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

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．

我々は過去にルールベース手法を提案してきた [1]．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している．本稿ではさらに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で検出し，ルールベース手法との比較を行う．

以上に向けて，第 2 節では，記載必要項目の詳細について示す．第 3 節では，人手ルールと機械学習を利用した論文の文章作成支援の実験を行う．第 4 節では，論文の自動修正に向けた人手による論文の記述パターンの分析を行う．最後に第 5 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 [2] 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．また，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．

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

3　人手ルールと機械学習を利用した記載不備の自動検出

3.1　問題設定

論文の文章作成支援の問題設定は以下の通りである．記載必要項目が欠落している (文章作成支援が必要である) と人手で判別した論文を自動で検出できた場合，文章作成支援に役立つとする．

3.2　提案手法

本稿ではルールベース手法 [1] と機械学習手法の 2 手法を提案する．以下に詳細を示す．

3.2.1　ルールベース手法

ルールベース手法 [1] では，記載必要項目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する．記載必要項目の検出に役立つ単語の詳細を表 2 に示す．

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
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

3.2.2　機械学習手法

記載必要項目が欠落していると人手で判別した論文と記載必要項目が欠落してい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 値分類を機械学習で行うことで，記載必要項目が欠落していると人手で判別した論文を自動で検出する．本研究の機械学習法は，最大エントロピー法 [3][4] を用いる[1]．また，機械学習の素性に

[1] 今回の実験では，SVM も試したが最大エントロピー法よりも精度が良くなかった．

は, 論文に出現する全ての単語とルールベース手法 [1] でルールとして用いた単語を用いる.

3.3 データ

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. データの詳細を表 3, 表 4 に示す.

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.

表 3: 2011 年度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: 2012 年度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

3.3.1 評価方法

本実験の評価方法は, 記載必要項目が欠落されていると自動で検出された論文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. その結果から再現率・適合率・F 値を算出し, 評価する.

3.4 実験結果

ルールベース手法による文章作成支援と機械学習手法による文章作成支援の結果を表 5 から表 8 に示す. また, 全ての論文を記載必要項目が欠落していると判別した場合の結果をベースラインとしている.

表 5: 「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
機械学習	0.61 (36/59)	0.21 (36/174)	0.31

表 6: 「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)	0.81 (70/86)	0.70
機械学習	0.69 (79/114)	0.47 (79/169)	0.56

表 7: 「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
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
機械学習	0.44 (41/94)	0.32 (41/127)	0.37

表 8: 「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
機械学習	0.33 (3/9)	0.02 (3/129)	0.04

3.5 考察

3.4 節の表 5 から表 8 を見ると, 全ての記載必要項目においてルールベース手法の F 値が最も高いことが分かる. 機械学習手法とベースライン手法の F 値を比較すると, 機械学習手法のほうが F 値が低くなっていることが分かる.

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.

また, 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. そのような曖昧な論文について過去の研究 [1] で判別基準を設定して判別を行ったが, 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.

4 記載不備の自動修正に向けた分析

4.1 概要

第 3 節での文章作成支援の結果から, 2 値分類ではなく更に細かい分類にする必要があると考察した. そこで, 本稿では論文の記載必要項目について分析を行う. 具体的には, 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. また, 論文の記載必要項目の記述パターンも調査し, 論文の自動修正にも役立てる. 具体的には, レベル 1 の論文に対してレベル 5 の論文の記述パターンを利用することで自動修正を行う.

本稿では, 記載必要項目「目的」「問題点」について 5 段階のレベル設定を行う.

4.2 データ

2011 年度の言語処理学会年次大会論文 (266 件) を分析に使用する. 266 件の中からランダムに 50 件選び, それらを分析し, 5 段階の判別レベルを設定する.

4.3 5 段階のレベルの定義

5 段階のレベルの定義を表 9 に示す．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いる．

表 9: 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

4.4 分析結果と具体例

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 10 に示す．また, 記載必要項目「目的」についてレベル 5 からレベル 1 の論文の具体例を図 1 から図 5 に示す.

表 10: 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	12	16	19
問題点	3	4	6	26	11

... 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 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 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.

図 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.

... 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 2: 「目的」についてレベル 4 であると判別した論文の一部

... 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 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 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.

図 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.

図 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

... 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 4: 「目的」についてレベル 2 であると判別した論文の一部

... 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: 「目的」についてレベル 1 であると判別した論文の一部

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.

記載必要項目「問題点」について分析した結果,「〜という問題がある」「〜することが困難である」「しかし, 〜といったことが生じる」などの記述パターンが見られた.

4.5 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, このようなレベル設定が新たな論文の文章作成支援になると考える. レベル 5 の論文が良い論文なので, レベルを自動判別し, レベルの低い論文の場合には, 何故そのレベルになったか理由を示し, レベル 5 の論文の文章例を示すことでレベル 5 の論文の文章のように書くように促せると考える. 完全な自動修正ではないが, このような修正方法が考えられ, 今回の分析はこの修正方法で扱うレベル 5 の論文の文章例の獲得に役立つと考える. また, 比較や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.

5 まとめ

本稿では, ルールベース手法による論文の文章作成支援と機械学習手法による論文の文章作成支援の 2 手法を提案して実験を行った. その結果, どの記載必要項目においてもルールベース手法の精度が一番高く,「比較」「問題点」「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. また, 論文の自動修正に向けて記載必要項目の分析を行った. その結果,「〜のために〜を行う」や「しかし, 〜といったことが生じる」などの記載必要項目の記述パターンを調査できた.

今後の課題として, 機械学習手法の精度向上を考えている. そのためには, 考察した原因の一つである素性について検討する必要があると考える. 具体的には第一章のみに出現する単語を素性として用いるなどが考えられる. それにより, 現状の機械学習手法の精度より高くなると考える. さらに, 今後の課題として, 論文の自動修正に向けた分析と記述パターンを利用した文章作成支援を考えている. 論文の文章を自動でレベル設定することで新たな文章作成支援ができると考える.

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] 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] 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.

[3] 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.

[4] 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.