論文における記載不備の自動修正に向けた分析

岡田 拓真 [*1]　村田 真樹 [*1]　馬 青 [*2]
[*1] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*2] 龍谷大学 理工学部 数理情報学科
[*1]{s112011,murata}@ike.tottori-u.ac.jp
[*2] 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・修正することで, 論文の文章作成支援を行うことを目的とする．

我々は過去に記載必要項目が欠落している論文の自動検出方法としてルールベース手法と機械学習手法を提案してきた [1][2]．ルールベース手法では, 論文の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し, その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ルールベースで自動検出する．機械学習手法では, 機械学習を用いて記載必要項目が欠落している論文を自動検出する．

過去の我々の研究 [2] では, 論文の自動修正に向けた分析も行っているが, 論文の自動修正を行うためには更なる分析が必要であると考える．そこで, 本稿では論文の自動修正技術の構築に向けた更なる分析を行い, 論文の修正に役立つような修正パターンを獲得する．

以上に向けて, 2 節では, 記載必要項目の詳細について示す．3 節では, 論文の自動修正技術の構築に向けた分析を行う．最後に 4 節でまとめを述べる．

2　記載必要項目

本稿では, 過去の我々の研究 [1] で決定した記載必要項目を用いる．以下に記載必要項目の決定方法と定義を示す．

2.1　記載必要項目の決定方法

まず, 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可能性が高いと考え, 論文に出現する単語の調査を行った．さらに, 意味ソート [3] を利用して論文に頻出している単語の類似単語についても調査を行った．以上の 2 つの調査を参考に人手で論文の記載必要項目の決定を行った．記載必要項目の決定方法の詳細については文献 [1] を参照のこと．

2.2　記載必要項目の定義

2.1 節の決定方法より決定した記載必要項目とその定義を表 1 にまとめる．表 1 以外の記載必要項目もあると考えられるが, 2.1 節の決定方法では表 1 で挙げた 4 項目が決定できた．

3　論文の自動修正技術の構築に向けた分析

3.1　分析方法

本研究では, 修正のために追加した文字列を修正文と定義し, その修正文に出現する文字列パターンを修正パターンとする．論文の修正に役立つような修正文や修正パターンを獲得するために分析を行う．本研究では, 4 種類の分析方法で分析を行う．

表 1: 記載必要項目と定義

項目名	定義
比較	先行研究との比較や実験結果の比較
問題点	世の中の問題 (研究背景) や先行研究の問題点
目的	その研究を行う理由
例	具体例

表 2: 5 段階のレベルの定義

レベル	定義
5	手がかり手法があり, 誰が読んでも記載必要項目について容易に理解できるもの
4	専門的な知識がなくても文脈から容易に予測でき, 記載必要項目について理解できるもの
3	文脈から予測することが少し難しいが, 考えて読めば記載必要項目について理解できるもの
2	専門的な知識と深い洞察により記載必要項目について理解できるもの
1	記載必要項目について全く理解できないもの

3.1.1　5 段階レベルを使った分類による分析

過去の研究 [1][2] では, 記載必要項目がある論文とない論文の 2 値分類で行っていた．しかし, 完全に記載必要項目がない論文は非常に少ないため, 2 値分類では曖昧性が生じていた．そこで, 本研究では 5 段階のレベルを設定し, レベルが高いほど記載必要項目について明瞭に書かれている論文であるとして分類を行う．5 段階のレベルの定義を表 2 に示す．

3.1.2　人手による論文修正

3.1.1 節で 5 段階レベルに分類した結果のうち, レベル 4 以下に分類された論文をレベル 5 に分類される論文になるように人手で修正する．修正前の論文と修正後の論文の差分 (修正文) を抽出し, 分析する．本研究では差分抽出に mdiff[6] を用いる．

3.1.3　修正文に出現する単語の頻度調査による分析

3.1.2 節で抽出した修正文に出現する単語の頻度を調査し, 分析する．本研究では 2 単語連続と 3 単語連続での出現頻度を調査し, 出現頻度の高い単語連続を人手で見て修正パターンを調査する．2 単語連続, 3 単語連続での出現頻度の調査とは, 例えば単語「A」「B」「C」「D」「E」の 5 単語で構成されている文「ABCDE」であるとき, 2 単語連続の場合は「AB」「BC」「CD」「DE」という形で頻度を調査し, 3 単語連続の場合「ABC」「BCD」「CDE」という形での頻度を調査する．

