医学生物学文献からの専門用語の抽出に向けて：タンパク質名の自動抽出

福田 賢一郎[†] 角田 達彦[†]
田村 あゆち[†] 高木 利久[†]

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．しかし専門用語はたえず新たに作られ続けるため，専用の辞書をあらかじめ用意できたとしても未知語に遭遇することは避けられない．また，専門用語には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，専門家が文献中で専門用語辞書の見出しに正確に一致するように言葉を選ぶことは少ない．このような理由により，専門用語を同定するために，優れた専門用語辞書をあらかじめ網羅的に作成することは困難である．我々は本報告で医学生物学分野を取りあげ，領域固有の辞書をあらかじめ用意することなく専門用語を抽出する手法を提案する．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[1)]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行い，適合率94.70%，再現率98.84%の結果を得た．

Extracting Technical Terms from Medical and Biological Articles

KENICHIRO FUKUDA,[†] TATSUHIKO TSUNODA,[†] AYUCHI TAMURA[†] and TOSHIHISA TAKAGI[†]

In processing documents of special fields, 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[1)]. We obtained the result of 94.70% precision and 98.84% recall.

1. はじめに

医学生物学のように広範かつ高度な専門知識が要求される分野では，文献検索・情報抽出などの文献処理に対する強い需要が存在する．また，ある分野に特化した知識を体系化するためには，膨大な数の文献に目を通す必要があり，従来は領域専門家の膨大な労力を必要とした．このため，辞書の全/半自動生成，複数の文献間でのリンクの自動生成などによって専門家にかかる負荷を減らし，かつ知識を網羅的に体系化することが望まれる．

ところが，専門分野の文献には他の分野とは異なった意味・用法で用いられる単語や言い回し，さらに一般用語にない単語が多く存在する．さらに，これらの単語および言い回しは多くの場合専門用語であり，文献の内容に関する多くの情報を含んでいる重要な用語である．このため専門分野の文献は一般に処理が困難である．

専門分野の文献に文献処理を正しく施すためには，文献に出現する専門用語を正確に同定することが大変重要である．専門用語を同定するためには，以下にあげる手法の適用が考えられる．

まず，領域専門家によって作成された用語辞書をあらかじめ用意し，これを文献中の単語と照応する方法がある．しかし各分野に固有の辞書を新たにつくるには大変な労力を費やさねばならず，さらに，あらかじめ辞書を用意できたとしても未知語の出現には対処できない．これに対して専門用語はたえず新たに作り出され続けるため，未知語の出現は不可避である．よっ

† 東京大学医科学研究所
Institute of Medical Science, University of Tokyo

て，未知語に遭遇するたびに辞書を人手に頼って更新しなければならない．また，この手法は文献中での表記が辞書の見出し語と異なる場合について対処不能である．

専門用語は複合語の形をよくとる．専門用語辞書を用いない複合語の自動認識手法には文献 2)，3) にみられるように統計的な手法を用いたいくつかの方法が提案されている．しかし，文献 2) においては bigram および trigram しか考慮されていない．これに対して，たとえば医学生物学文献では 6 単語以上からなる複合語が稀でない．さらに，専門用語には専門家の間でのみ通用する曖昧な表記が存在し，1 つの物質名や概念名について多様な表記をとりうる．このため複合語を構成する単語間には強い相関が現れにくく，統計量を用いた識別能の高い閾値を求めることは困難であると予想される．

次に，専門用語の同定を知識の自動抽出という観点から見る．情報抽出（Information Extraction，IE）システムについては Message Understanding Conference（MUC）[4] においてさかんに研究が行われている．IE システムの用いている代表的な手法は文献 5) にみられるように，固有名詞辞書とパターン辞書の併用である．パターン辞書とは，たとえば，

PATTERN	EXAMPLE
passive-verb ⟨ dobj ⟩	killed ⟨ victim ⟩

のような構造[6]を持ったテンプレートの集合である．このパターン辞書を用いたテンプレートマッチによって目的の知識が抽出される．しかし，パターン辞書を用いるには以下の問題が存在する．第 1 に，パターンを学習するためのコーパスが通常では存在しない．第 2 に，前処理が必要となる．専門用語，特に医学生物学分野の専門用語には非常に長い複合語が多く存在する．このためテンプレートマッチをする際に，複合語もテンプレートにマッチしうるように，あらかじめ複合語を認識しなければならない．

