにべた書きすることや，分かち書きや送り仮名などに関して曖昧さがあるので，解析が困難であることである。膨大な解析時間がかかることや辞書に登録されていない単語への対策なども考慮しなければならない。

真似した文章：
日本語の文章における形態素解析の目的は、辞書に登録されている単語の可能な組み合わせに文章を分割して、構文上における役割を決めることである。問題は、英語やフランス語などの言語とは異なり、日本語は単語間に空白を入れないでべた書きすること、あるいは分かち書きや送り仮名などに対して曖昧さがあるため、解析が難しいことである。解析に膨大な時間を要することや、辞書に登録されていない単語への対策なども考慮すべきである。

図 1　他人の文章を真似した例

人間が真似だと判断するような文章は，真似を作成する人がもとの文章に記述されているテーマを本質的に理解しておらず，表層的な文章表現の多様性を利用して作成されることが多い．そのため，図 1 で見て取れるように表面的な言い換えがなされているだけで，文単位での対応付けが可能であるものが数多く見られる[1]．しかし，より一般的にはもとの文章に書かれている内容について元々持ち合わせている知識や，表現を書き換える際に頭を使う程度によって，真似した際に生じる変化は様々である．これを表 1 に示す．

表 1　知識により分類される真似の手法

もとの文章に対する知識や頭を使う程度	使われる主な手法
知識が無い，もしくは頭をほとんど使っていない (レベル 1)	• ピリオド (.) を句点 (。) に • です／ます調をだ／である調に • 「」をはずす
文法レベルの知識を用いる (レベル 2)	• 文や文節の出現順序を換える • 一般的な単語を類義語・同義語へ変換 • 2 文を 1 文に • 冗長な語を追加／省略
単語レベルの知識がある，どれが重要な文であるかわかる (レベル 3)	• 専門的な単語を類義語・同義語へ変換 • 抄録 (重要な文の抜き出し)

[1] 実際には真似した文章の中には文単位で対応付けができないものも存在する．しかしそのような文章は人間が見ても真似かどうかの判断は困難であるため，本研究では処理の対象外とした．

表 1 では知識によって真似の手法を分類しているが，実際の真似した文章では複数のレベルにまたがってもとの文章の言い換えがなされる．つまり，ひとつの文章中でも理解しやすい個所とそうではない個所とでは使われる真似の手法は異なる．そのため，理解できる個所は自分の表現で書き，理解しがたい個所のみを真似するといった部分的な真似が行われることもある．

これらの真似した文章における特徴を踏まえ，真似の手法による表現の変化に対応できる類似度計算法を提案する．

3. 文章間類似度計算法

3.1 類似度計算の手順

文章の類似性を比較するにあたり問題となることは，文章の特徴としてどのようなものを用いるかということと，それらの特徴から類似性をどのように判断するかということである．文書クラスタリングなどの研究では，文章の特徴として，n グラム頻度や $tf \cdot idf$ ，エントロピーなどがよく用いられてきている．真似かどうかを判断する問題においては，表 1 に示されるレベルにより，その特徴が大きく変わらないものが望ましい．日本語は英語などの言語と比べ語順が比較的自由であり，真似においてもその特性に基づいて語順を入れ替えることが頻繁に行われる．また，付属語である助詞や助動詞は，文章の内容を決定付けるものとはなりにくく，複数の語が文法上の同じ役割を持つため言い換えの対象となりやすい．本手法では，それらの手法による表現の変化に左右されない文章の特徴として，名詞と動詞の頻度を用いることにする．

次に，その特徴から類似性をはかる方法について述べる．抄録の形で行われる真似や部分的に真似が含まれる場合，文章全体から得られる特徴同士の比較では，真似かどうかの判別は困難であると考えられる．そのため，文章全体よりも細かいレベル，すなわち文章を構成する段落や文に注目する必要がある．そこで本手法では文単位でまず類似性を判断し，文章全体における類似文の数や割合，出現位置から文章全体の類似性を判断することにする．

本研究では 2 つの文章 X，Y 間の類似度は以下の手順により計算することにする．

(1) 2 つの文章 X，Y を文に分割する．
(2) 各文に対し形態素解析を行い，各文に含まれる名詞と動詞を取り出す．
(3) 文章 X を構成する文と文章 Y を構成する文の全ての組み合わせに対し文間の類似性(距離)を計算する．
(4) 文間の距離から類似文を決定する．
(5) 類似文の数や割合，出現位置から文章全体の類似性を計算する．

