2M3-OS-34b-1

文章における書き方の特徴の抽出と理解

Features Extraction and Understanding in Ways of Document Writing

外村 耀平 *1 Youhei Tonomura　砂山 渡 *2 Wataru Sunayama　畑中 裕司 *2 Yuji Hatanaka　小郷原 一智 *2 Kazunori Ogohara

*1滋賀県立大学大学院工学研究科 Graduate School of Engineering, The University of Shiga Prefecture

*2滋賀県立大学工学部 School of Engineering, The University of Shiga Prefecture

A document has been used for the mutual understanding of each other in our daily life. People need to read the correct intention of a document and write a document that can express their intention. In this paper, we propose a system that helps the characteristics understanding of a document by using the interpretation of the elements obtained by the text mining techniques. Evaluation experiment showed the possibility that the system helps a user to understand the characteristics of a document.

1. はじめに

普段の生活において，我々は常に文章に触れている．情報を得るために新聞やニュースサイトを読み，コミュニケーションを取るためにブログや Twitter の SNS を利用し，娯楽や勉学のために書籍を手に取る．また，人に思いや情報を伝えるために，文章を書き起こす．

文章はお互いの意思疎通のために用いられている．自分が文章を書く場合においてはその意図が正しく伝わるように書き，読む場合には意図を正しく読み取る必要がある．

本研究では，文章の特徴をさまざまなテキストマイニングの技術によって抽出し，得られた結果の解釈をもとに書き方の特徴の理解を促す環境を構築する．これにより，文章の書き方の特徴を踏まえた上での文章作成や文章理解を促す．

小説やニュース記事，日記，ブログ，Twitter など，媒体とするものによって求められる文章というのは異なる．本稿では，論文を対象として，文章の特徴を抽出し，書き手の特徴理解を促す．

本システムによる文章の書き手側の利点として，自分の文章の書き方について，客観的に知ることができる機会は少ないと考えられることから，自身の文章の内省の支援となることが挙げられる．見た目ではわからない単語の頻度情報や文の長さなど，数値として示されて初めて意識できる指標がある．また無意識のうちに現れる文章の特徴として，よく使う表現や書き方の癖が提示されることで，自身の書き方の内省を行うことが可能になる．

また文章の読み手側の利点として，多くのレポートを読む必要がある教員や，論文や報告書の作成を指導する人にとっては，文章の客観的な特徴が提示されると，それらを参考に必要な指導を行うことが可能になる点が挙げられる．

加えて，これらの両者に共通する内容として，文章の特徴がただ列挙されるだけでは，逐次的な対応になり，根本的な解決につながらない可能性がある．そのため，得られた特徴の集合をまとめて抽象化する過程を設けることで，より根本的な文章の特徴を見いだす支援を行う．すなわち，文章の書き手としては，複数のことに気を配るよりも，根本的な 1 つあるいは少数のことに気を配って文章を書くことができ，また文章の読み手として文章作成の指導をする際にも，多くのことを語らずとも，根本的な少数のことを伝える指導が可能になる．

連絡先: 外村耀平，滋賀県立大学大学院工学研究科電子システム工学専攻，〒 522-8533 滋賀県彦根市八坂町 2500，ot23ytonomura@ec.usp.ac.jp

2. 関連研究

2.1 文章からの特徴抽出

ソフトウェアに文章を読み込ませて特徴を抽出，評価する取り組みとしては，E-rater[2] がある．英語の文章の採点用に用いられており，構造，組織化，内容の 3 つの観点から小論文を評価し，評価結果を 6 点満点で示す．実際の採点にも用いられているのだが，日本語の文章に対して利用することができない．また日本における取り組みとしては JESS[3] がある．JESS は文章作法を評価する「修辞」と，アイディアが理路整然と表現されていることを示す「論理構成」と，トピックに関連した語彙が用いられているかを示す「内容」の 3 つの観点から小論文を評価する．妥当な結果が得られるのが 800 字から 1600 字程度とされている．

2.2 文章作成支援

文章作成支援としては GUNGEN-Web[4] がある．機能としては次のような物である．テンプレートに沿ってアイデアを考え，他人にも理解しやすい具体的なアイデアを出す．出したアイデアを図解化・文章化するプロセスを KJ 法にしたがって作成できるインタフェースを提供する．KJ 法というのは，データをカードに記述し，カードをグループごとに整理し，情報を視覚化して文章にまとめる手法である．GUNGEN-WEB は Web 技術を用いるため，パソコンやタブレット端末など JavaScript が実行できる端末であれば利用できる．

