DEIM2021 B14-5

アブストラクトの定型性に着目した論文の構造推定に関する検討

玄道　俊†　松下　光範††

† 関西大学大学院総合情報学研究科　〒 569–1095 大阪府高槻市霊仙寺町 2–1–1
†† 関西大学総合情報学部　〒 569–1095 大阪府高槻市霊仙寺町 2–1–1
E-mail: †{k911867,t080164}@kansai-u.ac.jp

あらまし　本研究の目的は，論文の内容把握を支援するシステムの実現である．論文閲読者が論文の内容を的確に汲み取るためには，論文の意味的構造を理解し，その主張点を把握することが重要である．しかし，論文の記述方法は著者によって異なり，その構造を把握することは困難である．そこで，論文のアブストラクトを用いて本文の意味的構造を推定する手法を提案する．提案手法では，アブストラクトが定型的な要素列として表現されることに着目し，類似度に基づいてその各要素と対応づいた本文箇所を特定する．これにより，論文の意味的構造を推定し主張点の把握を容易にする．

キーワード　論文構造，構造推定，内容把握支援，アブストラクトの定型性，学術論文

1　はじめに

インターネットの普及に伴って多くの論文がオンラインで公開されるようになり，研究のサーベイを行う際に「関心のあるキーワードで検索して論文を探す」ことが一般的になってきている．オンライン化によって手軽に論文を検索・入手できるようになった反面，論文数の急激な増加 [1], [2] により，興味に合致した論文の収集や論文内容の効率的な把握はその困難さを増している．

論文を執筆する際，著者はその論文を執筆するに至った背景や，自身の提案に対する科学的根拠などを論理的に記述する．論文執筆の際の力点は必ずしもすべての論文で共通するわけではなく，新しい手法の提案に力点が置かれた論文や課題解決の着想に力点が置かれた論文，実験設計の緻密さに力点が置かれた論文など，論文によって様々である．

また，読者の関心によって同じ論文であっても価値を見出す箇所はしばしば異なる．例えば，論文読者が実験設計について関心を持っている場合と，新しいアルゴリズムに関心を持っている場合とでは，着目する論文の箇所は大きく異なってくる．

論文読者が自らの関心に沿った内容を効率よく把握し，自らの研究に役立てるには，論文著者の執筆の力点を勘案しつつ論文を構造化し，論文読者の関心に合致する箇所を特定してアクセス可能にすることが必要になる．

こうした背景の下，本研究では論文の内容把握を支援するシステムの実現を目指す．その端緒として，本稿では論文の構造を推定し，上述した論文の力点を論文の主張点と捉え，主張点の把握を容易にする．手段として構造が定型化されているアブストラクトを用いることとする．

2　アブストラクトの特徴

アブストラクトには，指示的アブストラクトと報知的アブストラクトの 2 つの見方が存在する [6]．指示的アブストラクトは元となった文章をより深く読むか選択する参照機能を提供している．対照に，報知的アブストラクトは元の文章の情報を代替えしている．また，報知的要約は指示的と報知的の両方の機能を果たしているとみなしており，したがって報知的要約は指示的要約の完全な部分集合とみなせる．

研究者が自身の研究に関係ある論文だと判断し，アブストラクトを頼りに，その研究の提案手法や実験結果を参照する場合，これはアブストラクトを「指示的側面」で用いているといえる．一方で，自身の研究に関係ある論文を網羅的に探す場合，アブストラクトのみを読んでその本文を把握しているのであれば，これは「報知的側面」として用いられているといえる．つまり，アブストラクトはそれ自体で内容を把握できると同時に本文への指示的な要約にもなっている．

アブストラクトの各文には要素が存在する．論文の提案や結果などが各文に割り当てられており，アブストラクトは主張点がまとめられているものであるといえる．このようにアブストラクトは要素という定型性を持つため構造化を行うことが容易であると考える．構造化を行ったアブストラクトを指示的に用いることで，本文の構造も行えると考える．

