社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－130 (5)
2003－ＦＩ－ 71 (5)
2003／5／22

ライフサイエンス向けテキストマイニングツール MedTAKMI

松澤 裕史[†] 長野 徹[†] 村上 明子[†] 浦本 直彦[†‡] 武田 浩一[†]

近年，ライフサイエンスやバイオインフォマティクス分野の研究が盛んに行われている．バイオメディカル系の研究者らにより MEDLINE と呼ばれる文献データベースが広く利用されており，彼らは文献データベースを検索し，様々な文献を参照して，彼ら自身の研究に役立てている．1,200 万件を超えるバイオメディカル文献を保持する MEDLINE は，大量の文書から知識発見を行うテキストマイニングにとって非常に興味深い対象である．筆者らは，バイオメディカル文献データベースを対象としてセレスター・レキシコ・サイエンシズ（株）と共同で，ライフサイエンス向けテキストマイニングシステム（本稿では，MedTAKMI システムと呼ぶ）を構築して MEDLINE への適用を行った．

MedTAKMI : a Text Mining Tool for Life Sciences

Hirofumi MATSUZAWA[†] Tohru NAGANO[†] Akiko MURAKAMI[†]

Naohiko URAMOTO[†‡] Kohichi TAKEDA[†]

MEDLINE is a famous bio-medical document database, in which over 12 million articles are registered. Bio-medical researchers often refer to MEDLINE for articles on established or emerging biotechnology. Due to its huge size, the MEDLINE database is an ideal candidate for the application of text mining techniques. We developed a text mining system for life science articles in cooperation with Celestar Lexico-Sciences, Inc., and applied it to the MEDLINE database.

[†]日本アイ・ビー・エム（株）東京基礎研究所
Tokyo Research Laboratory, IBM Research
[‡]国立情報学研究所
National Institute of Informatics

1 はじめに

近年，ライフサイエンスやバイオインフォマティクスが注目を浴びており，ヒトゲノムの解読が終了したことなどがニュースとしても大きく取り上げられている．バイオインフォマティクスなどの研究分野は，医学，医療，創薬といったバイオメディカル系の研究者だけでなく，データベース，データマイニングの研究者にとっても非常に興味深い研究分野である．これらの研究テーマとして，多種多様に蓄積されているバイオメディカル系データベースを対象とした異種統合データベースシステム，ゲノムのシーケンスを対象としたパターンを発見する技術 [2]，化合物の構造を対象としたデータマイニング技術 [1] などが挙げられる．

また，バイオメディカル文献から自動的に有益な情報を取得しようとする要望が高まっており，バイオメディカル系の文献が自然言語処理のコミュニティでも注目されている．自然言語処理を対象とした国際会議 ACL 2003[3] においてもバイオインフォマティクスをワークショップ[1] で取り上げるなど，重要な研究テーマとなっている．情報検索を対象とする国際会議 SIGIR 2003[4]，TREC 2003[5] などでもワークショップの開催が予定されている．

従来の自然言語処理技術が効果的であったのは，新聞記事のようにきちんと推敲された文書であり，使われる用語も辞書にあるものが多かった．しかしながら，バイオメディカル文献では次のような特徴があり，その処理は容易ではない．

- 通常の辞書に含まれない技術用語が非常に多く存在する．例えば，UMLS[6] には，200 万を超えるタームが登録されている[2]．
- 化学式，物質名，ゲノムのシーケンスなど，記号や数式などが，文献中に多数含まれている．
- 省略語（アクロニム）が多数存在する．例えば，DNA(deoxyribonucleic acid)，COPD (Chronic Obstructive Pulmonary Disease)[3] などである．
- 同じ物質を指す言葉が複数存在する．例えば，発見者の名前を由来とするターム，機能を由来とするターム，学名など複数の名称を一つの物質に対して付与しているような場合が存在する．
- 類似した振る舞いに対して，多種な表現が多い．例えば，reduce と downregulate など．