固有名詞の同定は専門用語同定と深く関連する研究であるが，知識抽出に近い観点からの研究として文献 7) による固有名詞の認識がある．彼らのシステムは手動生成による 100 個程度の CFG ルールを固有名詞の命名文法として持っている．この命名文法は，たとえば，“ORGAN_NP → NAMES_NP '&' NAMES_NP” のようなルールからなる．この例では，企業名自体が辞書になくても，企業名を構成する単語が人名辞典にあれば企業名が認識される．しかし文法が適用されるためには構成要素となる単語，たとえば上の例では人名のいずれかが辞書に存在しなければならない．このため，複合語の構成単語がすべて未知語であった場合の処理は難しい．

文献処理に際して専門用語の処理が難しい理由として以下の 3 点が考えられる．

- 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り，辞書に依存したシステムでは辞書を頻繁に更新する必要がある．
- 非常に長い複合語による表記が多い
  複合語の構成法は自由度が高く，専門用語の構成要素の組合せの数は膨大であるため，想定されうる語句をすべて用意することは困難である．さらに，ハイフネーションなどの用い方によって表記の曖昧性が生じる可能性がある．このため，複合語をあらかじめ用語辞書に登録していくことは，1 単語からなる用語の登録以上に困難である．
- 不統一な表記法
  たとえば，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．なぜならば，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．さらに，明示的な命名法がない場合には専門用語はまったく未整備のまま蓄積されることになる．このような場合，“正確な” 名称が明示的に存在しないことが考えられる．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．特に複合語の表記が多様になることは避けられない．

タンパク質名は医学生物学分野の中でも，上述の特徴をすべて合わせ持った用語の端的な例である．また後に例示するようにタンパク質名の表記は非常に自由である．つまり，たとえ専門用語辞書を用意できたとしても，未知語のみならず既知語の同定も困難であるという問題を持つ．このためタンパク質名を認識するためには，専門辞書を前提とせずに目的の用語をリアルタイムに認識する専門用語認識手法が必要となる．そこで我々は本報告で医学生物学分野を取りあげ，タンパク質名の抽出実験を行った．

本報告で我々が提案する専門用語抽出処理は表層的手がかりに着目した，領域固有の辞書を前提としないものである．専門辞書が有効に働かずかつ統計的手法によっても有意な特徴を見つけることが困難である分野の用語の同定には，表層的手がかりを十分に利用す

ることが重要となる．我々は，未知語/既知語，あるいは1単語からなる用語/複合語の区別なく専門用語を抽出することに成功し，また曖昧な表記や表記の多様性にも完全に対応した．本手法はある種の専門用語をうまく特徴付けているものであり，タンパク質のみならず，遺伝子名や細胞名，また他分野の一部の抽出が困難な用語に対しても有効であると考えられる．

我々は医学生物学分野の論文要旨に対して抽出実験を行い，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は様々なものがある．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目的物質名として抽出処理の対象とした．

- 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，motif名，配列）

タンパク質に関連した知見を体系化することはこの分野で急務となっている．医学生物学分野では，“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おり，DNA上に記述されている遺伝子とその機能に関する網羅的な解析が進んでいる．この結果，遺伝子の産物であるタンパク質どうしの相互作用に関する知見が急速に蓄積されている．しかし，これらの知見は生命現象の原理を解明するうえできわめて重要であるにもかかわらず，生物学やゲノム計画の研究者によって論文という形でしか蓄積されていない．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．このためには上述の物質名を文献中から特定することが重要となる．

3. タンパク質名の特徴

3.1 一般的なタンパク質名の命名法

タンパク質名を形のうえ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用語．

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)のような形の物質名が非常によく現れる．

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．これらの具体例は3.2.1項で示す．

- 新規語の問題
  ヒトの遺伝子だけで約十万個あるといわれていることからも分かるように医学生物学分野において蓄積されている知見は膨大であり，かつ今後さらに指数関数的に増えることが予想される．
- 長い複合語の問題
  医学生物学分野には6単語以上の非常に長い複合語が多く存在する．
- 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まったく未整備のまま蓄積されてきている．このため，“正式な名称”というものが定着せず，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．このため既知語の同定についても専門用語辞書が有効に働かない．

3.2.1 タンパク質名の多様な表記の例

以下に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節で説明したように大文字，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. 自由度の高い同値表現
以下は，GTP結合タンパク質Rasに結合したGDPをGTPに交換する反応を促進するタンパク質Sosの表記である．この例では物質名とその

