概要

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, 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, そのような記載不備のある論文に対して文章作成支援を行うことを目的とする.

記載不備論文の文章作成支援のプロセスには, 「1. 記載不備論文の検出」と「2. 記載不備論文の修正」が考えられる. 記載不備論文の検出を扱う研究として岡田ら [1] の研究が挙げられる. 岡田ら [1] は, 論文に記載すべき情報を「記載必要項目」と定義し, それらの情報が欠落している論文の自動検出法としてルールベース手法を提案している. 論文の記載必要項目と記載必要項目の検出に役立つ単語を決定し, そ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であるとして自動で検出を行っている. しかし, 先行研究 [1] では, 他の検出手法との比較を行っておらず, また「記載不備論文の修正」について扱っていないため, 記載不備論文の文章作成支援として不十分である. そこで，本研究では機械学習を利用した記載不備論文の自動検出と先行研究の手法の比較と記載不備論文の修正に向けた分析を行う.

記載不備論文の自動検出手法として先行手法のルールベース手法と比較手法として提案した機械学習手法の比較を行った結果, どの記載必要項目においても機械学習手法と比べてルールベース手法のほうが検出精度が高く, 先行手法であるルールベース手法の有効性を確認できた.

記載不備論文の修正に役立つ修正パターンを獲得するために論文の記載必要項目「目的」「問題点」について人手で修正を行い, 得られた修正文を単語連続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」「問題点」の修正パターンが獲得できた. 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.

目 次

表 目 次

図 目 次

第1章　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, 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, そのような記載不備のある論文に対して文章作成支援を行うことを目的とする.

記載不備論文の文章作成支援のプロセスには, 「1. 記載不備論文の検出」と「2. 記載不備論文の修正」が考えられる. 記載不備論文の検出を扱う研究として岡田ら [1] の研究が挙げられる. 岡田ら [1] は, 論文に記載すべき情報を「記載必要項目」と定義し, それらの情報が欠落している論文の自動検出法としてルールベース手法を提案している. 論文の記載必要項目と記載必要項目の検出に役立つ単語を決定し, そ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であるとして自動で検出を行っている.

先行研究 [1] の問題点として以下の 2 点が考えられる.

- 他の記載不備論文の検出手法との性能比較を行っていないため, 先行研究の提案手法の有効性が確認できない.
- 「記載不備論文の修正」について扱っていないため, 記載不備論文の文章作成支援として不十分である.

これらの問題点の解消に向けて, 本論文では, 以下の研究を行う.

- 機械学習を利用した記載不備論文の自動検出手法と先行研究 [1] の提案手法の比較を行い, 先行研究 [1] の有効性を確認する.
- 記載不備論文を人手で修正し, 修正した部分の分析を行う. 修正部分の分析を行うことで, 記載不備論文の修正に役立つ修正パターンを獲得する. 獲得した修正パターンを利用して, 記載不備論文の修正支援となる技術を構築する.

本研究の主な主張点は以下である.

- 記載不備論文の自動検出において, ルールベース手法と機械学習手法を比較すると, ルールベース手法のほうが検出精度が高く, 先行研究の有効性が確認できた.
- 単語連続の頻度調査と階層クラスタリングによる分析で, 多くの論文修正に役立つ修正パターンが獲得できることが分かった.
- 階層クラスタリングよりも単語連続の頻度調査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.
- 獲得できた修正パターンを使って, 記載不備論文の著者に対して修正パターンをヒントとして掲示することで論文の文章作成支援を促す方式を考案した.

第2章　関連研究

関連研究として以下の研究が挙げられる.

2.1　論文の閲覧支援やサーベイ支援の研究

村田ら [2] 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 [3] 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.

難波ら [4] の研究では, 論文の閲覧支援を目的として, 研究分野の動向を概観するのに必要不可欠である研究動向情報を論文から自動的に抽出し, 可視化する研究を行っている. 特定の分野の論文を収集し, それらの論文から可視化に必要な情報を抽出している. 抽出した情報を表にし, 可視化を行う.

村田ら [2] の研究と樫本ら [3] の研究と難波ら [4] 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.

2.2 欠落個所の指摘を目的とした研究

灘本ら [5] 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 [5] の研究と本研究を比較すると, 灘本ら [5] 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 [5] 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が存在している.

2.3 文章作成支援の研究

都藤ら [6] 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 [6] 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 [6] 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.

都藤ら [6] 以外にも数多くの研究で文章作成支援が行われている [9][10][11][12][13]. 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.

2.4 論文作成支援の研究

Ptaszynski ら [7] の研究では科学論文の作成支援システムの提案をしている. システムが科学論文を執筆するのに必要な精度やスコアなどの実験データを自動的に計算することで科学論文の作成支援を行っている.

Liu ら [8] の研究では学生が自分の論文を改訂するための支援の形として自動質問生成 (AOG) ツールの開発を行っている. 「この論文の著者は先行研究の考えに反対ですか？」「あなたはこの論文を読んで著者が反対であるという証拠を見つけることができますか？」などの質問を自動作成する. その質問を学生が参考にすることで論文の改訂支援を行っている. 論文を読んで自動質問生成ツールでは, 引用から取得した構文情報や意味情報を基に質問を生成している.

Ptaszynski ら [7] の研究や Liu ら [8] の研究は論文を対象とした作成支援を目的に行っている. 論文の作成支援のなかでも, Ptaszynski ら [7] は実験データの自動計算システムにより論文の実験データ作成支援を行っており, Liu ら [8] は自動質問生成ツールにより論文作成支援を行っている. それらの研究に対し, 本研究は, 論文の記載必要項目の欠落しているか否かを自動判別することで論文の文章作成支援を行っている. Ptaszynski ら [7] の研究と本研究を比較すると, 論文の実験データの作成支援と論文の文章の作成支援といった作成支援の対象が違っている. Liu ら [8] の研究と本研究を比較すると, 自動質問生成ツールを用いた論文作成支援と論文の記載必要項目を利用した論文作成支援といった論文作成支援へのアプローチ方法が違っている.

第3章　記載必要項目

3.1　概要

本研究では, 先行研究 [1] で決定した記載必要項目を用いる. 「記載必要項目」とは, 研究の目的などの論文に記載すべき情報であると定義している. 本章では, 先行研究 [1] で行った記載必要項目の決定方法と説明を記す.

3.2　記載必要項目と検出に役立つ単語の決定方法

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 [14] 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．先行研究 [1] 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 [14] を利用する．意味ソート [14] 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．

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

3.2.1 節で挙げられた方法で頻度調査を行った．全論文数は 393 件あり, その論文中に出現する単語の総数は 19,234 単語であった．その内の出現率の高い上位 100 単語までの結果をまとめて表 3.1 に示す．

表 3.1: 論文内で出現率が高い上位 100 単語

単語	出現率	単語	出現率	単語	出現率
示す	1.000	従事	0.901	工学部	0.831
研究	1.000	実験	0.898	作成	0.831
場合	1.000	提案	0.898	有効	0.828
用いる	0.997	工学	0.898	分類	0.828
考える	0.997	基づく	0.898	今後	0.828
結果	0.997	**比較**	**0.895**	呼ぶ	0.826
言語	0.997	博士	0.895	科学	0.826
処理	0.997	同様	0.895	一方	0.826
情報	0.997	大きい	0.895	**重要**	**0.811**
必要	0.994	学会	0.895	定義	0.808
に対して	0.994	一般	0.895	論文	0.805
述べる	0.977	に対する	0.895	行う	0.802
得る	0.971	情報処理	0.895	抽出	0.797
可能	0.971	存在	0.890	次に	0.797
以下	0.965	与える	0.881	名詞	0.791
含む	0.962	に関する	0.881	出現	0.791
自然	0.959	表す	0.878	同年	0.788
対象	0.956	卒業	0.875	程度	0.788
関係	0.953	解析	0.875	会員	0.788
異なる	**0.951**	修了	0.875	条件	0.785
以上	0.951	複数	0.869	求める	0.782
方法	0.948	全体	0.869	適用	0.779
現在	0.948	手法	0.866	単語	0.779
大学院	0.942	日本語	0.863	構造	0.779
課程	0.942	**例えば**	**0.858**	**目的**	**0.773**
高い	0.936	対応	0.858	特に	0.770
利用	0.933	考慮	0.858	少ない	0.767
表現	0.933	見る	0.855	精度	0.765
多い	0.930	多く	0.852		
意味	0.930	計算	0.852		
持つ	0.922	以外	0.852		
問題	**0.916**	実際	0.849		
同じ	0.913	構成	0.843		
評価	0.907	種類	0.840		
システム	0.904	説明	0.834		
部分	0.901	データ	0.834		

3.4.2 意味ソートの結果

論文での出現率の高い上位 500 単語を意味ソート [14] 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 [14] の結果の一部を図 3.1 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 3.1: 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「問題」「目的」などの出現率が高いことがわかった. 「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 3.4.2 節のような意味ソート [14] 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 3.2 に示す. 表 3.2 以外の記載必要項目もあると考えられるが, 3.2 節の決定方法では, 表 3.2 で挙げた 4 項目を決定した.