バイオメディカル分野の研究者らは，様々な目的で日常的にバイオメディカル文献を参照している．例えば，ウェットラボの研究者は自らの実験を行う前に，事前に似たような実験報告が掲載された論文がないか検索して目を通しておくことで，時間や資金を無駄に費やすことを避けたり，新しい実験の参考にすることができる．しかしながら，蓄積された文献数は，すでに膨大な量になっており，キーワードを入力して検索を行うだけでは，場合によっては読みきれないほど多くの文献が発見され，全ての文献に目を通すことが困難なほどである．

通常のリレーショナルデータベースでは，データマイニング技術によって，大量に蓄積されたデータから新たな規則性や知識の発見を支援する技術が確立され，多いに役立っている．同様に，大量に蓄積された文書の中から新たな規則性や知識を発見する技術としてテキストマイニングが注目を集めている．テキストマイニングは，文献を対象としており，文中から情報を抽出する技術，即ち，自然言語処理技術の寄与する部分が非常に大きいのである．

また，テキストマイニングでは，自然言語で記述された情報だけでなく，その他の情報を統合して活用することで，非常に有意義な規則性や知見の発見を行うことができる．そのためには，データマイニングの技術と組み合わせることが重要である．

1.1 データマイニング

文献 [7] によりデータベースからの相関ルールのマイニングが提案され，文献 [8] で改善されたアルゴリズム Apriori が示された．Apriori を基本とした多くの研究 [9, 10, 11] がなされている．

また，文献 [12] では，アイテムが階層構造を持つようなカテゴリー体系に属する場合の相関ルールのマイニング手法を提案している．階層を持つカテゴリー体系は多くの分野に存在している．特に，階層が細かく分類されるとリーフに相当するアイテムの件数が少なくなり，最小サポート値を超えるだけの件数が存在せずにルールとして発見できない場合でも，親の階層で集計することで有益なルールとして発見できることがある．

テキストを対象としてマイニングを行う場合，文章中から取り出した単語は，既存のオントロジーやタクソノミーに属する言葉であり，階層構造を考慮

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/ACL03/bionlp.htm
[2] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
[3] 慢性閉塞性肺疾患

したマイニングによって有益な情報を取り出すことが可能であると考えられる．

1.2 テキストマイニング

近年，グループウェアの普及，インターネットの普及などにより情報の取得，集積が容易になり，テキストで記述された文書データが多く蓄積されるようになった．大量に蓄積された文書データから知識発見を行うための技術としてテキストマイニングが注目されている [15]．

大量の文書の中から欲しい情報が簡単に見つかり，そこから知見を得ることができれば非常に有益であるが，目的の文書を探し出すことがそもそも容易ではない．通常，文書データはその有益性を判断する担当者が目を通して，キーワード等を付与し，これらを分類することで検索可能なデータベースの構築を行っている．また，キーワードを付与せずに，全文検索を行って，目的の文書を発見するという手法もあるが，表記の揺れがあった場合，発見できないなど問題も多い．

企業などでは，顧客からの問い合わせなどをコールセンターで対応し，蓄積している．コールセンターのコスト削減や将来の機能拡張，商品の不具合の早期発見などに役立てようと蓄積されたコールログを分析して，新たな規則性や知見の発見を目的として，テキストマイニングが活用されている．テキストデータから単純にキーワードを取り出してデータマイニングの手法を用いて相関ルールを求めるよりも，自然言語処理技術を用いることで，より効果的にテキストデータを分析できる技術が報告されている．筆者らは，自然言語処理技術を用いることで，テキスト中に含まれる類似内容文書を取り出す技術を開発し，コールセンターに寄せられる「よくある質問」を取り出すシステムを開発した [13]．また，コールセンター向けテキストマイニングシステム IBM TAKMI [14] では，自然言語処理技術を用いて，文書中から「何がどうした」という概念を取り出すことで，コールセンターにおける大量の文書を解析し，顧客からの問合せ内容分析の迅速化に貢献した．

