論文からの記載必要項目の抽出と文章作成支援

岡田 拓真 [*1]　村田 真樹 [*2]　徳久 雅人 [*2]　馬 青 [*3]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*3] 龍谷大学 理工学部 数理情報学科
[*1,*2]{s112011,murata,tokuhisa}@ike.tottori-u.ac.jp
[*3] qma@math.ryukoku.ac.jp

1　はじめに

論文において研究成果や研究の必要性・有効性などの記載すべき情報が記載されていない場合が存在する．その場合，研究の内容が読者に伝わり難いという問題が発生する．

本研究では，論文に記載すべき情報を「記載必要項目」と定義し，論文内で記載必要項目が欠落しているか否かを自動検出することで，論文の文章作成支援を行うことを目的とする．

2　関連研究

論文の閲覧支援やサーベイ支援として村田ら [1] の研究や樫本ら [2] の研究が挙げられる．村田ら [1] の研究では，論文の閲覧支援を目的として，論文アブストラクトから重要な情報を抽出し，その情報を表の形で可視化している．さらに重要な情報を抽出するための教師データを作成し，それを用いて教師あり機械学習により重要な情報を抽出している．樫本ら [2] の研究では，論文のサーベイを効率良く行うことを目的として，論文から表，図，脚注，参考文献の 4 つの論文構成要素をルール及び機械学習 (SVM) を用いて抽出を行っている．村田ら [1] の研究と樫本ら [2] の研究は，論文データの可視化を行うことで，論文のサーベイに役立てることを目的としている．それらに比べて本研究では，論文データの可視化を目的にしているのではなく，論文内に記載必要項目の欠落している文章が存在しているかを確認し，論文作成の際の文章作成支援を行うことが目的である．

欠落個所の認識として灘本ら [3] の研究が挙げられる．灘本ら [3] の研究では，SNS やブログのようなコミュニティ型コンテンツ内で議論が集中し，視点が狭くなる可能性があると述べている．これにより議論におけるテーマを多面的に捉えられなくなる危険性を指摘している．見落とされた視点をコンテンツホールと呼び，SNS やブログにおけるコミュニティ内の議論の履歴からコンテンツホールを抽出し，ユーザに提示している．灘本ら [3] の研究と本研究を比較すると，灘本ら [3] のユーザに見落された点を指摘するという目的と本研究の論文著者に欠落した文の存在を知らせ文章作成支援を行うという目的は類似していることがわかる．しかし，灘本ら [3] はコミュニティ型コンテンツ内での研究であり，本研究は論文内での研究であるという異なる点も存在している．

文章作成支援として都藤ら [4] の研究が挙げられる．都藤ら [4] の研究では，冗長な文章を自動検出することを目的としている．冗長な文章の自動検出の提案手法として，冗長度と機械学習を利用して実験を行っている．冗長度を素性にすることにより機械学習だけを利用して検出するより精度が向上することが確認されている．都藤ら [4] の研究と本研究を比較すると，どちらの研究も不適切な文章の検出を目的としているが，都藤ら [4] は機械学習を用いて冗長な文章の検出を行っている．一方，本研究では冗長な文章ではなく記載必要項目が欠落している文章の検出が目的であり，検出対象が異っていることがわかる．また，都藤ら [4] 以外にも数多くの研究で文章作成支援が行われている．しかし，その数多くの文章作成支援の研究の中でも，論文の記載必要項目を利用して文章作成支援を行っている研究はない．

3　記載必要項目と検出に役立つ単語の決定

3.1　問題設定

記載必要項目と記載必要項目を検出するのに役立つ単語を決定する．検出に役立つ単語が一つもない場合は記載必要項目が欠落している論文であると判別でき，記載必要項目が欠落している論文の検出に役立つ．

3.2　決定手順

記載必要項目とその項目の検出に役立つ単語の決定は以下の手順で行う．

1. 多くの論文に出現する単語を調査する (3.2.1 節)
2. 1 の結果から意味ソート [5] を利用して意味の類似している単語をまとめて表示させる (3.2.2 節)
3. 2 の結果を人手で検討して，記載必要項目とその項目の検出に役立つ単語を決定する (3.2.3 節)

手順の詳細を以下に示す.

3.2.1 頻度調査

多くの論文に出現する単語は論文の記載必要項目である傾向である可能性が高いと考えられる. 単語の出現した論文数を全論文数で割ることで単語の出現率を算出する. 例えば, 全論文 300 件中 250 件の論文に単語「Z」が存在している場合, 単語「Z」の出現率は 250/300 となる.