表 3.2: 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	定義
比較	比較　比べる	先行研究との比較 実験結果の比較
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	その研究を行う理由
例	例えば　　例　　具体	具体例

第4章　記載不備論文の自動検出

4.1　概要

本章では, 先行研究である岡田ら [1] が提案したルールベース手法と本研究で比較手法として提案する機械学習手法の比較を行い, 自動検出手法の有効性を検証する.

4.2　ルールベース手法 (先行手法)

表 3.2 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 3.2 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であると判別し, 自動で検出する.

4.3　機械学習手法 (比較手法)

記載不備論文であると人手で判別した論文と記載不備論文ではないと人手で判別した論文の 2 分類のデータに対して, 2 値分類を教師あり機械学習を利用して行うことで, 記載不備論文であると人手で判別した論文を自動で検出する.

機械学習手法では, 以下のものを素性とする.

- 平仮名のみの単語を除いた論文全体に出現する全単語
- 検出に役立つ単語 (表 3.2 参照)

4.4 最大エントロピー法

本実験では，教師あり機械学習法に，最大エントロピー法を使用する．最大エントロピー法の説明を記述する．

A, B を分類と文脈の集合とすると, 文脈 $b(\in B)$ で分類 $a(\in A)$ となる事象 (a, b) の確率分布 $p(a, b)$ を最大エントロピー法で推定する．文脈 b は k 個の素性 $f_j(1 \leq j \leq k)$ の集合で表す．

$$g_j(a,b) = \begin{cases} 1 & (exist(b.f_j) = 1 \ \&\text{分類} = a) \\ 0 & (\text{それ以外}) \end{cases} \tag{4.1}$$

式 (4.1) のような素性関数を定義する．式 (4.1) では, 文脈 b において, 素性 f_j が観測され, 分類が a になるときに 1 を返す．$exist(b, f_j)$ は, 文脈 b において素性 f_j が観測されるか否かで 1 あるいは 0 を返す関数である．

推定する確率分布 $p(a|b)$ による素性 f_j の期待値と既知データにおける確率分布 $\tilde{p}(a, b)$ による素性 f_j の期待値が等しいことを制約として, エントロピー最大化 (確率分布の平滑化) を行って, 出力と文脈の確率分布を求める (式 (4.2)).

$$\sum_{a \in A, b \in B} \tilde{p}(b) p(a|b) g_j(a,b) = \sum_{a \in A, b \in B} \tilde{p}(a,b) g_j(a,b) \quad for \quad \forall f_j (1 \leq j \leq k) \tag{4.2}$$

式 (4.2) の $\tilde{p}(b)$ を既知データにおける事象 b の出現頻度 $freq(b)$, $\tilde{p}(a, b)$ を既知データにおける事象 (a, b) の出現頻度 $freq(a, b)$ により以下のように推定する．

$$\tilde{p}(b) = \frac{freq(b)}{\sum_{b \in B} freq(b)} \tag{4.3}$$

$$\tilde{p}(a,b) = \frac{freq(a,b)}{\sum_{a \in A, b \in B} freq(a,b)} \tag{4.4}$$

よって, 式 (4.2) の制約を満たす確率分布 $p(a, b)$ のうち, 式 (4.5) を最大にする確率分布が推定したい確率分布である．

$$H(p) = -\sum_{a \in A, b \in B} \tilde{p}(b) p(a,b) log(p(a,b)) \tag{4.5}$$

最大エントロピー法は, この推定した確率分布にしたがって求まる各分類の確率のうち, 最も大きい確率値を持つ分類を求める分類方法である [15][16][17][18].

4.5 データ

本実験では, 2011 年度の言語処理学会年次大会論文 (266 件) を学習用データとして使用し, 2012 年度の言語処理学会年次大会論文 (305 件) を評価用データとして使用する. また, 記載必要項目が欠落していると人手で判別した論文を正解データ, 記載必要項目が欠落していないと人手で判別した論文を不正解データとしている. 機械学習手法では, 正解データと不正解データを同数にして学習を行っている. データの詳細を表 4.1, 表 4.2 に示す.

表 4.1: 2011 年の言語処理学会年次大会論文の詳細

項目名	正解	不正解	総数
比較	53	213	266
問題点	73	193	266
目的	83	183	266
例	7	259	266

表 4.2: 2012 年の言語処理学会年次大会論文の詳細

項目名	正解	不正解	総数
比較	59	246	305
問題点	114	191	305
目的	94	211	305
例	9	296	305

4.6 評価方法 (F値)

本研究では, 記載不備論文の自動検出の精度を再現率 (*recall*), 適合率 (*precision*), F 値 (*F-measure*) で評価する. 再現率と適合率は以下の式で算出される.

$$\text{再現率} = \frac{\text{システムの正解数}}{\text{テストデータ中の正解数}} \tag{4.6}$$

$$\text{適合率} = \frac{\text{システムの正解数}}{\text{システムの出力数}} \tag{4.7}$$

本研究では文章作成支援に役立っている論文を正解として式 (4.6) と式 (4.7) を算出した. また, 式 (4.6) と式 (4.7) の値の調和平均式 (4.8) を求めることで F 値を算出できる.

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{4.8}$$

4.7 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 4.3 から表 4.6 に示す.

表 4.3: 「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
ルールベース	0.58 (34/59)	0.60 (34/57)	0.59
機械学習	0.61 (36/59)	0.21 (36/174)	0.31

表 4.4: 「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
ルールベース	0.61 (70/114)	0.81 (70/86)	0.70
機械学習	0.69 (79/114)	0.47 (79/169)	0.56

表 4.5: 「目的」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
ルールベース	0.53 (50/94)	0.60 (50/84)	0.56
機械学習	0.44 (41/94)	0.32 (41/127)	0.37

表 4.6: 「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
ルールベース	1.00 (9/9)	0.75 (9/12)	0.86
機械学習	0.33 (3/9)	0.02 (3/129)	0.04

4.8 検出の成功例と失敗例

記載不備論文として自動で正確に検出できた論文例と誤って検出した論文例を図 4.1 から図 4.4 に示す. 本研究では, 記載不備がある論文を検出できた場合, 検出に成功したものとしている. 同様に, 記載不備のない論文を検出した場合, 検出に失敗したものとしている. また, 記載必要項目「比較」「例」は, 「比較しているか否か」「例があるか否か」という判別になっているため, 省略する.

> 質問応答システムは従来の Web 検索に比べ，より具体的にユーザが必要とする情報だけを提示するシステムである．
> 本研究では，factoid 型で回答できる質問文に対応し研究したものであり，non-factoid 型については対応していない．質問文に含まれる新情報と旧情報を利用して Yes-No を判断するシステムを構築する．旧情報に対して新情報が本当に正しいかを判断することで回答を決める． ...