筆者らは，セレスター・レキシコ・サイエンシズ（株）と共同で，バイオメディカル文献データベース MEDLINE を対象としたライフサイエンス向けテキストマイニングシステムを構築した．以下，本稿では，筆者らが構築したライフサイエンス向けテキストマイニングシステムを MedTAKMI システムと呼ぶ．

2 バイオメディカル文献データベースからの知識発見

2.1 MEDLINE

多くのバイオメディカル研究者によって MEDLINE と呼ばれるバイオメディカル文献データベースが広く使われている．MEDLINE とは，米国立医学図書館 [17] (NLM[4]) の米国立バイオテクノロジー情報センター [18] (NCBI[5]) により管理されている巨大なバイオメディカル文献アーカイブであり，1960 年代からのバイオメディカル文献が登録されており，2003 年の時点で 1,200 万件を超える文献が登録されている．

MEDLINE に蓄積されたバイオメディカル文献は，PubMed[19] と呼ばれる Web サイト上の検索システムで公開されており，誰でも自由に検索して，情報を得ることが出来る．PubMed には，タイトル，著者名，著者所属，提出日，掲載誌などのカテゴリーがあり，そのカテゴリーに属するキーワードが与えられている．論文のタイトルと概要が掲載されており，これらは自然言語による記述である．一般に，テキストマイニングシステムが，MEDLINE に対して自然言語処理の対象とするのは，これらタイトルと概要である．

2.2 バイオメディカル文献データベースからの知識発見

ライフサイエンス向けテキストマイニングシステムを構築するために自然言語処理技術が解決すべき課題として，頻出するバイオメディカル分野特有の固有表現の処理，多種の類義語の処理などが挙げられる．

また，ライフサイエンス向けテキストマイニングの機能として，NLM が MEDLINE に付与する階層付きカテゴリー MeSH Term や既存のオントロジーやタクソノミー体系の活用や物質と病気などの因果関係の発見などが，より有意義であると考えられる．

[4]United States National Library of Medicine
[5]National Center for Biotechnology Information

2.3 関連研究

PubMed の文書データから知見を得ようとする研究が近年，盛んに行われている．多くの場合，特定の物質に関する文献に絞り込むことで，小さなデータセットを用いており，特定ドメインに対する知識抽出が目的とされている．

例えば，蛋白質間の相互作用を自動的に抽出する研究が行われている [21]．文献 [21] では，事前に登録された蛋白質名辞書，及び相互作用を表す幾つかの動詞[6]を辞書に登録し，文字列のマッチングにより単語を抽出し，さらに，” 蛋白質 A －相互作用 (動詞) －蛋白質 B ” という 3 項関係を文中の区切り記号 (”,” ”.” ”;”) で区切られた断片中から抽出し集計するという方法を採用している．さらに，ショウジョウバエに関係する論文から蛋白質や細胞周期の相互作用に関する知識の発見を試みている．この方法は，ドメイン毎に辞書を変更することで，他の相互関係も調べることができる．しかしながら，自然言語処理技術を用いていないため，動詞の過去形と受動態の区別が付かなければ，本来の意味とは別の解釈による 3 項関係が抽出されている可能性があり，さらに，関係詞節を飛び越える 3 項関係を抽出することが出来ないなども問題もある．

文献 [16] では，バイオメディカル文献中で偏頭痛 (migraine) と伝搬性抑制 (spreading depression) が頻繁に共起するという事実と，伝搬性抑制 (spreading depression) とマグネシウム (magnesium) が頻繁に共起するという事実から，未知の知見として ”magnesium can inhibit *spreading depression* in the cortex, and *spreading depression* may be implicated in migraine attacks” という関係が発見できたという報告がなされている．これは，文献中のキーワードの相関を見ることで，未知の知見が得られた一例である．