役割を説明する表現が同値表現として並置されている．

- the Ras guanine nucleotide exchange factor Sos
- the Ras guanine nucleotide releasing protein Sos
- the Ras exchanger Sos
- the GDP-GTP exchange factor Sos
- Sos (mSos), a GDP/GTP exchange protein for Ras

この例では，“guanine nucleotide exchange factor” という機能についての説明が様々な形の表記をとっている．

c. 複合語を構成する単語は著者によってi. のように語順が入れ替わったり，ii. のように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，同一タンパク質が複数の文献において同じ形で記述される保証はない．

3.3 我々が着目した特徴

前節で見てきたようにタンパク質名は未整備に蓄積され非常に自由な表記が許されている．しかし，このような未整備な専門用語には，きわめて目立つ特徴が存在し，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．以下に下線で例を示す．

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

下線で示されている単語は特徴的な形をしていると同時に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．我々はこのような単語を “core-term” と呼ぶことにする．core-term の主だった特徴は，大文字で始まる，あるいは小文字で始まった後に数字が現れる，などである．すなわち “core-term” は目的物質名中に出現する生物学的に意味のある情報を持つ単語であり，英小文字以外の文字が含まれるという特徴がある．よって，“+/−” や “E.”（人名の頭文字）などの目的物質名外に出現する単語は core-term ではない．これらの特徴をどのようにして文字並びの情報から正確にとらえるかは，core-term 抽出処理法の節で論ずる．

前述の目的物質名の範囲を選定する際には，領域専門家からタンパク質（protein）名を同定したい，あるいはさらに受容体（receptor）名，タンパク質中の領域（domain）名を同定したい，といった要望が出される．このように領域専門家から指示される，同定すべき目的物質名のカテゴリ名（protein, receptor, domain など）を “f-term (feature-term)” と呼ぶ．f-term は，文中に存在する目的物質名表記の境界が core-term からの情報だけでは曖昧な場合の補助的な情報として用いられる．また，伸長連結ルールの適用条件としても用いられる．どのような単語が f-term であるかは目的物質名の範囲の選定に依存する．たとえば，細胞名を同定したいのであれば “cell” という単語が f-term に含まれるであろう．f-term となる単語の個数は量的に少ない．以下の例のように f-term は目的物質中にしばしば観察される．

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

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になる．

4. タンパク質名の抽出法

対象とした物質名は**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つの段階を経て抽出される．

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

対象とした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の処理は文字並びの情報から core-term と予想される候補単語を文章中からすべて抽出し，得られた結果を次の処理に渡す．第2から第5までの4つのフィルタは第1の処理で得られた候補単語からセマンティカルには

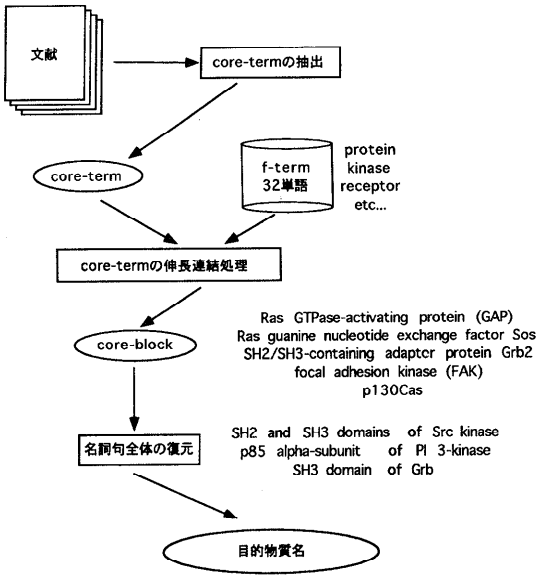


図1 目的物質名抽出処理の流れ
Fig. 1 Extracting target material names.

core-term でない例外的な単語を取り除いていく．以下が具体的な処理内容である．

(処理 a-1) 大文字や数字，記号文字が混在する単語を core-term 候補として正規表現を用いて抽出する．
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．

(処理 a-2) "-" と小文字からなる一定の長さ以上の単語を排除する．
これによって，たとえば，"full-length" や "dual-specificity" のように対象とした物質名以外でも使われる単語を排除する☆．