図 4.1: 「目的」についてルールベース手法で検出に成功した論文の一部

図 4.1 の論文は 1 節の「はじめに」の部分を全て引用している. 質問応答システムに関する研究であるが, 何のために質問応答システムの研究を行うか書かれておらず, 記載必要項目「目的」について理解ができない. よって研究の目的を明記する必要がある (記載必要項目を補う必要がある) と考え, 記載不備論文の検出に成功したと判別した.

... 現在, 機械翻訳の分野において, 対訳文対から自動的に翻訳規則を生成し翻訳を行う統計翻訳が注目され, 研究が盛んに行われている. 統計翻訳では, イタリア語-英語など文法構造が類似する言語対において翻訳精度が高くなり, 日本語-英語などの文法構造の異なる言語対においては翻訳精度が低くなる傾向がある. 別の翻訳手法にパターン翻訳がある. パターン翻訳では文パターン辞書と単語辞書を用いて翻訳を行う. 文パターンが有する大局的な文法情報を用いることで, 翻訳文全体の構造を保持した翻訳精度の高い翻訳文を生成出来る利点がある. **しかし, 従来, 文パターン辞書の作成は人手で行うため, 開発にコストがかかる欠点がある.**
そこで本研究では, 文パターン辞書を対訳文対から自動的に作成する手法を検討する. 文パターン辞書の自動作成により, 開発にかかるコストの削減が可能となる. ...

図 4.2: 「目的」についてルールベース手法で検出に失敗した論文の一部

図 4.2 の論文において, 読み手が想像する研究の目的として以下のものが挙げられる.

- 文パターン辞書の開発にコストがかかるという問題点を解消することを目的としている
- 自動生成された文パターン辞書と人手作成された文パターン辞書での翻訳精度調査を目的としている
- パターン翻訳の精度向上を目的としている

この例の研究では, 文パターン辞書の開発にコストがかかるという問題点を解消することを目的としていると考えられる. この論文内に「X という問題点がある. そこで本研究では Y を行う」という文章が存在していることがわかる. このような文章は前に記述された問題点の解消を行うことを目的として研究を行っていることが容易に理解できる文章が使われている論文であると考える.

... 近年, 評判分析の対象として, Twitter が注目されている. ○○は, 「:)」や「:(」などの emoticon を利用した訓練データの獲得と機械学習による極性判定 (肯定・否定判定) の手法について提案した. ××は, Twitter 特有の機能である「リプライ」などに着目し, 極性判定する手法について提案している. □△は, Twitter によく見られる「Cooool」のような繰り返し表現に着目し, その繰り返し表現の正規化や感情との関連性などについて検証している.

ここで, 我々は Twitter 上で用いられる文体に着目する. Twitter 上にはさまざまな文体が存在し, その特性が異なる. 例えば, 話し言葉に近い口語体では, 特徴的な文末表現や記号などによる顔文字の多用によってその感情を表現する傾向があるのに対し, 書き言葉的な文語体では, 言語表現そのものによってその感情が表されることが多い. このような文体の違いを考慮することは, さまざまな場面に有効であると考えられる. ...

図 4.3: 「問題点」について機械学習手法で検出に成功した論文の一部

記載必要項目「問題点」というのは先行研究の問題点や研究の背景を差している. 図 4.3 の論文では, 先行研究について述べられており, さらに研究の有効性も記述されている. しかし, 先行研究の手法の概要のみを記述しており, 先行研究で生じた問題についての記述が存在していない. 仮に先行研究で問題が生じていなかったと考えても, その場合は先行研究の手法の概要と先行研究との明確な違いを記述する必要があると考える.

研究の背景として「近年, 評判分析の対象として, Twitter が注目されている. 」とあるが, この例の文章であると何故 Twitter が評判分析の対象として注目されているかが理解し難いと考える.

... 近年, Facebook や Twitter などのマイクロブログが急速に普及し, **ユーザによるマイクロブログを用いた情報発信が活発化している.** 特に Twitter は, 140 文字という制限によりユーザの情報発信への敷居が大きく下がっており, 2011 年 3 月 11 日に発生した東日本大震災においては, リアルタイムに情報を伝える重要な情報インフラの 1 つとして活用された. **しかし, 安否情報などの重要な情報の共有・伝搬が行われた一方で, 多くの流言も拡散された. 流言は適切な情報共有を阻害し, 特に災害発生時には, 流言が救命のための機会損失を生む場合もある**ため, 流言の広がりにくい環境を作る必要がある. ...
(中略)
... そこで本研究では, 流言拡散を防ぐための仕組みとして, 流言情報クラウドを提案する. 流言情報クラウドは, リアルタイムに流言情報を蓄積し, その情報を提供することにより, 流言拡散を防止する. ...

図 4.4: 「問題点」について機械学習手法で検出に成功した論文の一部

図 4.4 の論文では, マイクロブログを用いた情報発信が活発化して様々な利点が生まれたが, その反面として多くの流言が発生してしまい, それが原因で適切な情報共有ができなくなるという問題点が説明されている. この論文では, 「ある事象 A があります. 事象 A により○○のようなメリットが存在します. しかし, その反面で××というデメリットも存在してしまいます. 」という文章が使われていることがわかる. 研究の背景としてある事象を例として利点を挙げ, その後に欠点を記述することにより何が利点で何が欠点なのかが容易に理解できる文章が使われている論文であると考える.

4.9 考察

表 4.3 から表 4.6 の結果より, 機械学習手法とベースライン手法の F 値を比較すると, ルールベース手法のほうが F 値が高くなっていることが分かる. このことより記載不備論文の検出においてはルールベース手法は有効であると考える.

機械学習手法の精度が低い原因として, 素性の数が考えられる. 機械学習手法では論文全体に出現した全ての単語を素性として利用している. その結果, 素性の数が多くなってしまい, 機械学習が文章作成支援の対象である論文を検出することができないという可能性があると考える. この原因については, 素性の再選定をする必要があると考える. 具体的には, 論文全体に出現した単語ではなく, 第一章に出現した単語のみを素性にするなどが考えられる.

機械学習手法の精度が低い原因として, 2 値分類による曖昧性が原因であると考えられる. 本実験では記載必要項目が欠落しているか否かの 2 値で分類しているが, 使用している論文データの中には, 記載必要項目について全く書かれていない論文や書かれているようであるが不明瞭な論文もある. また, その論文の読み手によって正解・不正解の基準が変わると考える. 専門家が読むことを想定すると, ある程度詳しく書かれていなくても記載必要項目について理解できるが, 専門家以外が読むことを想定すると, 詳しく書かなければ記載必要項目について理解できないと考える. そのような曖昧な論文については, 人手でも判別が難しく, 機械学習で判別するのはさらに困難なのではないかと考える. 分類を 2 値ではなく細かい分類にすることで機械学習手法の精度が高くなる可能性があると考える.

第5章　記載不備論文の修正に向けた分析

5.1　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. また, 修正には「文字列を追加だけする修正」「文字列を削除だけする修正」「文字列を変更する (削除して追加する) 修正」の 3 つが考えられるが, 本研究では, 「文字列を追加だけする修正」や「文字列を変更する修正」の際に追加した文字列について分析する.

5.1.1　5 段階レベルを使った分類

第 3 章では, 記載必要項目がある論文とない論文の 2 値分類で実験を行っていた. 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 5.1 に示す.

表 5.1: 5 段階のレベルの定義

レベル	定義
5	手がかり表現があり, 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

5.1.2 人手による論文修正

5.1.1 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.

5.1.3 修正文に出現する単語連続の頻度調査

5.1.2 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.

5.1.4 階層クラスタリングによる分類

R による階層クラスタリングを用いて 5.1.2 節で抽出した修正文を分類し, 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. また, 5.1.3 節の方法だけでも修正パターンを得られると考えられるが, 網羅性を向上させるために階層クラスタリングを用いる.

