自然言語処理技術に基づく意味構造を利用した 情報検索と情報抽出：

MEDIEとInfo-PubMed

大田 朋子¹

著者抄録：医学生物学の分野では、分野に関連する論文のアブストラクトが網羅的にMEDLINEというデータベースに収録されており、その数は現在1,500万件にも上る。実験室で日々量産される実験データを解析・評価していくためにも、関連する多数の文献に記述されている情報を統合していく必要があり、特に、分子間相互作用や遺伝子と疾患の関連といった関係概念の検索・抽出への自然言語処理技術の応用が切望されている。そこで本稿では、構文解析によって計算された意味構造を用いて、高精度かつ高速に関係概念を検索・抽出するシステムについて紹介する。

キーワード：自然言語処理, 情報検索, 情報抽出, HPSG構文解析, 拡張領域代数

Semantic retrieval for the accurate identification of relational concepts based on deep syntactic parsing:

MEDIE and Info-PubMed

OHTA Tomoko¹

Author Abstract: Recently, biomedical researchers have been facing the vast repository of research papers, e.g. MEDLINE. These researchers are eager to search biomedical correlations such as protein-protein or gene-disease associations. The use of natural language processing technology is expected to reduce their burden, and various attempts of information extraction using NLP has been being made. However, the framework of traditional information retrieval (IR) has difficulty with the accurate retrieval of such relational concepts. This is because relational concepts are essentially determined by semantic relations of words, and keyword-based IR techniques are insufficient to describe such relations precisely. Here, we propose an intelligent search engine for the accurate retrieval of relational concepts from MEDLINE, MEDIE, and a GUI-based efficient MEDLINE search tool, Info-PubMed.

Key words: NLP, information retrieval, information extraction, HPSG Parsing, region algebra

原稿受理 (2006-11-17)

(情報管理 49 (10): 555-563)

1. はじめに

近年では、インターネットの普及も手伝って、さまざまな知識がテキストという形で蓄積されるようになり、テキストから効率よく情報を獲得する手法が注目を集めている。これまでの検索では、主に文書単位での検索が多かったが、より詳細な情報単位の検索に対する要求も急激に増加してき

ているのが現状である。特に、学術分野においては、ある事柄に関連する文書のみならず、文書中に記述されている専門的な概念間の関係を検索することが求められている。例えば、医学生物学分野においては、MEDLINEという文献データベースにこの分野で発表された論文の抄録が網羅的に収録されており、その登録件数は1,500万件にも上る。また、その数は一年間に10%の割合で増加し続け

¹ 東京大学大学院情報理工学系研究科コンピュータ科学専攻 (〒113-8656 東京都文京区本郷7-3-1) Tel. 03-5803-1697
E-mail: okap@is.s.u.-tokyo.ac.jp

¹ Department of Computer Science, Graduate School of Information Science and Technology, University of Tokyo (7-3-1 Hongo Bunkyo-ku, Tokyo 113-8656)

図1 MEDLINE収録文献数の推移

ているのである(図1)¹⁾。この膨大な文献の中に記述されている、遺伝子と疾患の相関やたんぱく質の相互作用といった専門的な関係概念を検索することが必要とされている。当然のことながら、関連する文献を人間がすべて読むことは困難であり、一方で、これまでに普及しているような単純な文書検索ではこのような詳細な情報を高精度で特定することはできない。

本稿では、特にこの医学生物学分野に着目し、MEDLINEから関係概念を高精度かつ高速に検索するシステムMEDIEと、この分野における関係概念の中でも、近年注目を集めているたんぱく質の相互作用に関する情報をインタラクティブに抽出するシステムInfo-PubMedについて紹介する^{2,3)}。

2. 自然言語処理技術

本稿で紹介する2つのシステムでは、われわれがこれまでに研究・開発を進めてきた各種の自然言語処理技術を利用している。本章では、その概要を簡単に説明する。

2.1 HPSG構文解析

文は、単なる記号(単語・形態素)の並びではなく、それらが構造的に組み合わされて複合的な意味が表現されている。主語句、目的語句、前置詞句など、句の構造とその関係を把握することが文によって表現された知識を解釈する基礎となるのである。このような文中の単語間、句間の構造

