品詞の組合せの拡張による看護学分野での専門用語抽出性能の改善

木浪孝治[†]，池田哲夫[††]，高山毅[††]，武田利明[†††]

[†]岩手県立大学 ソフトウェア情報学研究科　〒020-0193 岩手県岩手郡滝沢村滝沢字巣子 152-52

[††]岩手県立大学 ソフトウェア情報学部　〒020-0193 岩手県岩手郡滝沢村滝沢字巣子 152-52

[†††]岩手県立大学 看護学部　〒020-0193 岩手県岩手郡滝沢村滝沢字巣子 152-52

E mail: [†]g231d010@edu.soft.iwate-pu.ac.jp, [††]{ikeda, takayama}@soft.iwate-pu.ac.jp, [†††]takeda@iwate-pu.ac.jp

あらまし　今日，大学は産学連携の一層の活性化が求められており，これを可能にするためには大学側のシーズを簡単に検索できるシステムが望まれる．そこで著者らは，産学連携の専門家が研究のシーズを専門用語によって簡単に検索することができるシステムの構築を目標とした研究を開始している．本研究では，本学に対応する学部が存在し協力を得易い看護学分野を対象に専門用語抽出の研究を行った．予備研究によって，病気の症状や治療法を表す専門用語が情報検索分野における代表的な専門用語の抽出方法では抽出が難しいことが判明した．そこで，専門用語になりうる品詞の組合せを拡張することで専門用語抽出の性能改善を図った．研究途中ではあるが，特定のデータセットにおいて平均再現率が 77.0%から 90.7%と従来手法より 13.7%向上した．

キーワード　用語抽出，専門用語，看護学

Performance Improvement of Technical Term Extraction in the Nursing Domain by Enhancing Permissible Combinations of Word-class

[†]Koji KINAMI, [††]Tetsuo IKEDA, [††]Tsuyoshi TAKAYAMA, [†††]Kazuaki TAKEDA

[†]Graduate School of Software and Information Science, Iwate Prefectural University, 152-52 Takizawa-aza-Sugo, Takizawa-mura, Iwate-gun, Iwate, 020-0193 Japan.

[††]Department of Software and Information Science, Iwate Prefectural University, 152-52 Takizawa-aza-Sugo Takizawa-mura, Iwate-gun, Iwate, 020-0193 Japan.

[†††]Faculty of Nursing , Iwate Prefectural University, 152-52 Takizawa-aza-Sugo Takizawa-mura, Iwate-gun, Iwate, 020-0193 Japan.

E mail: [†]g231d010@edu.soft.iwate-pu.ac.jp, [††]{ikeda, takayama}@soft.iwate-pu.ac.jp, [†††]takeda@iwate-pu.ac.jp

Abstract　This paper presents our ongoing research for term extraction from documents in the nursing domain. An exploratory study showed that a well-known term extraction method, which has proven to be effective in extracting term specific to the computing domain, cannot effectively extract words representing symptoms or treatments of diseases. We, therefore, propose a new term extraction method to improve extraction performance. Its main characteristics is enhancing word-class combinations which can be constituents of technical term. Experimental results showed that our extraction method attained 90.7% recall, which is 13.7% better than the recall attained by the base method.

Keyword　*Term recognition, Domain specific terms, Nursing domain*

1　はじめに

今日，大学は社会に貢献することが求められているようになっている．特に，産業界と関係の深い学部においては産学連携が強く求められるようになってきている．そのような産

学連携を活性化するためには大学側のシーズを専門用語によって簡単に検索できるシステムが望まれる．そこで，昨年度から産学連携マッチングを支援する研究情報検索システムの研究を開始している．研究の手始めとして，専門用語の抽出に取り組んでいる段階である．対象分野としては専門用語による研究情報検索システムのニーズが高く，かつ対応する学部が著者らの所属する大学に存在し協力を得易い看護学分野を選択した．

専門用語抽出の研究は，情報処理分野を対象にした研究は盛んに行われている．しかしながら，一部の医学・基礎医学分野以外には他分野の専門用語抽出の研究は見当たらない．予備研究によって，病気の症状や治療法を表す専門用語が情報検索分野における代表的な専門用語の抽出方法では抽出が難しいことが判明した．そこで，専門用語になりうる品詞の組合せを拡張することで専門用語抽出の性能改善を図った．