バイオメディカル文献からのキーワード抽出では，未知語や複合語などの単語認識が問題となる場合が多い．まず，同義表現の問題が挙げられる．バイオメディカル文献は生物学，医学などさまざまな分野にまたがっており同一の物質に対して異表記が多数存在する．これらを自動的に発見し，統一された表現に定めることは困難である．次に，未知語の問題がある．物質名などでは，複数の単語や記号などを用いて表現することが多く，そのうちの大部分は複合語として表される．バイオメディカル文献に対して専門に作られたものでない限り，自然言語処理パーザーが自動的に単語境界を判断するのは，ほとんど不可能である．自然言語処理パーザーが処理を行う前に，複合語の特定作業をする手法も研究されている．文献 [22] では，バイオメディカル論文の中から蛋白質の名前を自動的に特定するという問題について議論している．

[6]過去形，三人称単数など語尾変化した語も含む．

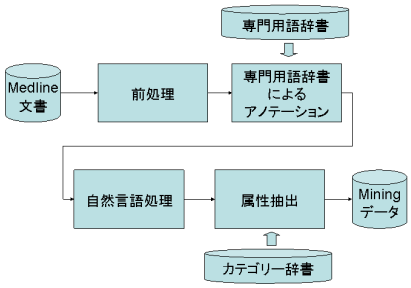

図 1: MedTAKMI Preprocess

我々が開発した MedTAKMI システムの概要については，文献 [23] にあるので，本稿においては主に，自然言語処理について記述する．

3 MedTAKMI システムにおける自然言語処理

3.1 概要

図 1 は，MedTAKMI システムの前処理について示している．MedTAKMI システムの前処理では，マイニングエンジンが計算対象とする定型データの抽出とテキストから自然言語処理によって文中のキーワード及び，2 項関係，3 項関係の抽出を行う．ここで，2 項関係，3 項関係とは,「A が B した」や「C が D に E した」というような複数の単語からなる概念を “A...B” や “C...E...D” というキーワードとして抽出したものを表す．

3.2 定型データの抽出

MEDLINE には，論文のタイトル，概要の他に著者名，掲載誌などの事前に決められた項目に関する情報が多く含まれている．これらの定型の項目を定型カテゴリーと呼ぶ．図 2 の左の大きな枠中には，MEDLINE に含まれる定型カテゴリーのデータが示されている．MedTAKMI システムの前処理では，最初に MEDLINE 文献から定型カテゴリーとその値を抽出する．また，MEDLINE に付与さ

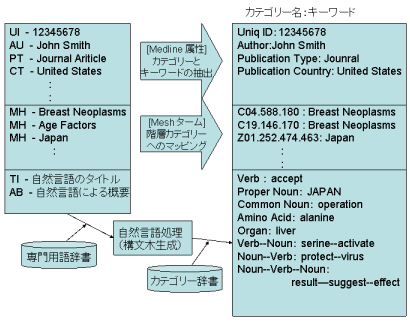

図 2: データ生成

れている MeSH Term と呼ばれる階層構造を持つカテゴリーのデータなどについても定型データとして抽出する．さらに，論文の概要とタイトルも自然言語処理の対象として取り出しておく．

テキストマイニングを行うためには，概要だけでなく文献の本文を対象にすべきであるのかもしれないが，MEDLINE から取得できるのは概要だけであり，本システムでは対象外とした．

3.3 専門用語辞書処理

MedTAKMI システムの前処理では，定型データの抽出で取り出した概要とタイトルに対して構文解析を行う．我々が自然言語処理で用いたパーザーは，新聞記事に人手によってつけられた構文解析結果から学習した統計的手法を用いたパーザー（以後，統計パーザー）であり，分野の違うバイオメディカル系の文書に対しては高い精度を得ることが困難である．従って，専門用語辞書による単語境界情報，品詞情報を事前に与えた後，統計パーザーを用いて構文木の生成を行った．事前に辞書を用いてアノテーションを行うことで，新聞記事で学習した統計パーザーでもバイオメディカル系の文章に対して言語解析の精度を向上させることができる．

