

# 初学者の論文閲覧支援のための日本語論文からの専門用語抽出の一手法

高橋 春成<sup>†</sup> 金澤 輝一<sup>††</sup> 上野 史<sup>†††</sup> 太田 学<sup>†††</sup>

<sup>†</sup> 岡山大学大学院自然科学研究科 〒700-8530 岡山県岡山市北区津島中 3-1-1

<sup>††</sup> 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

<sup>†††</sup> 岡山大学学術研究院自然科学学域 〒700-8530 岡山県岡山市北区津島中 3-1-1

E-mail: <sup>†</sup>ps4v8vmm@s.okayama-u.ac.jp, <sup>††</sup>tkana@nii.ac.jp, <sup>†††</sup>{uwano, ohta}@okayama-u.ac.jp

**あらまし** 初学者が学術論文のような専門性の高い文書を読む場合、専門用語のような未知の単語に遭遇する可能性が高く内容を理解するのが難しい。そのため、論文中の専門用語などの重要語をあらかじめ自動抽出し、ユーザにその解説を提示する学術論文閲覧支援インタフェースが開発されている。本稿では、このような論文閲覧支援のために日本語論文中の専門用語を自動抽出する手法を提案する。提案手法では、日本語論文中的名詞、複合名詞の中から、大辞泉を用いて専門用語を抽出する。実験では、DEIM2022 の日本語の論文から専門用語を抽出し、専門用語としての適切性、解説の必要性などについて評価する。

**キーワード** 情報抽出, 専門用語抽出, 学術論文, 閲覧支援

日本語論文から専門用語を抽出する手法について述べる。4 節で評価実験を行い、5 節でまとめを述べる。

## 1 はじめに

初学者が学術論文のような専門性の高い文書を読むことは難しい。初学者は専門家と比べて専門知識が乏しく、未知の専門用語に遭遇する可能性が高いためである。論文を理解しようとすれば未知の単語を調べる必要があるが、未知の単語に遭遇するたびに辞書を引く、Web サイトを検索するといった行為は手間である。そのため、予め論文中の未知の単語に解説が付与されていれば、論文の内容を理解する助けになる。筆者らは、日本語論文から専門用語を自動抽出し、ユーザにその解説を提示する学術論文閲覧支援インタフェースの開発を目指している。また本稿で日本語論文を対象にするのは、英語論文を対象とした研究開発に比べ日本語を対象とした研究が少ないと考えたからである。

専門用語抽出手法としては、FLR 法 [1] や MC-value 法 [1], TF-IDF 法 [2] などが提案されている。これらの手法は、専門用語を抽出する際に有用であるが、単語の出現回数に基づいた手法であるため、文章中に多く出現する一般的な単語を専門用語として抽出することがある。

本稿では、専門用語はある分野や業界内においてのみ使用され通用する単語で、専門外の人たちにとってはなじみのない言葉であると定義する。一方、一般的な単語はある分野や業界内においてのみ使用され通用する単語ではないため、一般的な単語は専門用語ではない。一般的な単語は、文章中の出現回数が多くても専門用語でないことは明らかなので、一般的な単語を除外することで専門用語抽出の適合率を改善できると考えた。

本稿では、複合名詞の抽出を工夫し、一般的な単語と数字を含む単語を除外することで、日本語論文中の専門用語を自動抽出する手法を提案する。提案手法では、日本語論文中的名詞、複合名詞の中から、大辞泉を用いて専門用語を抽出する。

以下に本稿の構成を示す。2 節で関連研究を紹介し、3 節で

## 2 関連研究

### 2.1 論文閲覧支援

阿辺川ら [3] は、論文中のキーワードに対する説明文や、本文に関連する情報を提示する脚注表示機能を備えた論文閲覧支援システム SideNoter を開発した。このシステムは Web ブラウザで動作し、表示されている論文の左右の脚注部に Wikipedia をリソースとした補足情報を表示する論文読解支援機能を有している。また、本文検索、専門用語の Web 検索、連続ページめくり、書き込み機能などの機能もある。

前野ら [4] [5] は、英語論文 PDF 内の任意の一単語に対して、単語の重要度などの解析結果や Web 上の関連情報をユーザに提示する学術論文閲覧支援インタフェースを開発した。このインタフェースでは、ユーザが論文中の任意のテキストを選択すると、Web の情報を得たり、そのページに移動したりできる。また、TF-IDF 値の大きい上位 30 語が表示されるため、ユーザは論文を読む前にその論文の重要語を知ることができる。