3　関連研究

3.1　論文の構造に関する研究

科学論文で主に使用される構造である IMRAD 形式は「序論」(Introduction),「方法」(Methods),「結果」 (Results),「討論 」(Discussion) で構成されている [5]．Lin らは，従来の IMRAD 形式では十分に説明することのできないセクションがあるとし，工学，応用科学，社会科学，人文科学の 39 分野，433 論文を対象とし調査した [3]．その結果，従来の「序論」(Introduction),「方法」(Methods),「結果」 (Results),「討論」(Discussion) に加え，「文献レビュー」(Literature review) と

表 1　学術論文の構造を明示する語の例 (文献 [7] より表引用)

	英語	日本語
I	Introduction Background	はじめに 序論
M	Material and Method Methodology	試料と方法 実験
R	Results Finding	実験結果 結果
D	Discussion Implication	考察 議論
L	Literature Review Related Research	文献レビュー 先行研究
C	Conclusion Summary	おわりに 結論

「結論」(Conclusion) が存在していた．また，最も頻繁に使用される構造パターンは「結果」と「討論」が結合した，序論-文献レビュー-方法-結果と討論-結論 (ILM[RD]C) であった．

また，石田らは英語論文だけでなく，日本語論文も IMRAD 形式であるか調査を行った [7]．英語論文 682 本，日本語論文 490 本，計 1,172 本の論文を収集し，表 1 に示した語によって明示的に構造が表現されている「見出し」のみにラベルを付与している．その結果，Web 上に存在する学術論文の 40%程度が IMRAD 形式を採っていることが示唆された．

3.2　アブストラクトに関する研究

橋本らは，抄録を構造化することで閲読者が論文の情報を効率的に得ることが可能な Structured Abstract に着目し，深層学習からこれらを自動生成する手法を提案している [8]．具体的には，自然言語処理で用いられる BERT モデルを科学技術分野に特化した SciBERT を利用することで，Structured Abstract の各見出しに適合する文を抽出している．結果，従来のモデルより上回る精度となり，生理学・医学系論文において対応できるモデルとなった．

柏木らはアブストラクトを用いて論文分類システムを提案している． 原子分子物理学分野の論文を分類する手法として Learning Vector Quantization(LVQ) を適応した．理化学辞典の用語と 8 種類の化学式の出現頻度を基準に特徴ベクトルを作成した場合，認識率 95%，再現率 80%，適合率 15%という結果を得ることができ，提案手法の有効性を示した．

4　デザイン指針

本稿では，アブストラクトを指示的に用いることで，論文本文の構造化を行うことを目的とし，以下の 3 項目に取り組む．

（1） アブストラクトに対し要素ラベルを付与

（2） アブストラクトと論文本文のベクトル値を求める．

（3） アブストラクトと論文本文のベクトルから類似度を算出する．

（1）ではアブストラクトの構造化を行うこととする．（2）では，アブストラクト各文と本文各文の数値化を行う．（3）では，（2）で求めた数値から，アブストラクト各文に対して類似度が最も高い本文中の文を算出する．（1）でアブストラクトに要素を付与しているため，類似度が高い本文中の文の要素を特定できる．要素を特定した文が本文中の出現場所から，論文本文の構造化を推定することが可能だと考えた．

5　提 案 手 法

5.1　データの収集

本稿ではコミック工学分野の論文を対象として分析を行う．これらの論文を用いる理由として，コミック工学分野の研究領域が広い [10] 点が挙げられる．例えば，画像処理技術の観点からマンガを識別する研究 [12] や，言語処理技術の観点からマンガ内のセリフを日本語能力試験と比較する研究 [11] など，アプローチは多岐にわたる．そのため自身に必要なキーワードを入力し検索した場合でも，自身に必要な論文であるか判断しづらいと考え，コミック工学分野をデータとして用いる．データに用いたコミック工学分野の選定方法として，「マンガ」，「漫画」，「コミック」，「アニメ」の 4 つの単語いずれかがタイトルに含まれているジャーナル論文 6 本と，コミック工学研究会の論文 14 本，合計 20 本を用いることとする．なお，すべての論文においてアブストラクトが日本語で記述されているものを選定した．