3.2.2 意味ソート

記載必要項目の検出に役立つ単語に類似している単語も記載必要項目がの検出に役立つ単語である可能性があると考える. 例えば「手法」という単語が記載必要項目の検出に役立つ単語である場合, その単語に類似している「方式」などの単語も記載必要項目の検出に役立つ単語である可能性がある. 本稿では, 記載必要項目の検出に役立つ単語に類似している単語を調査するために意味ソート [5] を利用する. 意味ソート [5] は意味の類似している単語をまとめて表示させることができる. これにより出現率の低い単語も参考にでき, より詳細な記載必要項目とその項目の検出に役立つ単語が決定できると考える.

3.2.3 人手での検討

3.2.2 節の結果を参考にして, 人手で記載必要項目とその項目の検出に役立つ単語を検討し決定する.

3.3 データ

記載必要項目の決定を行う際に使用した実験データは, 1994 年から 2013 年の言語処理学会論文誌 (393 件) である.

3.4 決定結果

3.4.1 頻度調査の結果

本稿では, 3.2.1 節で挙げられた方法で頻度調査を行った. その結果のうち, 比較的出現率が高く, 重要な表現である可能性が高いものを抜粋したものを表 1 に示す.

表 1: 論文内に出現する単語の出現率の調査結果 (抜粋)

単語	出現率	単語	出現率
必要	0.994	重要	0.811
異なる	0.951	違う	0.631
比較	0.895	比べる	0.727
例えば	0.858	目的	0.773
問題	0.931		

3.4.2 意味ソートの結果

論文での出現率の高い上位 500 単語までの単語を意味ソート [5] を使ってソートし, 意味の類似している単語をまとめて表示させた. 意味ソート [5] の結果の例を図 1 に示す.

(数量)
{ 量 } 出力 入力 総数 数値 頻度 番号 関数
{ 数 } 多く 多数 多い 大量 十分 少ない
{ 値・額 } 長い 短い 尺度 高い 低い 深い 近い 距離
(関係)
{ 因果 } 条件 有効 前提 原因 要因 結果 効果 影響
{ 理由・…} 理由 目的 実用
{ 異同 } 相対 相互 応じる 対応 相当 **比べる** **比較**
{ 相対 } 同じ 似る 同様 類似 **異なる** 含む
含める **違い** 区別
{ 有無 } 存在 既存
{ 出現 } 現れる 出現 実現 提案 提示 示す 出す

図 1: 意味ソートの結果 (一部)

3.4.3 記載必要項目と検出に役立つ単語の決定結果

3.4.1 節から研究の必要性・有用性を表す可能性がある「必要」や新規性を表す可能性がある「異なる」などの出現率が高いことがわかった. 研究の必要性や新規性が存在しない論文は研究の内容が理解し難くなる可能性が高いので, 「必要」「異なる」なども記載必要項目である可能性が高いと考えられる.

また, 「問題」「目的」などが存在しない論文は何が問題で何を目的にしているかを理解できなくなる可能性が高いと考えられる. さらに, 「例えば」などが存在しない論文でも理解しやすい具体例などがない可能性があり, 論文の内容の理解が難しくなる可能性があると考える. 従って, 「目的」「問題」「例えば」なども記載必要項目である可能性が高いと考えられる.

以上で記載必要項目である可能性が高いとされた単語と 3.2.2 節で述べた意味ソート [5] の結果を比べ, その単語に類似した単語を人手で検討し, 記載必要項目とその項目の検出に役立つ単語を決定した. 結果を表 2 に示す. 検出に役立つ単語が一つもない論文を記載必要項目が欠落している論文として自動検出でき, 論文の文章作成支援に役立つ.