3.3.1 専門用語辞書によるアノテーション

筆者らは，外部リソースなどから専門用語となりうるタームを収集し，およそ 200 万語以上のタームを含む専門用語辞書を用意し辞書によるアノテーションを行った．アノテーションとして，文書に対して辞書に含まれるターム（文字列）を探し出し，タームの境界情報，品詞情報，さらに正書形についてタグ付けを行った．

このアノテーションが解決する問題として，以下ようなものが挙げられる．

固有表現の判定

バイオメディカル文献では非常に長い単語長の固有の物質名や現象の名前が存在する．例えば，“1,2-dihydroxy-1,2-dihydroxynaphthalene dehydrogenase” といった物質名や “repetitive sequence-based polymerase chain reaction” という現象名がテキスト中に頻出する．このような文章から正しい固有表現を自動的に抽出することは，困難である．したがって，あらかじめ単語境界と品詞を文書に与えておくことにより，これを解決する．

同義語の検出

バイオメディカル文献に含まれる固有表現は複数の表記を持つものが少なくない．たとえば，物質名などは，人名から付けられた名前，機能から付けられた名前，学名など，複数の別名（同義語）を持つ場合が多く見受けられる．バイオメディカル文献データベースは複数の分野にまたがっていること，また過去の文献に対しては修正を行うことが無いことなどから様々な同義語が論文データベースに含まれてしまっている．これらの同一の単語として扱うためには，これらの同義語を 1 つの表記に統一する必要がある．例えば，“DNA” と “deoxyribonucleic acid” が別々の単語として集計されてしまと，本来同一のものが別々に集計されるという問題がある．従って，専門用語辞書には，“DNA” と “deoxyribonucleic acid” の正書形として，それぞれに “deoxyribonucleic acid” を登録しておくことで，これらを同一視する．

3.4 自然言語処理

専門用語辞書でアノテーションされた文書に対し，情報抽出に必要な言語情報を得るために構文解析を行う．バイオメディカル文献のような未知語，複合語を多く含んだ文章には頑健性が重要であるため，浅い解析が最適である．そこで，我々はマルコフモデルに基づいた統計パーザー[7]を用いて構文解析を行った．このパーザーは Penn Treebank corpus[8]の Wall Street Journal に人手で品詞をつけたものから品詞統計情報を取り，それを学習データとして構

[7]この統計パーザーは，IBM Watson 研究所で開発された．
[8]http://www.cis.upenn.edu/ treebank/

文解析を行うものである．次に, この構文解析結果を用い，主辞決定規則 [24] に基づいて依存構造木を作成した．

3.5 属性の抽出

MEDLINE の定型カテゴリーとして抽出される情報に加えて，浅い構文解析の結果から名詞，動詞，名詞－動詞の 2 項関係，名詞－動詞－名詞などの 3 項関係をキーワードとして抽出する．名詞 human を抽出する時にはカテゴリーが名詞で，その値が human として抽出する．2 項関係，3 項関係については， 係り受け関係に基づき抽出可能である．

筆者らは専門用語辞書とは別に，カテゴリー辞書を用意した．カテゴリー辞書には，特定のカテゴリーに属するタームが登録されている．例えば，“serine”, “tyrosine”, “alanine” などをアミノ酸カテゴリーのタームとして登録し， “brain”, “liver”, “lung” などを臓器カテゴリーのタームとして登録しておく．構文木から取り出したキーワードは，事前に登録したカテゴリー辞書と照合し，特定のカテゴリーに属するタームであれば，これを抽出する．

登録するキーワードは，専門用語辞書に登録された正書形であり，“DNA” と “deoxyribonucleic acid” は，同じカテゴリーに属するキーワードとして，その正書形 “deoxyribonucleic acid” に変換されて抽出される．