5.2　要素ラベル付与

本文の構造を推定するため，まずアブストラクトに要素のラベル付与を行った．ラベル名は「現状」，「問題」，「提案」，「結果」とした．これらのラベルをアブストラクトの各文に対し，付与した．ラベル付の手がかりとして，規定を制定した．表 2 の単語や語尾表現が出現する場合にそのラベルを付与することとした．手がかり語は「現状」，「問題」，「提案」，「結果」の順の優先度で付与した．例えば，「現状」と「提案」で出てくる単語が存在していた場合，「現状」を付与する．全てのラベルにおいて手がかり語が頻出しない場合，表 3 を確認する．これは，継承をあらわす単語であり，これらの語が頻出した場合は，1 文前のラベル名を継承することとする．これらのルールに全て当てはまらない文には「その他」のラベルを付与した．これらの作業は人手で行った．

5.3　アブストラクトと本文の類似度算出

本節ではアブストラクト各文と本文各文の類似度を算出する方法としてベクトル値を用いる．単語ベクトルを獲得するために，まずアブストラクトと本文の文章を形態素解析器 MeCab(ver. 0.996)[1]を用いることにより「名詞，動詞」の抽出を行う．次に学習のための辞書には，固有表現に強い辞書 mecab-ipadic-neologd[2]を用いた．この辞書を用いた理由として論文には固有表現が多数含まれているためである．

単語分散表現の学習では Python ライブラリである gensim[3] の Word2Vec を用いた．このベクトル生成には Skip-gram [4]

1：https://taku910.github.io/mecab/

2：https://github.com/neologd/mecab-ipadic-neologd

3：https://radimrehurek.com/gensim/

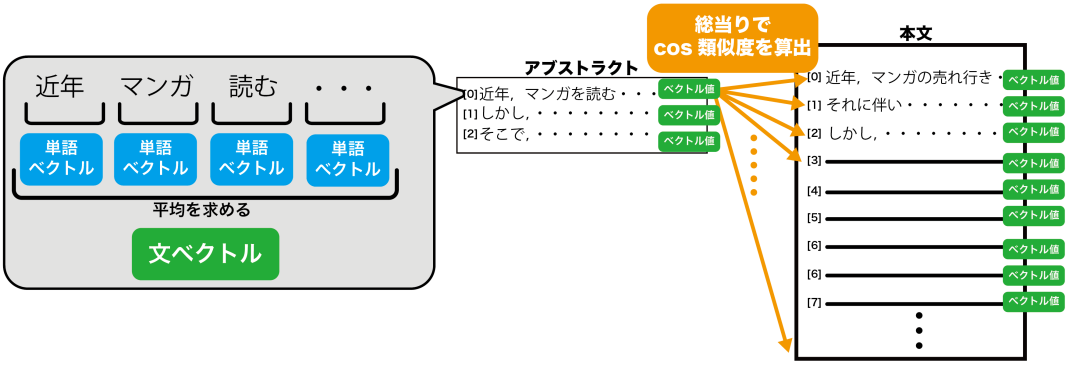

図 1　論文の本文箇所の特定フロー

表 2　手がかり語一覧

ラベル	文末表現	手がかり語
現状	[進行] いる.，している.	近年，現在，普及，需要
問題	[否定] ない.	しかし，難しい，必要
提案	[未来] する.	手法，開発，目的， 提案，ために，そこで， 本研究，本論文，本稿，議論
結果	[過去] した	結果，傾向，知見， 得，[数値]+倍，[数値]+%

表 3　前文継承語一覧

継承語	
[前文の固有名詞]+から	これは
さらに	具体的に

を用いた．パラメータは次元数う 100 次元，ウィンドウサイズ 8 単語に設定した．1 文に含まれる各単語ベクトルを平均した値を取ることにより，各文のベクトル値を算出する．アブストラクトの各文のベクトル値と本文の各文のベクトル値のコサイン類似度を算出し，最も値が高いものを特定する．アブストラクトを用いて，本文箇所を特定するまでの流れを図 1 に示す．