階層クラスタリングのアルゴリズムとしては, 最短距離法や群平均法といった様々なアルゴリズムが提案されている. 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. 最後にベクトルの全ての要素を足したものを各要素で割って大きさを 1 にし, 類似度を算出する. ベクトルの例を表 5.2 に示す.

表 5.2: ベクトルの例

文番号	要素								
	問題	が	という	…	問題が	という問題	…	という問題が	…
文 1	1	0	1	…	0	2	…	0	…
文 2	1	1	1	…	2	2	…	3	…
文 3	1	0	1	…	2	0	…	3	…

5.2 mdiff

本研究では, mdiff[19] を用いて差分の抽出を行い，記載不備論文の修正文を獲得する.

UNIX に標準で搭載されているコマンドとして diff コマンドがある. この diff コマンドは一般には差分の検出に利用され, 与えられた二つのファイル間の違いを探し, 順序情報を保持したまま結果を出力する. 例えば,

今日
学校へ
行く

ということが書いてあるファイルを

今日
大学に
行く

という文に修正して, 書き換えたファイルがあるとする. これらの 2 つのファイルに対して diff コマンドを行うと差分の部分が

< 学校へ
> 大学に

のような形で出力される.

「今日」と「行く」という部分は修正前のファイルにも修正後のファイルにも存在しており, 書き換えられていない. これは修正前後での共通部分として扱い, 共通部分の間にある書き換えられた「学校」と「大学」が差分として出力される仕組みとなっている.

これに加えて diff コマンドには-D オプションという便利なオプションがあり, これをつけて diff コマンドを使うと差分部分だけでなく共通部分も出力される. つまりファイルのマージが実現される. しかし ifdef という機械的な記号が含まれてくるため人間の目で認識しづらい面もある. そこで差分部分の始まりに▼, 二つのデータの境界に●, 差分部分の終わりに▲を用いて出力結果を表すことにする. 実際の形式としては,

▼▼▼▼▼
(一つめのファイルにだけある部分)
●●●
(二つめのファイルにだけある部分)
▲▲▲▲▲

となる. これをマージを行う diff として mdiff[19] と呼ぶ. そこで先程のデータに対して mdiff[19] を行うと以下のような結果になる.

今日
▼▼▼▼▼
学校へ
●●●
大学に
▲▲▲▲▲
行く

5.3 ウォード法

本研究では, 階層クラスタリングのアルゴリズムとしてウォード法を用いる. ウォード法は, 1 つのクラスタに含まれる要素とクラスタ間に含まれる要素間の分散の比を最小化するように 2 つのクラスタの距離を近似する手法である. 他の手法に比べて分類感度が高いため, クラスタが肥大化しにくいという長所があり, 均等なクラスタの生成ができる.

階層クラスタリングにおいて, クラスタ C_i とクラスタ C_j が統合されて, クラスタ C_{ij} が形成される場合, クラスタ C_{ij} に属していないクラスタ C_k との距離が統合前の距離との比較で, そのクラスタを統合するか否かを決定する.

クラスタ C_i とクラスタ C_j の距離を距離 d_{ij}, クラスタ C_i とクラスタ C_k の距離を距離 d_{ik}, クラスタ C_j とクラスタ C_k の距離を距離 d_{jk}, クラスタ C_{ij} とクラスタ C_k の距離を距離 $d_{(ij)k}$ とすると, 階層クラスタリングの式は以下のようになる.

$$d_{(ij)k} = \alpha \cdot d_{ik} + \alpha \cdot d_{jk} + \beta \cdot d_{ij} + \gamma \ |d_{ik} - d_{jk}| \tag{5.1}$$

ここでのパラメータ α, β, γ は, 以下のウォード法のパラメータ式により決定する.

$$\alpha = \frac{|C_i| + |C_k|}{|C_i| + |C_j| + |C_k|},\ \beta = \frac{|C_k|}{|C_i| + |C_j| + |C_k|},\ \gamma = 0 \tag{5.2}$$

5.4 データ

2011 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.

5.5 分析結果

5.5.1 5 段階レベルを使った分類: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 5.3 に示す.

各レベルの論文の一部を図 4.1 から図 4.10 に示す.

表 5.3: 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

... 個人の知識レベルや学習段階に応じた語彙・辞書資源の整備は、専門知識を基盤とするコミュニケーションの円滑化を支援するために有効な方策の一つである。そのためにはまず、知識レベルや学習段階に応じた形で現実に存在する語彙の特徴を把握しておく必要がある。そこで本研究では、中学・高校・大学の教科書における知識の構成を専門語彙のネットワーク構造として分析・比較することで、学校段階に応じた語彙体系の特徴を明らかにすることを 目的とする。 ...

図 5.1: 「目的」についてレベル 5 であると判別した論文の一部

図 4.1 の論文では, 「目的とする」といった表現があり, 誰が読んでも容易に目的が理解できる論文であることが分かる. また, 目的の記述パターンとして「〜を目的とする」「〜を目標にして〜」「〜のために〜を行う」といった表現が多く見られた.

... 上記で示したサービスでは，機械学習に基づく文書分類の技術を用いている．たとえば，ナイーブベイズ識別器やサポートベクターマシン (SVM) のような識別器が著名である．このとき，高い分類精度を実現するためには，大量の学習データから構築されたコーパスを用意しなければならない．このようなコーパスは大量のラベル無しデータに対し，人手によるラベル付与 (アノテーション) を行う作業を通して実現される．このとき，アノテーションの量が多くなるに連れ，人的コストと時間が増大することが 課題となる.
そこで, 上記で述べたサービスを対象とした文書分類用コーパスを構築する際に, アノテーションの量を減らす手法として，クラスタリングに基づく能動学習を用いた文書分類用コーパスの 構築手法を提案する.....

図 5.2: 「目的」についてレベル 4 であると判別した論文の一部

図 4.2 の論文では, 問題点の記述のすぐあとに「そこで本稿では〜を行う」といった表現が書かれており, 文脈から問題点解消が目的であることが理解できる. 問題点のあとに「そこで」などの表現を使うことにより, 考えなくても直感的に目的が理解できる.

... テレビ番組の字幕（closed captions）は、聴覚障害者への情報保障の１つの大きな柱である。近年は生放送でない番組にはほとんど字幕が付くようになった。総務省の視聴覚障害者向け放送普及行政の指針では、字幕付与の対象となる番組を、生放送番組（一部を除く）を含めた全ての番組まで拡大し、平成 29 年度（2017 年度）までに実施することを目標としている。
これを受けて、現在では、スポーツ番組などの生放送番組へのリアルタイムでの字幕の付与 が実施されるようになってきている。本稿は、テレビのスポーツ生放送番組（サッカーと大相撲の番組）に実際にリアルタイムで付与された字幕に関して、音声の書き起こしデータとの比較を行いながら、その文字数、固有表現の頻度、表示速度を中心とする 基本的な調査を行った。...

図 5.3: 「目的」についてレベル 3 であると判別した論文の一部

図 4.3 の論文では, 問題点が記述のすぐあとや「そこで」などの表現がないため, 論理的に考える必要がある. 文脈だけでなく内容を把握し考えることで論理的に目的が理解できる.

... 近年，Twitter をはじめとするマイクロブログサービスが急速に普及してきており，日々多数の投稿がなされている．マイクロブログは従来のブログサービスと比べて非常に高速な情報伝達スピードを持つ情報発信ツールである．
本研究は マイクロブログ特有のリアルタイムな投稿を活用し，ユーザに対して効果的な情報推薦を行う 手法を提案する. 過去の投稿を分析することでユーザの嗜好を推測し，実際の商品データを用いて効果的なタイミングで情報推薦を行うことは実用的であり 利用価値は高い と言える．...

図 5.4: 「目的」についてレベル 2 であると判別した論文の一部

図 4.4 の論文では, 研究を行う背景 (問題点) が見当たらないが, 研究の有効性が書かれていることが分かる. 深い洞察をしなければ, 目的が理解できない論文である.