3.6 3 項関係の抽出

従来のテキスト分析システムでは，一般的に，単語ごとに頻度または重要度を調べ，頻度順に表示したり単語間の相関 (例えば，単語 “蛋白質” と “脂質” は非常に相関が高い，など) をグラフ化することが多い．単語単位の集計を行うことで，データセット全体の頻出単語の傾向を把握することが出来る．しかしながら，これは，実際のテキストにどのようなことが書かれているのか，を把握するための役には立たない．

例えば,『単語 “smoking” が頻出している』という集計結果からでは，“smoking” が何に対してどのような影響を及ぼすかを知ることは出来ない．次に，“smoking” を含む文書集合に絞込み，その中で，共起するキーワードを調べることで，“smoking” に関係ありそうなキーワードを知ることが出来る．例えば， “smoking” と “risk” というキーワードの共起が仮に判明したとしても，実際に文章中では，同じ文にあるかどうかわからないし，文書を読まないと，“smoking” と “risk” の関係が何かまではわからない．また，従来手法のキーワードを一つのアイテムとして集計を行うテキストマイニングシステムでは，受身形であるか，否定形か，などを考慮することができない．例えば，*“Smoking increase risk of lang cancer.”* という文章と *“Risk of lang cancer has increase smoke”* という文章は同じ単語を含むため，これらの文章を区別することが出来ない．

コールセンターを対象とするテキストマイニングシステムでは，文章中から「何がどうした」という情報を取り出すだけで，大まかな問題を把握することができたのに対し，バイオメディカル文献を対象とする場合には,「何がどうした」よりも「何が何にどう作用したか」を取り出すことがより重要である．2.3 節の関連研究でも 3 項関係の抽出が対象となっている．

例えば，“物質 A が物質 B に作用 C する” という関係を取り出すことで，“物質 A が作用 C する” や “物質 B に作用 C する” という関係よりもより詳細な情報の抽出を行っていることは明らかであろう．また，そのためには，“物質 A が物質 B に作用 C する” と “物質 B が物質 A に作用 C する” という関係を明確に区別する必要もある．そこで，単語だけでなく，文章内で係り受けを持つ 2 項・3 項関係 (例 “smoking ... increase ... risk”) をアイテムとして抽出を行った．

表 1 は，MEDLINE データベース中から抽出した高頻度の “名詞－動詞－名詞” という 3 項関係の例である．英語の（主語，述語，目的語）の関係の取得を目的として取得したものであるが,（主語を修飾する単語，主語，述語）や（述語，目的語，補語）の取得を目的とした “名詞－名詞－動詞” “動詞－名詞－名詞” という 3 項関係も同時に別のカテゴリーとして取得した．この表では stop words を取り除いた後の全ての名詞と動詞からなる 3 項関係を取り出しているが，特定のカテゴリーに属する名詞と特定の動詞（例えば，作用を表す動詞）の組合せに限定することで，より有益な情報を取り出すことが可能である．

4 データ分析例

筆者らは，実際の MEDLINE のバイオメディカル論文を MedTAKMI システムに適用した．本章

表 1: 3 項関係の例

reaction ... see ... text
result ... suggest ... role
study ... investigate ... effect
result ... suggest ... function
result ... suggest ... effect
study ... provide ... evidence
study ... provide ... effect