3.1.4　階層クラスタリングを用いた分析

R による階層クラスタリングを用いて 3.1.2 節で抽出した修正文から似ている修正文を調査する．この結果で得られた似ている修正文同士から共通している部分を抽出する．本研究では, この共通部分が論文の修正パターンになると考える．本研究ではウォード法を利用し, 修正文に出現する単語 1 語と 2 単語連続と 3 単語連続が出現したか否かをベクトルの要素として類似度を算出し, 似ている修正文を調査する．単語 1 語のものが出現して

表 3: 5 段階のレベルの頻度

	レベル 1	レベル 2	レベル 3	レベル 4	レベル 5
目的	1	2	11	12	24
問題点	7	2	2	12	27

ウイグル語の再帰代名詞の人称の示し方自体は、当然すでに知られているが、これまでの研究は、比較的最近のものでも、筆者らが知る限り、データの提示にとどまっていて、メカニズムの分析・記述はほとんどなされていない。特に、普通名詞や人称代名詞とのふるまいの違いについての分析はほとんどない。そこで、本研究では、ウイグル語のメカニズムの解明を目的としてウイグル語の再帰代名詞の人称を、普通名詞や人称代名詞の人称との違いに着目し分析・記述する。分析は、語彙主義に立ち、HPSG（主辞駆動句構造文法）の枠組みで行う。そして、ウイグル語の再帰代名詞の人称が素性の単一化にもとづいて分析されることを示す。

図 1: 「目的」についてレベル 4 からレベル 5 に修正した論文の一部

いる場合, ベクトルの要素の値は 1 としている．2 単語連続が出現している場合, ベクトルの要素の値は 2 とし, 3 単語連続が出現している場合, ベクトルの要素の値は 3 としている．最後にベクトルごとに正規化を行い, 類似度を算出する．

3.2 データ

2011 年度の言語処理学会年次大会論文 (266 件) の中からランダムに 50 件選び, 分析に使用する．記載必要項目「比較」「例」については, 比較しているか否か, 例があるか否かの 2 値になり, 細かい分析が困難だと考えたため, 今回は記載必要項目「問題点」「目的」についてのみ分析を行う．

3.3 各分析の結果

3.3.1 5 段階レベルの分類結果

記載必要項目「目的」「問題点」について 50 件の論文データに対して 5 段階のレベル設定を行った．それぞれのレベル設定の頻度を表 3 に示す．レベル 5 からレベル 1 に分類した論文の具体例については文献 [2] を参照のこと．

3.3.2 人手による論文修正結果

人手でレベル 4 以下の論文をレベル 5 の論文になるように修正した．修正で得られた修正文を表 4 と表 5 に示す．表 5 の「////」の部分には元の文 (差分ではない文) が入る．

記載必要項目「目的」について修正した論文の一部を図 1 から図 4 に示す．修正箇所の 下線部 は追加した部分を意味しており, [[二重括弧]] は削除した部分を意味している．

3.3.3 単語頻度調査による分析結果

3.3.2 節で得られた修正文に出現する単語の頻度を調査し, 分析を行った．頻度調査により得られた頻度の高い単語連続の一部を表 6 に示す．表 6 の結果から人手で得られた修正パターンを表 7 に示す．

3.3.4 階層クラスタリングによる分析結果

3.3.2 節で得られた修正文を R を用いて階層クラスタリングし, 分析を行った．図 5 に記載必要項目「目的」についての階層クラスタリング結果, 図 6 に記載必要項目

表 4: 記載必要項目「目的」についての修正文 (26 文)

文番号	修正文
文 1	既存の類似尺度よりも良い精度を示す尺度を発見を目的として
文 2	国語辞典から情報の獲得を目的として
文 3	の問題点の解消を目的として 2 つのアプローチ
文 4	日本語の変化を調査するために
文 5	メカニズムの解明を目的としてウイグル語の
文 6	言語構成の解明を目的として
文 7	英語によるプレゼンテーションを支援するシステム開発を目的として
文 8	を目的として
文 9	本研究では，そのようなコストの問題を解消するために日英特許データベースからシソーラスを自動的に構築する手法を提案する．
文 10	本稿では、そのようなコストの問題を解消するために特許データベースと特許検索履歴からシソーラスを自動的に構築する手法を提案する。
文 11	本研究では，辞書見出し語の決定を目的として対訳表現の抽出手法を提案する．
文 12	獲得を目的として言い換えの
文 13	知的生産能力を持った人物の育成を目的として
文 14	本稿では、生放送番組への字幕付与を目的として
文 15	この課題に対して
文 16	誰でも容易に Web 情報を選り分けられる社会を目指し,
文 17	マイクロブログサービスの利便性の向上を目的として
文 18	ユーザをカテゴリごとに分けるため
文 19	そのような問題を解消するための
文 20	展開語リストを仮定しない略語展開を目的として
文 21	プレゼンテーション発表支援を目的として
文 22	同義語関係の把握を目的として
文 23	外国語話者を対象とした日本語ニュースの平易化を目的として
文 24	大量の言い換え知識獲得を目的として
文 25	そこで本研究では, 医療文書中に出現する略語の完全語を特定を目的とし, 大規模コーパスを用いて略語の復元を行う．
文 26	日本人と外国人の両方が理解しやすい日本語ニュースの作成を目的として

