自然言語処理 121－15
情報学基礎 47－7
（1997. 9. 12）

医学生物学文献からの専門用語の抽出

福田 賢一郎[1]　角田 達彦[2]　田村 あゆち[2]　高木 利久[2]

[1]早稲田大学大学院理工学研究科

〒169 東京都 新宿区　大久保 3-4-1

[2]東京大学医科学研究所

〒108 東京都 港区 白金台 4-6-1

e-mail: ichiro@ims.u-tokyo.ac.jp

要 旨

専門分野の文献処理では、専門用語の処理が重要な位置を占めることになるが、専門用語は絶えず新たに発生し続けるため、未知語に遭遇することは避けられない。また、領域専門家の間でのみ通用するあいまいな表記が存在する。このため、専門用語辞書をあらかじめ用意できたとしても十分な有効性が発揮できないことが予想される。我々は本報告で医学生物学分野をとりあげ、背景知識、すなわち領域固有の辞書をあらかじめ用意することなく的確に専門用語を抽出する手法を報告する。我々の手法は未知語・既知語の区別なく適用でき、さらに表記の多様性にも対応している。我々はMEDLINE[6]に登録されている論文要旨に対してタンパク質名の抽出実験をおこない、適合率94.70%, 再現率 98.84% の結果を得た。

キーワード　医学生物学文献, 専門用語, 未知語、情報抽出, 表層的手がかり

Extracting Technical Terms from Medical and Biological Articles

Kenichiro FUKUDA[1]　Tatsuhiko TSUNODA[2]　Ayuchi TAMURA[2]　Toshihisa TAKAGI[2]

[1]Graduate School of Science & Engineering, Waseda University

3-4-1 ookubo, shinjuku-ku, Tokyo 169, Japan

[2]Institute of Medical Science, University of Tokyo

4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

e-mail: ichiro@ims.u-tokyo.ac.jp

Abstract

In processing documents of special field, adequate processing of technical terms is important. However, technical terms are generated everyday and one cannot avoid encountering words unknown to the system. Moreover, vague expressions which are used only among the area experts exist. Therefore, in some fields, a technical term dictionary prepared beforehand may not work effectively. In this report, we propose a technique by which special terms are extracted adequately without background knowledge. Our technique can be applied to unknown words as well as already-known words and is robust against the variety of expressions. We implemented and evaluated our technique against abstracts of medical and biological articles which were retrieved from MEDLINE[6]. We obtained the result of 94.70% precision and 98.84% recall.

Keywords　medical biological articles, technical terms, unknown words, information extraction, surface clues

1 はじめに

医学生物学のような専門分野では文献検索・情報抽出のような、文献処理の強いニーズが存在する。このような専門分野の文献処理では、専門用語あるいは物質名に代表されるような固有名詞の処理が重要な位置を占めることになる。

専門分野の文献を処理する際の問題点としては、他の分野とは異なった意味・用法で使われている単語の処理、および固有名詞などの一般用語にない単語の処理が考えられるが、このうち、後者はとくに大きな問題となる。各分野に固有の辞書をつくるには大変な労力を費やさなければならないし、また、専門用語は絶えず新たに発生し続けるものであり、専門用語辞書をあらかじめ用意できたとしても未定義語に遭遇することは避けられないからである。

これに対し、文献中の専門用語の同定の典型的な手法は、対象とする用語をあらかじめ辞書の形で用意し、文中で照応することによって特定するものであり、辞書にない用語の処理について有効な手法は少ない。

固有名詞の同定についてはMUC[3]などでもさかんに行なわれているが、New York Universiry[2]のシステムやLaSIE[4]にみられるようにほとんどが用語辞書を用いた手法を採用している。また、Wakao等[5]のように固有名詞の命名文法を持ったシステムもあるが文法を適用するための芯となる単語には辞書を用いている。