... 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 5.5: 「目的」についてレベル 1 であると判別した論文の一部

図 4.5 の論文では, 手法のみ書かれており, 目的が理解できない論文である. 問題点などについても言及しておらず, 何のためにその手法で研究を行うのかが理解できない.

... しかしながら，そのような大規模コーパスによる言語モデルには以下に挙げる 2 つの問題点がある.
[構築時の問題点] 大規模なコーパスから N-gram を集計する処理には，莫大な計算とメモリを必要とする．また，そもそも大規模コーパスを 1 台のコンピュータに保存することは難しい.
[利用時の問題点] 言語モデルを実際に利用するには，検索などの必要な操作をリアルタイムに行える必要がある．そのためには，言語モデルをメモリ上に読み込めることが望ましいが，大規模な言語モデルでは現在のコンピュータのメモリ搭載量を上回ることが多い.
これらの問題はデータ量と性能とのトレードオフであり，完全に解決する方法は見つかっていない．そのため，言語モデルの応用を考える上でそのトレードオフについて知ることは重要である． ...

図 5.6: 「問題点」についてレベル 5 であると判別した論文の一部

図 4.6 の論文では,「以下に挙げる 2 つの問題点がある」という表現で研究の問題点を指していることが分かる.

... これまで研究・実用化されてきた音声対話システムはおおむね 2 種類に分類される。フライト情報案内やバスの運行案内などの明確なタスクを定義し、関係データベース（RDB）をバックエンドとした枠組みは、タスク達成に必要な意味表現の定義や対話のフローの記述が容易であった反面、Web などの大規模なテキスト情報に対して適用することが困難であった。それに対して、一般的な文書検索を用いた対話システムの研究も行われてきたが、表層的なキーワードや係り受け関係、質問タイプなどのみに着目し、深い言語的解析や対話処理は扱われていない。その結果、対話の文脈やユーザの要求とは無関係な、不自然な応答が生成されることがあった。これに対して本研究では、述語項構造に着目した情報抽出を行うことで、RDB のような構造を持たない Web 文書を扱いながら、その意味表現を扱えるシステムを構築する。 ...

図 5.7: 「問題点」についてレベル 4 であると判別した論文の一部

図 4.7 の論文では,「困難であった」「〜は扱われていない」といった表現で直接的ではないが問題点について記述していることが分かる.

... Web 上に存在する情報は，ブロードバンド化の進展やブログ等の普及に伴い，爆発的に増加し続けている．これらの情報の中には，出所が不確かな情報 や利用者に不利益をもたらす情報 などが含まれており，信頼できる情報を利用者が容易に得るための技術に対する要望が高まっている．そこで，我々は，利用者の信憑性判断を支援する技術の実現に向けて研究を行っている．...

図 5.8: 「問題点」についてレベル 3 であると判別した論文の一部

図 4.8 の論文では,「出所が不確かな情報」や「不利益をもたらす情報」といったマイナスの表現を使って問題点を間接的に表現していることが分かる.

... 言語研究のデータとして、近年コーパスが選択されることが多くなっている。市販のコーパスでは、『CD-ROM 版新潮文庫の 100 冊』が文法研究の用例抽出によく使われてきた。また、毎日新聞・日経新聞が全ての情報にアクセスできる言語研究用の記事データ集の発売を初めて開始したのは、1995 年である。大量のデータを収録し、一般に頻度の低い用例も検出できることから、新聞コーパスから用例を採集する研究も多く行われるようになってきている。

本研究の目的は、新聞の記事データ集は、言語資料として新聞紙面にどれほど忠実なのか、明らかにすることである。当然、新聞紙面と新聞記事データ集では、紙と CD というようにメディアが異なり、さらに写真・広告・文字装飾・段組の有無も異なる。しかし、新聞紙面と新聞記事データ集の間にはどのような違いが見られ、それが文法研究にどのような影響を与えるか、考察することで新聞記事データ集の適切な使用法が明らかになると思われる。 ...

図 5.9: 「問題点」についてレベル 2 であると判別した論文の一部

図 4.9 の論文では, 「新聞コーパスから用例を採集する研究も多く行われるようになってきている」という部分が背景に当たると考えられるが, 「研究が多く行われるようになってきている」が問題点に繋がるのは少し考えにくいため, レベル 2 としている.

... 我々は文章中に現れる比喩表現，その中でも直喩・隠喩的な比喩について，その認識・抽出を目的として研究を進めている.“ 名詞のような名詞 ”表現の比喩性判定モデルを提案し，現在は“ 名詞のように ”表現を対象としている.

“ 名詞のように ”表現は後に用言がつづくことが普通であるが，その用言には動詞，形容詞，形容動詞などがあり，それらは活用もするため，表現のパターンは多種多様である．手始めに，用言を形容詞連体形に限定し,「名詞＋のように＋形容詞＋名詞」表現を対象として，形容詞の情報（属性）を比喩性判定に利用することを試みる．...

図 5.10: 「問題点」についてレベル 1 であると判別した論文の一部

図 4.10 の論文は, 記載必要項目「問題点」について全く理解できなかった論文である. 論文のどの箇所を見ても記載必要項目「問題点」に当たる部分やヒントがなく, レベル 1 に分類した.

5.5.2 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. 記載必要項目「問題点」のレベル 1 の論文は文章から問題点を予測するのが困難であったため, 今回は修正していない. 修正で得られた修正文を表 5.4 と表 5.5 に示す. 表 5.4 の「////」の部分には元々の論文に書かれていた文字列 (差分ではない文字列) が入る.

表 5.4: 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

表 5.5: 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

記載必要項目「問題点」のレベル 1 の論文以外の各レベルの修正例の一部を図 4.11 から図 4.17 に示す. 修正例の 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.

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 5.11: 「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

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 5.12: 「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

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 5.13: 「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 5.14: 「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

... フライト情報案内やバスの運行案内などの明確なタスクを定義し、関係データベース（RDB）をバックエンドとした枠組みは、タスク達成に必要な意味表現の定義や対話のフローの記述が容易であった反面、Web などの大規模なテキスト情報に対して適用することが困難であった。それに対して、一般的な文書検索を用いた対話システムの研究も行われてきたが、表層的なキーワードや係り受け関係、質問タイプなどのみに着目し、深い言語的解析や対話処理は扱われていない。その結果、対話の文脈やユーザの要求とは無関係な、不自然な応答が生成される [[こと]]という問題点があった。これに対して本研究では、述語項構造に着目した情報抽出を行うことで、RDB のような構造を持たない Web 文書を扱いながら、その意味表現を扱えるシステムを構築する。 ...

図 5.15: 「問題点」についてレベル 4 からレベル 5 に修正した論文の一部

Web 上に存在する情報は，ブロードバンド化の進展やブログ等の普及に伴い，爆発的に増加し続けている．[[これらの情報の中には]]情報の増加によって，出所が不確かな情報や利用者に不利益をもたらす情報などが [[含まれており，]]出回るという問題が発生する．そのため，信頼できる情報を利用者が容易に得るための技術に対する要望が高まってきている．そこで，我々は，利用者の信憑性判断を支援する技術の実現に向けて研究を行っている． ...

図 5.16: 「問題点」についてレベル 3 からレベル 5 に修正した論文の一部

昨今では趣味の多様化，ニーズの多様化が唱えられている．そうした要求に対して応えるためには，ユーザーの趣向に見合ったものを提示できると望ましい．そのような背景からジャンル推定というタスクは重要であると考えられる．
ジャンル推定において文章の特徴を抽出するということは重要であり，これまで LSA(LatentSemantic Indexing ： 潜在的意味索引) や LDA(Latent Dirichlet Allocation：潜在的ディリクレ配分法) が用いられてきた．...