(処理 a-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．

(処理 a-4) 単位など数に関する単語を排除する．
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" として登録された単語とこれらを語尾として持つものを排除した☆☆．

(処理 a-5) リファレンスに出現する語を core-term から排除する．
正規表現を用いてリファレンスの表記法に関する template を用意し，これにマッチする候補単語を排除する．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 と f-term の伸長連結処理

次に，抽出された core-term をテキスト文中に印付けする．この印を適切に連結伸長することで，"and" 等の接続詞や "of" などの前置詞を含まない単純な構造の名詞句（以下 core-block と呼ぶ）を復元する．その後，このような名詞句間に存在する前置詞・接続詞を接続することで対象とした物質名を複合語の形で抽出する．f-term への印付けもこの段階で施す．

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印を伸長連結するものであり，2つ目は品詞情報をもとに印を伸長するものである．

以下の各例において下線はすでに印付け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

(規則 b-1) core-term，f-term が隣接している場合には印を単純につなぐ．
Src SH3 domain → Src SH3 domain
以下のように括弧も印に含める．
(SH3) → (SH3)☆☆☆

(規則 b-2) ギリシャ文字や大文字1文字が右にあったときに印を右に伸ばす．
p85 alpha → p85 alpha

(2) 品詞情報を用いたルール
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った．

(規則 b-3) 印間の単語の品詞がすべ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印を連結する．
Ras guanine nucleotide exchange factor Sos →
Ras guanine nucleotide exchange factor Sos

(規則 b-4) 冠詞や前置詞が印の左にあった場

☆ 今回の実験では9文字以上の単語を排除することとした．
☆☆ 今回の実験では次の8個の単語，aa, AA, fold, bp, nM, microM, %, UV.
☆☆☆ このような処理は "Src homology 3 (SH3) domains" のような複合語を印付けするために必要となる（下線は core-term）．

表1 対象とした物質名に含まれるcore-termとf-termの割合
Table 1 The rate of core-term and f-term in target material names.

文献	対象物質名	含core-terms	割合	含f-term or core-term	割合
SH3	689	653	94.78%	677	98.23%
SGN	710	596	83.94%	650	91.54%

合，印を冠詞，前置詞の直前まで左に伸ばす．

the focal adhesion kinase (FAK)

→ the focal adhesion kinase (FAK)

4.2.2 名詞句全体の復元

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，実際の係り受けの関係を考慮することなく，表層的な手がかりのみによって復元ルールを実現した．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．

(規則 c-1) "A, B, ... core-term and D f-term"

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

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

(規則 c-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

(規則 c-3) "A of B, C and E"

SH2 domains of Abl, Lck, Fyn, and p85

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

(規則 c-4) "A f-term core-term and core-term"

GTP-binding proteins Rac1 and Cdc42

→ GTP-binding proteins Rac1 and Cdc42

(規則 c-5) "A of B"

p85 alpha subunit of PI 3-kinase

→ p85 alpha subunit of PI 3-kinase

(規則 c-6) "A, B"

the Src-related tyrosine kinase, Hck

→ the Src-related tyrosine kinase, Hck

ルールがどのように適用されるかをc-2について示す．この例では"Src"，"SH2"，"SH3"，"Vav"はcore-termであり，"domains"がf-termである．まず，SH2およびSH3にcore-block復元ルールb-1が適用され，次にSrc homology 2 (SH2)にルールb-3が適用される．同様に3 (SH3) domainsもcore-blockとして復元され，最後に前置詞・接続詞結合ルールc-2が適用される．

4.2.3 誤った印の修正

以上の処理のみによって十分に高い再現率を得ることができるが，これらの処理は誤った印を修正するためのルールを持っていない．我々は誤った印の修正ルールとして以下の2つのルールを加えた．

1つ目はマークしたf-termが最終的に1単語のまま伸長せずに残った場合に印を修正する．たとえば，"protein"のようにf-termは非常にありふれた単語である．このため文献中で物質名の表記以外の箇所でもf-termが単独で用いられ，このような誤りが生じる．

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ものについて，印を修正する．たとえば，"Src-related"のようにcore-termが必ずしも名詞でないことがこのような誤りの原因である．