谷尻ら [6] [7] は、英語論文を対象に、タブレット端末による学術論文閲覧支援のためのインタフェースを開発した。谷尻らのインタフェースは、iPad のカメラ機能を用いて紙媒体の英語論文の画像を撮影し、OCR によって取得した画像内のテキストに対して重要度などの解析結果や Wikipedia, Weblio, Google の検索結果を表示する。前野らのインタフェース [4] [5], iBooks [8] と比較して、専門用語の意味を検索する際のページ間、複数文書やメディア間での移動が少ないという特徴を持つ。

岩本ら [9] [10] は、谷尻らのインタフェースを拡張し、論文中の文章に線を引くマーカ機能、論文中的画像の選択や論文画像と閲覧情報の保存、保存した論文画像を識別するためのタグの付与機能を実装した。また、谷尻らのインタフェースで提示さ

れる検索結果は、必ずしもユーザの興味に沿った情報とは限らないとして、検索結果を個人化する手法を提案し、Google の検索結果や他の個人化手法と比較して提案手法の有効性を示した。

## 2.2 専門用語抽出

専門用語抽出手法として、FLR 法 [1] や MC-value 法 [1], TF-IDF 法 [2] などが提案されている。FLR 法は、単語の接続頻度と文章中の出現回数に基づいて専門用語らしさを数値化する手法であり、スコアが閾値を超えた単語を専門用語として抽出する。単語の接続頻度は、ある単語の右方と左方に接続する単語の出現回数のことである。MC-value 法は、複合名詞を構成する形態素数と複合名詞の文章中の出現回数、複合名詞を含むより長い複合名詞の出現回数と種類に基づいて専門用語らしさを数値化する手法であり、スコアが閾値を超えた単語を専門用語として抽出する。

小山ら [11] は、FLR 法や MC-value 法などの方法では文章中の出現頻度が低い専門用語の抽出に限界があるとして、出現頻度の情報を用いない専門用語抽出法を提案した。まず、専門用語は文章中で提題的な用いられ方をしているとして、単語の前後の接続関係に関する制約を設定し、制約を満たさない候補を除外した。ここでいう提題的とは、単語の表す概念を単独で取り上げる形での記述であり、典型的には文節の先頭から始まり、「は」、「が」、「を」などの助詞に接続する形である。その後、複合語を構成する形態素による制約を設定し、制約を満たさない候補を除外する。例えば数字で始まる複合語の場合、2つの形態素で構成されている複合語を除外する制約がある。NTCIR-I 学会発表コーパス [12] の情報処理学会研究発表抄録 26,803 件を対象に専門用語抽出実験を行い、130,876 語の抽出結果を得た。抽出結果から 500 語をランダムに選び出した結果、84.6% が広く考えれば情報処理分野に関連する概念を表していると判断できる単語であった。この手法は、抽出する専門用語の適合率を改善する代わりに、提題的な用いられ方をしていない専門用語を抽出できない問題がある。

土田ら [13] は、論文からの専門用語抽出と専門用語の説明文を Web ページから生成する手法を提案した。専門用語抽出については、MC-value 法の問題点を挙げ、それを改善する手法を提案した。具体的には、MC-value 法では専門用語抽出に複合名詞の形態素数を利用するところを、複合名詞の文字数を利用する。また専門用語は文の中心になったり動作の目的になるという仮定の下、「X は」、「X とは」、「X を」というテンプレートを作成し、文中の主語や目的語となる単語のみを抽出した。第 18 回人工知能学会全国大会論文集の CDROM の 68 個の PDF ファイルを用いて専門用語抽出実験を行い、それぞれの PDF ファイルのスコア上位 30 語を抽出結果とした。複合名詞としての意味は持つが国語辞典に載っていないものを専門用語とする評価方法では、提案手法が 412 語、MC-value 法が 267 語の専門用語を抽出し、提案手法の有効性を示した。この手法は、抽出する専門用語の適合率を改善する代わりに、テンプレートに当てはまらない専門用語を抽出できない問題がある。専門用語の説明文の生成では、被説明用語に関連した単語を抽出して、