以下，2 章で関連研究とアプローチについて述べ，3 章で提案手法，4 章で実験及び評価，5 章で考察と今後の課題について述べる．

2　関連研究とアプローチ

2.1　関連研究

用語には1単語から構成されるものもあれば複数の単語から構成される複合語のものも存在する．例えば「専門用語抽出」は「専門」「用語」「抽出」の 3 つの単語から構成されている．多くの専門用語は，例のように複合語で構成されていることが多い．

このような複合語を考慮した情報処理分野の専門用語抽出の研究の代表的なものに中川らの研究[1]がある．中川らは，名詞と一部の特殊な形容詞を単名詞として扱い，それら単名詞の出現頻度と連接頻度を用いた専門用語抽出方法とスコア付け方法を提案している．また，中川らはこれらの手法を実装したシステム「言選 Web」[5]を構築・公開している．このシステムは日本語/中国語/英語などの多くの言語からの専門用語抽出が可能である．このシステムに加えて，言選 Web の機能を Perl モジュール化した TermExtract モジュール[6]を提供している．

複合語を構成する品詞の組合せに着目した関連研究として，連接対象に接頭語・接尾語を複合語の一部として扱っている研究[2]もある．この研究では，品詞の組合せを用いて専門用語を構築する場合に，名詞だけではなく接頭語・接尾語も対象とすることが適切であるという結論を導き出している．

上記研究から，複合語を専門用語の候補とみなし，出現頻度と連接頻度を用いたランキング手法を用いた抽出方法が有力であることがわかる．但し，これらはいずれも情報処理分野を対象としたものであり，他分野への適用可能性は不明である．

2.2　アプローチ

著者らは，上記研究での手法が優れた抽出性能(再現率・適合率)を有することに着目し，上記研究の手法をベースに看護学分野の専門用語抽出方法を考案することとした．

昨年度は TermExtract モジュールを用いて幾つかの看護学分野の文献から専門用語抽出を試行した．その結果，日本語文中の英語の専門用語を正しく抽出できないという問題があることが判明した．そこで昨年度はこの問題を解決する研究として，TermExtract をベースに英語の専門用語も抽出可能にする方法を考案し，試作システム[4]を構築した．その結果，平均適合率，平均再現率，F 値において従来手法よりも性能が向上したことを確認している．

研究の次のステージとして，今年度は試作システム[4]を用いて看護学分野の文献から専門用語を抽出する予備実験を行った．その結果，看護学分野において関連研究で前提条件としている品詞の組合せでは抽出できない専門用語が多数存在することが判明した．

そこで今年度は，専門用語の候補になりうる品詞の組合せを拡張することにより性能改善を図ることとした．

3　提案手法

昨年度の研究を元に，新たなルールを導出することで性能の改善を図る．本論文では研究の初期段階として，まず再現率の向上を性能改善の主要な目的とした．

なお，3 章以降で述べる「ルール」とは，専門用語になりうる品詞の組合せと，それら品詞を連接する条件を表す．

3.1 前提条件

本章以降で用いるデータセットの提供，正解セットの作成は看護学の専門家である本学看護学研究科の社会人大学院生に依頼した．

提案手法の検討に用いた計算環境は以下の通りである．

- 形態素解析器
  - 日本語　：茶筌 2.3.3[7]
  - 英語　　：BrillsTagger 1.14[8]
- 辞書
  - 日本語　：ipadic2.6.3-20
  - 英語　　：BrillsTagger 標準辞書

専門用語の抽出は形態素解析結果とルールを用いて行う．専門用語抽出処理の流れと実行例を図 1 に示す．

図 1 の専門用語抽出処理の流れを説明する．文章入力(図 1(a))として「看護学分野での専門用語抽出」という文章が入力されたとする(図 1(a'))．次に形態素解析(図 1(b))が実行される．例では 8 つの形態素と品詞情報が得られる(図 1(b'))．その次に形態素解析の結果として得られた品詞情報と専門用語を抽出するためのルールを用いて専門用語抽出を行う(図 1(c))．例では「看護」「学」「分野」の組合せである「看護学分野」(図 1(c1'))と，「専門」「用語」「抽出」の組合せである「専門用語抽出」(図 1(c2'))の 2 つが得られる．最後に専門用語が出力される(図 1(d)，(d'))．