を把握する処理を構文解析という。構文解析技術の中核は、いかにして単語の間の文法的に可能な構造をすべて洗い出すか、また、膨大な数の構造の中から正しい構造を認定するかにあるのである。

HPSGとは、文とその意味との対応関係を系統的に記述する現代文法理論の一つである主辞駆動型句構造文法(Head-driven Phrase Structure Grammar)のことである。従って、このHPSGに基づく構文解析により、文中の句構造のみならず述語項構造と呼ばれる意味の構造を計算することができる。述語項構造とは、文中の単語間の意味的關係を表現する意味構造で、HPSG構文解析器⁴⁾の出力では、句構造に加えて述語と項の関係を解析することができる(図2)。このHPSG構文解析器は、出力の情報量が豊富であるがゆえに計算量が多く、これまで実世界の大規模なテキストを解析するには計算機への負担が大きすぎると考えられていた。しかしわれわれは、高精度かつ高速な構文解析器を開発し、また、近年のコンピュータの仕様の飛躍的な進歩も手伝って、MEDLINE全文を現実的な時間で解析することに成功した^{5,6)}。

2.2 用語認識

近年、特に医学生物学分野では、大規模なオントロジや辞書の開発が進んでいる。これらは、ある分野の用語や知識を系統的に整理し、テキスト中の表現をそれが指し示す実世界の实体へ写像するためのものと考えられることができる。

例えば、遺伝子・たんぱく質名に関する用語辞

図2 HPSG構文解析器の出力

書では、既存のデータベースで用いられるシンボル、遺伝子・たんぱく質の一般名、別名、生物種名等が整理されている⁷⁾。この辞書を用いて、テキスト中に現れる用語に対してIDを付与することで、テキスト表現の差異を標準化し、用語の示す意味を表現することができる。

また、関係概念を記述するためには、動詞的な表現も非常に重要である。例えば、たんぱく質の活性化を意味する表現としては、“activate”や“enhance”といった動詞がよく用いられる。このようなイベントを表す表現の種類は、専門用語に比べればはるかに少ないが、オントロジや辞書は存在しない。そこでわれわれは、MEDLINEのアブストラクト500件を分析し、頻出する167のイベント表現を18のイベント型に分類して、独自にイベント表現オントロジを作成した。

本稿で紹介するシステムでは、この遺伝子・たんぱく質名辞書と、疾患名に関する用語辞書を用いて、辞書ベースの用語認識アルゴリズム⁸⁾を適用し、テキスト中の専門用語(遺伝子・たんぱく質・疾患名)に対して辞書IDを付与するとともに、イベント表現オントロジを用いて、イベント表現に対してもオントロジIDを付与した。

2.3 拡張領域代数

本稿で紹介するシステムでは、タグのネストおよび交差を含むような複雑な構造化テキストの検索のために拡張領域代数を提案し、これを基に開発した高速な検索アルゴリズム⁹⁾を用いて検索を行っている。拡張領域代数は、テキスト上の単語

表1 拡張領域代数の演算子

[tag]	“<tag>” タグで囲まれた領域
$A > B$	Bを含む領域 A
$A >> B$	Bを含む領域 A (Aはネストしない)
$A < B$	Bに含まれる領域 A
$A << B$	Bに含まれる領域 A (Aはネストしない)
$A - B$	Aで始まり Bで終わる領域
$A \& B$	Aと Bを含む領域
$A B$	Aまたは Bを含む領域

列を領域としてとらえ、これを引数とする演算子で定義されている(表1)。AとBは領域を表し、演算の結果も領域として返される。4つの包含演算子($>$, $>>$, $<$, $<<$)はXMLにおける先祖／子孫の関係を表している。例えば、“ $A > B$ ”はBを子孫に持つような領域Aを表す。これらの演算子を用いて記述する拡張領域代数クエリによる検索では、最初の解を検索するには定数時間、全解検索はクエリ中の単語の最低頻度に比例する時間で抑えることができ、高度に構造化されたテキストに対して、効率よく検索を行うことができる。

3. MEDIE：意味構造を利用した関係概念検索システム

(<http://www-tsujii.is.s.u-tokyo.ac.jp/medie/>)