略語を正しく復元することは非常に重要であるが，多くの略語は複数の完全語を持っているため，そして対象となる略語は大量にあるため，人手で復元することは困難である．そのため，計算機により自動的に略語の復元をすることが必要である．そこで本研究では, 医療文書中に出現する略語の完全語を特定を目的とし, 大規模コーパスを用いて略語の復元を行う．

図 2: 「目的」についてレベル 3 からレベル 5 に修正した論文の一部

「目的」についての階層クラスタリング結果を示す．図 5 と図 6 に書かれている番号は 3.3.2 節の修正文の番号を示している．

図 5 と図 6 の結果と 3.3.2 節の修正文から似ている修正文同士の共通部分を調査する．例えば, 記載必要項目「目的」であれば図 5 で隣り合っている文 5「メカニズムの解明を目的としてウイグル語の」と文 6「言語構成の解明を目的として」を見ると「〜の解明を目的として〜」という共通部分があることが分かる．同様に, 記載必要項目「問題点」であれば文 9「〜出回るという問題が発生する．〜」と文 14「という問題が発生する」を見ると「〜という問題が発生する」という共通部分があることが分かる．このようにして隣り合っている文番号の修正文の共通部分を調べて修正パターンを獲得する．獲得した修正パターンを表 8 に示す．

表 5: 記載必要項目「問題点」についての修正文 (16 文)

文番号	修正文
文 1	という問題点
文 2	その結果，上位下位の関係も漠然としたものになってしまうという問題が発生する．
文 3	という問題がある
文 4	そのような有効性に基づいて、
文 5	しかし，////という問題がある
文 6	の問題点として////という点が挙げられる
文 7	という問題がある
文 8	おり，操作性の面で問題があった
文 9	増加によって////出回るという問題が発生する．そのため////高まってき
文 10	という問題がある
文 11	ないという問題があっ
文 12	が////という問題があり
文 13	そのような背景から////と考えられる
文 14	という問題が発生する
文 15	問題点////の問題点////挙げ
文 16	という問題がある

本研究ではユーザをカテゴリごとに分けるため[[今回]]，一般ユーザに対して，より有効と考えられる SNS 機能としてのマイクロブログサービス [[に注目する．]] 特に，ユーザが行う返信行動に着目し，その元投稿及び返信内容から投稿内容ベクトルを生成した上で，ユーザのクラスタリングを行う．そして，獲得クラスタと現実生活に存在するコミュニティとの比較を行うこととする．

図 3: 「目的」についてレベル 2 からレベル 5 に修正した論文の一部

言葉の意味と表現形式は多対多の関係であることが多く、ゆえに自然言語処理は challenging な領域であると考えられる。言い換え知識獲得は、同じ意味を持つ複数の異なる表現形式を認識／生成するための知識を獲得する技術である。本稿では 大量の言い換え知識獲得を目的としてWeb 上の定義文からの言い換え知識獲得法を提案する。「言い換え」は両方向の含意関係が成立する表現対と定義する。同じ概念を定義する文は Web に大量に存在し、それらは言い換え関係にある場合が多く、それゆえ言い換え知識の宝庫と考えられる。

図 4: 「目的」についてレベル 1 からレベル 5 に修正した論文の一部

3.4 考察

論文の記載必要項目「目的」「問題点」について論文を人手で修正し, 修正結果から分析を行った．3.3.2 節の表 4, 表 5 と 3.3.3 節の表 7, 3.3.4 節の表 8 から, 記載必要項目「目的」であれば「〜を目的として〜」「〜という問題を解消する〜」という修正パターンが得られ, 記載必要項目「問題点」であれば「〜という問題がある (発生する)」という修正パターンが得られた．この結果から, 記載必要項目「目的」「問題点」について書く場合「目的」「問題点」という単語を使うことが一番内容が理解しやすい書き方であると考えられる．先行研究 [1][2] で記載必要項目が欠落している論文の自動検出の際に，「目的」「問題点」という単語が書かれていない論文を記載必要項目「目的」「問題点」について不明瞭な論文であると判断するルールベース手法を用いていた．今回の分析結果は, 記載必要項目が欠落している論文の自動検出において, 機械学習手法と比べてルールベース手法のほうが精度が良い結果になる理由の一つにもなると考える．