基本となるルールは，中川らの研究で述べられている品詞に昨年度の研究成果で得られた「英語の専門用語」を追加したものを用いる．ルールを表 1 に示す．例における下線部は対応する品詞を示す．なお，ここで用いる品詞形態は IPA 品詞形態に準拠している．

以下，連接条件について説明する．「無条件連接」とは「連接品詞一覧のいずれかの品詞が連続して現れなくても連接可能である」ことを表す．無条件連接以外の連接条件においては，条件を満たす場合に形態素の前(あるいは後)の形態素と連接されて用語を構成する．条件を満たさない形態素は破棄される．以下に例を示す．

例1) 無条件連接：名詞-一般，名詞-サ変接続と連続した場合

> 用語(名詞-一般)　抽出(名詞-サ変接続)
> →どちらも無条件連接なので「用語抽出」という用語が構成される．

例2) 条件付連接：名詞-一般，名詞-形容動詞語幹と連続した場合

> 円錐(名詞-一般)　小体(名詞-形容動詞語幹)；
> → 名詞-形容動詞語幹の次には連接対象が必要だが，連続していないため「小体」が破棄され「円錐」までが用語として抽出される

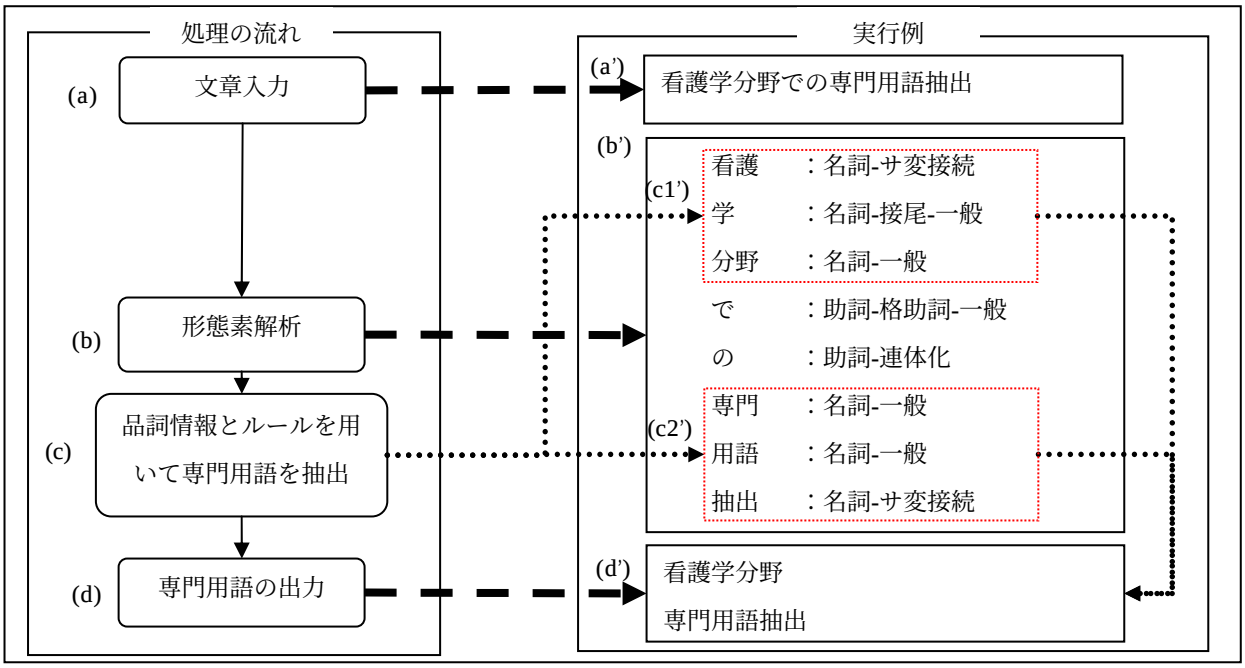

図 1：専門用語抽出処理の流れと実行例

表 1：ルール一覧

品詞	例	連接条件
名詞-一般	**用語**抽出	無条件連接
名詞-サ変接続	情報**検索**	無条件連接
名詞-接尾-一般	修正**法**	無条件連接
名詞-接尾-サ変接続	**設計**知識	無条件連接
名詞-形容動詞語幹	**帰納的**推論	連接品詞一覧のいずれかの品詞が次に連続した場合のみ連接
名詞-ナイ形容動詞語幹	**問題**解決手法	連接品詞一覧のいずれかの品詞が次に連続した場合のみ連接