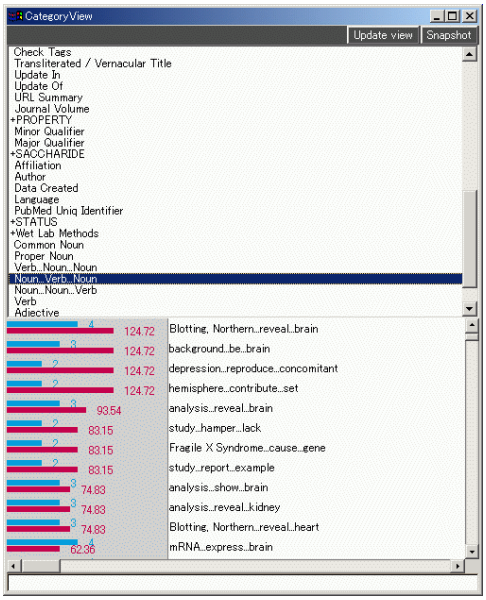

図 3: brain を含む文献に含まれる 3 項関係

では，自然言語処理の関係する部分を中心に，その例を示す．

本稿での分析で用いているデータは，2001 年 11 月から 2002 年 6 月までのデータを PubMed から取得して，作成した約 7 万件のデータである．

図 3 には，7 万件の中から brain という語を含む 557 件のデータセットに対して，“Noun...Verb...Noun” という 3 項関係について集計した結果である．

一つのキーワードに対して，二つの棒グラフが表示されている．上の棒グラフは，条件 (brain) で絞り込んだデータ (557 件) 中での出現数が表示されており，“Blotting, Nothern...reveal...brain” が 4 文献中に出現していたことがわかる．その下の棒グラフは，相対頻度と呼ばれる指標で，557 件中での出現割合（4/557）を全件約 7 万件中での出現割合で除算した数値が示されている．図 3 は，相対頻度の高い順にソートされた結果であり，上位ほど brain という条件として与えたキーワードと相関が強いことを意味する．

図中下部の 3 項関係のリストの中で，最上位の “Blotting, Nothern...reveal...brain” は全部で 4 件存在した．これらの 4 件中で，この 3 項関係を抽出した元の文は次のようなものであった．

- Northern blotting reveals ubiquitous distribution of Nrdp1 in human adult tissues, but message is particularly prominent in heart, brain, and skeletal muscle.
- Northern blot analysis revealed that magphinins are expressed in brain, ovary, testis, and epididymis
- Northern blot analysis revealed that hCERK mRNA expression was high in the brain, heart, skeletal muscle, kidney and liver.
- Northern blot analysis and RT-PCR revealed that rat calcyon mRNA was expressed only in the brain.

この 4 文中に含まれる “Northern blot analysis” と “Northern blotting” はそれぞれ異表記のタームであるが，専門用語辞書により正書形として “Blotting, Nothern”[9]が登録されているため，正書形で同一視された結果である．

第 3 項の箇所に現れる “brain” については，直接，第 2 項の “reveal” と係り受けを持つものではないので，構文木生成についてはまだ，改良の余地が残されている．

筆者らは医学の専門家ではないので，これらの文を含む 4 件の文献から得られる知識の有用性について説明できないが，MedTAKMI システムでは対話的に操作することにより，このような情報を適宜取り出して，MEDLINE 文献からの知識発見を支援することができる．

5 おわりに

本稿では，大量のバイオメディカル文献データベースを対象としたライフサイエンス向けテキストマイニングシステム MedTAKMI システムについて述べた．特に，バイオメディカル分野のテキストを対象とした自然言語処理技術の問題について触れ

[9](RNA) ノーザンブロット法 (発現された塩基配列を分析する実験方法の一つ)

た．今後，自然言語処理技術，マイニング技術に対して更なる改良を行って，MEDLINE 以外の外部文書データベースとの統合や，テキストマイニング技術と他のバイオメディカル系データベースとの統合などを目的とした情報統合技術の研究を行っていく予定である．

謝辞

MedTAKMI システムの共同開発先であるセレスター・レキシコ・サイエンシズ株式会社に感謝します．また，TAKMI システムの開発に貢献された那須川哲哉氏，多くの助言，協力を頂いた堤泰治郎氏，本プロジェクトに貢献された竹内広宣氏，坪井祐太氏に感謝します．

参考文献

[1] Akihiro Inokuchi, Takashi Washio, Hiroshi Motoda, “An Apriori-based Alogorithm for Mining Frequent Substructures from Graph Data”, In proceeding of the 4th European Conference on Princeples and Practice of Knowledge Discovery in Databases, pp. 13-23, 2000.