前者，後者の場合ともに正規表現によるパターンマッチで印を取り除いている．

我々は以上のようなルールによって再現率，適合率ともに高い結果を得た．次章でこの方法によって得られた結果を示す．

5. 実　　験

5.1 実験対象

我々はSrc相同3領域に関する30本の論文要旨（以下SH3）と細胞内信号伝達全般に関する50本の論文要旨（以下SGN）を用いて前章までに述べた手法を評価した．論文要旨はすべてMEDLINE[1]に登録されているものである．また正解語は，専門家が上記の要旨文中で1つずつ印付けすることで用意した．

f-termは領域専門家の要求する認識対象範囲の選定に従い32単語を用いた．core-term抽出処理については以下のとおりである．処理a-2では，いくつかの文字数について予備実験を行った結果，9文字以上の単語を排除することとした．また処理a-4では，次の8個の単語，aa，AA，fold，bp，nM，microM，%，UVを単位としてあらかじめ登録した．

表1はf-termと正解として用意したcore-termが対象とする物質名中にどれだけ含まれるかを表している．この表が示すように対象とする物質名の90%以上がcore-termかf-termのどちらかを必ず含んでおり，我々の着目した表層的手がかりが目的物質名中によく観察されることが分かる．残りの用語は"insulin"，

表2 core-term 抽出処理の結果
Table 2 Result of core-term extraction phase.

文献	rc-term	抽出結果	抽出洩れ	抽出誤り	適合率	再現率
SH3	196	208	3	15	92.79%	98.47%
SGN	225	230	6	11	95.22%	97.33%

表3 伸長連結処理の結果
Table 3 Result of concatenation & extension phase.

文献/評価	正解用語数	抽出用語数	f-p	unsu	disag	p-f	適合率	再現率
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%

"dynamin" のような単語である．

5.2 core-term 抽出処理の評価

表2 は core-term 抽出処理の結果を評価したものである．rc-term（real core-term）は領域専門家が用意した正解語から人手によって選定した core-term を表す．我々の着目した特徴が実際に高い精度で，目的物質名外に出現する単語から識別できていることが分かる．抽出洩れはすべて，処理 a-2 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" などがこの抽出洩れに含まれていた．抽出誤りには細胞名や virus 名が含まれていた．

5.3 伸長連結処理の評価

我々は誤りを次のように分類した．E1 と E2 は抽出誤りであり，E3 は抽出洩れである．

(E1) 誤った箇所への印付け（unsu: unsuitable）
- **(E1-a)** "NINS"（the 258-bp novel insert の略）のようにタンパク質名でないもの
- **(E1-b)** "PC12 cell" のように物質名を表すが今回は対象とする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．
- **(E1-c)** 特定の物質名を表さないもの
  "a novel protein"

(E2) 伸長・連結処理の失敗（disag: disagreement）
- **(E2-a)** 伸長しきらなかったもの
  "interleukin 1 (IL-1) -responsive kinase"
- **(E2-b)** よけいに伸長してしまったもの
  "*same* proline-rich region of FAK (APPKPSR)"
- **(E2-c)** 連結がうまくいかなかったもの
  "p80 and p85 (p80/85)"

(E3) まったく印付けがされなかったもの（f-p: fasle-positive）
"insulin"，"adenylyl cyclase"

以上の分類に従った結果，伸長連結処理の評価は**表3** のようになり，適合率 90.24～93.51%，再現率 93.32～99.85%という結果が得られた．ここで誤りは延べで数えている．たとえば，ある印付けの誤りがあって，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．表3 における f-p（false-positive）は抽出洩れを表し，p-f（positive-false）は抽出誤りを表し unsu と disag が相当する．評価1では全対象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っている．評価3はさらに 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．

誤りの分類 E1-b に含まれる誤り（cell 名やファージ名）はその周辺の単語（"in PC12 cell"）を調べることにより目的物質名と明確に区別できる．よって，これらの誤りが今後の知識抽出処理においてノイズとなるとは考えにくい．

我々の評価は前置詞・接続詞の接続の失敗（誤りの分類 E2-c）を厳しく数え上げている．たとえば，以下では core-block はすべて正しく印付けされているが，名詞句全体の復元に失敗しているために p-f を5と数えている．

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

名詞句の復元は係り受けの曖昧性などがあるために，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，上の例のようにすべてを解決できたわけではない．

誤りの分類 E3 に含まれるものを削減するためには新たなルールを導入することが必要となる．**表4** においては以下のルールを core-term 抽出ルールに加えている．

表4 追加ルールを用いた場合の伸長連結処理の結果
Table 4 Result with an additional rule.