名詞-接尾-形容動詞語幹	論理**的**知識	連接品詞一覧のいずれかの品詞が前後に連続した場合のみ連接
名詞-固有名詞	**Disc Array** システム	無条件連接
記号-アルファベット	ビタミン **D**	無条件連接
未知語	**クラスタリング**手法	無条件連接

3.2 ルールの導出手順

ルールの導出は図 2 のサイクルの繰り返しを試みる. このサイクルは基本的に全ての専門用語を抽出可能なルールが導出されたときに終了する.

図 2 について説明する. まず専門用語抽出システムにより専門用語抽出を行う(図 2(a)). 次に抽出できなかった専門用語の人手による抽出を行う(図 2(b)). その次に抽出できなかった専門用語を抽出するルールの導出・洗練を行い, ルールの更新を行う(図 2(c)). これらの処理は全ての専門用語が抽出されるまで繰り返し行われる(図 2(d)).

なお, 現段階では再現率向上につながるルールの導出を行っている. 適合率の向上につながるルールの導出・洗練(専門用語と誤って抽出されないようにするルールの導出・洗練)は今後取り組む予定である.

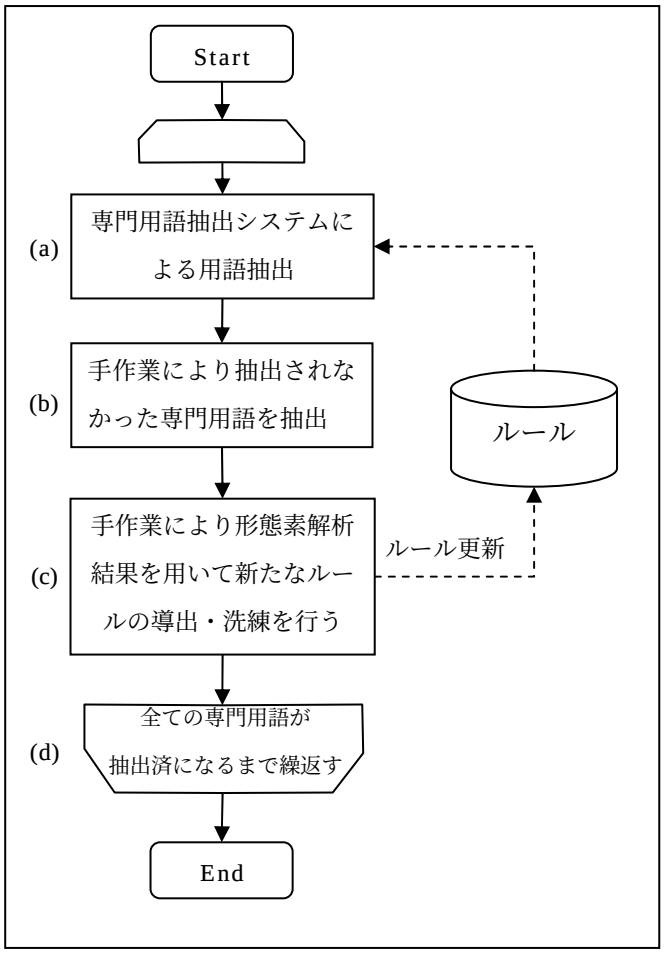

図 2：ルール導出のサイクル

3.3 データセット

医学用語辞書[9][10]及び本学看護学研究科から提供された看護学分野の文献 5 つを用いてルールの導出と洗練を行った. 以下にデータセットとして用いたデータの詳細を示す.

- 医学用語辞書
  - 辞書数 ：2 つ
  - 単語数(専門用語) ：55533 語
- 看護学分野の文献
  - ドキュメント数 ：5 つ
  - ドキュメントあたりの単語数 ：約 9000 語
  - 正解単語数(専門用語) ：664 語

3.4 導出したルール

大きく分けて 4 つのルールを導出した.

