2006－DBS－140（Ⅱ）（71）
2006／7／13

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再現率の改善

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．前回の発表においては，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%向上したという結果を報告した．今回は更なる再現率の向上を狙いとした抽出方法改善研究を行ったので提案手法と評価を報告する．提案手法は①英字あるいは片仮名からなる語の形態素解析手法の改良，②品詞組合せ規則の改良を特徴とする．

キーワード　用語抽出，専門用語，看護学

Recall 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-Village, Iwate-gun, Iwate, 020-0193 Japan.

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

[†††]Faculty of Nursing , Iwate Prefectural University, 152-52 Takizawa-aza-Sugo Takizawa-Village, 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 propose a new term extraction method to improve extraction recall ratio. Its main characteristics are enhancing permissible combinations of word-class; and enhancing morphological analysis logic which is used to analyze the sentences containing sequences of alphabets or katakana.

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．

研究情報検索システムにおいては，専門用語を索引語として登録し，質問と検索対象の双方に該当する専門用語が含まれる場合は，ランキング計算に強く寄与するような専門用語の使い方を想定している．

このことより，専門用語抽出の性能指標としては，もれなく専門用語を抽出できること，すなわち再現率が重要となる．これまでの研究[5]では平均再現率が 77.0%から 90.7%と従来手法より 13.7%向上したが，約 10%向上の余地が残っている．そこで，本研究では更なる再現率向上を狙いとした抽出方法改善研究を行うこととした．

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」[6]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[7]を提供している．

著者らはこれまでに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，従来手法では対応していない形容詞や動詞を専門用語の一部とする品詞組合せが存在することが判明した．そこで，新たな品詞の組合せを拡張し専門用語抽出の性能を向上させる研究として，従来研究をベースに品詞の組合せを拡張することで新たな品詞の組合せを含む専門用語の抽出を可能したシステムを構築した[5]．その結果，前述したように平均再現率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．

2.2　アプローチ

これまでの研究には再現率向上の余地が残っていることから，これまでの研究で提案した手法をベースに更なる専門用語抽出方法の改善を行うこととした．

抽出に失敗した要因を分析した結果，以下の 2 つに大別されることがわかった．1 つ目の要因は，形態素解析誤りに起因する．辞書に登録されていない語の解析に失敗するため，形態素解析結果に誤りが生じている．実際に形態素解析に失敗している形態素を調査した結果，英字のみから構成される語及び片仮名のみから構成されるものが殆どであることが判明している．2 つ目の要因は，形態素の連接の失敗に起因する．実際の専門用語の中には，これまでに得られた品詞の組合せでは連接可能としていない形態素が連接されているものがあった．

要因別にアプローチを検討し，以下のアプローチを取ることとした．1 つ目の要因に対しては，英字のみからなる語及び片仮名のみからなる語の形態素解析の改善を図る．2 つ目の要因に対しては，連接可能な品詞の追加および連接可能な特定の語の追加によって形態素の連接の改善を図る．

3　提案手法

これまでの研究を元に形態素解析手法の改善とルールの改善を行うことで再現率の改善を図る．

なお，本章以降で述べる「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[8]
  - 英語　　：BrillsTagger 1.14[9]
- 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これまでの研究成果で得られたルールを追加したものを用いる．ルール一覧を表 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．以下に連接条件の例を示す．