6　結　　果

6.1　アブストラクトの構成

本節ではラベル付けを行ったアブストラクトの構成の特徴を見ていく．

アブストラクトの構成において，連続して同じラベルが付与されている場合，前文の内容を継承していると判断し統合し，カテゴライズする．例えば，「現状+現状+問題+提案」とラベルを付与した場合は，連続している「現状」を統合し，「現状+問題+提案」とする．ただし，「その他」においては統合しないものとする．その結果，20 本の論文中 15 種類のラベルとなった．全ての論文において，「現状」よりも前文に「問題」，「結果」のラベルが付与されることはなかった．

また，最も多かったラベルは「現状+提案」であり，4 本の論文がこれに該当した．次に，多かったらラベルは「現状+問題+提案+結果」と「提案+問題+提案」でありどちらも 2 本のラベルがこれに該当した．「提案」がアブストラクトの先頭にくる論文は 8 本存在した．8 本の論文全てにおいて，「本研究」「本稿」「本論文」という言葉から書き始めていた．

6.2　類似度の結果

アブストラクトと本文の類似度が最も高い値から，本文箇所を確認する．論文の本文数は論文によって異なるため，割合を求めることで正規化した．正規化したデータから度数分布表を作成し，ヒストグラムを作成した (図 2 参照)．10%未満の割合が最も多く，「現状」が 11 件，「問題」が 6 件，「提案」が 8 件含まれていた．この階級が最も「現状」と「問題」が頻出していた．また 90%以上，100%未満の割合が 2 番目に割合が多く，「提案」が 11 件，「結果」が 5 件含まれていた．この階級が最も「提案」が頻出していた．そして，「結果」が最も頻出していた階級は 60%以上，70%未満の 6 件であった．

7　議　　論

10%未満の階級で本文箇所の件数が多くなった理由として，本文の冒頭に Introduction が書かれる傾向にあると考えられる．20 本の論文全てにおいて冒頭は「はじめに」や「まえがき」といった章で構成されていた．Introduction では研究背景とその学術的問いを記述し，それに対しての提案を記述している．また，90%以上，100%未満の階級で 2 番目に件数が多くなった理由として，本文の末尾には Conclusion が書かれる傾向にあるためだと考えられる．20 本中 19 件の論文の末尾に「おわりに」，「終わりに」，「まとめ」，「むすび」といった章で構成されていた．Conclusion では自身が提案したことを振り返りその結果を述べている．そのためアブストを構成する際，それらの部分を参照してくることが考えられる．

また，「結果」が 60%以上 70%未満の階級にもっとも多く出現した理由として，本文箇所に Result や Discussion が存在すると考えられる．Conclusion で記述した，「結果」ではなく，Result や Discussion で記述したより詳しい「結果」を書いていることになる．