1) 名詞-副詞可能

品詞が「名詞-副詞可能」で連接条件が「無条件連接」であるルールを新たに導出した. 表 2 に主な形態素を示す. ただし, 連接対象として不要と判断した形態素は連接対象から除外した. 除外した形態素を表 3 に示す.

以下に名詞-副詞可能を含む専門用語の例を示す. 下線部が名詞-副詞可能の形態素である.

急性腎**前**性腎不全, 鼓室形成**術後**後遺症, **産後**脚気,
絶対好気性菌, **前後**十字靱帯損傷, **時間**薬理学

2) 形容詞-自立

品詞が「形容詞-自立」, 細分類が「アウオ段-ガル接続」で連接条件が「無条件連接」であるルールを導出した. 表 4 に主な形態素を示す.

以下に形容詞-自立を含む専門用語の例を示す. 下線部が形容詞-自立の形態素である.

暗視野照明, 炎症性**硬**結, **緩**速導入,
狭隅角緑内障, **硬**膜下出血, **多**剤耐性

3) 動詞-自立

品詞が「動詞-自立」, 細分類が「五段・ラ行体言接続特殊」で連接条件が「無条件連接」であるルールを導出した. 表 5 に主な形態素を示す.

以下に動詞-自立を含む専門用語の例を示す. 下線部が動詞-自立の形態素である.

下顎**切**創, 外旋**拘**縮, 眼位性眼**振**, **駆**散薬,
散腫, **殺**真菌薬, **粘**膿性, **破**骨細胞

4) そのほかのルール

4-1) 名詞-接尾, 接頭詞

関連研究[2]より連接対象に含めることが適切であることが判明していることから, 名詞-接尾の連接条件を「無条件連接」へ変更, 接頭詞を「無条件連接」として追加した.

以下に名詞-接尾,接頭詞を含む専門用語の例を示す. 下線部が名詞-接尾,接頭詞の形態素である.

膝神経**節**, 膝**部**皮下膿瘍, S 字結腸**炎**, くも膜**下**,
てんかん**重**積, シナプス**前**線維, **異**染**体**, 胃**全**摘

4-2) 名詞-形容動詞語幹, 名詞-ナイ形容詞語幹

次に連接対象が続かない専門用語が存在したため, 連接条件を「無条件連接」へ変更した.

以下に名詞-形容動詞語幹, 名詞-ナイ形容詞語幹を含む専門用語の例を示す. 下線部が名詞-形容動詞語幹, 名詞-ナイ形容詞語幹の形態素である.

看護**問題**, 眼部外傷性色素**沈着**, 胃腸機能**異常**,
喀痰喀出**困難**, 膵硬**変**, アウエル**小体**

4-3) 名詞-数

単体では専門用語としての意味を成さないため, 連接条件を「前または後に連接対象が続いた場合のみ連接」であるルールを導出した.

以下に名詞-数を含む専門用語の例を示す. 下線部が名詞-数の形態素である.

二次性高血圧症, **四**段脈, 一次性脳幹外傷,
膝蓋骨一次中枢若年性骨軟骨症, 三色性色覚

4-4) 名詞-接尾-一般の形態素「ごと」

専門家の判断に基づき, 名詞-接尾の中で専門用語として不要と判断された形態素「ごと」を除外した.

表 2：名詞-副詞可能

術後	前	時間	絶対	直接
多数	短時間	全部	全体	早朝
偶然	長日	後半	瞬間	長時間
直後	現在	満目	毎時	今
常時	一番	終夜	一時	病後
夜	毎秒	冬期	最近	将来
前後	同期	昼	産後	結果
一部	生涯	死後	同年	太古
時期	晩	朝	半分	程
昼間	前日	長年	隔離	

表 3：専門用語に不要な名詞-副詞可能

その後	それぞれ	以後	今後	１月～１２月
うち	現在	近年	以降	一月～十二月

表 4：形容詞-自立　アウオ段-ガル接続

硬	痒	暑	疎	堆
暗	濃	臭	悪	異
多	鋭	粘	醜	脆
遠	甘	速	眠	赤
浅	乏	強	薄	若
固	早	短	稚	汚
遅	深	寒	鈍	幼
細	淡	拙	青	すい
狭	緩	貴	赤	眩
近	広	長	白し	良