説明としての情報の量と長さを考慮して説明文にスコアを付与し、スコアが高い説明文を抽出した。

## 2.3 固有表現抽出による専門用語抽出

固有表現抽出とは、文章中から組織や地名、人名などの固有表現を抽出することである。固有表現抽出の手法を応用して特定の分野の専門用語を抽出する研究が行われている。

Luo ら [14] は、Bi-LSTM-CRF に注意機構を加えた化学用語抽出モデルを用いて、CHEMDNER コーパス [15] と CDR コーパス [16] から化学用語を抽出した。化学用語抽出の F 値は、CHEMDNER コーパスでは 0.9114, CDR コーパスでは 0.9257 であった。

崔ら [17] は、Bi-LSTM-CRF を用いた化学用語抽出モデルに自己学習を取り入れる手法を提案した。まず、手順 1 として化学用語ラベルが付与されている CHEMDNER コーパスの訓練データを用いて化学用語抽出モデルを学習する。手順 2 として、化学用語抽出モデルを用いて化学用語ラベルが付与されていない MEDLINE コーパス [18] に疑似的な正解ラベルを付与する。手順 3 として、CHEMDNER コーパスの訓練データと疑似的な正解ラベルが付与された MEDLINE コーパスのデータを併せて新しく化学用語抽出モデルを学習し、CHEMDNER コーパスのテストデータを用いてこのモデルの精度評価を行う。その後、手順 2 と手順 3 を化学用語抽出モデルの精度向上が見られなくなるまで繰り返す。実験の結果、自己学習なしで F 値が 0.824, 自己学習 1 回で F 値が 0.839, 自己学習 2 回で F 値が 0.841, 自己学習 3 回で F 値が 0.839 となり、自己学習を行うことでモデルの性能が向上することが示された。

## 3 日本語論文からの専門用語抽出

### 3.1 提案手法の概要

本稿では、複合名詞の抽出を工夫し、一般的な単語と数字を含む単語を除外することで、専門用語を抽出する手法を提案する。提案する手法の処理の流れを図 1 に示す。まず MeCab [19] を用いて形態素解析を行い、論文本文から名詞と複合名詞を抽出する。その後、Weblio のデジタル大辞泉 [20] を用いて一般的な単語を除外し、数字を含む単語も除外する。名詞と複合名詞の抽出方法が異なる 2 種類の手法を提案する。

学術論文の中には、学会名やツール名などの、ある分野や業界内において重要である固有名詞が存在する。これらが専門用語であるかは議論の余地があるが、提案手法では論文中のこれらの固有名詞の一部が抽出できるようにしている。

### 3.2 名詞と複合名詞の抽出

専門用語の多くは複合名詞である [1]。そのため本稿では、名詞と複合名詞を日本語論文から抽出する。また MeCab ではアルファベットが名詞として扱われないことがあるが、日本語の論文に出現するアルファベットは会議名やツール名のような、固有名詞の可能性はある。そのため、アルファベットの単語も名詞として扱う。