MEDIE(図3)は、構文解析や用語認識によって意味構造を付与されたテキストを対象とし、拡張領域代数の検索アルゴリズムを用いて効率よく関係概念を検索するシステムである²⁾。

図3 MEDIE (<http://www-tsujii.is.s.u-tokyo.ac.jp/medie/>)

図4 関係概念検索システムのアーキテクチャ

3.1 システム概要

本システムは、テキストの意味構造を計算するオフライン処理と、ユーザの入力を受け付けて意味構造を検索するオンライン処理から構成されている（図4）。

オフライン処理：2章で紹介したHPSG構文解析器と用語認識器を適用し、2004年度末の時点でMEDLINEに収録された全文献約1,500万件に対して、述語項構造と辞書・オントロジIDが付与された意味構造付きテキストを作成する。表2は、意味構造を付与する処理前後のデータサイズをまとめたものである。

オンライン処理：ユーザの検索要求を拡張領域代数のクエリに変換し、拡張領域代数の検索アルゴリズムを用いてクエリの構造と一致する意味構造を検索する。

例えば、“CRP excludes something” という意味構

表2 意味構造付きテキストのサイズ

文献数	14,785,094
アブストラクト数	7,291,857
文数	70,935,630
単語数	1,462,626,934
構文解析に成功した文の数	69,243,788
述語項関係の数	1,510,233,701
句構造の数	3,094,105,383
用語数（遺伝子）	84,998,621
用語数（たんぱく質）	27,471,488
用語数（病名）	19,150,984
用語数（イベント表現）	51,810,047
MEDLINE のサイズ	9.3 GByte
意味構造付きテキストのサイズ	292 GByte
インデックスのサイズ	954 GByte

```
[sentence] >>
([word arg1="$subject" arg2="$object" base="v"] &
 ([phrase id="$subject" > x] &
 ([phrase id="$object" > y]))

[sentence] >>
([word arg1="$subject" arg2="$object" base="v"] &
 ([phrase id="$subject" > x] &
 ([phrase id="$object" > y]))
```

図5 拡張領域代数のクエリ

造を検索するクエリを図5の上部に示す。このクエリは、文が単語“exclude”を含み、かつその第一引数(“arg1”)つまり意味上の主語にあたる句が単語“CRP”を含むことを表している。しかし、この図で示したようなクエリをユーザが直接入力するのは困難なため、本システムではユーザが主語x、目的語y、動詞vを入力すると、この各語を辞書またはオントロジ中のIDと置き換えた上で、図5の下部のように拡張領域代数クエリを自動生成し、検索を行うことができる。

3.2 評価実験

本節では、MEDIEシステムの検索速度と精度の評価実験について述べる。

実験に使用した検索要求を表3に示す。“something”は任意の語を、“disease”は疾患名を表す。“(negative)”は否定文を検索することを意味している。これらは生物学者によって選択された、医学生物学分野における典型的な検索要求で、特定の遺伝子を制御するものの検索や特定の病気を引き起こす要因の検索などを意味している。

検索件数、検索時間、検索精度を表4に示す。実験では、述語項構造を利用するか、辞書・オントロジIDを利用するか、の2点を変化させ、それぞれが検索結果に与える影響を検討した。“キーワード検索”とは、入力単語の共起のみの検索、“意味構造検索”とは、述語項関係を利用した検索であり、この比較によって述語項構造を利用する効果を検証することができる。表4におけるクエリ番号の接尾辞は、辞書・オントロジIDを利用するかどうかを表しており、X-1はIDを利用せず、X-2は、遺伝子、タンパク質、疾患名の辞書IDのみを利用、X-3はすべてのIDを利用した検索結果である。検索精度は、適合率とキーワード検索の結果に対する相対再現率を測定した（適合率とは、検索結果のうちで正解したものの割合、相対再現率とは、キーワード検索によって検索された正解のうち、意味構造検索によって検索された正解の割合のことで

表3 評価実験に用いた検索要求

番号	検索要求
1	<i>something inhibit</i> ERK2
2	<i>something trigger</i> diabetes
3	adiponectin increase <i>something</i>
4	TNF activate IL6
5	dystrophin cause <i>disease</i>
6	macrophage induce <i>something</i>
7	<i>something</i> suppress MAP phosphorylation
8	<i>something</i> enhance p53 (negative)