例：「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．

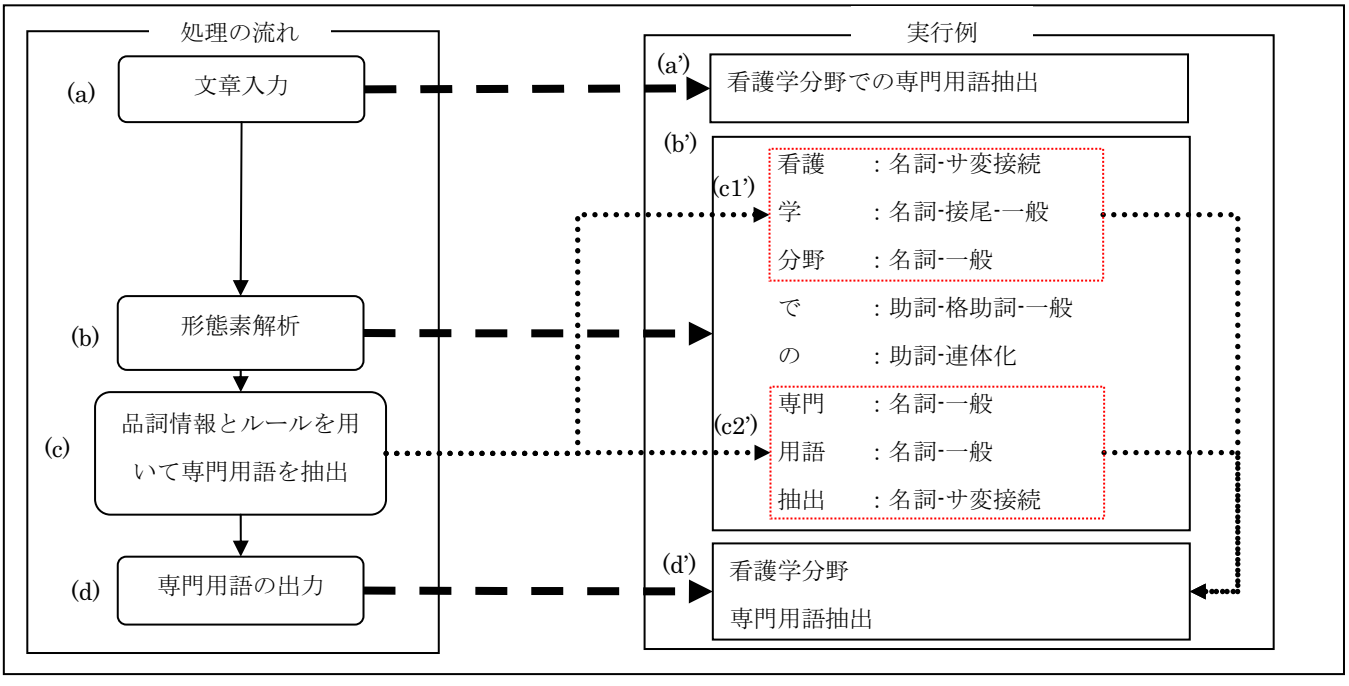

図 1：専門用語抽出処理の流れと実行例

表 1：ルール一覧

品詞	例	連接条件(※1 中川らのルール[1]と※2 著者らの研究[5]で得られたルール)
名詞-一般	**用語**抽出	無条件連接※1
名詞-サ変接続	情報**検索**	無条件連接※1
名詞-接尾	修正**法**	無条件連接．ただし，形態素「ごと」は除外する※2
名詞-形容動詞語幹	**帰納的**推論	無条件連接※2
名詞-ナイ形容動詞語幹	**問題**解決手法	無条件連接※2
名詞-固有名詞	**Disc Array** システム	無条件連接※1
記号-アルファベット	ビタミン **D**	無条件連接※1
未知語	**クラスタリング**手法	無条件連接※1
接頭詞	てんかん**重**積	無条件連接※2
名詞-副詞可能	**絶対**好気性菌	無条件連接．ただし，表 2 に示す語は除外する※2
動詞-自立	**破**骨細胞	品詞分類が「動詞-自立-五段・ラ行-体言接続特殊 2」の場合に連接※2
形容詞-自立	**多**剤耐性	品詞分類が「形容詞-自立 アウオ段 ガル接続」の場合のみ連接※2
名詞-数	**二**次性高血圧症	前または後に連接対象が続いた場合のみ連接※2

表 2：除外する名詞-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
文字通り	連日	当日	すべて	以前
倍量	近く	今回	従来	適宜
自ら	当時	各々	はじめ	以上
昨年	多く	後半	初頭	ひとつ
以来	当たり	過去	結果	同年
最近				

3.2 抽出方法の改善

3.2.1 形態素解析手法の改善

形態素解析の結果得られた英字または片仮名に関して，連続する英字または片仮名を連接して1語とすることで形態素解析手法の改善を図る．形態素解析手法の流れと実行例を図 2 に示す．

図 2 について説明する．まず文章入力(図 2(a))として「コンプライアンスに影響される」という文章が入力されたとする(図 2(a)')．次に形態素解析(図 2(b))が実行される．例ではカタカナ語が間違った形態素解析が行われ，結果として 8 つの形態素が得られる(図 2(b)')．その次に形態素解析の結果からカタカナあるいはアルファベット部分だけを連結処理する(図 2(c))．例では「コン」「プライア」「ン」「ス」を組み合わせた「コンプライアンス」が得られる(図 2(c)')．