複合名詞を生成する手法として、2 種類の手法を提案する。

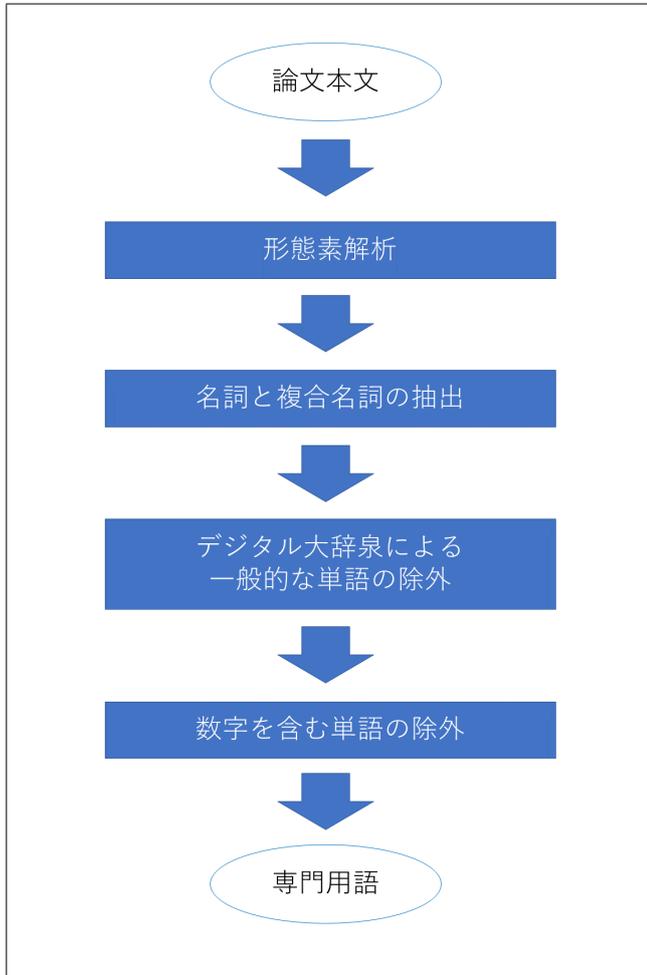


図1 提案手法の処理の流れ

一つ目の手法は、名詞とアルファベットを用いて複合名詞を生成する手法である。図2にこのようにして複合名詞を生成する手順を示す。まず形態素を取得し、名詞もしくはアルファベットであるかを調べる。名詞もしくはアルファベットであれば右隣の形態素を取得し、名詞もしくはアルファベットであるかを調べる。名詞もしくはアルファベットであればそれらの形態素を連結し、さらに右隣の形態素を取得する。違うなら、単語を記録する。この複合名詞生成手法を用いた専門用語抽出を提案手法1とする。

二つ目の手法は、名詞とアルファベットに加えて接尾辞を用いて複合名詞を生成する手法である。図3にこのようにして複合名詞を生成する手順を示す。一つ目の手法とほとんど同じだが、形態素を結合する際に接尾辞を考慮する点が異なる。この複合名詞生成手法を用いた専門用語抽出を提案手法2とする。

### 3.3 一般的な単語の除外

一般的な単語を除外するために、デジタル大辞泉を用いる。デジタル大辞泉は、小学館が発行する国語辞典である大辞泉のデジタル版で、2023年2月時点で約30万語を収録している。本稿では、Weblioに収録されているデジタル大辞泉に問い合わせを行い、掲載されている単語をすべて一般語とみなして専門用語候補から除外する。