図 5.17: 「問題点」についてレベル 2 からレベル 5 に修正した論文の一部

5.5.3 修正文に出現する単語連続の頻度調査:結果

表 5.4 と表 5.5 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を表 5.6 と表 5.7 に示す．表 5.6 と表 5.7 の結果から人手で得られた修正パターンを表 5.8 に示す．

表 5.6: 「目的」の修正文に頻出する単語連続

2 単語連続	出現した修正文数	3 単語連続	出現した修正文数
を目的	19	**を目的として**	17
目的として	18	本研究で	3
では	5	獲得を目的	3
するため	4	**するために**	3
ような	3	**手法を提案**	3
本研究	3	そのような	3
獲得を	3	を提案する	3
提案する	3	研究では	3
手法を	3	解消するため	3
そのよう	3	**問題を解消**	3
の問題	3	を解消する	3
問題を	3		
を解消	3		
ために	3		
を提案	3		
研究で	3		
解消する	3		

表 5.7: 「問題点」の修正文に頻出する単語連続

2 単語連続	出現した修正文数	3 単語連続	出現した修正文数
問題が	11	**という問題が**	8
という問題	11	**問題がある**	7
がある	5	**問題が発生**	3
問題点	3	が発生する	3
が発生	3	そのような	2
発生する	3	**の問題点**	2
ような	2	問題があっ	2
そのよう	2		
があっ	2		
の問題	2		

表 5.8: 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～

5.5.4 階層クラスタリング:結果

表 5.4 と表 5.5 より得られた修正文を R を用いて階層クラスタリングし, 分析を行った. 図 4.18 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 4.19 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を示す. 図 4.18 と図 4.19 に書かれている番号は表 5.4 と表 5.5 の修正文の番号を示している.