ある)。精度測定には、各検索要求に対する検索結果100件を上限として、すべての検索結果をランダムに結合したもの（1,839文）を生物学者が正解判定したものをを用いた。

実験結果から、意味構造検索によって適合率が有意に改善したことがわかる。多くの検索で適合率は80%を超え、検索要求4と5についてはほぼ100%の適合率を実現している。これは、述語項構造を利用することで関係概念が正確に特定できていることを示している。また、検索件数の増加から、辞書やオントロジを利用した検索要求の拡張が有効であることが示された。特に、イベント表現に関するオントロジの利用が有効であることがわかる。

また、検索結果のエラー解析を行ったところ、キーワード検索で検索されたが意味構造検索では検索できなかった文（false negative）では、動詞を名詞化して表現している例や、複合動詞の例、共参照を用いている例などがあった。また、意味構造検索で検索されたもののうちで不正解だった文（false positive）では、名詞句内の解析が必要な例とともに用語認識の誤りなどであった。

4. Info-PubMed：たんぱく質相互作用情報のインタラクティブな抽出システム

(<http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/>)

近年、医学生物学分野への情報抽出技術の適用が試みられることが多いが、実用化までできていない例は多くない。これは、これまでの情報抽出技術はまだ精度が十分ではなく、また、より詳細な分野への適用には多くの人手による労力を必要とするためである。

これまでに実用に適した精度を得られている既存のシステムの多くは、人手によって書かれた情報抽出ルールを用いたものであり、他の細分野に

表4 評価実験の結果

クエリ 番号	キーワード検索			意味構造検索			
	検索件数	時間(first/all)	適合率	検索件数	時間(first/all)	適合率	相対再現率
1-1	252	0.00/1.5	74/100(74%)	143	0.01/2.5	96/100(96%)	51/74(69%)
1-2	348	0.00/1.9	61/100(61%)	174	0.01/3.1	89/100(89%)	42/61(69%)
1-3	884	0.00/3.2	50/100(50%)	292	0.01/5.3	91/100(91%)	21/50(42%)
2-1	125	0.00/1.8	45/100(45%)	27	0.02/2.9	23/27(85%)	17/45(38%)
2-2	113	0.00/2.9	40/100(40%)	26	0.06/4.0	22/26(85%)	19/40(48%)
2-3	6259	0.00/12.1	42/100(42%)	662	0.01/1527.4	76/100(76%)	8/42(19%)
3-1	287	0.00/1.5	20/100(20%)	30	0.05/2.4	23/30(80%)	6/20(30%)
3-2	309	0.01/2.1	21/100(21%)	32	0.10/3.5	26/32(81%)	6/21(29%)
3-3	338	0.01/2.2	24/100(24%)	39	0.05/3.6	32/39(82%)	8/24(33%)
4-1	4	0.26/1.5	0/4(0%)	0	2.44/2.4	0/0(—)	0/0(—)
4-2	195	0.01/2.5	9/100(9%)	6	0.09/4.1	5/6(83%)	2/9(22%)
4-3	2063	0.00/7.5	5/100(5%)	94	0.02/10.5	89/94(95%)	2/5(40%)
5-2	287	0.08/6.3	73/100(73%)	116	0.05/14.7	97/100(97%)	37/73(51%)
5-3	602	0.01/15.9	50/100(50%)	122	0.05/14.2	96/100(96%)	23/50(46%)
6-1	10698	0.00/42.8	14/100(14%)	1559	0.01/3014.5	65/100(65%)	10/14(71%)
6-3	42106	0.00/3379.5	11/100(11%)	2776	0.01/5100.1	61/100(61%)	5/11(45%)
7	87	0.04/2.7	34/87(39%)	15	0.05/4.2	10/15(67%)	10/34(29%)
8	1812	0.01/7.6	19/100(19%)	84	0.20/29.2	73/84(87%)	7/19(37%)

移植するためには膨大な労力を費やさなければならぬ。

一方、機械学習によって抽出ルールを構築するシステムでは、学習データが大量に必要であったり、抽出ルールが十分に一般化されていないなどの問題点が多く、実用化には至っていない。