上記の手順を以下で詳しく述べる．

3.2 文への分割

2 つの文章 X，Y を句点 (。) やピリオド (．) により文に分割する．真似する際に使われる手法として，2 文で構成される部分を，例えば接続詞を用いて 1 文に書き直す (表 1 のレベル 2) ということが行われうるため，本来ならば複文や重文で構成される文は単文に分割するようにして，文章全体を単文に分割することが望ましい．しかし，学生レポートのような文章を対象とした場合，文章としての程度が低い (文法的に誤りを含む) 可能性が高く，正しく分割されることを期待できないことが多い．さらに処理が複雑化することもあり，本手法では最も単純な方法で文を分割することにした．

3.3 文間の距離計算

基本的には文間の距離は 2 つの文に含まれる名詞と動詞の頻度が一致する度合いを見ることで計算する．ただそのままでは真似が作成される際に，表 1 で挙げた類義語・同義語への変換する手法が頻繁に行われると，文間の距離は大きくなり似ていないと判断されることになる．したがって真似かどうかの判断にあたっては，類義語・同義語同士は同一のものであると見なす方がよいと考えられる．2 つの単語が類義語・同義語であるかどうかを判断する方法は次節に述べる．以下では類義語・同義語を考慮した文間の距離計算法を説明する．

まず 2 つの文A，Bに対しJUMAN Ver.3.61[5]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A，文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A，文Bのうち少なくともどちらかに現れる単語のリスト $W_A \cup W_B$ を作成する．そしてこの単語リストに対し，類義語・同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・同義語のクラスタ $c_1, c_2, \cdots, c_n$ を求める．具体的には，単語リストに含まれる単語を w_i $(i = 1,2,\cdots,k; k \geq n)$ とするとき，類義語・同義語のクラスタの集合 $W_A \cup^* W_B$ を以下のアルゴリズムで求める．

[1] $c_i = \{w_i\}$ ， $w_i \in W_A \cup W_B$ $(i = 1,2,\cdots,k)$
$W_A \cup^* W_B = \{c_1, c_2, \cdots c_k\}$
[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・同義語である 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

次に 2 つの文 A，B に対し，求めた類義語・同義語のクラスタ $c_1, c_2, \cdots, c_n$ に含まれる単語について，各文

内における頻度を求める．そして文 A，B の特徴ベクトル f_A， f_B を次のように定義する．

$$f_X = (h_X(c_1), h_X(c_2), \cdots, h_X(c_n)) \quad (1)$$

ここで $h_X(c)$ は文 $X(=A,B)$ における類義語・同義語のクラスタ c に含まれる単語の頻度の和を表す．この f_X を用いて，文 A と文 B の間の距離 $Dis(A,B)$ は f_A， f_B の差の 1 ノルムで定義することにする．すなわち，

$$Dis(A,B) = |f_A - f_B|_1 = \sum_{k=1}^{n} |h_A(c_k) - h_B(c_k)| \quad (2)$$

と計算される．$\sum_{k=1}^{n} h_X(c_k) = 1$ であるため，$0 \le Dis(A,B) \le 2$ である．この文間の距離 $Dis(A,B)$ は文 A と文 B で同一の名詞と動詞が同じ回数出現しているか，もしくは一方が他方の単語を単純な類義語・同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異なる場合に対しては最大値 2 となる．

例えば次の文を考える．

文 A: **機械と違い，人間は真似を判別できる．**

文 B: **人間は機械と異なり，真似の文章と真似でない文章を判別できる．**

文 A，B に含まれる名詞と動詞の集合 W_A， W_B は，

W_A = {機械, 違う, 人間, 真似, 判別}

W_B = {人間, 機械, 異なる, 真似, 文章, 判別}

である．「違う」と「異なる」は同義語であるので，同一の類義語・同義語クラスタの要素となる．よって類義語・同義語のクラスタの集合 $W_A \cup^* W_B$ は，

$W_A \cup^* W_B$ = {{機械}, {違う, 異なる}, {人間}, {真似}, {判別}, {文章}}

となる．各文について類義語・同義語クラスタごとに頻度を求める．文 A，B の特徴ベクトル f_A， f_B はその頻度を要素に持つので，

$f_A = (0.2, 0.2, 0.2, 0.2, 0.2, 0)$

$f_B = (0.125, 0.125, 0.125, 0.25, 0.125, 0.25)$

文間の距離 $Dis(A,B)$ は f_A， f_B を用いて(2)式により，

$$\begin{aligned} Dis(A,B) &= |f_A - f_B|_1 \\ &= 0.075 + 0.075 + 0.075 + 0.05 + 0.075 + 0.25 \\ &= 0.6 \end{aligned}$$