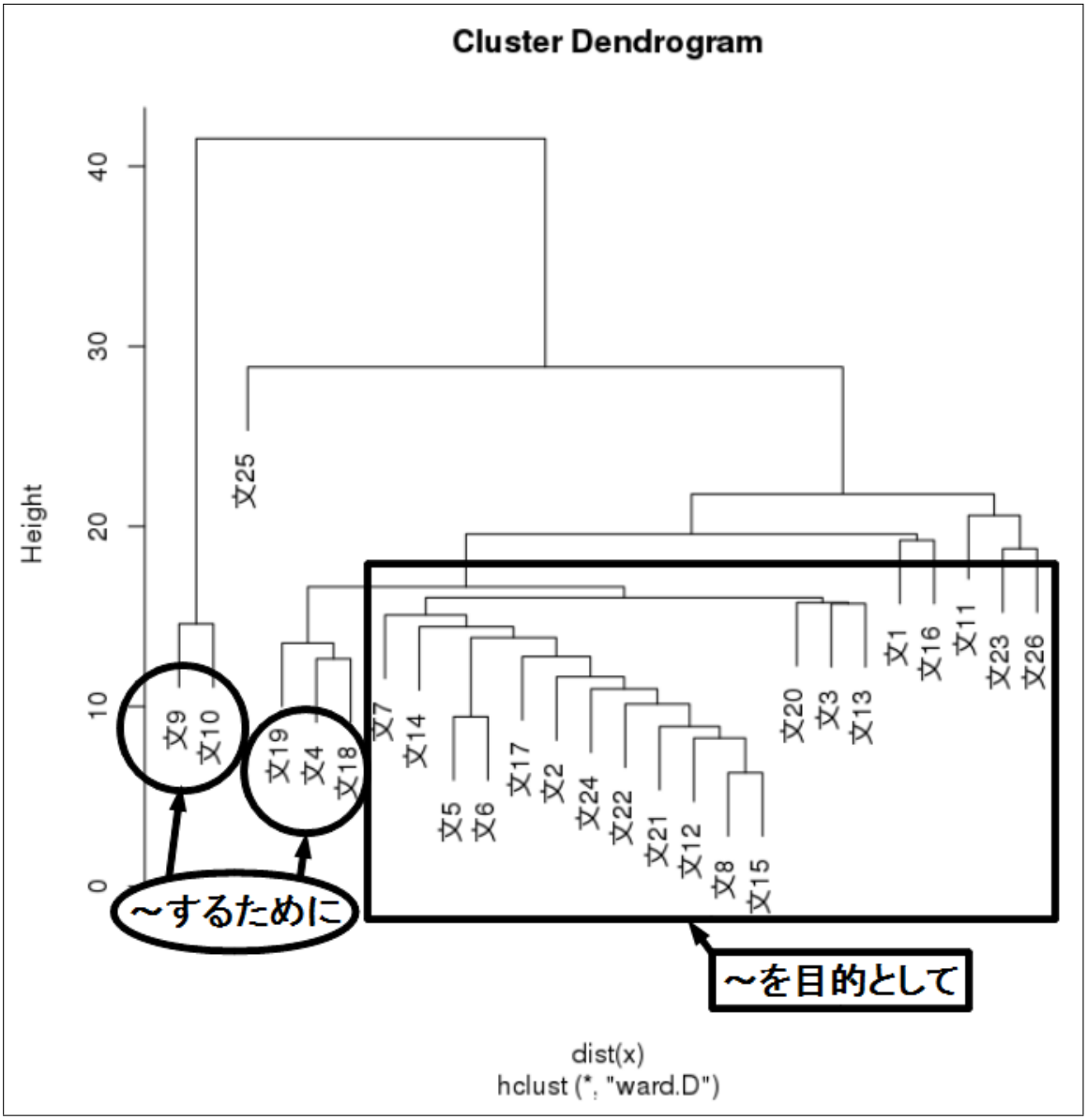

図 5.18: 「目的」についての修正文の階層クラスタリング結果

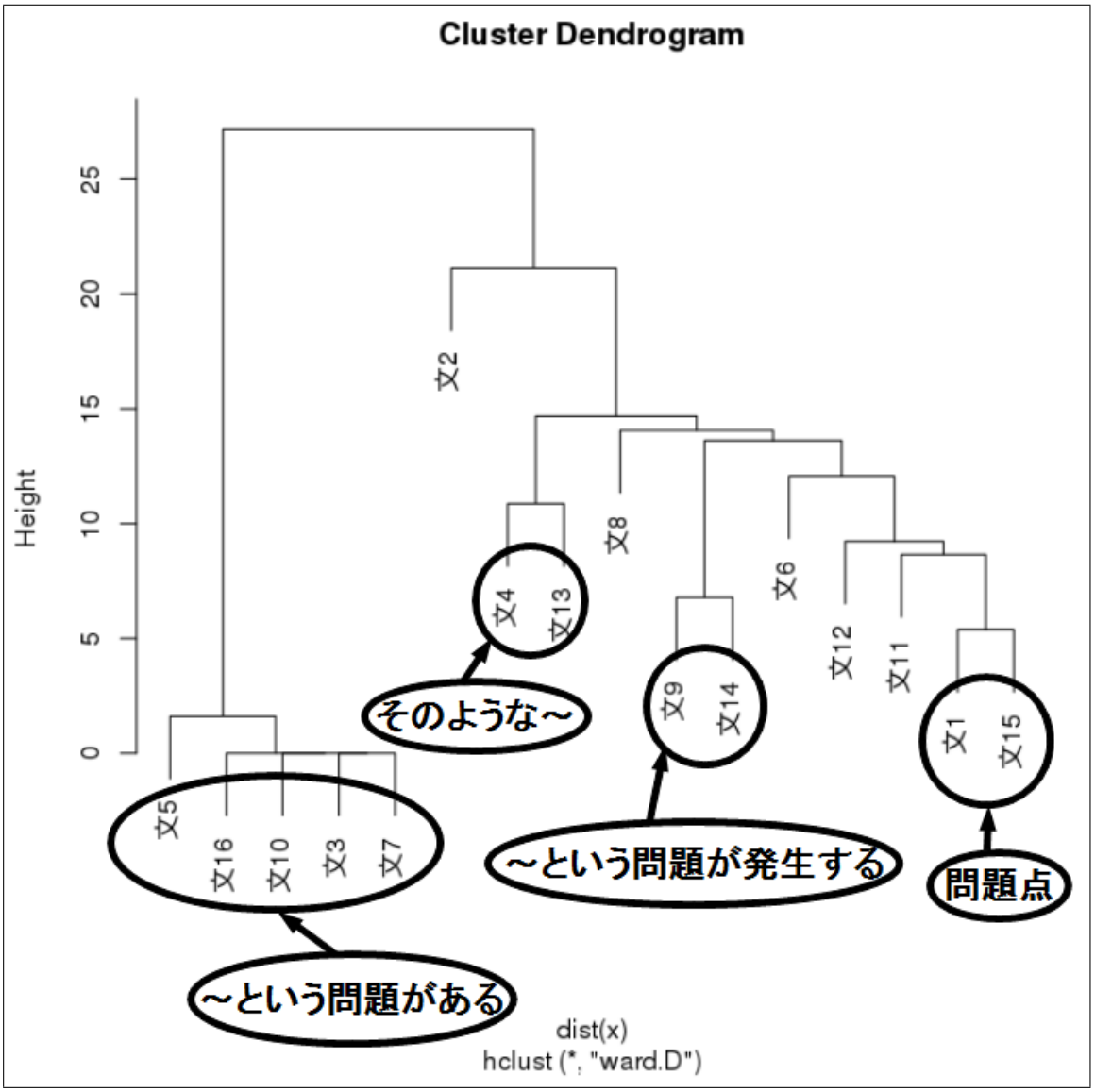

図 5.19: 「問題点」についての修正文の階層クラスタリング結果

図 4.18 と図 4.19 の結果と表 5.4 と表 5.5 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 4.18 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」であれば図 4.19 で隣り合っている文 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 5.9 に示す．

表 5.9: 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〜

5.6 考察

5.6.1 5段階レベルの分類からの考察

論文の記載必要項目「目的」「問題点」について分析し, 5段階のレベル設定を行った. その結果, それぞれ項目ごとに多く見られる傾向 (パターン) があることが分かった. 2値分類と比べると, より定義が詳細になり曖昧性は解消されたと考えられる. しかし, 論文の中には問題点 (背景) が存在しないものもあり, そういった論文を問題点が書かれていないからといってレベル1にするのか例外としてまた違う分類にするのかといった更に細かい定義がまだできていないので完全に曖昧性が解消されたと言い切れないと考える. 今後はさらに分析や定義の検討・設定を行い, 判定レベルの自動推定などによる文章作成支援なども試みたい.

また, 記載必要項目「比較」「例」についても分析を行ったが, 比較がされているか否か, 例が書かれているか否かという判別になるため, 5段階のレベルだとレベル5とレベル1の論文の2値になってしまうため, 5段階ではなく2値で評価しても変わらないと考える. そのため, 記載必要項目「比較」「例」は人手では予測が困難であり, 人手による修正が困難であると考える.

5.6.2　獲得した修正文・修正パターンからの考察

論文の記載必要項目「目的」「問題点」についてレベル 4 以下の論文を人手でレベル 5 の論文になるように修正した. その結果, 記載必要項目「目的」であれば 26 文, 記載必要項目「問題点」であれば 16 文の合計 42 文の修正文が獲得できた. また, 獲得できた修正文から, 単語連続頻度調査と階層クラスタリングを用いて記載必要項目「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. 第 4 章で, 「目的」「問題」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」について記載不備がある論文であると判別し自動検出するルールベース手法の有効性を確認した. 今回獲得した修正文・修正パターンの結果は, 記載不備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くなった理由の一つになると考える.

単語連続頻度調査で得られた修正パターン (表 5.8) と階層クラスタリングで得られた修正パターン (表 5.9) を比較すると, 単語連続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, 単語連続頻度調査での分析のほうが階層クラスタリングでの分析に比べて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 4.18 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視覚的に分かる. 同様に, 図 4.19 の結果から, 記載必要項目「問題点」の修正文の傾向も「〜という問題がある」「そのような〜」「〜という問題が発生する」「問題点」の 4 種類だということが視覚的に分かる. このことから階層クラスタリングは修正文全体の傾向から視覚的に修正パターンを獲得するのに適していると考える.

第6章　分析結果を使った文章作成支援方式

第 5 章により記載必要項目の修正文と修正パターンが得られた. 獲得できた修正パターンを使った論文の文章作成支援方式を検討する.

6.1　修正のヒント出力方式

分析で得られた修正パターンを使った論文の文章作成支援方式として修正のヒント出力方式を提案する. 記載必要項目が欠落している記載不備論文を自動検出し, 論文著者に対して今回得られた修正パターンを修正ヒントとして提示することで論文の文章作成支援が可能であると考える. 記載不備論文の自動検出は第 4 章のルールベース手法により可能であると考える. 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ヒントが掲示されるため, 論文著者の確認作業や修正作業の負担が軽減されると考える. 修正のヒント出力方式の例を図 5.1 に示す.