3. 文章の書き方の特徴を抽出して理解するための枠組み

文章の書き方の特徴を抽出して理解するための枠組みを図 1 に示す．入力された文章から，さまざまな特徴量を取り出して利用者に提示し，利用者がその特徴量の意味を解釈して列挙し，最終的に解釈を一つ（あるいは少数）にまとめて知識とする．

3.1 TETDM

本研究では，テキストマイニングのための統合環境 TETDM[1] をベースに環境を構築する．TETDM はテキス

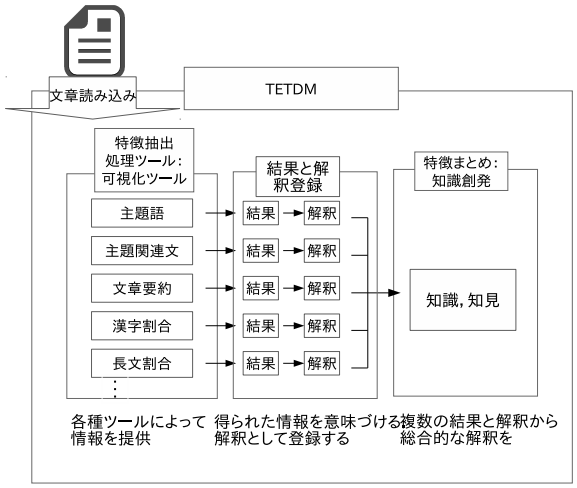

図 1: 文章の書き方の特徴を抽出して理解するための枠組み

トマイニングためのオープンソースソフトウェアで，一般に誰でも利用できる [*1].

TETDM には，文章の特徴を捉えるための多くの処理ツール，ならびに処理結果を出力する可視化ツールが実装されている．また分析を行うだけでなく，分析結果として分かったことを集めるための「結果と解釈」を登録する機能，ならびに，集めた「結果と解釈」をまとめて新たな知識を得るための「知識創発インタフェース」が備わっている．そのため，本研究で目指す，文章の特徴を抽出して理解するための環境づくりに，適したソフトウェアとなっている．

3.1.1 文章の特徴量を抽出するツール

文章の特徴を抽出するためのツールとして，テキストマイニングツールの TETDM[1] を用いる．TETDM は，複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し，社会的創造的活動を支援できる環境としての提供を目指している．TETDM に実装されているテキストマイニングツールや，新たに実装するツールによって文章の特徴理解を促す．ある文章を入力として与えた時に，章（文章の各部分）ごとに，長文の数や，重複した文，失礼な単語を含む文の割合，漢字の割合，冗長な文の数などの数値データや，文章の主題語，主題語に関連する文の位置，主題語に関連する単語の情報，使用されている単語の数などの，文章の特徴を捉える助けとなる情報を利用者に示す．

3.1.2 結果と解釈の登録機能

TETDM は，利用者が各ツールによって提示された情報を「結果」として，それによって得られる知見を「解釈」としてシステムに入力，登録できる機能を有する．例えば，100 字以上の長文が 10 文存在するという出力が得られた時，結果として「100 字以上の長文が 10 文あった」，また解釈の欄に「長い文を書く傾向がある」または「説明がまわりくどい」などの解釈を記述する．この機能を，自分なりの解釈を記述していくことで自分の文の特徴の把握に役立てる．

3.1.3 知識創発インタフェース

TETDM には，自身の積み重ねた解釈をもとに，汎用性が高い新たな知識を発見するための知識創発インタフェースがある．例えば，「説明がまわりくどい」「漢字の使用率が高い」という複数の解釈をもとに，「文章が難しくて読みづらい」という知識を得ることができ，「より平易な文を書くこと」を作文時に意識することができるようになる．この機能を，新たな文章作成や文章理解の指針を得るために役立てる．

*1 TETDM サイト：(URL)http:/tetdm.jp

表 1: 文章の書き方の特徴

番号	文章の特徴
1	主題一貫性 (主題に関係のある話でまとめられているか)
2	段落内一貫性 (段落の主張点が明確か，その主張点に対して一貫性があるか，適切な接続詞が使用されているか)
3	段落分けの妥当性 (適切な箇所，バランスで段落が分けられているか)
4	段落間の論理構造 (段落間のつながりが明確になっているか)
5	日本語の表現 (主語や目的語の欠落，意図や可能性を限定できない曖昧さ，内容や表現の重複，違和感のある表現，がないか)
6	日本語の表記 (文法ミス，誤字，脱字がないか)