文献/評価	正解用語数	抽出用語数	f-p	unsu	disag	p-f	適合率	再現率
SH3/4	689	698	8	11	21	37	94.70%	98.84%

- '.*"子音"(in | ase | ol)(s | ε)' というパターンにマッチする名詞は core-term である

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．表4内の評価4は，評価1と同様に対象物質名すべてを評価の対象とし，さらに細胞名やファージ名は抽出されても構わないとして解釈を拡張したものである．

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 p-f が22減少している．これは，名詞句全体の復元に成功した例が増えたためである．以下の例では印付けが "3-kinase" から "phosphatidylinositol" まで伸長しなかった．その結果 core-block 接続ルールがマッチせず，p-f は4つと数えられる．これに対して上の追加ルールを用いると印付けが "phosphatidylinositol" まで適切に伸長するため，係り受け復元ルールが適用でき全体が正しく認識されている．

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

⟶

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

6. 考　　察

我々の用意した core-term 抽出ルールは表層的な手がかりを用いたものであり，医学生物学分野の知識に特化したものではない．すなわち，我々のルールは小文字と数字の組合せのように，通常の英文中では見られない単語の形をルールとして記述しているだけであり，たとえば，"c-で始まれば細胞由来（cell-derived）のタンパク質を表す" というような分野固有の知識は用いていない．このような形態の用語は医学生物学分野に固有のものではない．

2章において医学生物学分野の例を示したように専門用語には様々なものが存在する．これらは

- 命名法が存在するかどうか，
- 新しい用語がたえず出現するのか，あるいは稀にしか出現しないのか，
- 一度出現した用語は，その後定着するのか，

という観点から分類できる．新規語が頻出しかつ明示的な命名法なしで命名される用語は多くの分野で観察されることが想像される．本報告で我々が取りあげたタンパク質名ではそのような専門用語の特徴が端的に示されていると考える．我々の手法はタンパク質のみならず遺伝子名や細胞名に対して有効であり，また，他分野の用語についても，上記の分類に従った一部のものについて有効であると期待される．

一般的に専門的な分野の文献を処理する際には必要とする語彙情報が不足する．なぜならば，一般用語辞書にない単語やいい回しが多く存在するからである．特に学術分野の文献は，新聞記事などの標準的な文章に対して，用いる語彙と文体の両面で大きくかけ離れている．このような場合，利用可能な情報は非常に限定される．表層的手がかりはつねに利用可能な数少ない情報の1つである．このため，我々は品詞情報などを極力利用せずに表層的手がかりを用いた．伸長連結ルールの適用条件に必要な品詞情報は大まかなものでよく，形態素解析の精度に大きくは依存しない．

実験の結果，我々の手法は表層的手がかりの乏しい一部の語句を除いたすべての表記に対応できることが分かった．すなわち，従来の手法が苦手とする新規語や造語，長い複合語，表記の多様性に同時に対応している．

表層的手がかりの乏しい用語の抽出処理について，抽出誤りと抽出洩れの観点から見た2つの問題点を以下に示す．

まず，抽出誤りについての問題である．

(a) focal adhesion kinase

(b) major tyrosine-phosphorylated protein

上の例では（a）は抽出したいタンパク質名であるのに対して，(b) は "チロシン残基がリン酸化されたタンパク質" といっているだけであり抽出されるべきではない．しかし，(a), (b) はともに小文字のみからなり core-term を含まず，"kinase" や "protein" を f-term として含んでいる．このため前章で述べた伸長連結ルールによって（a）を抽出すれば（b）も同様に抽出されてしまう．このように表層的手がかりの乏しい目的物質名を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，再現率と適合率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)" という表記がある．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シノニム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と考えられる．

次に抽出洩れの問題についてである．"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，上述のような言い替え表現がされることは少ない．このような単語の洩れを削減する方法として以下の2つが考えられる．

第1の方法は形態的な手がかりを用いるものである．"insulin" のような単語には子音・母音の使われ方や語尾の形など，一般の用語にはあまり見られない特徴が存在する．このような文字単位の表層的手がかりを用いたcore-term抽出ルールを追加することで我々はこれらを抽出し，その結果適合率94.70%，再現率98.84%を得た．

しかし，このような語句には新規語があまり登場せず，さらに表記の揺れもほとんどない．このためもう1つの方法としては，優れた辞書が入手可能であり，かつその利用によってただ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．