表 6: 修正文に出現する頻度の高い単語 (一部)

項目名	2 単語連続	3 単語連続
目的	を目的	を目的として
	目的として	獲得を目的
	するため	するために
	提案する	手法を提案
	問題を	問題を解消
	解消する	解消するために
問題点	問題が	という問題が
	という問題	問題がある
	問題点	問題が発生
	そのよう	が発生する
	発生する	そのような
	ような	問題があっ

表 7: 修正文に出現する単語の頻度調査で得られた修正パターン

項目名	修正パターン
目的	〜を目的として〜
	〜するために〜
	〜手法を提案する〜
	〜問題を解消する〜
問題点	〜という問題が〜
	〜問題が発生する〜
	〜の問題点が考えられる〜

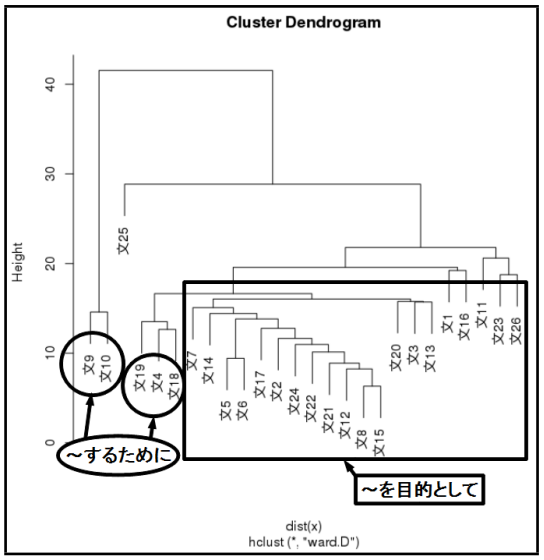

図 5: 「目的」についての修正文の階層クラスタリング結果

表 7 と表 8 を比較すると, 単語頻度調査で得られた記載必要項目「目的」の修正パターンの中には「〜するために〜」があるが, 階層クラスタリングで得られた記載必要項目「目的」の修正パターンの中にはないことが分かる．「〜するために〜を行う」といった表現は目的を示す文として考えられるので, 単語頻度調査での分析のほうが単純な方法でより良い修正パターンを獲得できると考える．しかし, 階層クラスタリングの図 5 の結果から, 記載必要項目「目的」の修正文の傾向として大きく分けて「〜するために」と「〜を目的として」の 2 種類だということが分かる．階層クラスタリングは修正文全体の傾向から修正パターンを獲得するのに適していると考える．

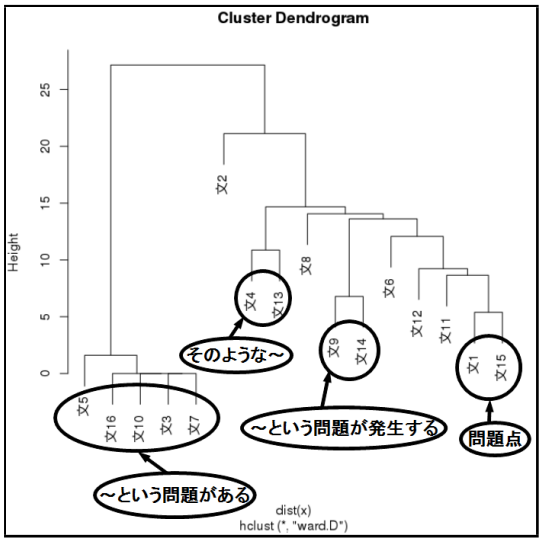

図 6: 「問題点」についての修正文の階層クラスタリング結果

表 8: 階層クラスタリングで得られた修正パターン

項目名	比較した文番号	修正パターン
目的	文 3-文 13	～の～を目的として
	文 5-文 6	～の解明を目的として
	文 9-文 10	～では，そのようなコストの問題を解消するために～特許データベース～からシ ソーラスを自動的に構築する手法を提案する．
	文 23-26	～と～し～日本語ニュースの～を目的として
問題点	文 1-文 15	～問題点～
	文 4-文 13	～そのような～
	文 9-文 14	～という問題が発生する～
	文 3-文 7-文 10-文 16	～という問題がある～