このため以下のような特徴を持つ分野の専門用語の処理は考慮すべき問題として残されている。

- 未知語が頻出する
  "もの"がたえず発見され続けるような分野では当然発見されたものの名前もたえず生成され続ける。このため、網羅的な辞書をあらかじめ用意することが不可能である。
- 非常に長い複合語による表記が多い
  専門用語の構成要素の組合せの数は膨大であり、単純に想定され得る語句を全て用意することは困難である。また、ハイフネーションなどの表記の曖昧性が生じる可能性がある。
- 不統一な表記法
  例えば、研究の細分化の進んだ専門分野では固有名詞や専門用語の表記に注意がはらわれなくなることが想像される。何故ならば、同じ研究に携わる研究者の間では、あいまいな表記を用いても相手に自分の主張が伝わるからである。その結果、同一物を指す用語が文献や著者の間で正確に一致しなくなる。

このような特徴を持った分野の端的な例として医学生物学分野があげられる。

従って、先に上げた特徴を持つ分野の文献から専門用語を特定するためには、新規の用語を特定する仕組みと表記の多様性に対応する方法を考えることが必要となる。

我々は本報告で、上述の特徴を持つ専門分野として医学生物学分野をとりあげ、背景知識、すなわち領域固有の辞書をあらかじめ用意することなく的確に専門用語を抽出する手法を報告する。我々の手法は未知語・既知語の区別なく適用でき、さらに表記の多様性にも対応している。我々は医学生物学分野の論文要旨に対して抽出実験をおこない、本手法の有効性を示す結果を得た。

2 研究の背景と目的

医学生物学分野における専門用語にも様々なものがある。例をあげると、実験動物の系統名、実験手法名、試薬名、遺伝子名、タンパク質名、遺伝子の部分構造の名前、タンパク質の部分構造の名前、反応名などとなる。本研究で我々は以下のものを抽出処理の"目的物質名"として抽出した。

- タンパク質名 (キナーゼ・レセプター・リガンド・エンザエム・複合体を含む)
- タンパク質の領域名 (domain 名)

医学生物学分野では、いわゆる"ゲノム計画"の進展にともない様々な種のDNA配列が次々に決定されている。この結果、個々の遺伝子の解析が進むことで遺伝子やタンパク質の間の相互作用に関する知見が急速に蓄積されている。その更新量は目覚しいものがあり、生命現象のパズルを解くために大変有用である。しかし、このような知見は生物学やゲノム計画の研究者によって日々論文という形で蓄積されるものである。各研究者が世界中の論文を網羅的に読むことは時間と労力の点で不可能であるため、物質間の相互作用に関する知見を、文献の中から質量ともに高い水準で抽出し、万人が利用できるデータベースにすることが

望まれる。そのためには上述の物質名を特定することが極めて重要となる。

3 タンパク質名の特徴

3.1 タンパク質名の命名法

タンパク質名を形の上から分類するとおよそ以下のようになる。

1. c-Myc、 p53、 Nef のように大文字や小文字、数字、記号文字が混在した単語
2. interleukin 1 (IL-1)-responsive kinase のように大文字や小文字、数字、記号文字が混在した複合語
3. actin,tublin,insulin のように小文字のみからなる単語。

医学生物学の分野は物質や物質の機能が新しく発見されることが非常に多い。研究者はそれらを世に発表するとき、他の物質や概念と明確に区別できる用語を自分で作る。このため、実際の文献中では 3 のようなものは相対的に少なく、1,2 の様な物質名が非常によく登場する。

3.2 タンパク質名同定処理上の問題点

前節の 1、 2 の用語は研究者が新しく導入する新規単語・複合語といった造語であることが多く、日々更新されている。これらを文献中で同定するためには以下の問題を解決しなければならない。それぞれの具体例は 3.2.1 で示す。

- 新規語の問題

辞書に依存したシステムの場合新しい文献の入手とともに辞書の更新が必要となるが、それは極めて time-consuming であり、誤りも多い。よって、新規語・既知語の区別なくタンパク質名を特定できる必要がある。