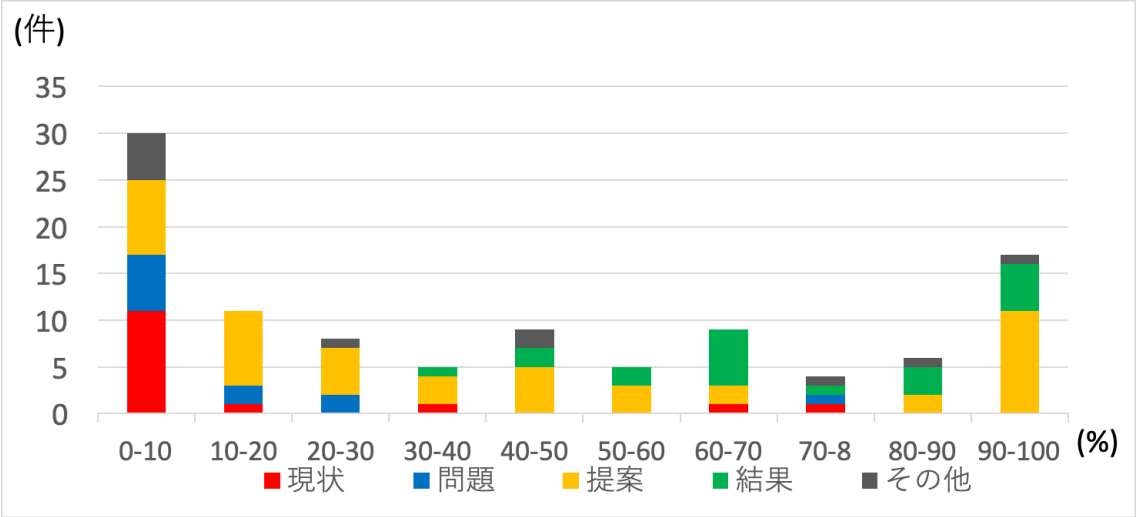

図 2　本文における各要素の分布

アブストラクトからその論文の「結果」を把握したい場合，Conclusion にも記述してあるが，Result や Discussion で記述してある「結果」のほうが論文著者の主張点である可能性が高いといえる．

Introduction 部分に「現状」,「問題」,「提案」が記述されており，Conclusion 部分には「提案」,「結果」が記述されている．これらは論文全体の情報が一定のレベルで記述されているといえる．つまり冒頭部分と末尾部分は報知的側面の可能性が考えられる．一方で Result や Discussion で記述されている「結果」はより詳しく記載されているため，ここから論文の情報全体の把握は困難である．そのため，指示的側面であると考える．

課題として，ラベル付けのルールのより明確化する必要がある．本稿では独自のルール付けを行ったが，今後はより客観性をもたせるため，TF-IDF 法といった機械的にルールを設定する．

8　展　　望

本稿で提案した手法を用いることで，研究者自身が保有している論文を整理するためのフレームワーク作成の一助になると考えられる．アブストラクトはその論文における問題解決のための作業（以下，問題タスク）に関係する単語（e.g., 構築, 抽出, 拡張）が記載されている．一方で，アブストラクトは簡潔かつ端的に記述されているため，問題タスクに関係する単語が指示する技術等の方法は詳細に記載されていない．そのため本稿で提案した手法を用い本文を参照することにより，指示する箇所の特定を行うことで論文の問題タスクを整理することが可能であると考える．

例えば，アブストラクトに「生成」という問題タスクの単語が用いられている場合，指示語として「GAN」といった具体的な技術が記述された方法と紐付けることが可能となる．これらの工程を踏むことで，各論文のフレームワークの作成が可能となり，論文同士を比較し差異や一致する項目を理解することができる（図 3 参照）．

また，フレームワークを作成することで同じ方法でクラスタリングすることが可能となる．これにより，研究者自身が保有している関連する論文を整理し多様な用途が想定できる．例えば，クラスタリングされたフレームワーク全体を俯瞰することで関連研究の項目を執筆することや，サーベイ論文の執筆を行うことが可能である．一方で，1 つの方法に着目することで，自身の研究にその方法を用いることを検討し，参照することが可能である（図 4 参照）．

9　おわりに

本稿では，論文の内容を効率的に把握するために，論文の構造を推定することで主張点の把握を容易にすることを試みた．手段として，定式化しているアブストラクトを用い，各要素と対応づいた本文箇所を特定した．結果として，Introduction と Conclusion 部分にアブストラクトと対応づいた本文箇所がより多く出現した．一方で，「結果」のラベルを付与した箇所は，Result や Discussion が記述されていると考えられる箇所に多く出現した．このことから報知的側面から Introduction と Conclusion から構成され, 指示的側面から Result や Discussion は構成されていると示唆された．

また，展望として研究者における思考の整理を行なうことが可能なフレームワークの作成構想を述べた．今後の方針として，アブストラクトを本稿で提案した 4 つの要素よりさらに詳細化した要素に分解し，アブストラクトの形式化を試みる．

文　　献

[1] Ordunña-Malea, E., Ayllon, J. M., Martín-Martín, A. and López-Cózar, E. D.: About the size of Google Scholar: playing the numbers, *arXiv-1407* (2014).

[2] Bornmann, L. and Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, *Journal of the Association for Information Science and Technology*, Vol. 66,No. 11, pp.2215-2222 (2015).