7. 結　　論

我々は，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．本手法は自動抽出されたcore-termに対して伸長連結処理を施すことで対象とした物質名を抽出する．

この手法によって我々は，領域固有の辞書を用意することなしに医学生物学文献から専門知識，すなわち物質名を網羅的に獲得することに成功した．

本手法は従来の手法が苦手とする新規語や造語，長い複合語，表記の多様性に対応し，98.84%の再現率と94.70%の適合率を得た．専門分野の文献から情報抽出を行うための前処理として十分に実用にたえるものである．

我々のとら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期待される．

抽出した用語から言い替え表現などの知識を抽出しテキスト中にフィードバックするほかに，抽出した用語（複合語）から部分構造を取りだしフィードバックすることも適合率を高めるために有効であると考えられる．また，既存辞書を抽出結果に融合していくことで再現率をさらに向上することが可能である．本手法をbootstrapに用いることでタンパク質名の出現環境に関する情報を獲得することは，抽出ルールを拡充していくのに有効であると期待される．SGNの結果がSH3についてのものより低かった原因としてはルールが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．

また，本報告では論文要旨を用いて実験を行ったが，論文全文についても同様の手法がとれることを確認する必要がある．

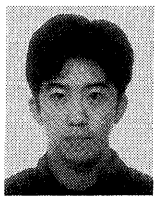
情報抽出処理によって分野固有の知識を獲得し，さらに知識の体系化を行うためには同一物を指している用語の認識が不可欠である．前述の手法による略語辞書生成は同義語認識に有効であると期待される．また，表記の多様性に対処するための同義語認識手法の構築が必要となる．

謝辞　この研究の一部は文部省科学研究費補助金「ゲノムサイエンス」の援助を受けている．

参 考 文 献

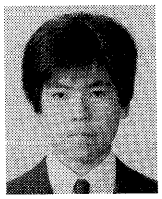
1) Internet Grateful Med Development Team, National Library of Medline: MEDLINE (1996).
2) K-Su, M.W. and Chang, J.: A Corpus-based Approach to Automatic Compound Extraction, *32th Annual Meeting of the Association for Computational Linguistics* (*ACL'94*) (1994).
3) Smadja, F.: Retrieving Collocations from Text: Xtract, *Computational Linguistics*, Vol.19, No.1, pp.143–177 (1993).
4) *Proc. 6 Message Understanding Conf.* (*MUC-6*), Morgan Kaufmann (1995).
5) Grishman, R.: The NYU System for MUC-6 or Where's the Syntax, *6 Message Understanding Conf.* (*MUC-6*), pp.167–176, Morgan Kaufmann (1995).
6) Riloff, E.: Automatically Generating Extraction Patterns from Untagged Text, *13th National Conference on Artificial Intelligence* (*AAAI-96*) (1996).
7) Wakao, T., Gaizauskas, R. and Wilks, Y.: Evaluation of an Algorithm for the Recognition and Classification of Proper Names, *16th International Conference on Computational Linguistics* (*COLING 96*) (1996).

(平成9年10月7日受付)
(平成10年6月5日採録)



福田賢一郎（学生会員）

1972年生．1996年早稲田大学情報科学科卒業．1998年同大学大学院理工学研究科卒業．1998年東京大学大学院理学系研究科情報科学専攻博士課程入学，現在に至る．同大学医科学研究所において医学生物学データからの知識獲得に関する研究に従事．



角田　達彦（正会員）

1967年生．1989年東京大学理学部物理学科卒業．1995年同大学工学系大学院博士課程修了．工学博士．同年京都大学工学研究科助手．1997年東京大学医科学研究所特別研究員．1998年同大学医科学研究所助手，現在に至る．IJCNN'93 Student Award受賞．1994年情報処理学会学術奨励賞受賞．岩波ソフトウェア科学第15巻「自然言語処理」（岩波書店）．言語処理学会，日本神経回路学会，電子情報通信学会，人工知能学会，日本認知科学会，分子生物学会各会員．

田村あゆち

1972年生．1994年ミシガン大学卒業分子生物学専攻．1997年東京大学大学院医学系研究科修士課程修了．同年同大学医科学研究所勤務．

高木　利久（正会員）

1976年東京大学工学部計数工学科卒業．九州大学を経て，現在，東京大学医科学研究所ヒトゲノム解析センター教授．工学博士．ゲノム情報処理，データベース等の研究に従事．人工知能学会，生物物理学会等各会員．