と計算される．

3.4 類義語・同義語の判別

ここでは 2 つの単語が類義語・同義語である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・同義語の判定を行う．この処理では，EDR 概念辞書[7]を利用する．この EDR 概念辞書はシソーラスの一種である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[6]．すなわち，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 R_{ij} を，

$$R_{ij} = \max_{x,y}\left(\frac{d_{xy} \times 2}{d_{ix} + d_{jy}}\right) \quad (3)$$

と定義する．この類似度は 2 つの単語が同じ単語の場合に 1，全く無関係の場合に 0 になる．図 2 に EDR 概念辞書の一部と「食べる」と「食う」の単語の辞書中における位置を示す．(3)式により「食べる」と「食う」の類似度は 0.91 となる．本研究では経験的に $R_{ij} > 0.85$ となる単語同士を類義語・同義語と判定することにした．

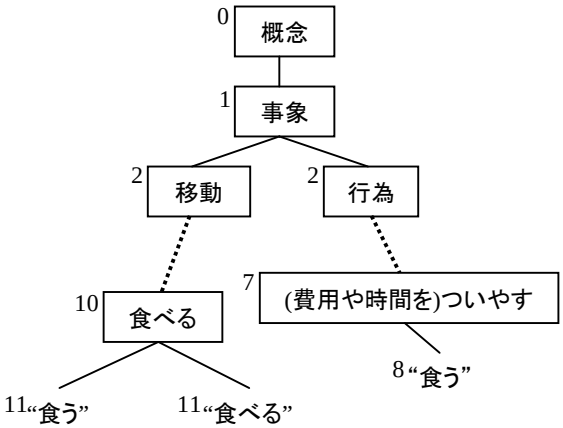


図 2　EDR 概念辞書の一部(数字はルートからの深さ)

3.5 文章間における類似文の決定

文章 X，Y を文に分割した後，文章 X，Y の間で作られる全ての文の組み合わせに対し，(2)式により文間の距離が求める．そして後の処理のために，どの文の組が類似文同士であるかを決定する．基本的には閾値より小さい値になる組を類似文の組とするが，1 つの文が複数の文と閾値以下の距離となることも考えられるため，文章 X，Y を構成する文をそれぞれ $X_1, X_2, \cdots, X_m$， $Y_1, Y_2, \cdots, Y_n$，類似文の判断基準となる閾値を T とするとき，次の手順により 1 対 1 に対応付けを行う．

[1]　$C_X = \{X_1, X_2, \cdots, X_m\}$， $C_Y = \{Y_1, Y_2, \cdots, Y_n\}$

[2]　$\min_{X_i \in C_X, Y_j \in C_Y} Dis(X_i, Y_j) < T$ となる X_i， Y_j を類似文とする．存在しなければ終了

[3]　X_i， Y_j を C_X， C_Y から取り除き，[2]へ

3.6 文章間類似度

文章 X と文章 Y の類似度の計算方法について述べる．まず初めに，文章 X を注目文章，文章 Y を比較文章とする．比較文章 Y が注目文章 X の単純な真似である場合，比較文章 Y には注目文章 X に含まれる文に類似の文が複数存在することが考えられる．類似文が数多く(高い割合で)存在するほどこれらの文章は類似している(近い)と考えられるが，単純に類似文の数(割合)のみで文章間類似度を決めてしまうと，真似ではないが類似文が偶然複数含まれている場合と，抄録の形で真似されていた場合もしくは部分的に真似されている場合を区別することができない．しかし，単に類似文が偶然複数含まれている場合ならば比較文章 Y 中では類似文は分散して存在するが，抄録や部分的な真似の場合では類似文がある程度近い場所に存在すると考えられる．そこで本手法では，比較文章 Y 中で類似文がどのように存在しているかを考慮に入れることにする．具体的には，比較文章 Y における(注目文章 X との)類似文の文章中での出現位置の標準偏差で評価する．

注目文章 X と比較文章 Y の間に類似文の組が n 組あるとし($n \geq 2$)，比較文章 Y に含まれる全ての文の数を n_Y ，比較文章 Y における類似文の割合を $r = n / n_Y$ とする．また，比較文章 Y における類似文の位置の重心からの標準偏差を σ とする．これらを用いて，注目文章 X に対する比較文章 Y の類似度 S_{XY} を，以下のように計算する．

$$S_{XY} = n^2 \cdot \frac{1}{\sigma} \cdot r \tag{4}$$

この類似度 S_{XY} は，類似文の組が多いほど(n^2)，類似文の数が同じでもそれらがまとまって存在しているほど($\frac{1}{\sigma}$)，比較文章 Y における類似文の割合が大きいほど(r)大きな値となる．