最後に形態素解析結果が出力される(図 2(d)，(d)')．

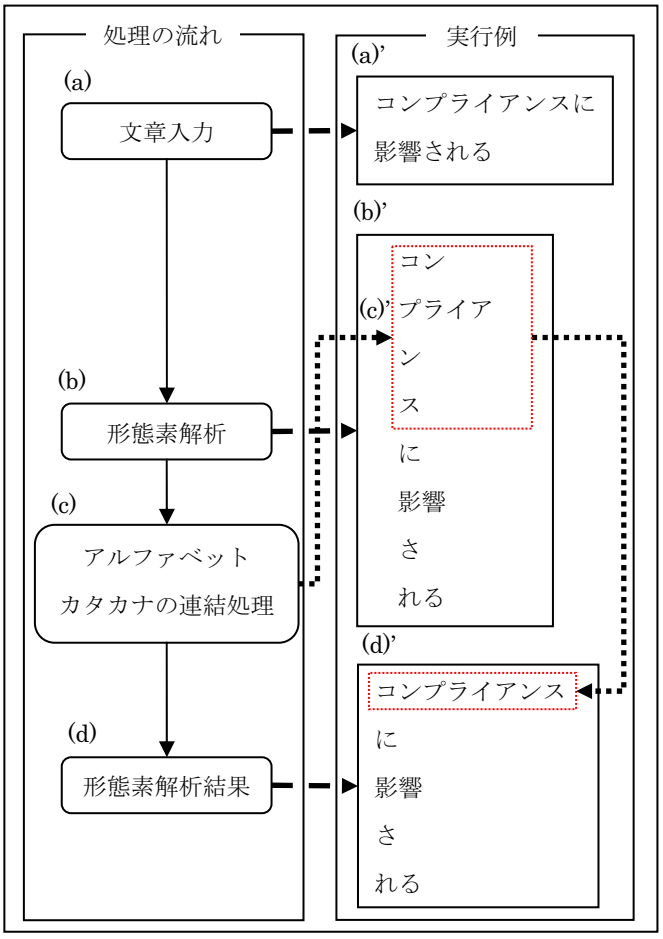

図 2：形態素解析手法の流れと実行例

3.2.2　ルールの改善

ルールを改善するために図 3 のサイクルを繰り返す. 図 3 のサイクルは全ての専門用語を抽出可能になるまで繰り返す．ただし，「うつ病」など形態素解析器の限界(同綴異品詞を正しく形態素解析できないことがある)によって抽出が難しいものは除外する.

図 3 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 3(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 3(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・洗練を行う. 最後に，ルールの妥当性を評価し問題がなければルールの更新を行う(図 3(d))．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 3(e)).

ルールの妥当性は以下の手順で評価する.

1)　品詞レベルでルールを適用した結果，再現率が向上し適合率が低下しないか評価する．再現率が向上し，適合率が低下しない場合，このルールは妥当なルールであると評価する.

2)　適合率が低下した場合，品詞レベルでルールを適用するのではなく特定の語を連接対象とする．特定の語は，その意味分類が看護学分野に関連した分野に属するものを選択する．具体的には，分類語彙表[12]の「医療・看護」「生理・病気など」「救護・救援」など看護学分野に関連した分類に属するものを選択する.

3.3 データセット

本学看護学研究科から提供された看護学分野の文献 16 ドキュメントを用いてルールの導出と洗練，妥当性の評価を行った．以下にデータセットとして用いたデータと語彙分類に用いた分類語彙表の詳細を示す.

- 看護学分野の文献
  - ドキュメント数　: 16 ドキュメント
  - ドキュメントあたりの単語数　: 約 6000 語
  - 正解単語数(専門用語)　: 2630 語
- 分類語彙表

  大日本図書，分類語彙表　増補改訂版[12]