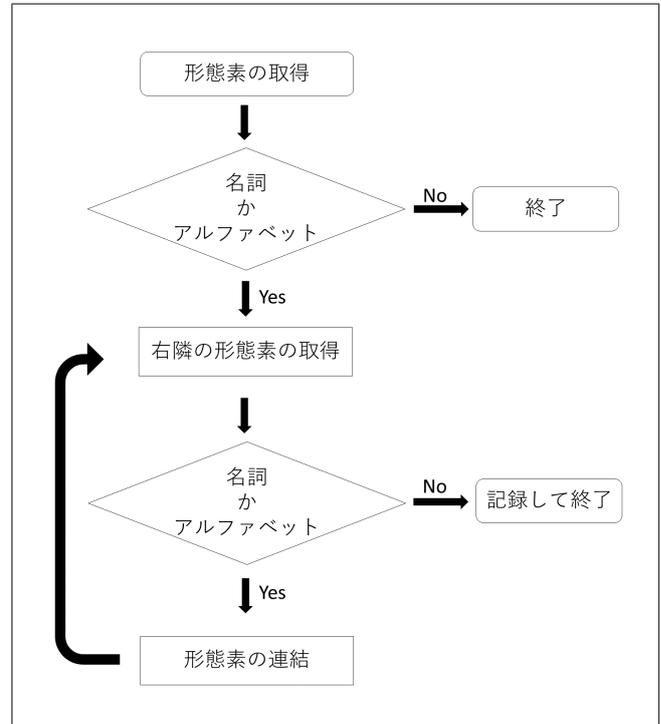


図2 複合名詞の生成（提案手法1）

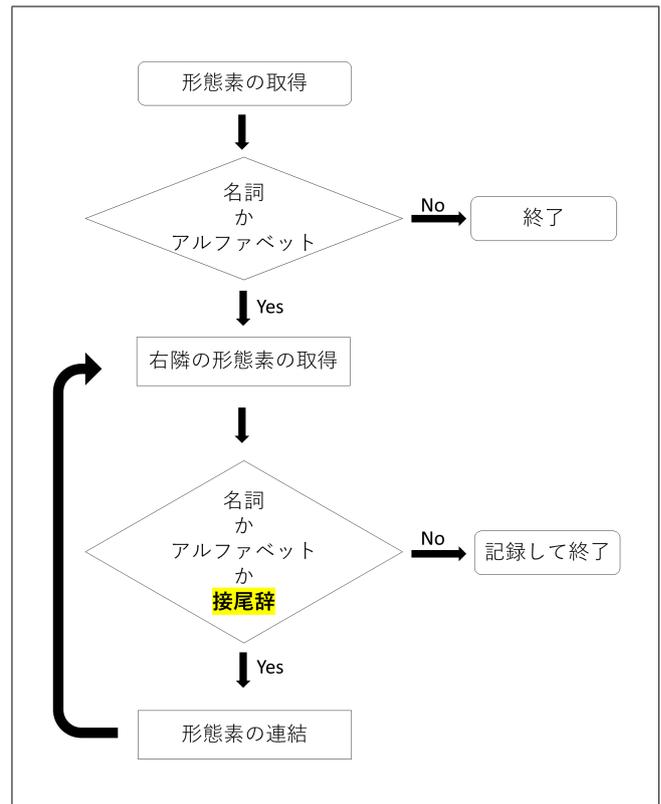


図3 接尾辞を考慮した複合名詞の生成（提案手法2）

### 3.4 数字を含む単語の除外

本稿で使用する形態素解析器のMeCabは、数字を名詞と判定する。しかし、数字を名詞として扱っていると図2や図3の処理によって「2万語」や「1000種類」といった単語が抽出されるが、これは専門用語ではない。そのため、数字を含む単語を除

表 1 専門用語抽出を行う論文

1	福島航平, 大島裕明, “チラシ媒体通販における商品の販売数予測,” DEIM2022, H23-1, 2022.
2	吉岡弘隆, 合田和生, 喜連川優, “揮発性メモリ性能測定のためのマイクロベンチマークの設計と実装,” DEIM2022, K34-4, 2022.

外する.

## 4 専門用語抽出実験

### 4.1 実験方法

専門用語抽出を行う論文は, 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM2022) の 2 つの日本語論文である. 使用する論文の書誌情報を表 1 に示す. これら 2 つの論文の本文から提案手法 1 と提案手法 2 を用いて専門用語を抽出する.

専門用語抽出の精度評価を行うにあたって, 正解データとなる専門用語を決定する必要がある. 正解データを作成する手順を説明する. まず, 第一著者が 2 つの論文から人手で専門用語候補を抽出した. その後, 抽出した専門用語候補についてアンケートで「専門用語」, 「知らない単語」, 「専門用語でない」から 1 つを選ばせた. アンケートの被験者は, 岡山大学大学院自然科学研究科の大学院生 5 名である. 本研究では, 専門用語をアンケートの結果の多数決で決定する. そのため, まず過半数の 3 人以上が専門用語だと回答した単語 245 語を正解として精度評価を行う. また, 論文閲覧者にとっての未知の単語が専門用語である可能性は高いため, 過半数の 3 人以上が専門用語もしくは知らない単語だと回答した単語 377 語を正解とした精度評価も行う.

さらに, 専門用語辞書の索引語を正解とする精度評価も行う. 使用する専門用語辞書は, IT 用語辞典 e-Words [21] である. e-Words の 12,562 語の索引語を専門用語として論文から抽出した単語 297 語を正解とした精度評価を行う.

精度評価には再現率 (Recall), 適合率 (Precision), F 値 (F-measure) を用いる. それぞれ以下の式で定義する.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{F-measure} = \frac{2\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

これらの式において各記号の意味は以下ようになる.

- $TP$  は専門用語として抽出された専門用語の数
- $FP$  は専門用語として抽出された専門用語でない単語の数
- $FN$  は専門用語として抽出されなかった専門用語の数

よって再現率は, 論文中の全ての専門用語に対し, 実際に抽出できた専門用語の割合である. 適合率は, 専門用語として抽出された単語が実際に専門用語であった割合である. F 値は,

表 2 3 人以上が専門用語だと回答した単語を正解とした時の精度評価

	再現率	適合率	F 値
提案手法 1	0.527	0.281	0.366
提案手法 2	0.551	0.260	0.353

表 3 3 人以上が専門用語または知らない単語だと回答した単語を正解とした時の精度評価

	再現率	適合率	F 値
提案手法 1	0.560	0.460	0.505
提案手法 2	0.581	0.422	0.489

表 4 e-Words に掲載されている単語を正解とした時の精度評価

	再現率	適合率	F 値
提案手法 1	0.074	0.048	0.058
提案手法 2	0.084	0.048	0.061

再現率と適合率の調和平均である.