表 2: 決定した記載必要項目と検出に役立つ単語

項目名	検出に役立つ単語	記載必要項目の説明
必要性	必要　重要	研究の必要性
新規性	異なる　違う　違い	研究の新規性
比較	比較　比べる	先行研究との比較 精度の比較実験
問題点	問題	先行研究の問題点 研究の背景
目的	目的　目標　目指す	研究の目的
例	例えば　例　具体	具体例

4 文章作成支援

4.1 問題設定

表 2 の結果を基にルールベースで記載必要項目が欠落している論文を検出する．記載必要項目を補う必要がある論文が自動検出できれば, 論文の文章作成支援に役立つとする．

4.2 記載必要項目が欠落している論文の検出方法

表 2 の検出に役立つ単語をルールとしてルールベースを利用し論文の検出を行う．表 2 の検出に役立つ単語が一つも出現していない論文を記載必要項目が欠落している論文として検出する．

4.3 データ

文章作成支援の実験を行う際に, 2011 年度の年次大会論文 (266 件) を学習データとして使用し, 2012 年度の年次大会論文 (305 件) をテストデータとして使用した．また, 学習データを 4.4.2 節の判別基準の設定に利用し, テストデータを評価に利用する．

4.4 評価方法

4.4.1 評価の手順

ルールベースの評価は, 以下の手順で行う．

1. 4.2 節で記載必要項目が存在しないと判断されたものが文章作成支援に役立っている (その記載必要項目を補う必要がある) かを人手で判別する．
2. 1 の結果から提案手法であるルールベースの再現率・適合率・F 値を算出する．
3. 全ての論文をシステムの出力 (記載必要項目が欠落している) にした場合をベースラインとして, ベースラインの再現率・適合率・F 値を算出する．
4. 2 と 3 で算出した結果から提案手法であるルールベースとベースラインの精度を比較し評価を行う．

4.4.2 人手での判別基準

それぞれの項目の人手での判別でばらつきが生じないように項目ごとに基準を設定した．何故なら, 曖昧な判別を行い, 判別結果がばらつくとそれだけ再現率・適合率・F 値が正確でなくなるからである．より正確な再現率・適合率・F 値を求めるために 2011 年度の年次大会論文 (266 件) のデータを使用し, 提案手法で処理した結果を人手で評価した．そこでの評価を参考にして, 人手での判別基準を細かく設定した．また, 判断基準を設定する際に人手判定の κ 値 (一致率) 利用して評価を行った．κ 値は 4.5 節でテストデータを人手で判別する人物一人 (人物 A とする) と人物 A とは別の人物一人 (人物 B とする) の合計二人で一致率を算出した．また, 人物 A が学習データにおいて人手判別したものからランダムに文章作成支援に役立っているものと役立っていないものをそれぞれ 12 件ずつ取り出した．その合計 24 件の論文を人物 B が判別基準を参考にして人手判別し, その結果の一致率で κ 値を算出した．本稿で設定した判別基準での κ 値は 0.67 であった．

記載必要項目の判別基準の一部を表 3 に示す．また, 表 3 では, 文章作成支援に役立つと判別したものは○, 文章作成支援に役立たないと判別したものは×としている．

表 3: 各記載必要項目の判別基準 (一部)

項目名	判別	基準
比較	○	先行研究との比較や実験結果の比較が行われていない論文
	×	「…という手法が先行研究で挙げられている．それに対して我々は…」などの文で先行研究との比較が行われている論文
問題点	○	世の中の問題 (研究背景) や先行研究の問題点についての説明が不足している (説明が不明瞭な) 論文
	×	研究背景や先行研究の問題点について詳しく説明されている (説明が明瞭な) 論文
目的	○	詳しく読まないと何を目的に研究を行っているのかが理解できない論文 (一読するだけでは研究の目的が理解できない論文)
	×	詳しく読まなくても, 一読すれば研究の目的が何であるかが理解できる論文
例	○	具体的な例がない論文 (図中も含む)
	×	具体的な例がある論文 (図中も含む)

4.5 文章作成支援の実験結果

2012 年度の年次大会論文 (305 件) をテストデータとして実験を行った．結果を表 4 から表 7 に示す．

表 4: 「比較」について文章作成支援の評価結果

手法	再現率	適合率	F 値
ベースライン	1.00 (59/59)	0.19 (59/305)	0.32
ルールベース	0.58 (34/59)	0.60 (34/57)	0.59

表 5: 「問題点」について文章作成支援の評価結果

手法	再現率	適合率	F 値
ベースライン	1.00 (114/114)	0.37 (114/305)	0.54
ルールベース	0.61 (70/114)	0.80 (70/87)	0.69

表 6: 「目的」などについて文章作成支援の評価結果

手法	再現率	適合率	F 値
ベースライン	1.00 (94/94)	0.31 (94/305)	0.47
ルールベース	0.53 (50/94)	0.60 (50/84)	0.56

表 7: 「例」について文章作成支援の評価結果

手法	再現率	適合率	F 値
ベースライン	1.00 (9/9)	0.03 (9/305)	0.06
ルールベース	1.00 (9/9)	0.75 (9/12)	0.86

「重要」「異なる」などを学習データでルールベースで処理し, 人手での判別基準を検討したが, ルールベースの性能が低いことがわかった. これは「重要」「異なる」などの人手の判別基準の設定が困難であったことが原因であると考える. これについては 4.6.2 節で考察する.