3.2 文章の書き方の特徴

文章の書き方の特徴には，一貫性，論理構造，表現に関わる表 1 の項目が考えられる．(1)1 と 2 の項目は，文章の一貫性に関わる話として，文章が主題に関係のある話でまとまっているか．各段落内で主張は明確化，その主張に対して一貫性があるかを測る．(2)3 と 4 の項目は，文章の論理構造に関わる話として，適切な箇所，バランスで段落がわかれているか，段落感のつながりが明確になっているかを測る．(3)5 と 6 の項目は，文章の表現に関わる話として，文の長さや語句の長さ，単語の多様性，冗長な文の割合など，文章の読みさすさを測る．

現時点では，主に文章の形式的な特徴を対象としているが，文章の意味的な内容や，書き手や読み手の感情に関わる要素など，将来的には幅広い特徴量を扱えるようにしていきたいと考えている．

3.3 文章の書き方の特徴の抽出

文章の書き方の特徴は，TETDM に実装したツールにより抽出する．例えば現時点で，表 1 の 1 と 2 に関わるツールとして，TETDM には主題語を抽出するツールや主題語が文章中でどれだけ使用されているか，主題語に関係する文や単語は何であるか示すツールが実装されている．また，3 と 4 に関わるツールとして，TETDM には段落順序を評価するツールや各段落の類似度を評価するツールが実装されている．5 と 6 に関わるツールとしては，単語の頻度情報，長文や漢字の使用状況，などを評価するツールが実装されている．

各ツールが抽出する特徴について，利用者が特に着目すべきと判断した結果について，TETDM の「結果の解釈」機能を用いて登録する．

またユーザの助けとして，一部の特徴については，ツールによって自動的に「結果と解釈」機能を用いて，結果を登録する機能を設ける．すなわち，各ツールが出力する文章の特徴を表す指標について，他の多くの文章の平均値と比較して，十分に大きいまたは小さい値をとる指標を，文章の特徴として自動的にに登録する．例えば，各文の長さをカウントするツールにおいては，一文の長さが 100 文字を超える長い文があると「100 文字以上の長文がある」と「結果と解釈」に自動的に登録される．なお，登録したあるいは自動的に登録された結果と解釈は，いつでも削除することが可能になっている．

3.4 文章の書き方の特徴の抽出のためのインタフェース

文章の特徴抽出を行う TETDM の初期画面を 2 に示す．この図は 4 つのパネルから構成されている．左端のパネルには，分析対象とする文章を選択するための，「文章全体」と各段落

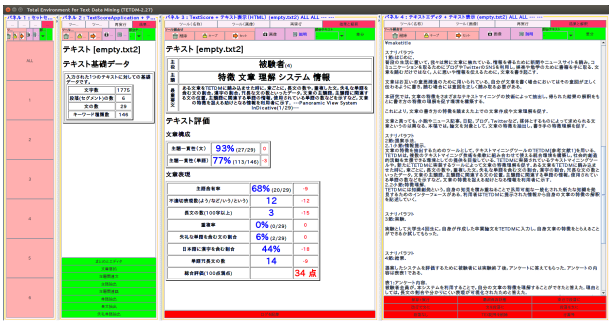

図 2: 文章の特徴抽出を行う TETDM の初期画面

図 3: 文章の特徴をまとめる知識創発インタフェース

番号を表す「1」「2」「3」といったボタンが配置されている．左から 2 番目のパネルには，分析するツールを選択するためのボタンが並べられており， 表 1 の特徴に対応したツールを選択できるようになっている．また，残りの右側のパネルにおいては，選択されているツールによる結果が表示される．

各パネルの上部には，「結果と解釈」ボタンがあり，文章について気になる特徴があれば，そのボタンを押した後に結果と解釈を登録する．

3.5 文章の書き方の特徴の理解のためのインタフェース

集めた文章の特徴をまとめ，汎用的な知識を生み出すための知識創発インタフェースを，図 3 に示す．知識創発インタフェースは，それまでにユーザが登録した解釈がノードとして並べられた状態で起動する．ユーザはこのインタフェース上で，関連があると思う複数の解釈を選択し，それらをまとめた，より抽象的，汎用的な解釈を，インタフェース下部のテキストフォームに入力する．その後，最下部の「結合」ボタンを押すことで，選択された複数のノードが 1 つの新たなノードに置き換わる．