次に，先ほどとは逆に文章 X を比較文章，文章 Y を注目文章として類似度 S_{YX} を求める．文章 X と文章 Y の最終的な類似度 S は，2 つの類似度 S_{XY} と S_{YX} のうち大きい方とする．

$$S = \max(S_{XY}, S_{YX}) \tag{5}$$

文章 X と文章 Y の間で類似文が 1 つも存在しない場合($n = 0$)，もしくは類似文が 1 つのみ存在する場合($n = 1$)では，その文章同士が真似の関係である可能性が低いかまたは人間が見ても真似の判断が困難であると考えられる．したがって，類似度 $S = 0$ としてそれらの場合を表すことにする．

4. 類似度計算法の評価実験

4.1 評価用データについて

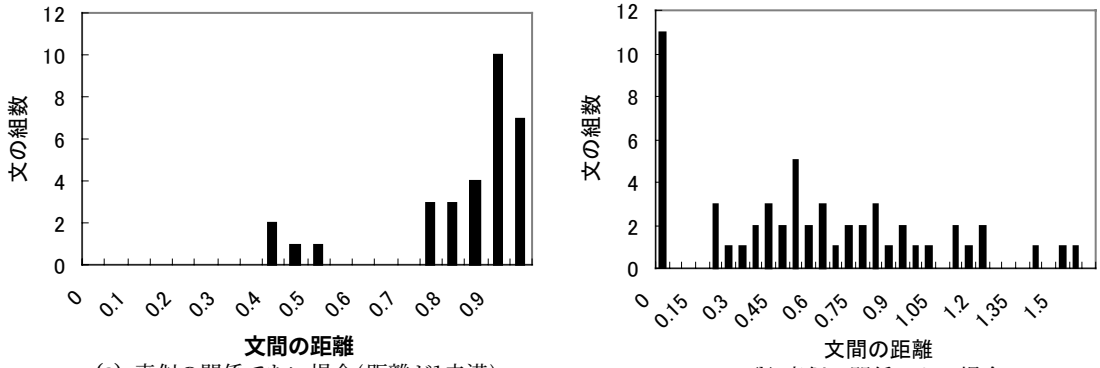
真似した関係にある文章の組と，同一のテーマで記述されているが真似ではない関係にある文章の組について提案手法を適用し，その有効性を評価する．そのためにまず，同一のテーマで記述されている文章データとして学生レポートを用意した．「日本語文における形態素解析の目的と問題を述べよ」という課題に対するレポートを 10 人に作成してもらい，10 部のオリジナルレポートを得た．そして，別の 8 人にその 10 部のレポートをもとに真似したレポートを 1 人 1 部ずつ作成してもらい，合計 8 部の真似したレポートを得た．図 1 に示されている文章はそれらのレポートの 1 つである．さらに，「目的」について述べる部分は他人のレポートの真似であるが，残りの「問題」について述べる部分は自分の言葉で書いた部分的な真似のレポートを 8 部用意した．そのため，部分的な真似のレポートには全て真似した場合のおよそ半分の類似文が存在する．これら用意したレポートは 3 文から 10 文によって構成される．

4.2 類似文判定における閾値

類似文判定における閾値 T は，真似の関係にある文と，そうでない文を最もよく分離するように定める必要がある．したがって，上記のデータに含まれる文の組み合わせに対し文間の距離計算を行い，適切な閾値を定めることとする．真似の関係ではない文の 5103 組の組み合わせに対して文間の距離計算を行った結果，1 未満の距離となった組の数を図 3(a)に示す．図 3(a)に表示されていない，1 以上の距離となった文の組は 5072 組ある．併せて，真似の関係となる 54 組の組み合わせに対する距離計算を行った結果を図 3(b)に示す．

8 部の真似したレポートのうち 6 部に，オリジナルで 2 文からなる部分を 1 文にする手法(もしくはその逆)が 1 部につき 1 回ずつ見られた．すなわち，真似の関係となる 54 組のうち 12 組は，その手法によって作り出される部分的な真似となっている組である．図 3(b)において，文間の距離が 1 を超える文の組は，多くが部分的な真似である組であった．よってそのような場合を除いては，真似の関係である文の組とそうではない組は概ねよく分離される．