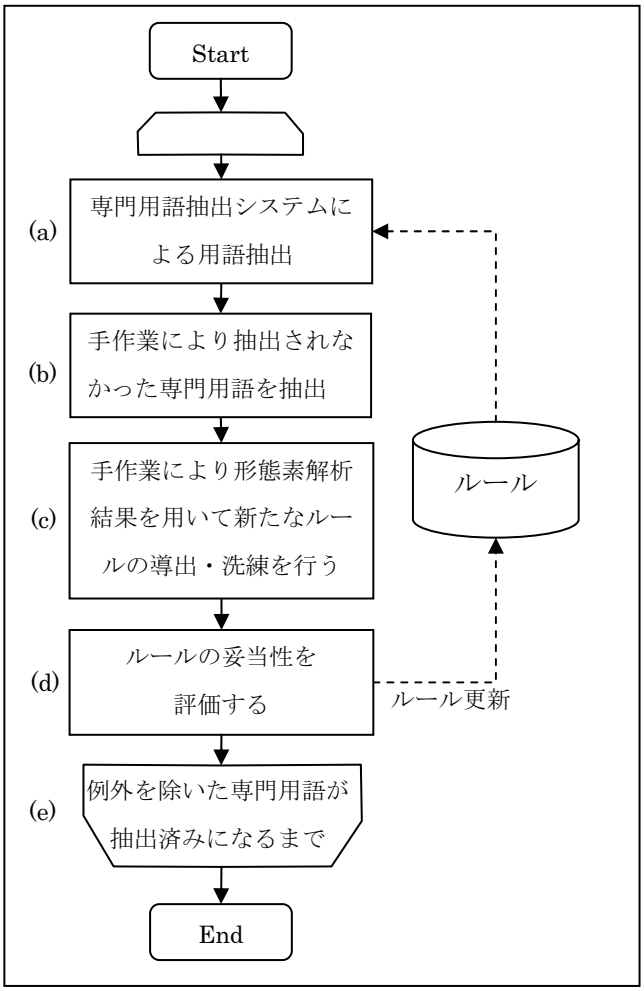

図 3：ルール改善のサイクル

3.4 改善したルール

大きく分けて 4 つのルールを導出した．以下で説明する.

1)　動詞-自立

品詞が「動詞-自立」，細分類が「一段　連用形」で連接条件が「前または後ろ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．ただし，語彙分類が「医療」「救護・救援」に分類されている語のみを連接対象とした．連接対象となる語およびこれらの語を含めることによる再現率，適合率の変化を表 3 の No3 に示す.

以下に品詞が「動詞-自立　一段　連用形」を含む専門用語例を示す．下線分が「動詞-自立　一段　連用形」の形態素である.

病**診**連携，脈**診**，**視**紫紅，**視**束前核，低拍**出**性，拍**出**量

2) 副詞-一般

品詞が「副詞-一般」で連条件が「無条件連接」であるルールを導出した．ただし，語彙分類が「生理・病気など」に分類されている語と「的」で終わる語を連接対象とした．連接対象となる語およびこれらの語を含めることによる再現率，適合率の変化を表 3 の No5 に示す．

以下に品詞が「副詞-一般」を含む専門用語例を示す．下線分が「副詞-一般」の形態素である．

極低産体重児，早発症，早成，漸深帯，漸加

3) 名詞-非自立

品詞が「名詞-非自立」，細分類が「副詞可能」で連接条件が「無条件連接」であるルールを導出した．ただし，連接対象は「間」とした．

ここで「間」の語彙分類が看護学分野に属するものではないにも関わらず連接対象とした理由を説明する．辻川らの研究[2]によって接辞を専門用語の構成要素としてみなすのが妥当であることが判明していることと，「間」は接辞としての性質を有する語であることから「間」を連接対象とした．連接対象となる語およびこれらの語を含めることによる再現率，適合率の変化を表 3 の No8 に示す．

以下に品詞が「名詞-非自立　副詞可能」を含む専門用語例を示す．下線分が「名詞-非自立　副詞可能」の形態素である．

間質性肺炎，間入性，間擦疹，間擦性湿疹

4) その他

4-1) 連接対象から除外する語の追加

専門用語の一部になりえない語を連接対象から除外した．各項目に例を示す．下線部が除外対象となった語である．

- 専門用語の一部になりえない特定の未知語を連接対象から除外した．

①急性盲腸炎，②異染体，Ⅲ破骨細胞，ⅳ膵硬変

- 小数点を含む数値のみを除外した．

10，0.1，1999，20060714

- 年代や区間を示すものを除外した．

0.01-9.95，1999-2006

- 数値と特定の単位で構成されている語を除外した．

50kg，24 歳，20 回

- 数式を表す語を除外した．

0.1 < x < 9.1，y＝2x+b

- 図表番号を除外した．

図 1，表 2-1，Fig.a，Table.b-1

- 1 文字で構成されている語は専門用語ではないとして除外した．

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

これまでの研究で得られた従来手法と，ルールの改善を行った本提案手法を用いて部分一致，完全一致における平均再現率，平均適合率，F 尺度を用いて評価する．

ここで部分一致とは，提案手法によって抽出された専門用語候補の文字列の一部分に正解セット中の専門用語の文字列が一致することを言う．

4.3 実験結果

表 4 に各手法の部分一致，完全一致の平均再現率，平均適合率，F 尺度を示す．全ての評価指標において提案手法が従来手法を上回っている．部分一致の平均再現率は 0.957 から 0.993 となり，ほぼもれなく専門用語を抽出可能となったと