4.6 考察

4.6.1 文章作成支援の実験考察

表 4 から表 7 を見るとそれぞれベースラインよりもルールベースのほうが F 値が高いことがわかる. また, 表 7 の「例」についての結果を見ると, ベースラインと提案手法であるルールベースの F 値の差は 0.80 もあることもわかる. さらに表 4 の「比較」についての結果もベースラインと提案手法であるルールベースの差が約 0.3 あることもわかる. しかし, 表 5 と表 6 の「問題」や「目的」についての結果は, ベースラインと提案手法であるルールベースの差が約 0.10 ほどであり, これは他の表 4 や表 7 の結果と比べると, ベースラインとの性能差があまりないように見える. これは「問題」や「目的」などの項目のルールを詳細化することでベースラインとの性能差が生まれると考えられる. 今後は本稿の提案手法だけでなく, 更に新しい手法を取り入れ, 記載必要項目を決定することでルールが詳細化されると考える. また, 本稿では情報抽出の基本的な手法しか使っておらず, 論文の記載必要項目という特性をほとんど利用できていなかった. したがって, 更なる手法の改良が今後の課題であると考える.

4.6.2 「必要」「異なる」などについての考察

まず今回の実験で, 「必要」「異なる」などが欠落しているだけでは記載必要項目が欠落しているという判別はできないということがわかった. 「必要」「異なる」などが欠落している論文と欠落していない論文を見比べたところ, 「必要」「異なる」などの有無に関係なく, 研究の必要性や新規性が述べられている論文が散見された. これにより, 3.4.3 節で述べた「必要」「異なる」などが研究の必要性や新規性を表しているという考えが間違っていると考えられた. しかし, 必要性や新規性などは研究にとって不可欠であり, それらの項目が存在しないと研究の詳細が理解しづらくなるのは安易に想像できるだろう. 必要性や新規性の欠落は今回の手法で検出するのは困難であると考えられるが, 検出すべき項目であり, 今後扱っていく課題である.

5 まとめ

本稿では, 論文の文章作成支援を行うことを目的に, 論文の記載必要項目を調査し, その結果を基にルールベースによって記載必要項目が欠落している論文を検出した. また, 本稿で決定した記載必要項目のうち研究の必要性・新規性以外の項目を検出することができた. さらに「比較」「問題」「目的」は F 値が 0.6 から 0.7 で検出でき, 「例」は F 値が 0.86 で検出できた.

今後の課題として, 今回の実験で使用したルールベースの性能向上を考えている. そのために, 更なるルールの詳細化が必要であると考えられる. それにより, 本稿で検出できなかった論文の記載必要項目であろう研究の必要性や新規性を検出できるようになり, 更なる論文の文章作成支援に繋がると考える. また, 論文の記載必要項目という特性をうまく利用できるように手法を改良することも今後の課題である.

謝辞

本研究は科研費 (26330252) の助成を受けたものである.

参考文献

[1] 村田真樹, Stijn De Saeger, 橋本力, 風間淳一, 山田一郎, 黒田航, 馬青, 相澤彰子, 鳥澤健太郎: “論文データからの重要情報の抽出と可視化”, 2009 年度人工知能学会全国大会 (第 23 回) 論文集, pp.1-4, 2009.

[2] 樫本達矢, 太田学, 高須淳宏: “学術論文からの構成要素抽出の一手法”, 第 12 回日本データベース学会年次大会, C5-2, 2014.

[3] 灘本明代, 阿辺川武, 荒巻英治, 村上陽平: “コミュニティ型のコンテンツホール抽出手法の提案”, 日本データベース学会 Letters, 6 巻, 2 号, pp.29-32, 2007.

[4] 都藤俊輔, 村田真樹, 徳久雅人, 馬青: “機械学習と冗長度を用いた冗長な文章の検出”, 言語処理学会第 20 回年次大会発表論文集, pp.939-942, 2014.

[5] 村田真樹, 神崎享子, 内元清貴, 馬青, 井佐原均: “意味ソート msort 意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例”, 自然言語処理, Vol.7, No.1, pp.51-66, 2000.