- 表記の多様性の問題

タンパク質名にはもともと明示的な命名法がなく、これらの専門用語は全く未整備のまま蓄積されてきている。このため、既知のタンパク質について言及している場合でも著者間あるいは文献間で表記の対応がとれていないことがある。このため、タンパク質名辞書を用意しても見出し語と実際の文献中の表記が一致するという保証はない。特に複合語の表記が多様になることは避けられない。

3.2.1 多様なタンパク質名表記の例

以下に遺伝子・タンパク質に関連した物質の表記の多様性を示す代表的な例をあげる。

a. 不統一な単語表記

タンパク質名を表す単語は 3.1 にあるように大文字、小文字、 "-" や "/" などの記号文字、数字からなる。しかし、小文字が大文字になったり、あるいは "-" や "/" の省略・追加または混同が起こるというような表記の揺れが存在する。

- c-Jun または c-jun または c jun
- cyclin D1-cdk4 complexes または cyclin D1-Cdk4 complexes

b. 自由度の高い説明的語句

以下は物質名がその役割を説明している例である。研究者によってタンパク質をとらえる視点が異なるために、同一の物質であっても、説明が異なることがある。このため、用語の表記は大変な多様性を備えている。

- the Ras guanine nucleotide exchange factor Sos

c. 複合語を構成する単語は著者によって語順が入れ替わったり、単語長が変化することがある。

i. tumor suppressor protein p53

i'. p53 tumor suppessor protein

ii. interleukin 1-responsive K protein kinase

ii'. interleukin 1 (IL-1) -responsive kinase

d. 新規の用語はもともと、複合語の頭文字などをとった略語であるものが多い。しかし、実際の文献では、分かりやすくするために著者が意図して略語の一部または全部を復元して用いることがある。

- epidermal growth factor receptor
- EGF receptor
- EGFR

e. b. の多様性に加えてさらに係り受けによって物質の特定部位について記述している場合がある。

- p85 alpha subunit of PI 3-kinase
- carboxy-terminal SH3 domain of Vav

このように物質名の表記は論文の著者のスタイルに依存する面が強く、同一タンパクが複数の文献において同じ記述で出現する保証はない。

3.3 我々が着目した特徴

我々はこの分野の専門用語には、極めて目立つ特徴があることに気がついた。論文著者が扱う物質の用語には以下のように一般的な語とは明確に区別できる特徴的な単語が含まれている。

例えば次のタンパク質やタンパク質の領域名には下線のように大文字や数字、記号文字が混在する特徴的な単語が多く出現する。

- SH2-containing protein Grb2
- Src homology (SH) 2 and SH3 domains
- p54 SAP kinase

これらの単語は読み手に与える情報量が多くその物質名の中核をなしている。この意味で、我々は目的物質中に出現するこのような単語を "core-term" と呼ぶことにする。

また、次の例のように、複合語がどのような機能や性質を持つかを示す key-word が含まれていることがある。

- EGF receptor
- Src-homology 3 domain
- Ras GTPase-activating protein (GAP)

これらの単語を "f-term(feature-term)" と呼ぶことにする。文献中から何を特定するかによって f-term には多少の違いがあるので、今回の実験では given なものとした。これらの特徴を利用すれば、新しく現れた用語も含めて、専門用語の候補をみつけることは大変容易である。

4 タンパク質名の抽出法

目的物質名は図 1 に示すように core-term の抽出と core-term の伸長連結処理という 2 つの phase によって抽出される。

以下これを詳しく説明する。

4.1 核となる単語 "core-term" の抽出法

目的物質名に出現する特徴的な単語である core-term を抽出するためにはこれを他の語句から区別する工夫が必要である。我々は、まず形態的に core-term と予想される語を全て列挙し、次に core-term でない例外的な語を排除する方法をとった。

我々は直列につながった 5 つの処理によって core-term を抽出した。第 1 の処理は core-term の候補となる単語を文章中から全て抽出し、得られた結果を次の