この操作を繰り返すことで，得られた解釈を，1 つあるいは少数にまとめることができ，まとめられた内容を新たな知識として活用する．

4. 文章の書き方の特徴を捉える予備実験

本研究で構築する文章の書き方の特徴を捉えるために必要な要素を確認するための予備実験を行った．すなわち，文章の特徴を捉えるために必要なツール，あるいは機能を探ることを目的とした実験を行った．理系の大学生，大学院生 8 名に，2017 年 2 月までに作成した自身の卒業論文あるいは修士論文をシステムに入力してもらい，その文章の特徴を捉えてもらう実験を行った．

表 2: 実験に用いた文章の特徴を抽出するためのツール

番号	ツール
1	文章情報 (長文や漢字の割合)
2	文章要約
3	主題抽出
4	主題関連文
5	主題関連語
6	単語抽出
7	長文抽出
8	失礼単語抽出

表 3: 実験用アンケート

番号	質問内容
Q1	システムを使って自分の文章の特徴を理解することができましたか？ (5 段階評価と理由)
Q2	システムが出力した特徴の中には，自分が気が付かなかった特徴がありましたか？
Q3	今回まとめて得られた自分の文章の特徴は，すでに自身が知っていたものですか？
Q4	今回まとめた内容以外に，自分の文章には，どのような特徴があると思いますか？
Q5	もっと文章の特徴を知るためには，どのようなツールがあればよいと思いますか？

4.1 実験準備と実験手順

被験者が書いた論文をシステムに入力してもらい，自身の文章表現についての特徴をまとめるよう指示を与えた．実験に使用したシステムは，前章で述べた図 2 や図 3 のインタフェースを含むものとした．また，被験者が実際に利用できるツールを表 2 に示す．

「文章情報」は長文や漢字の割合，冗長な文の数などを表形式でまとめた情報を出力する（図 2 の左から 3 番目のパネル）．「文章要約」は文章から重要文を抽出するシステム，「主題抽出」は文章の主題を表す単語（主題語）を抽出する．「主題関連文」は主題語に関係する文の文章内の位置情報を視覚的に出力し，「主題関連語」は主題語に関連する単語を出力する．「単語抽出」は指定した単語の文章中での出現位置を視覚的に表示することができ，「長文抽出」は 100 文字を超える文の出現位置を表示する．「失礼単語抽出」は文章表現において失礼な可能性になる単語を出力する．

すなわち，表 2 の 2 から 5 のツールは，分析対象とする文章や段落を切り替えながら用いることで，一貫性と論理構造に関する確認を行うことができ，6 から 8 のツールにおいて，日本語の表現を確認することができる．また，1 のツールはこれらすべての内容を含んでいる．

被験者にはシステムを利用しながら，自分が入力した文章に関する特徴を，「結果と解釈」機能を用いて登録するよう指示を与えた．また，30 分間または 10 個以上の特徴を集めた後，「知識創発インタフェース」を用いて，自身の文章の特徴をまとめるように指示を与えた．実験の評価は，実験時の被験者の使用ログ，ならびに表 3 に示すアンケートの結果をもとに行った．

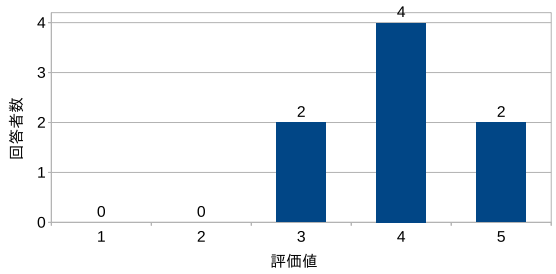

図 4: アンケート Q1:文章の特徴を理解できたか，の結果（数値が大きいほど高評価）

4.2 実験結果

アンケート Q1「文章の特徴を理解できたか」の結果を図 4 に示す．多くの被験者は，提案システムによって文章の特徴を理解することができた，と回答しており，その理由としては，長文の割合や分かりにくい文章表現が可視化されたためとの回答があった．このことから，文章作成時には認識することが難しい文章の特徴を提示する，本システムの基本的なプロセスには大きな問題がないと考えられる．