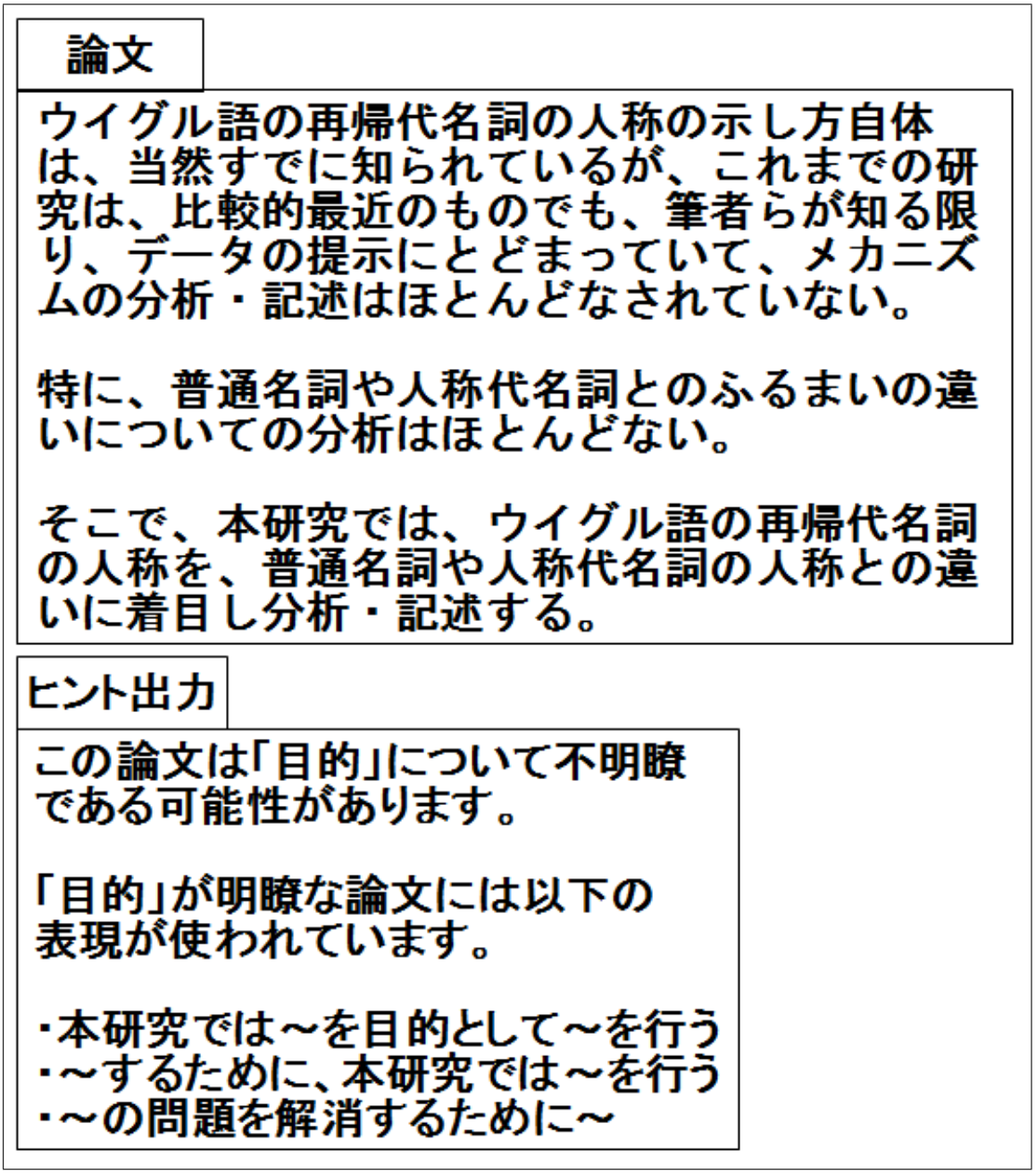

図 6.1: 修正のヒント出力方式の例

6.2　文章作成支援方式についてのアンケート

6.1 節の論文の文章作成支援方式が役立つか否かを被験者 20 人に対してアンケートを実施して調査する．アンケートには, 図 5.1 のような例と 6.1 節で記述した説明を載せる．被験者は, 「この方式は, 論文を作成・修正する際に役立つと思いますか？」という質問に対して「1. とても役立つ」「2. 役立つ」「3. やや役立つ」「4. どちらでもない」「5. あまり役立たない」「6. 役立たない」「7. 全く役立たない」の 7 段階で評価する．最後に, この方式についてのコメントを自由記述形式で回答できるようにする．

6.3　文章作成支援方式についてのアンケート結果

被験者 20 人に 6.1 節の論文の文章作成支援方式が役立つか否かをアンケート形式で調査した．表 6.1 に結果を示す．表 6.1 は, アンケートの結果「5. あまり役立たない」「6. 役立たない」「7. 全く役立たない」に回答する人がいなかったので省略している．

表 6.1: アンケート結果

1. とても役立つ	2. 役立つ	3. やや役立つ	4. どちらでもない
2 人	10 人	6 人	2 人

自由記述欄に書かれたコメントを以下に示す．

- 不明瞭である箇所を指摘できたらより多くの人に役立つ. (2 人)
- 色々な分野に対して正しい書き方を提示できるか分からない. (1 人)

6.4 考察

被験者 20 人に 6.1 節の論文の文章作成支援方式が役立つか否かをアンケート形式で調査した. その結果, 役立たないと回答した人はおらず, 「2. 役立つ」と答えた人が 10 人となり, 一番多いことが分かった. しかし, 「不明瞭である箇所を指摘できたらより多くの人に役立つ」といったコメントが見られた. 人で見て修正するよりは役立つと考えられるが, 欠落箇所の指摘ができていない現状の文章作成支援方式では, 限られた人にしか役立たないと考える. この結果より, 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でき, より多くの人に役立つ方式になることが考えられ, 今後の課題とする.

第7章　今後の課題

本研究の今後の課題として以下のものが考えられる.

- 論文の文章作成支援の網羅性向上
- 文章作成支援方式の利便性向上

7.1　論文の文章作成支援の網羅性向上

先行研究 [1] と本研究では, 記載必要項目「目的」「問題点」「比較」「例」を使っている. それらの項目が欠落している論文に対して文章作成支援を行うといった目的で研究をしている. しかし, 現状の記載必要項目では, 項目が少ないという問題があることが分かる. 論文の記載必要項目はその 4 項目以外の記載必要項目があると考えられる. 例えば, 記載必要項目「考察」「結果」が考えられる. 記載必要項目の項目を増やすことでより細かい論文の文章作成支援が可能になると考えられる.

7.2　文章作成支援方式の利便性向上

本研究では, 記載不備論文の分析を行うことで論文の修正パターンを獲得した. さらに, 記載不備論文の論文著者に対して修正ヒントとして獲得した修正パターンを提示するといった新たな文章作成支援方式を提案した (6.1 節). しかし, 現状の方式では, 記載不備のある箇所の指摘までできていないという問題がある. 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でき, より多くの人に役立つ方式になることが考えられる.

第8章　おわりに

本研究では, 記載不備論文の自動検出手法として先行手法のルールベース手法と比較手法として提案した機械学習手法の比較を行った. その結果, どの記載必要項目においても機械学習手法と比べてルールベース手法のほうが検出精度が高く, 先行手法であるルールベース手法の有効性を確認できた.

さらに, 記載不備論文の修正に役立つ修正パターンを獲得するために論文の記載必要項目「目的」「問題点」について人手で修正を行い, 得られた修正文を単語連続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 26 文の修正文から「〜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 16 文の修正文から「〜という問題がある」「〜という問題が発生する」などの修正パターンが獲得できた.

単語連続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.

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

本研究を進めるに当たり, 終始に渡り研究の進め方や本論文の書き方など, 細部に渡る御指導を頂きました, 鳥取大学工学部知能情報工学科自然言語処理研究室の村田真樹教授, 村上仁一准教授に心から御礼申し上げます. また, ご多忙の中, 助言を頂きました徳久雅人講師に厚く御礼申し上げます. 加えて, 種々の御助言を龍谷大学理工学部数理情報工学科の馬青教授に頂きました. ここに深く感謝致します. その他様々な場面で御助言を頂いた自然言語処理研究室の皆様に感謝の意を表します.

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, pp.980-991, 2015.

[2] 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.

[3] 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.

[4] 難波英嗣, 谷口裕子: “学術論文データベースからの研究動向情報の抽出と可視化”, 言語処理学会第 12 回年次大会発表論文集, pp.35-38, 2006.

[5] 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.

[6] 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.

[7] Michal Ptaszynski, Fumito Masui: “SPASS: A Scientific Paper Writing Support System”, ICIEIS2014, pp.1-10, 2014.

[8] Ming Liu, Rafael A.Calvo: “An Automatic Question Generation Tool for support of Sourcing and Integration in Students Essays”, Australasian Document Computing Symposium (ADCS), pp.45-54, 2009.

[9] 菅沼明, 牛島和夫: “テキスト処理による推敲支援情報の抽出”, 人工知能学会誌, 23 巻, 1 巻, pp.25-32, 2008.

[10] Masaki Murata, Hitoshi Isahara: “Automatic detection of mis-spelled Japanese expressions using a new method for automatic extraction of negative examples

based on positive examples", IEICE Transactions, VOL.E85-D, No.9, pp.1416-1424, 2002.

[11] 村田真樹, 井佐原均: "自動言い換え技術を利用した三つの英語学習支援システム", 情報科学技術 Letters, 3 巻, pp.85-88, 2004.

[12] 内元清貴, 村田真樹, 馬青, 関根聡, 井佐原均: "コーパスからの語順の獲得", 言語処理学会論文誌「自然言語処理」, Vol.7, No.4, pp.163-180, 2000.

[13] 村田真樹, 馬青, 井佐原均, 内元清貴: "日本語文と英語文における統語構造認識とマジカルナンバー７±２", 言語処理学会論文誌「自然言語処理」, Vol.6, No.7, pp.61-73, 1999.

[14] 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-", 言語処理学会論文誌「自然言語処理」, Vol.7, No.1, pp.51-66, 2000.

[15] Eric Sven Ristad: "Maximum Entropy Modeling for Natural Language", ACL/EACL Tutorial Program, Madrid, 1997.

[16] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均: "種々の機械学習手法を用いた多義解消実験", 電子情報通信学会言語理解とコミュニケーション研究会, 2001.

[17] Masao Utiyama. Maximum entropy modeling package: http://www.nict.go.jp/x/x161/members/mutiyama/software.html, 2006.

[18] 内元清貴, 村田真樹, 関根聡, 井佐原均: "日本語係り受け解析に用いる ME モデルと解析精度", 言語処理学会第 5 回年次大会併設ワークショップ論文集,

[19] 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", 言語処理学会論文誌「自然言語処理」, Vol.9, No.2, pp.91-110, 2002.

[20] A.L.Berger, S.A.D.Pietra, V.J.D.Pietra: "A Maximum Entropy Approach to Natural Language Processing", Computational Linguistics, Vol.22, No.1, pp.39—71, 1996.

[21] N.Cristianini, J.Shawe-Taylor: "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.