### 4.2 実験結果

3 人以上が専門用語だと回答した単語を正解とした時の精度評価を表 2 に示す. また, 3 人以上が専門用語または知らない単語だと回答した単語を正解とした時の精度評価を表 3 に示す. どちらの場合も提案手法 1 は適合率と F 値で提案手法 2 を上回り, 提案手法 2 は再現率において提案手法 1 を上回った.

e-Words に掲載されている単語を正解とした時の精度評価を表 4 に示す. 提案手法 2 が再現率と F 値において提案手法 1 を上回っている. しかし, どちらの提案手法も再現率, 適合率, F 値全てにおいて 0.1 に満たない結果となった.

### 4.3 考察

表 2, 表 3 に示すように, アンケートから正解となる専門用語を決定した実験では, 提案手法 1 は提案手法 2 に F 値において上回った. これは, 提案手法 2 では「週-毎」, 「家庭-内」, 「スリッパ-等」のように, 専門用語でない単語に接尾辞を連結することで, デジタル大辞泉に掲載されていない単語になっていることに起因する. 一方, 表 4 に示す e-Words に掲載されている単語を正解とする実験では, 提案手法 2 は提案手法 1 に F 値において上回った. これはそもそも抽出できた専門用語が少ないため, 「揮発-性-メモリ」, 「永続-化」, 「データ-型」のように, 接尾辞を連結することで抽出できた専門用語の影響が大きくなっているためである. 接尾辞を含む複合名詞にも専門用語が存在するが, 専門用語でないものが相当数存在して適合率を低下させているため, 複合名詞の生成方法をさらに検討したい.

e-Words に掲載されている単語を正解とする実験において, 提案手法 1 の再現率は 0.074, 適合率は 0.048, 提案手法 2 の再現率は 0.084, 適合率は 0.048 で, 正解データ内の専門用語をほとんど抽出できていなかった. 一方, 専門用語が「関数」で, 提案手法で抽出したのが「損失関数」であるときのように, 専門用語を含む単語を抽出した場合でもよいとすると, 提案手法 1 の再現率は 0.616, 適合率は 0.492, 提案手法 2 の再現率は 0.657, 適合率は 0.482 となる. そのため, 提案手法は e-Words の専門用語をそのまま抽出できてはいないものの, その文字列

を含む単語でもよいなら再現率は0.6以上、適合率は約0.5であるといえる。ただし詳しく見ると、e-Wordsの専門用語には、論文中で使用されている単語とは意味的には無関係で、文字の並びが部分的に一致しただけのものもある。例えば、e-Wordsには「Tor」「ピング」「ログ」という専門用語があるが、これらは論文中の「PyTorch」「メモリマッピング」「プログラミングモデル」という単語の部分文字列となっている。このような専門用語は論文中に文字として出現しているものの、論文中で使用されている専門用語とは言えないため、本来抽出すべきものではない。

アンケートで3人以上が専門用語だと回答しなかったものの、3人以上が専門用語もしくは知らない単語だと回答した単語について考察する。これらのうちの91.67%がアルファベットを含む単語であり、被験者にとって未知の単語である可能性が高かった。例えば、x86系プロセッサ固有の情報を格納したレジスタである「MSR」、機械学習で複数の予測モデルを組み合わせた結果の平均をとる予測方法である「Model Fusion」、不揮発性メモリを意味する「NVM」等の単語がある。これらは本稿における専門用語の定義に当てはまる単語である。そのため、被験者にとって未知の単語に専門用語が含まれている可能性が示された。

提案手法1と提案手法2はどちらも、論文中の固有名詞28語中15語抽出できていた。抽出できなかった固有名詞13語中9語が数字を含む固有名詞であったため、抽出できる固有名詞ではなかった。また、ストレージの業界団体である「SNIA」という固有名詞が抽出できたが、アンケートの結果3人以上が専門用語だと回答しなかったものの、3人以上が専門用語もしくは知らない単語だと回答している。論文中の未知の単語に解説などの情報を付与する論文閲覧支援の観点からは、このような未知の単語や固有名詞の抽出についても検討する必要がある。

## 5 まとめ

本稿では、複合名詞の抽出を工夫し、一般的な単語と数字を含む単語を除外することで日本語論文から専門用語を抽出する手法を提案した。まず、論文の本文から名詞と複合名詞を抽出し、デジタル大辞泉を用いて一般的な単語を除外する。その後、数字を含む単語を除外した。