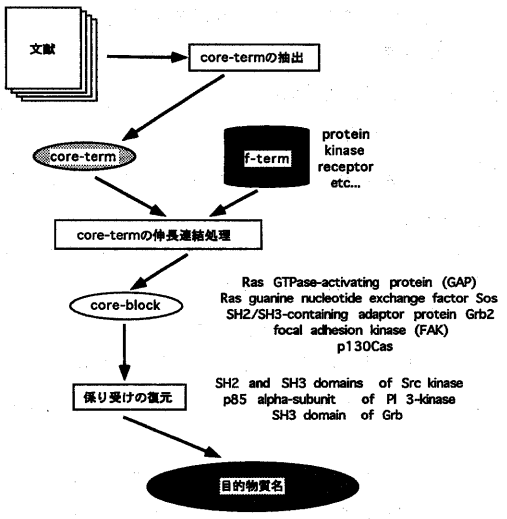

図 1: 目的物質名抽出処理の流れ

処理に渡す。第 2 から第 5 までの 4 つのフィルターは 1 で得られた候補単語から core-term でないものを取り除いていく。以下が具体的な処理内容である。

1. 大文字や数字、記号文字が混在する単語を core-term 候補として抽出する。
   ただし、大文字 1 文字のものは冠詞 "A" のように目的物質名以外でも出現するため候補に含めない。
2. "-" と小文字からなる 9 文字以上の単語の排除
   これによって、例えば "full-length" や "dual-specificity" のように目的物質名以外でも使われる単語を排除する。
3. 文字列の半分以上が記号のみからなる単語を取り除く
   これにより "+/−" のような記号列を排除する。
4. 単位など数に関する単語の排除
   候補単語には "50-aa" や "258-bp" などの数・単位に関する単語や "microM"(マイクロモル) などのように単位自体が候補として含まれている。第 4 の処理では、あらかじめ "単位" として登録された 8 個の単語 (aa, AA, fold, bp, nM, microM, %, UV) とこれらを語尾として持つものを排除した。
5. あらかじめ用意したリファレンスの表記法の template に当てはまる語を core-term から排除する。
   これにより、論文中の参考文献にある人名やジャー

ナル名が core-term 候補から排除される。

例：(Z. Weng, J. A. Taylor, C. E. Turner, J. S. Brugge, and C. Seidel-Dugan, J. Biol. Chem. 268 :14956-14963 , 1993)

4.2 core-term の伸長連結処理

次に抽出された core-term をテキスト文中に annotate する。この annotatation を適切に連結伸長することで、"and" 等の接続詞や "of" などの前置詞による係り受けを含まない名詞句 (以下 core-block と呼ぶ) を復元する。その後、このような名詞句間の係り受けを復元することで連語の形をした目的物質名を抽出する。f-term への annotation もこの段階で施す。

4.2.1 core-block の復元

我々が用いた core-term と f-term の伸長連結ルールは 2 種類に大別される。1 つ目は表層的手がかりから annotation を伸長連結するものであり、2 つ目は品詞情報をもとに annotation を伸長するものである。

以下の各例において下線は既に annotate された単語を表している。

1. 表層的手がかりを用いたルール

(a) core-term、f-term が隣接している場合には annotation を単純につなぐ

Src SH3 domain → Src SH3 domain

以下のように括弧も annotation に含める。

(SH3) → (SH3)

(b) ギリシャ文字や大文字 1 文字が右にあった時 annotation を右に伸ばす。

p85 alpha → p85 alpha

2. 品詞情報を用いたルール

我々は品詞情報を得るツールとして Brill tagger[1] を使用した。品詞情報を用いることで、以下の場合について伸長連結処理を行なった。

(a) annotation 間の単語の品詞が全て名詞か形容詞または数詞である場合に隣接していない annotation を連結する。

Ras guanine nucleotide exchange factor Sos →

Ras guanine nucleotide exchange factor Sos

(b) 冠詞や前置詞が annotation の左にあった場合、annotation を左に伸ばす。

the focal adhesion kinase (FAK) →

the focal adhesion kinase (FAK)