そこでわれわれは、前章で紹介したMEDIEシステムによる検索結果に加え、述語項構造を用いて一般化した抽出ルールを用いて、インタラクティブにたんぱく質相互作用情報を抽出するシステム、Info-PubMedを構築した。

4.1 システム概要

Info-PubMed (図6) は、主に辞書検索モジュールと相互作用検索モジュール、オブジェクトの詳細情報表示モジュールから成り、それぞれがサイト内のサブウィンドウとして表示されている。図7に示すように、辞書のエントリや相互作用情報、相互作用が記述されている文章などはすべてオブジェクト化されており、各サブウィンドウ間では、オブジェクトをドラッグ・ドロップによって取り扱うことができる。また、ユーザ用のフォルダが用意されており、興味のあるオブジェクトはこのフォルダに整理して保存したり読み込んだりすることができる。

図6 Info-PubMed

(<http://www-tsuji.is.s.u-tokyo.ac.jp/info-pubmed/>)

4.2 遺伝子辞書の検索

辞書の検索は、“Gene Dictionary”モジュールで行うことができる。遺伝子辞書には、公共データベースで用いられる遺伝子シンボル、一般名、別名、産物名、生物種名とともに、データベースへのリンク情報が整理されている。辞書検索モジュールでは、ユーザからの検索要求がシンボル、一般名、別名、産物名のいずれかのフィールドと一致

図7 Info-PubMedの基本動作

した辞書エントリをリストとして返す。この検索対象フィールドは、検索オプションとして任意に設定することが可能である。また、この辞書には、ヒト、マウス、ラット、ショウジョウバエ、分裂酵母、出芽酵母、線虫の7生物種の遺伝子が収録されており、初期設定では全生物種を検索対象としているが、フィールドと同様に検索オプションとして検索対象生物種を任意に選択することが可能である。また、各辞書エントリは、遺伝子オブジェクトとしてユーザ用フォルダに保存することができる。

4.3 相互作用情報の検索

辞書検索の結果得られたリストから、興味のある遺伝子オブジェクトを、相互作用検索モジュールにドラッグ・ドロップすることによって、相互作用情報を検索することができる。検索対象は、3.1節および表2で説明した2004年度末の時点でMEDLINEに収録された全文献約1,500万件である。

図8 述語項構造に基づいたパターンの抽出

この相互作用情報の検索では、3章で説明した述語項構造を用いた意味検索に加え、図8に示するような述語項構造に基づくパターンを抽出し、そのパターンを機械学習することによって、特徴的な表現を一般化して検索することが可能になっている¹⁰⁾。

検索結果は、MEDLINE中で二つの遺伝子名が共起している文章数と、本システムの検索アルゴリズムによって相互作用を意味していると判定され

た文章数とともに、遺伝子のリストの形で表示される(図9)。相互作用オブジェクト中の数字は、分母は共起している文章数、分子は相互作用と判定された文章数を示しており、分子の数が大きい順に並び替えられている。

4.4 詳細情報の表示

詳細情報表示モジュールでは、遺伝子オブジェクトや相互作用オブジェクトなどの詳細情報を参照することができる(図10, 11)。遺伝子オブジェクトの詳細情報表示では、図10に示すように各辞書エントリの詳細が参照できると同時に、リンク

情報のボタンをクリックすることによって、その遺伝子についての情報を得ることができる公共データベースを参照することができる。また、同様に相互作用オブジェクトの詳細情報表示では、その相互作用を記述していると予測される文の一覧を参照することができ、それぞれの文オブジェクトに添付されたMEDLINEへのリンクから、その文章を含んでいるアブストラクトを参照することが可能である。

4.5 ユーザ用フォルダ

ユーザ用フォルダには、辞書検索や相互作用検索で得られた各オブジェクトを整理・保存することができる。このフォルダは、フォルダ名を任意の名前に変えることもでき、また、複数作成することも可能なので、ユーザの興味があるテーマごとに遺伝子オブジェクトや相互作用オブジェクトを整理・保存することができる。さらに、各オブジェクトには簡単なメモを書き込むこともできるので、遺伝子オブジェクトのシンボルから一般名を連想しにくい場合などに、利用すると便利である。