表 5：動詞-自立　五段・ラ行　体言接続特殊

尖	頻	絞	辷	降
振	退	すべり	腐	送
散	破	懸	擦	ち
切	陥	駆	阿	停
走	殺	嵌	覆	出
拘	滞	誤	障	疑
湿	濁	焦	遮	去
漏	浸	触	捻	煎
遺	粘	放	坐	凝
織	鳴	蒙	映	採
還	炙	困	与	曇り
承	軋	遡	捕	貼
齧	炒	あぶ	くすぶり	撓
ダブ	吃	間切	診	載

表 6：実験結果

	平均再現率	平均適合率	F 値
①提案手法	0.90688	0.34488	0.49972
②従来手法	0.76977	0.37391	0.50333
① － ②	+0.13711	-0.02903	-0.00361

4　実験及び評価

4.1　データセット

本学の看護学研究科の社会人大学院生から提供された看護学に関する１6 文献をデータセットとした．正解セットは上記社会人大学院生が手動で作成した．

作成したデータセットは以下の通りである．

- ドキュメント総数　：16 ドキュメント
- 1 ドキュメントあたりの単語数　：約 6000 語
- 全正解単語数(専門用語)　：2630 語

4.2　評価方法

昨年度作成した用語抽出システムである①従来手法と，ルールを拡張した用語抽出システムである②提案手法の 2 つを用いて平均再現率，平均適合率，F 値の 3 つの観点から評価を行った

4.3　実験結果

表 6 に各手法の平均再現率，平均適合率，F 値を示す．

平均再現率においては，①提案手法が②従来手法に対して +0.13711 と大きく向上している．平均適合度は-0.02903 と僅かに低下する結果となった．F 値は-0.00325 とごく僅かに低下しているが提案手法とほぼ変わらなかったと言える．

5　考察と今後の課題

5.1　考察

①提案手法と②従来手法を比較する実験を行った結果，平均再現率 0.76977 から 0.90688 と 0.13711 向上し，ルールの拡張が有用であることがわかった．

平均適合率が低下したのは，拡張したルールにより専門用語ではない語が抽出されるためである．ルールに対応する品詞ごとに専門用語ではない用語が抽出される原因の分析結果を説明する．

- 名詞-副詞可能

時制を表す意味で用いられ，専門用語の一部とはなりえないものが誤って専門用語の一部と判断されて連接されたためである．誤抽出の具体例として「**はじめ**感染管理」「**連日**胸腔ドレーン」のような用語が挙げられる．誤抽出された用語の一部に含まれていた形態素は 24 種類であり，用語例に

あげた「はじめ」「連日」に加えて「以前」「当時」「すべて」「過去」「多く」などがあった.

・　名詞-数

症例数，図表番号を表す意味で用いられ，専門用語の一部とはなりえないものが誤って専門用語の一部と判断されて連接されたためである．誤抽出の具体例として「食道癌 **42** 例」「てんかん症候群 **15** 機会」「図 **1**」「表 **1**」のような用語が挙げられる.

・　名詞-接尾

個数，尊称，時間を表す意味で用いられ，専門用語の一部とはなりえないものが誤って専門用語の一部と判断されて連接されたためである. 誤抽出の具体例として「幼児期 2 **名**」「無気肺 5 **例**」「山田太郎**様**」のような用語が挙げられる.

誤抽出された用語に含まれていた名詞-接尾の形態素は 70 種類であり，用語例に加えて「%(全角文字)」や年月日などがあった.

・　未知語

「①」などの機種依存文字や「±」などの特殊記号,「%(半角文字)」「mg」などの単位,「.(半角ドット)」が誤って専門用語の一部と判断されて連接されたためである．誤抽出の具体例として「**①**虚血性心疾患」「**.**PCI」のような用語が挙げられる．誤抽出された用語に含まれていた形態素は 62 種類であった.

なお，平均再現率は 13.7%向上しているものの，再現率にかかわる問題として拡張したルールでは抽出できない用語が約 10%存在するという問題が存在することが判明している．原因としては，形態素解析誤りによるもの，拡張したルールでは抽出できない品詞の組合せが残存していることなどが挙げられる.