ルールがどのような時に適用されるか、以下に例を示す。

the interleukin 1 (IL-1) -responsive K protein →

the interleukin 1 (IL-1) -responsive K protein

この例ではまず、ルール 1.(a) が適用され (IL-1) が (IL-1) となる。次に 2.(a) が適用されてタンパク質名が特定される。

4.2.2 係り受けの復元

名詞句内の係り受けのみを特定するので、係り受けの曖昧性は文全体の場合に比べると、組合せの数が少ない分だけ減る。このため、我々は復元ルールを単純なパターンで実現した。以下に我々の用意したパターンの例をあげる。各例において A,B,C,D,E は core-block を表す。

1. "A, B, ... C and D f-term"

   Src , Fyn , Lyn , Yes , and PI3K SH3 domains →

   Src , Fyn , Lyn , Yes , and PI3K SH3 domains

2. "A, B, ... C and D of E"

   Src homology 2 (SH2) and 3 (SH3) domains of Vav

   → Src homology 2 (SH2) and 3 (SH3) domains of Vav

3. "A of B, C and E"

   SH2 domains of Abl , Lck , Fyn , and p85

   → SH3 domains of Abl , Lck , Fyn , and p85

4. "A f-term core-term and core-term"

   GTP-binding proteins Rac1 and Cdc42

   → GTP-binding proteins Rac1 and Cdc42

5. "A of B"

   p85 alpha subunit of PI 3-kinase

   → p85 alpha subunit of PI 3-kinase

6. "A , B"

   the Src-related tyrosine kinase , Hck

   → the Src-related tyrosine kinase , Hck

4.2.3 誤った annotation の修正

以上の処理のみによって充分に高い recall を得ることができるが、これらの処理は誤った annotation を修正するためのルールを持っていない。我々は以下の 2 種類の誤った annotation について修正処理を施した。1 つは、マークした f-term が最終的に 1 単語のまま伸

長せずに残るものであり、f-termが非常にありふれた単語であることが原因である。2つ目は伸長連結処理をして得られた語句の右端の単語が名詞でない場合であり、"Src-related" のようにcore-termが必ずしも名詞でないことが原因である。

1、2の場合ともに正規表現によるパターンパッチでannotationを取り除いている。

我々は以上のようなルールによって recall、precision共に高い結果を得た。次節でこの方法によって得られた結果を示す。

5 実験

我々はSrc相同3領域に関する30本の論文要旨と細胞内信号伝達全般に関する50本の論文要旨(以下SGN)を用いて前節までに述べた手法を評価した。論文要旨はすべてMEDLINE[6]に登録されているものである。

5.1 core-term抽出処理の評価

表1はf-termと実際に抽出されたcore-termが目的物質名中にどれだけ含まれるかを示している。この表から目的物質名の95%前後がcore-termかf-termのどちらかを必ず含むことがわかる。残りの5%は"insulin", "adenylyl cyclase", "dynamin" のような単語である。次の表2はcore-term extraction phaseの結果を評価したものである。false-positiveはすべて、2番目の処理で生じている。この処理は"full-length" のように目的物質名以外でも使われる単語をcore-term候補から排除するものである。"interleukin-beta" がこのfalse-positiveに含まれていた。

positive-falseには細胞名やvirus名が含まれていた。

5.2 伸長連結処理の評価

我々は誤りを次のように分類した。1と2はpositive-falseであり、3はfalse-positiveとなる。

1. 誤った箇所へのanotation(unsu:unsuitable)
   (a) "NINS" (the 258-bp novel insertの略) のようにタンパク質名でないもの
   (b) "PC12 cell" のように物質名を表すが今回はたまたま目的物質名から除いていたもの。このような例として他に"filamentous bacteriophage fuse5"、"pre-T cell line" などがあげられる。
   (c) 具体的な物質名をあらわさないもの
   "major tyrosine-phosphorylated protein"
2. 伸長・連結処理の失敗(disag:disagreement)
   (a) 伸長しきらなかったもの
   " interleukin 1 (IL-1) -responsive kinase"
   (b) よけいに伸長してしまったもの
   " *same* proline-rich region of FAK (APPKPSR)"
   (c) 連結がうまくいかなかったもの
   " p80 and p85 (p80/85)"
3. 全くanotationがつかなかったもの
   (f-p:fasle-positive)。
   "insulin"、"adenylyl cyclase"