言える．

5　考察と今後の課題

5.1　考察

従来手法と提案手法を比較する実験を行った結果，平均再現率，平均適合率，F 尺度において提案手法が上回っていることから，今回のルール改善が有効であることがわかった．

完全一致における再現率は 0.957 と改善の余地が残されているものの，部分一致における再現率は 0.993 と高い値となった．これにより，専門用語をほぼもれなく抽出可能になったと言える．

残された問題として，平均適合率が完全一致では 0.405，部分一致では 0.432 と低い値であるという問題が残されている．適合率が低い原因の分析結果として，以下の 2 点が挙げられる．

1) 連接頻度と出現頻度が高いという条件さえ満たせば，一般的な語も抽出多雨賞となってしまう．以下に具体例を示す．

規制緩和，民営化，研究会，シンポジウム

2) ルールの拡張を行ったことで形態素が冗長に連接された専門用語が増加し，完全一致における適合率の低下を招いている．そのため，専門用語をより正確に抽出可能とするルールの改善が必要である．以下に具体例を示す．下線部が冗長に連接された形態素である．

高齢者<u>自身</u>，血小板<u>数</u>，<u>1</u>疾患治療完結 感染管理<u>全般</u>，基礎疾患手術<u>後</u>

次に今後新たに出現する専門用語の抽出可能性に関する考察を述べる．

一般的には新たに出現する用語は，形態素解析に用いる辞書に登録されていない語すなわち未知語であるため正確な形態素解析が不可能である．ここで，看護学分野で新たに出現する用語の殆どは片仮名のみからなる語あるいはアルファベットのみからなる語であると我々は予想している．逆に漢字と平仮名からなる新語が現れる確率は低いと予想している．この予想が正しいとするならば，本研究における連接する片仮名あるいはアルファベットを専門用語の構成要素とする拡張により，今後新たに出現する専門用語の殆どは抽出可能になると予想できる．

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．

一般的な語が多数含まれてしまう問題に関しては「日本語の語彙特性」[12]に含まれている一般名詞を用いて一般的な語を除外するなどのアプローチが考えられる．

形態素が冗長に連接されてしまう問題に関しては，抽出された語の特徴を調べ，連接条件を改善するなどのアプローチがあると考えられる．

		完全一致		部分一致	
No	ルール	適合率	再現率	適合率	再現率
1	拡張無し	0.4041	0.9291	0.4316	0.9909
2	動詞-自立 一段 連用形を全て	0.3909	0.9276	0.4191	0.9927
3	動詞-自立 一段 連用形で「看，診，視，出，射，ひきつけ，動機付け」のみ	0.4048	0.9305	0.4323	0.9923
4	副詞-一般を全て	0.3918	0.9247	0.4210	0.9915
5	副詞-一般で「極，逐次，抑，早，漸」と「的」で終わるもののみ	0.4041	0.9297	0.4316	0.9915
6	副詞-一般 助数詞類接続を全て	0.4041	0.9291	0.4316	0.9909
7	名詞-非自立 副詞可能を全て	0.3927	0.9232	0.4225	0.9913
8	名詞-非自立 副詞可能で「間」のみ	0.4043	0.9294	0.4317	0.9913

表 3：妥当性の評価

	完全一致			部分一致		
	平均再現率	平均適合率	F 尺度	平均再現率	平均適合率	F 尺度
従来手法	0.907	0.373	0.529	0.957	0.359	0.522
提案手法	0.932	0.405	0.565	0.993	0.432	0.602
性能差	+0.025	+0.032	+0.036	+0.036	+0.073	+0.080

表 4：評価結果

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．これまでの研究によって得られた専門用語抽出のルールに，連続する英字または片仮名を連接して 1 語とする形態素解析手法の改善と連接可能な品詞の追加および特定の語の追加によって看護学分野における専門用語抽出の平均再現率が 99.3%とほぼ全ての専門用語を抽出可能となった．

今後の課題として，一般的な語の除外や形態素が冗長に連接されてしまう問題の解決を行うことで適合率向上を行う予定である．そのほかに，データセットの追加を行い更なる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] 木浪孝治，池田哲夫，高山毅，武田利明，“品詞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”，電子通信学会第 17 回データベース工学ワークショップ， 1B-i8，2006.

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

[7] 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.

[8] 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/.

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

[10] 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.

[11] 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/.

[12] 国立国語研究所:国立国語研究所資料集 14「分類語彙表 増補改訂版」，大日本図書，2004.

[13] 天野成昭，近藤公久：NTT データベースシリーズ日本語の語彙特性，三省堂，2005.