アンケート Q2「自分が気づかなかった文章の特徴を発見できたか」では，被験者全員が「はい」と答えており，システムを利用することで自分の文章の新たな特徴を発見できる可能性が確認された．

アンケート Q3「まとめた結果は自身にとって既知の内容だったか」の回答結果について，具体的に挙げてもらった未知の内容には以下のものがあった．

- 主語が足りていない部分があった。
- 冗長な表現が多かった。
- 同じ単語を多く使っている。

これらの内容は，文章を目で追って意識的に確認するのは困難，あるいは見落としやすい特徴と考えられる．

これら，Q2,Q3 の結果より，被験者は自分の書く文章の特徴を捉えることができ，システムを利用して自身の文章の解釈を促すことができる可能性が確認された．

アンケート Q4「回答した以外の自分の文章の特徴」についての回答には，以下のものがあった．

- 言葉足らずな表現が多い
- 接続詞が適切でなく，不完全な可能性がある．
- 文の区切りが悪く読みにくい構成になっている．

主に文の表現，あるいは論理構造の明確化に関わる内容と考えられる．また現在は，日本語の表現としてまとめている項目について，特に文章の読みやすさ，に関わる項目を重視することが考えられる．

アンケート Q5「特徴を捉えるどのようなツールを利用したいか」についての回答には，以下のものがあった．

- 句読点などが適切に使われているかを知るために，句読点の割合や場所がわかるツール
- 図表の多さや挿入のタイミングが適切かどうか判定するツール

これらの回答は，今回入力対象としたものが論文であったことと関係しており，これらをより一般的に解釈すると，前者は文章を読む際のリズムや内容理解の助けとするもの，後者は説明のわかりやすさと冗長さに関係するもので，これらも広くは，文章の読みやすさとして捉えることができる．

そのため，特に論文という文章においては，専門的な知識を持つ人にはもちろん，専門的な知識を持たない人を含め，できるだけ多くの人に意味が伝わる文章の作成が重要であると考えられる．

最後に，被験者が「知識創発」によってまとめた，自身の文章の特徴の例を，表 4 に示す．被験者は各自で自分自身の文章特徴を捉え，反省や改善例を含む知識を得ることができていた．

以上の結果から，提案する枠組みが，文章の特徴の理解につなげられる可能性が確認された．

表 4: 知識創発でまとめた自身の文章に関する知識の例

被験者	まとめた内容
A	(原因) 長文など文章を読みにくくする点や，論文に不適切な言葉がある．(結果) 全体的に文章を短くし，不適切な言葉を正しい物に変える必要がある．
B	(原因) 曖昧な単語が使われている反面，主題となる単語の出現率が低い． (結果) 曖昧な単語を減らし，主語となる単語を増やすべき．
C	研究テーマに沿った文が多いが，もっとバランスよく主題を意識した文を書くべき．

5. 結論

本研究では，文章の特徴について，システムからの出力をもとに解釈を与え，また得た解釈をまとめる過程を経て，文章の特徴の理解を促すシステムを提案した．また予備実験により，提案するシステムが文章の特徴の理解に役立てられる可能性を検証した．今後はよりさまざまな特徴量を対象としたツールを実装し，文章の読みやすさに関する特徴量を重視しつつ，文章の特徴を幅広く捉えられる枠組みへと拡張していきたいと考えている．

参考文献

[1] 砂山渡，高間康史，徳永秀和，串間宗夫，西村和則，松下光範，北村侑也：統合環境 TETDM を用いた社会実践，人工知能学会論文誌，Vol.32，No.1，NFC-A，pp.1-12，(2017)

[2] Yigal Attali & Jill Burstein : Automated Essay Scoring With e-rater V.2,The Journal of Technology, Learning, and Assessment, Vol.4，No.3，(2006)

[3] 石岡 恒憲，亀田 雅之：コンピュータによる小論文の自動採点システム Jess の試作 JESS，AN AUTOMATED JAPANESE ESSAY SCORING SYSTEM，計算機統計学，Vol.16，No.1，pp.3-19，(2003)

[4] 阪本浩基，伊藤淳子，宗森純：意見や意図の伝わりやすさに重点をおいた発想一貫支援システムの開発，マルチメディア、分散協調とモバイルシンポジウム 2014 論文集，Vol.2014，pp.445-452，(2014)