以上の分類に従った結果、concatenation phaseの評価は表3のようになった。ここで誤りは延べで数えている。例えば、あるanotationの誤りがあって、それが1つの文献中で3回出現すれば誤りは3個と数えた。評価1では全目的物質名を評価の対象に含めている。評価2と3では小文字のみからなる目的物質名を除いて評価を行なっている。評価3はさらに他の物質と明確に区別できるcell名やファージ名を抽出しても構わないものとした評価である。

我々の評価は係り受けの失敗(誤りの分類2.(c))を厳しくcountしている。例えば、以下ではcore-blockは全て正しくanotationされているが、係り受けの復元に失敗しているためにp-fを5とcountしている。

Grb2, Crk, Abl, p85 phosphatidylinositol 3-kinase, and GTPase-activating protein SH2 domains

自然言語処理の問題の中で、係り受けは、その曖昧性のため、一般的に非常に難しい問題である。前述のように範囲を名詞句内に限ることで曖昧性は抑制できるが、上の例のように全てを解決できたわけではない。

5.3 前節の誤りの分類の3について

評価1と評価2のrecallの比較から我々の手法が表層的手がかりのない目的物質名の抽出に向いていないことがわかる。これについてはdiscussionで論じる。このようなfalse-positiveをなくすためにはcore-term extractionのための新しいルールが必要である。表4における評価4では以下のルールをcore-term抽出処理に加えている。

文献	目的物質名	including core-terms	ratio	including f-term of core-term	ratio
SH3	689	623	90.42%	661	95.93%
SGN	749	653	87.20%	705	94.12%

表 1: 目的物質名に含まれる core-term と category-name の割合 ; c-term:core-term;cat&c-term:category-name and core-term

文献	core-term	extracted	f-p	p-f	precision	recall
SH3	198	208	3	15	92.74%	98.48%
SGN	231	230	6	11	95.22%	97.40%

表 2: result of core-term extraction phase ; f-p:false-positive,p-f:positive-false

- .*'consonant'(in | ase | ol)s{0,1} というパターンにマッチする名詞は core-term である

1 同様全目的物質名を評価の対象をしているが、さらに、細胞名やファージ名は抽出されても構わないものとした。

このルールは core-term を拡張したものであり、"insulin" や "phspoliphase" などが新たに core-term に追加されることになる。

SH3/1 と SH3/4 の比較から f-p が減ったことが分かる。また、unsu と disag が合計 18 減少しているのに対して unsu+disag が 32 減少している。これは、係り受けの復元に成功した例が増えたためである。

Src, Fyn, Lyn, and phosphatidylinositol 3-kinase (PI3K) SH3 domain

⟶

Src, Fyn, Lyn, and phosphatidylinositol 3-kinase (PI3K) SH3 domain

6 考察

我々の手法は、一部の表層的手がかりの全くない語句を除いた、すべての表記に対応している。すなわち、従来の辞書ベースの手法が苦手とした新規語や造語、表記の多様性にも対応している。表層的手がかりの乏しい用語は以下の 2 つに分類される。我々が抽出に失敗したのは主に "adenylyl cyclase" のように小文字のみからなる用語であり、これらは後者に分類される。

1. 文献中に言い替え表現が定義されているもの

(a) focal adhesion kinase

(b) major tyrosine-phosphorylated protein

(a) は core-term を含んでいないが、f-term を含んでいるため前節で述べた伸長連結ルールによって抽出される。しかし、(b) も同様に抽出されてしまう。(a) をこのような具体的でない物質名と形態的に区別することは難しいため、recall と precision の間のトレードオフが生じる。