一方，真似ではないが文間の距離が近くなってしまった場合でも，そのような文の出現する位置には偶然性が仮定できるため，後の文章間類似度計算を考慮に入れるとそれほど大きな問題とはならない．そのため，真似ではない文を真似であるとしてしまうリスクを多少負っても，より多くの真似の関係にある文を適切に判断する方がよいと考えられる．



(a) 真似の関係でない場合(距離が1未満)　(b) 真似の関係である場合

図 3　文間の距離計算結果

以上のことを踏まえ，類似文判定における閾値は $T = 0.9$ とし，それに満たない距離となるものを類似文とする．

4.3 文章間類似度の評価

文章間類似度の評価実験は，①オリジナルのレポート 10 部によって作られる 45 組の文章の組，②全てを真似したレポートとそのもととなったレポートによる 8 組の文章の組，③部分的に真似したレポートとそのもととなったレポートによる 8 組の文章の組に対して行った．それら①，②，③のデータを類似度計算した結果を表 2 に示す．

表 2　類似度 S によるデータ別の文章の組数

	データ①	データ②	データ③
$S = 0$	45	0	1
$0 < S \leq 5$	0	0	2
$5 < S \leq 10$	0	1	3
$10 < S \leq 15$	0	2	2
$15 < S$	0	5	0

オリジナルレポートの組み合わせであるデータ①については，類似度は全て 0 となった．このことはすなわち，$S > 0$ となる文章の組を真似であると判断してよいことを示す．また，データ②，③の結果から，類似文の数が多いほど類似度は大きくなり，人間の感覚に近いものとなった．その一方で，データ①，③の結果には (1 例を除いて) 明確な差が生じた．

4.4 考察

オリジナルのレポート同士の組み合わせであるデータ①の中にも，類似文と判断される文の組はいくつか存在した (図 3(a))．それにもかかわらず，全ての類似度が 0 となったことは，オリジナルレポートの間で類似文が偶然に複数存在することはめったに無いとした仮定が正しかったためであると言える．

また，文間の距離計算において概ね真似とそうでないものを分離できたことは，文の特徴として単語の頻度を用いたことや類義語・同義語判定の有効性を示すものだと言える．ただし，真似の関係である場合でも，比較的大きな距離となる場合が存在した (図 3(b))．以下に示す文 D は文 C の真似であるが，その間の距離は 1.06 となる．

文 C: **まず日本語を形態素に分解するとき，分解の仕方が文脈に依存するため，１通りではないことである。**

文 D: **最初の問題点として日本語の文章を形態素に分割するとき，その結果が文脈に依存してしまう為に一意に定まらないということが挙げられる。**

文 D は，「まず」を「最初の問題点として」に換えるという，文節レベルの言い換えがされている巧妙な例である．そのため，類義語・同義語判定ではそれらの対応をとることはできない．このような「まず」と「最初の問題点として」とを同一と見なすには，文脈に対する知識が必要とされるが，その実現は困難である．

5.おわりに

本研究では人が文章を真似する際の言い換えの手法を説明し，その手法に対応する文章間の類似性計算法を提案した．その計算法の主な特徴は次のとおりである．

- 抄録の形の真似や部分的な真似に対応するため，文単位による照合をする
- 単語を類義語・同義語へ変換する手法に対応するため，概念辞書を用いる

そして真似して書かれた文章にその類似性計算法を適用し，その有効性を示した．しかし，現在では以下のような問題点があると考えられる．

本研究で用いたデータでは各文は最大 10 文で構成され，文章数も 26 部と比較的小さいため，文章のサ

イズ，数ともにより大きな規模の文章集合での実験が課題である．現在，他人の文章の真似が含まれていると考えられる実際の学生レポートでの評価実験を検討している．

本手法は文単位での類似性に基づいて文章の類似性を評価しているため，2 文で構成される部分を 1 文にする言い換え手法が頻繁にされると，本質的に対応することができない．そこで，類義語・同義語判定における文節レベルでの言い換えの問題と併せて，文より細かいレベルである文節に注目することが考えられる．

参考文献

[1] 村田哲也，黒岩丈介，高橋勇，白井治彦，小高知宏，小倉和久: “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 (CD-ROM)，E-10，2002

[2] 小川貴博，岩堀祐之，岩田彰: “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] 谷村正剛，田中 (石井) 久美子，中川裕志: “ 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

[4] Lewis,D.D, Schapire,R.E., Callan,J.P., Papka,R: ”Training Algorithm for Linear Text Classifiers”, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-96), pp.298-306, 1996

[5] 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

[6] 長尾真: “ 自然言語処理 ”，岩波書店，1996

[7] 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