5. おわりに

本稿では、自然言語処理技術に基づく意味構造を用いた情報検索・情報抽出システムについて紹介した。本システムでは、HPSG構文解析器と用語認識器を適用することで意味構造付きのデータを作成し、その意味構造を検索対象とすることで、関係概念が高精度かつ高速に検索できることを示

図9 相互作用検索結果

図10 辞書エントリ詳細表示

図11 相互作用詳細情報表示

した。本システムでは、構文解析器の出力のうち、主語－動詞－目的語の関係のみを扱ったが、意味構造はその他の関係も含んでおり、実験環境や様相などといったさまざまな関係概念を検索対象としていくことも可能である。

これらのシステムは、われわれがこれまでに研究・開発を進めてきた自然言語処理の基礎技術が、現実世界の大規模なテキストに対して適用可能であることを示すために実装したものであり、改良の余地は多く残されている。特に、辞書は既存のリソースを利用したにすぎず、辞書データの質の改善によって用語認識の質を高めることができるに違いない。加えて、機械学習アルゴリズムを適用することにより、あいまい性の解消などを期待することもできる。また、今回利用したイベント

表現オントロジは、手作業でアブストラクトを解析して作成した規模の小さなものであるが、より大規模なオントロジを構築することによって、検索精度を改良することが可能である。

しかし、現段階の実装でもデモサイトに対する反響は大きく、単純な文書検索ではない意味構造を用いた情報検索や情報抽出といった関係概念の検索に対する要求と期待を実感している。

また、本稿で紹介したシステムは、医学生物学分野に特化したものであるが、これらのシステムに用いた自然言語処理の基礎技術は、分野に依存するものではなく、関係概念の検索を必要とするさまざまな分野に応用が可能であると期待されるものである。

参考文献

- 1) National Library of Medicine. "Fact Sheet MEDLINE". (online), available from <<http://www.nlm.nih.gov/pubs/factsheets/medline.html>>, (accessed 2006-11-08).
- 2) Miyao, Y.; Ohta, T.; Masuda, K.; Tsuruoka, Y.; Yoshida, K.; Ninomiya, T.; Tsujii, J. "Semantic retrieval for the accurate identification of relational concepts in massive textbases". Proc. of COLING/ACL 2006. Sydney, 2006-07, Association for Computational Linguistics. p.1017-1024.
- 3) Ohta, T.; Miyao, Y.; Ninomiya, T.; Tsuruoka, Y.; Yakushiji, A.; Masuda, K.; Takeuchi, J.; Yoshida, K.; Hara, T.; Kim, J. D.; Tateisi, Y.; Tsujii, J. "An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing". Proc. of COLING/ACL 2006. Sydney, 2006-07, Association for Computational Linguistics. p.17-20.
- 4) Miyao, Y.; Tsujii, J. "Probabilistic disambiguation models for wide-coverage HPSG parsing". Proc. of ACL 2005. Michigan, 2005-06, Association for Computational Linguistics. p.83-90.
- 5) Ninomiya, T.; Tsuruoka, Y.; Miyao, Y.; Taura, K.; Tsujii, J. Fast and scalable HPSG parsing. Traitement automatique des langues (TAL). Vol.46, No.2, 2006.
- 6) Taura, K. "GXP: An interactive shell for the grid environment". Proc. IWIA2004. 2004-01, IEEE Computer Society. p.59-67.
- 7) Koike, A.; Takagi, T. "Gene/protein/family name recognition in biomedical literature". Proc. Biolink 2004. 2004-05. p.9-16.
- 8) Tsuruoka, Y.; Tsujii, J. Improving the performance of dictionary-based approaches in protein name recognition. Biomedical Informatics. Vol.37, No.6, 2004, p.461-470.
- 9) Masuda, K.; Ninomiya, T.; Miyao, Y.; Ohta, T.; Tsujii, J. Nested region algebra extended with variables. In Preparation, 2006.
- 10) Yakushiji, A.; Miyao, Y.; Ohta, T.; Tateisi, Y.; Tsujii, J. "Automatic Construction of predicate-argument structure patterns for biomedical information extraction". Proc. EMNLP 2006. Sydney, 2006-07, Association for Computational Linguistics. p.284-292.