しかし多くの場合、複合語による物質名表記は文献の先頭に近い位置で言い替え表現を定義しており、この例では "focal adhesion kinase (FAK)" である。我々の手法はこのような言い替え表現を抽出しているので、これらの言い替え表現から synonym を抽出しテキスト文にフィードバックすることでこのトレードオフは解消できる。

しかし、残念ながら辞書・synonym の自動生成、およびそのフィードバックルールを適用しての評価はまだできていない。

2. 言い替え表現が定義されていないもの

"adenylyl cyclase" や "insulin" などの語句は 1 単語か 2 単語程度の単語長で使用されることが多く、1.(a) の例のような言い替え表現がされることは少ない。これらには子音や母音の使われ方や語尾の形など、文字単位の表層的手がかりは存在する。実際に我々はルールを追加するしてこれらの語を抽出し、適合率 94.70%, 再現率 98.84% の結果を得ている。しかし、このような語句には新規語の登場する可能性があまりなく、さらに表記の揺れもほとんどない。このため優れた辞書が入手可能であり、かつその利用によって直ちに精度の向上がもたらされるのであれば、辞書を併用することも可能である。

文献 / 評価	ANO	AUTO	f-p	unsu	disag	unsu + disag	precision	recall
SH3/1	689	683	40	26	24	59	91.90%	93.32%
SH3/2	646	679	1	26	24	59	91.31%	99.85%
SH3/3	646	663	1	10	24	43	93.51%	99.85%
SGN/2	669	666	19	17	43	65	90.24%	97.16%

表 3: result

文献 / 評価	ANO	AUTO	f-p	unsu	disag	unsu + disag	precision	recall
SH3/4	689	698	8	11	21	37	94.70%	98.84%

表 4: best result

このように抽出した知識をテキスト文へフィードバックすることと、あらかじめ用意した辞書を上手に我々の手法に融合していくことで、今回我々の手法がカヴァーしきれなかった用語も抽出可能である。

また、今回の実験では残念ながら SGN の結果が SH3 についてのものより低かった。原因として用いたルールが SH3 関連の文献に over-fitting していることが考えられる。ルールの一般性を高めるために今後さらに文献数を増やして実験をすることが必要である。

7 結論

我々は、core-term を中心に連結伸長処理をすることで目的物質名を抽出するという表層的な手がかりを用いた専門用語抽出手法を初めて提案した。

この手法によって我々は、医学生物学文献から専門知識、すなわち物質名を背景知識なしに網羅的に獲得することに成功した。我々の捉えた表層的手がかりは医学生物学分野の物質名表記によく見られるものであり、本報告で実験した文献以外にも広く適用可能である。また、同様の手法は特徴的な表記を含む他分野の複合語の抽出にも適用できると予想される。我々の手法は 95.93% の recall と 94.70% の precision を達成しており、文献からの情報抽出のための前処理として十分に実用にたえるものである。

(1) E. Brill, "Some Advances in Transformation-Based Part of Speech Tagging," in *Proc. of the Twelfth National Conference on Artificial Intelligence.*

(2) R. Grishman, "The NYU System for MUC-6 or Where's the Syntax?," in *Proc. of sixth Message Understanding Conf.(MUC-6).*

(3) MUC-6, "The NYU System for MUC-6 or Where's the Syntax?," in *Proc. of sixth Message Understanding Conf.(MUC-6).*

(4) R. Gaizauskas, T.Wakao, K. Humphreys, H. Cunningham, Y. Wilks, "UNIVERSITY OF SHEFFIELD: DESCRIPTION OF THE LaSIE SYSTEM AS USED FOR MUC-6," in *Proc. of sixth Message Understanding Conf.(MUC-6).*

(5) T.Wakao, R. Gaizauskas, Y Wilks, "Evaluation of an Algorithm for the Recognition and Classification of Proper Names," in *Proc. of the 16th International Conference on Computational Linguistics(COLING 96).*

(6) "MEDLINE," *Internet Grateful Med Development Team,National Library of Medline, 8600 Rockville Pike,Bethesda,MD 20894 USA,1996*