[2] Rigoutsos, I., Floratos, A. “Combinatorial pattern discovery in biological sequences: The Teiresias algorithm.” Bioinformatics vol.14, No.1, pp. 55-67, 1998.

[3] http://www.ec-inc.co.jp/ACL2003/

[4] http://www.sigir2003.org/

[5] http://trec.nist.gov/

[6] http://www.nlm.nih.gov/research/umls/

[7] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, “Mining association rules between sets of items in large databases”, Proceedings of the ACM SIGMOD Conference on Management of Data, Wasington D. C., pp.207-216, 1993.

[8] Rakesh Agrawal, Ramakrishnan Srikant, “Fast Algorithms for Mining Association Rules”, Proceedings of International Conference on Very Large Data Bases, pp. 487-499, 1994.

[9] Jong Soo Park, Ming-Syan Chen, Philip S. Yu, “An Effective Hash-Based Algorithm for Mining Association Rules”, Proceedings of the ACM SIGMOD Conference on Management of Data, San Jose, California, pp.175-186, 1995.

[10] Sergey Brin, R. Motowani, Jeffrey Ullman, S. Tsur, “Dynamic Itemset Counting and Implication Rules for Market Basket Data”, Proceedings of the ACM SIGMOD Conference on Management of Data, Tucson, Arizona, pp.255-264, 1997.

[11] Roberto J. Bayardo Jr., “Efficiently Mining Long Patterns from Databases” Proceedings of the ACM SIGMOD Conference on Management of Data, Seattle, Washington, 1998.

[12] Ramakrishnan Srikant, Rakesh Agrawal, “Mining Generalized Association Rules”, Proceedings of the 21st International Conference on Very Large Data Bases, pp. 407-419, 1995.

[13] 松澤裕史, “自然言語処理技術と構造化パターンマイニングを用いた FAQ 作成支援システム”, 情報処理学会 FIT2002 論文集, 情報技術レターズ, Vol.1, pp. 69-70, 2002.

[14] Tetsuya Nasukawa, Tohru Nagano, “Text analysis and knowledge mining system”, IBM Systems Journal, Vol. 40, No. 4, pp. 967-984, 2001.

[15] Marti A. Hearst, “Untangling text data mining”, Proceedings of Association for Computational Linguistics, pp. 3-10, 1999.

[16] D. R. Swanson, N.R. Smalheiser, “An interactive system for finding complementary literatures: A stimulus to scientific discovery”, Artificial Intelligence, 91(2):183-203, 1997.

[17] NLM
http://www.nlm.nih.gov/

[18] NCBI
http://www.ncbi.nlm.nih.gov/

[19] PubMed
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[20] MeSH (Medical Subject Headings)
http://www.nlm.nih.gov/mesh/meshhome.html

[21] Chiristian Blaschke, Miguel A. Andrade, Christos Ouzounis, Alfonso Valencia, “Automatic extraction of biological information from scientific text: protein–protein interactions”, Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, Heidelberg Germany, pp. 60-67, 1999.

[22] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, “Toward information extraction: identifying protein names from biological papers”, Pacific Symposium on Biocomputing vol. 3, pp.705-716, 1998.

[23] 松澤 裕史, 長野 徹, 村上 明子, 竹内 広宜, 武田 浩一, 神田 靖, “バイオメディカル文献データベースを対象とするテキストマイニングシステム MedTAKMI”, 第 3 回データマイニングワークショップ予稿集, 日本ソフトウェア科学会データマイニング研究会, 2002.

[24] Michael Collins, ”Head-Driven Statistical Models for Natural Language Parsing”. PhD Dissertation, University of Pennsylvania, 1999.