前者に関しては，使用している形態素解析器の能力に起因するため，あるいは辞書に存在しない新語が現れたためと考えられる．能力に起因する形態素解析誤りとして，１つの形態素に複数の品詞が存在する際に必ずしも適切な品詞が選択されるとは限らないことに起因する解析誤りがある．例えば「うつ病」の「うつ」は名詞として解析されるべきだが動詞の「打つ」と解析される．辞書に存在しない新語が現れた場合の形態素解析誤りとしては，適切な形態素解析が行われるとは限らないことに起因する解析誤りがある．例えば「コンプライアンス」という用語の形態素解析を行った場合「コン」「プライア」「ン」「ス」という解析結果になり解析誤りとなる.

後者(拡張したルールでは抽出できない品詞の組合せが残存している)に関しては, ルールを導出する際に用いたデータセットが不足していたために，網羅的にルールを導出することができなかったためと考えられる.

5.2　今後の課題

5.1 で挙げた問題点を解決することが今後の課題となる.

平均適合率に関しては，5.1 の品詞ごとの考察に基づき適合率を向上させるルールの改良が必要である.

形態素解析誤りに関しては, 山本らの研究[3]のように形態素解析の誤りに対応する必要があると考えられる.

山本らの研究では，形態素を単位とする単語 n-gram の統計，表層情報を利用し用語の獲得を試みる研究を行っている．単語 n-gram の対象は中川らの研究[1]と同様の名詞に加えて「ファイルの削除」のように名詞句の形を持つ特徴的な検索に役立ちうる特徴的な文字列の 2 つを対象としている．単語 n-gram で構成された用語うち先頭または最後尾に助詞，副詞を持つ n-gram は削除している．その結果，名詞句，複合名詞，特徴的な文字列から始まる用語，解析誤りによって分解された用語などを専門用語として抽出することができる.

拡張したルールでは抽出できない品詞の組合せが存在する問題に関しては，さらに多くのデータセットを追加して実験を行う必要があると考えている.

6　まとめ

本論文では，専門用語候補になりうる品詞の組合せを拡張することにより専門用語抽出の性能改善を図った．既存の専門用語抽出のルールに新たに導出したルールを拡張することで，看護学分野における専門用語抽出の平均再現率が 13.71%向上し改善がみられた.

今後の課題として，今回の研究では行うことができなかった適合率向上のためのルールの導出・改善を検討する予定である．そのほかに，再現率を向上させるために形態素解析誤

りへの対応方法の検討，データセットを追加しての実験などを予定している．

参考文献

[1] 中川裕志，森辰則，湯本紘彰，“出現頻度と連接頻度に基づく専門用語抽出”，自然言語，Vol.10，No.1，pp.27-45，2003.

[2] 辻河亨，吉田稔，中川裕志，“語彙空間の構造に基づく専門用語抽出”，情報処理学会 研究報告，自然言語処理研究会，Vol.159，No.22，pp.155-162，2003.

[3] 山本英子，池野篤司，濱口佳孝，井佐原均，“検索支援に向けた Web 文書集合からの用語獲得”，情報処理学会研究報告，自然言語処理研究会，Vol.164，No.29，pp.171-176，2004.

[4] 木浪孝治，池田哲夫，高山毅，“産学連携マッチング支援システムの研究 -日英二ヶ国語から構成される専門用語の抽出-”，情報処理学会第 67 回全国大会，講演論文集(3)，5Q-5，pp.145-146，2005.

[5] 中川裕志，森辰則，”言選 Web“，http://gensen.dl.itc.u-tokyo.ac.jp/index.html.

[6] 中川 裕志，前田 朗，小島 浩之：専門用語自動抽出用 Perl モジュール TermExtract，http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html.

[7] 奈良先端科学技術大学院大学自然言語処理学講座，“日本語形態素解析器 ChaSen”，http://ChaSen.naist.jp/hiki/ChaSen/.

[8] Eric Brill，“英語形態素解析器 Brill's Tagger”，http://research.microsoft.com/%7Ebrill/.

[9] 京都大学学術情報メディアセンター，“ライフサイエンス辞書プロジェクト”，http://lsd.pharm.kyoto-u.ac.jp/ja/index.html.

[10] 大久保クリニック研究所，“Yo-Nagisa 版一万語医学辞書”，http://hp.vector.co.jp/authors/VA003305/.