アンケートから正解となる専門用語を決定した評価実験では、提案手法1は提案手法2にF値において上回った。また、e-Wordsに掲載されている単語を正解とした評価実験では、提案手法2は提案手法1にF値において上回ったが、どちらの手法も再現率、適合率ともに0.1に満たなかった。そのため、提案手法はe-Wordsの専門用語をそのまま抽出できてはならないものの、その文字列を含む単語でもよいなら再現率は0.6以上、適合率は約0.5であった。

課題としては、複合名詞の生成方法の検討が挙げられる。今後は、抽出した専門用語を利用した学術論文閲覧支援について検討したい。

## 謝 辞

本研究の一部は、科学研究費補助金基盤研究(B)(課題番号22H03904)、同基盤研究(C)(課題番号18K11989)、および2022年度国立情報学研究所共同研究(22FC01)の援助による。

## 文 献

- [1] 中川裕志, 湯本紘彰, 森辰則, “出現頻度と接続頻度に基づく専門用語抽出,” 自然言語処理, Vol. 10, No. 1, pp. 27-46, 2003.
- [2] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” Journal of documentation, Vol. 28, No. 1, pp. 11-21, 1972.
- [3] 阿辺川武, 相澤彰子, “内部構造解析機能と脚注表示機能を備えた論文閲覧システム,” 人工知能学会第二種研究会資料, Vol. 2014, No. AM-07, pp. 13-18, 2014.
- [4] 前野朋子, “電子書籍閲覧端末による学術論文閲覧支援インタフェースに関する研究,” 岡山大学大学院自然科学研究科修士論文, 2015.
- [5] 前野朋子, 太田学, 高須淳宏, “学術論文閲覧支援インタフェースのための頭字語の活用,” 情報処理学会研究報告, Vol. 2014-DBS-160, No. 16, pp. 1-8, 2014.
- [6] 谷尻淳貴, “タブレット端末による学術論文閲覧支援に関する研究,” 岡山大学大学院自然科学研究科修士論文, 2018.
- [7] 谷尻淳貴, 太田学, 高須淳宏, 安達淳, “タブレット端末によるカメラ機能を用いた学術論文閲覧支援の一手法,” 第10回データ工学と情報マネジメントに関するフォーラム(DEIM2018), E3-5, 2018.
- [8] <https://www.apple.com/jp/apple-books/>
- [9] 岩本拓実, “論文ブラウザの操作履歴を利用した閲覧支援に関する研究,” 岡山大学大学院自然科学研究科修士論文, 2021.
- [10] 岩本拓実, 金澤輝一, 上野史, 太田学, “ユーザの興味を利用した学術論文閲覧支援の一手法,” 第83回全国大会講演論文集, vol. 2021, no. 1, pp. 535-536, 2021.
- [11] 小山照夫, 竹内孔一, “候補の接続関係を考慮した複合語用語抽出,” 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2009, No. 13, pp. 1-6, 2009.
- [12] N. KANDO eds., “Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition,” Proc NTICIR Workshop I, 1999.
- [13] 土田正明, 松井藤五郎, 大和田勇人, “論文からの専門用語抽出とウェブを用いた用語説明生成,” 日本ソフトウェア科学会大会講演論文集, Vol. 21, 2004.
- [14] L. Luo et al., “An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition,” Bioinformatics, Vol. 34, No. 8, pp. 1381-1388, 2018.
- [15] M. Krallinger et al., “The ChEMDNER corpus of chemicals and drugs and its annotation principles,” Journal of Cheminformatics, Vol. 7, No. 1, pp. 1-17, 2015.
- [16] J. Li et al., “BioCreative V CDR task corpus: a resource for chemical disease relation extraction,” Database, Vol. 2016, 2016.
- [17] 崔一鳴, 西川仁, 徳永健伸, 吉川和, 岩倉友哉, “自己学習による化学文書中の専門用語抽出,” 人工知能学会全国大会論文集, Vol. JSAI2019, 2019.
- [18] [https://www.nlm.nih.gov/databases/download/pubmed\\_medline.html](https://www.nlm.nih.gov/databases/download/pubmed_medline.html)
- [19] <https://taku910.github.io/mecab/>
- [20] <https://www.weblio.jp/cat/dictionary/sgkdj>
- [21] <https://e-words.jp/>