3.5 修正のヒント出力

3.1.3 節と 3.1.4 節の分析より表 7 と表 8 のような修正パターンが得られた．そこで, 図 7 のように記載必要項目が欠落している論文を自動検出し (先行研究 [1][2] の利用), 論文著者に対して今回得られた修正パターンを掲示する方式で文章作成支援が可能であると考える．この方式を修正のヒント出力方式と呼ぶ．修正のヒント出力方式は直接的な修正や完全な自動修正ではないが, 記載必要項目が欠落しているか否かとその修正方法が掲示されるため, 論文著者の確認作業や修正作業の負担を軽減できると考える．また, 修正のヒント出力方式ではピンポイントで論文の欠落箇所を指摘することで更なる作業負担の軽減が考えられる．現状では論文の欠落箇所の指摘まではできていないので, 今後の課題とする．

4 おわりに

論文の記載必要項目「目的」「問題点」について単語頻度調査と階層クラスタリングを用いて分析を行った．その結果, 記載必要項目「目的」であれば「～を目的として～」「～という問題を解消する～」, 記載必要項目「問題点」であれば「～という問題がある」「～という問題が発生する」などの多くの修正パターンが獲得できた．

単語頻度調査により得られた修正パターンと階層クラスタリングにより得られた修正パターンを比較すると, 単語頻度調査により得られた修正パターンにのみ「～するために～」があることが分かった．このことから, 階層クラスタリングによる分析より単語頻度調査による分析のほうが単純でより良い修正パターンが獲得できることが分かった．また, 階層クラスタリングによる分析は, 修正文全体の傾向を把握し, 修正パターンを獲得するのに適していることが分かった．

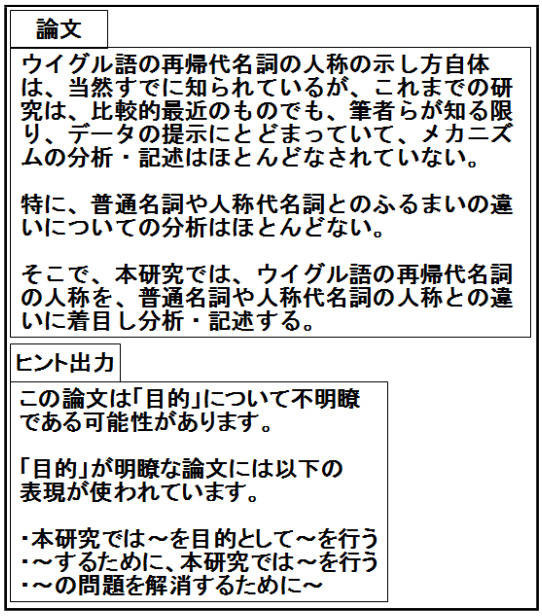

図 7: 修正のヒント出力例

今回の分析で得られた結果を使って, 記載必要項目が欠落している論文の著者に対して修正パターンを掲示する修正のヒント出力方式を考案した．この方式により, 記載必要項目が欠落しているか否かとその修正方法が掲示されるため, 論文著者の確認作業や修正作業が軽減できると考える．修正のヒント出力方式をより良くするためには, 論文の欠落箇所をピンポイントで指摘する技術が必要であると考える．現状では, 論文の欠落箇所の指摘できる段階ではないので, 今後の課題として取り組む．

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] 岡田拓真, 村田真樹, 徳久雅人, 馬青: “論文からの記載必要項目の抽出と文章作成支援”, 言語処理学会第 21 回年次大会, P4-25, pp.980-991, 2015.

[2] 岡田拓真, 村田真樹, 馬青: “論文における記載不備の自動検出と自動修正に向けた分析 ”, 言語処理学会第 22 回年次大会, P7-2, pp.176-179, 2016.

[3] 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.

[4] 掛谷英紀: “最大エントロピー法の解析的解法”, 言語処理学会第 16 回年次大会, PB1-13, pp.470-473, 2010.

[5] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均: “最大エントロピーモデルと書き換え規則に基づく固有表現抽出”, 自然言語処理, Vol.7, No.2, pp.63-90, 2000.

[6] 村田真樹: “diff を用いた言語処理-便利な差分検出ツール mdiff の利用”, 自然言語処理, Vol.9, No.2, pp.91-110, 2002.