[3] Lin, L. and Evans, S.: Structural patterns in empirical re-

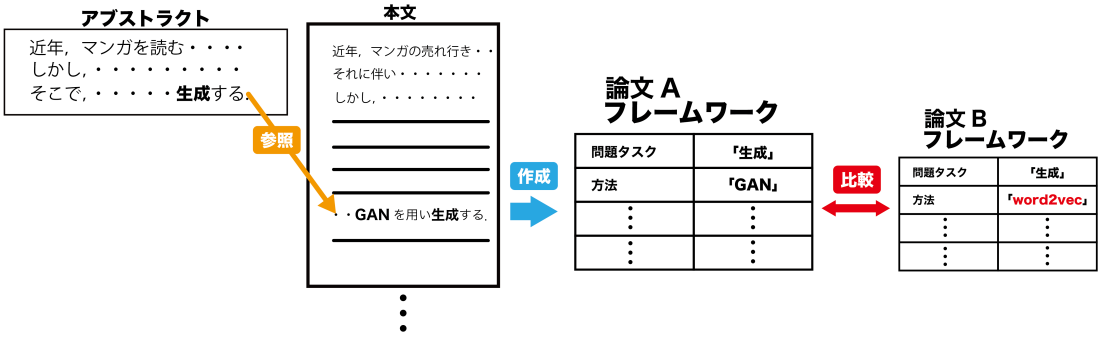

図 3　アブストラクトと本文を用いたフレームワーク化までの工程

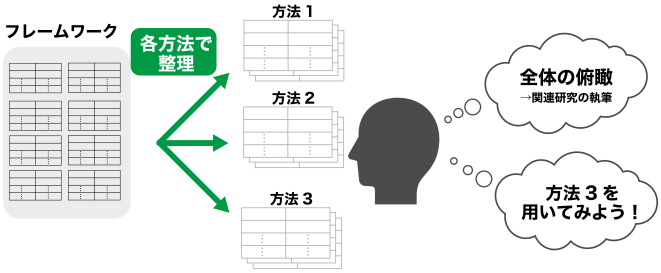

図 4　論文のフレームワークの活用例

search articles: A cross-disciplinary study, *English for Specific Purposes*, Vol. 31, No. 3, pp.150–160 (2012).

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S.and Dean, J.: Distributed Representations of Wordsand Phrases and their Compositionality,*Advances in neural information processing systems*, Vol. 26, pp.3111-3119 (2013).

[5] Day, R. A. and Gastel, B.: 世界に通じる科学英語論文の書き方: 執筆・投稿・査読・発表, 美宅成樹 (訳), 丸善 (2010).

[6] 奥村学, 難波英嗣:テキスト自動要約: オーム社, pp. 12-13, (2005).

[7] 石田栄美, 安形輝, 宮田洋輔, 池内淳, 上田修一: 構造と構成要素に基づく学術論文の自動判定, 日本図書館情報学会誌, Vol. 60, No. 1, pp. 18–34 (2014).

[8] 橋本快生, 井上潮: 深層学習による学術論文からの Structured Abstract 自動生成, 第 12 回データ工学と情報マネジメントに関するフォーラム, G5-3 (2020).

[9] 柏木裕恵, 髙田雅美, 佐々木明, 城和貴: アブストラクトを用いた論文分類システムの設計と実装, 情報処理学会 研究報告 (MPS), Vol. 2006, No. 95, pp.33-36 (2006).

[10] 山西良典, 松下光範, 上野未貴: コミック工学と AI, 人工知能学会誌, Vol. 33, No. 6, pp.819–825 (2018).

[11] 西原陽子, Shan, J., 山西良典, 福本淳一: 日本語学習 支援を目的とした漫画の台詞の難度の判定, 第 30 回 人工知能学会全国大会, pp. 2J4OS08a2–2J4OS08a2 (2016).

[12] 石井大祐, 山崎太一, 渡辺裕: マンガ固有の特徴を利用したマンガ登場人物識別に関する一検討, 研究報告オーディオビジュアル複合情報処理, Vol. 2013, No. 1, pp. 1–4 (2013).