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

2003－DBS－130 (4)
2003－ＦＩ－ 71 (4)
2003／5／22

ライフサイエンス分野における テキストマイニング技術適用の動向

浦本 直彦[†‡] 松澤 裕史[†] 猪口 明博[†] 武田 浩一[†]

テキストマイニング (Text Mining) は，大量の構造化されていないテキスト情報を様々な観点から分析することにより，価値のある情報やパターンを発見するための技術であり，自然言語処理，情報検索，情報抽出，データマイニングなどの要素技術を組み合わせたものである．近年，テキストマイニング技術をライフサイエンス (バイオや医療分野を含む) 文献に適用し遺伝子やタンパク質の関係に関する新たな知見を得ようとする研究が注目されている．本論文では，テキストマイニング，特に情報抽出技術が，ライフサイエンス分野においてどのように用いられているかを，この分野の主要な国際会議やジャーナルで発表された論文を中心にサーベイする．

Recent Trends in Text Mining Techniques for Life Sciences

Naohiko URAMOTO[†‡] Hirofumi MATSUZAWA[†]
Akihiro INOKUCHI[†] Kohichi TAKEDA[†]

This paper surveys recent trends in promising text mining technologies for discovering new knowledge from unstructured text in the life sciences area. Recently, text mining techniques have bee applied to life sciences literature collections such as MEDLINE, a large repository of abstracts in this area. This paper introduces various methods proposed by many papers appearing at major conferences (e.g., PSB, ISMB, and GIW) and in major journals (e.g., Bioinformatics).

[†]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

テキストマイニング (Text Mining) は，大量の構造化されていないテキスト情報を様々な観点から分析することにより，価値のある情報やパターンを発見するための技術であり，自然言語処理，情報検索，情報抽出，データマイニングなどの要素技術を組み合わせたものである [15, 39, 40]．コールセンターの通話記録や特許文書などに適用されてきたテキストマイニング技術が，近年，ライフサイエンス分野での適用が注目を集めている．テキストマイニング分野のキーパーソンの一人であるカリフォルニア大の Marti Hearst は 1998 年の ACL 招待論文の中で”text data mining” としてテキストマイニングについて解説している [15]．この初期の論文のなかで，すでに医療文献を用いて既知の遺伝子に関連するキーワード情報を元に，未知の遺伝子の機能を予測するシステムが紹介されている．これはライフサイエンス分野が，テキストマイニングの有望なターゲットであることは早くから自然言語処理研究者たちにも認識があったことを示している．理由を考えてみよう．

(1) 大量のテキスト情報が入手できる．例えば，米国 National Library of Medicine (NLM) が提供する Pubmed データベースは，1960 年代からのライフサイエンス分野の文献抄録を検索可能にしており，エントリ数は，1200 万件にも及ぶ (最近では一年に約 50 万件が新規に登録されている)．

(2) テキスト情報量の増加のスピードは，遺伝子の配列情報の公開スピードより早いことが報告されており [31]，テキストマイニング技術を使って有用な情報を自動的に取り出すための仕組みが重要視されている．

(3) テキスト情報を処理するためには，分野依存の辞書やオントロジが必要不可欠であるが，NLM をはじめとする様々な研究機関が，様々なリソースを提供している．

(4) 遺伝子やタンパク質に代表されるライフサイエンス分野の研究において，もっとも信頼できるのは実験結果であるが，結果が何を意味するかを判断するのが困難な場合がある．たとえば，マイクロアレイや DNA チップを用いた遺伝子発現の実験結果として得られるのは，遺伝子名の集合 (クラスタ) である．研究者は，そのクラスタに含まれる遺伝子間の関係は何かを見出す必要があるが，数値データだけではなく MEDLINE のようなテキストデータベースから得られる情報を組み合わせることで，より直感的な理解が可能になる．

本論文では，テキストマイニング技術が，ライフサイエンス分野においてどのように用いられているかをこの分野の主要な国際会議やジャーナル (PSB: Pacific Symposium on Biocomputing, ISMB: International conference on Intelligent Systems for Molecular Biology, GIW: Genome Informatics Workshop, Bioinformatics など) の論文を元に概観する．特に，MEDLINE に代表されるテキストデータベースからの情報抽出に関する技術を取り上げる．なお，PSB2001 では辻井・Ananiadou による情報抽出のチュートリアルがあったが，その時の発表資料およびアノテーションつき参考文献リストが入手可能である [44]．

情報抽出の技術を用いて，MEDLINE に代表される文献 (抄録) 集合から，必要な情報を抽出し，関連付ける研究が数多くなされている．具体的には以下のタスクに分別される．

- 単語および固有表現 (named entity) の抽出: 遺伝子，タンパク質，疾病，化合物名の抽出
- 省略語解析: 省略形とその元表現の抽出
- 関係抽出: 遺伝子，タンパク質，疾病，化合物間の関係の抽出
- パスウェイ構築と: 関係によるネットワーク構造の抽出と可視化

次節以降，それぞれのタスクの代表的なアプローチについて述べる．

2 単語・固有表現抽出

ライフサイエンス分野の文献は，他の自然科学の文献と同じように，分野に依存した単語 (遺伝子，タンパク質，化合物名など) とその同義語が頻出する．近年，遺伝子名については，その語彙が同定されつつあるが，化合物名など多くの未知語が存在する．数字や非アルファベット文字 (“-” や “()” など) が含まれるのが特徴のひとつである．複合語も頻出するので，いかに正しく単語を認識し，さらに，文献からタンパク質や遺伝子といったある概念クラスに属す語を取り出すことが，重要な研究テーマのテーマの１つとなっている [11, 21, 5, 14]．

Fukuda[11] らは，単語の表層的な特徴を手がかりにタンパク質などの物質名 (material word) の認識を行うプログラム PROSPER を提案している．処理の手順は以下の通りである．

1. 単語分割された文から，中心となる単語 (core term) と core term に関する機能を表す単語 f-term (例. “EGP recepter” における “receptor”) を同定する．ここで，core term は，大文字，数字，非アルファベットを含む単単語 (例. “Nef”, “p53”) あるいは複合語 (例. “interleukin 1 (IL-1)-response kinase”)，小文字

の英字だけを含む単単語から選ばれる．次に選ばれた単語に単純な規則を適用し候補を絞る．例えば，特殊文字が単語を構成する文字の半分以上を占める場合，core term としないという規則を適用することで，単語 “+/-” が core term として選ばれないようにする．

2. 隣接する core term を結合する．ある core term の前方および後方にある core term を，あらかじめ規定したルールに従って検査する．この段階で，単語連続や括弧を含む表現などが，より大きな一つの core term として認識される．各単語は Brill の品詞タガーを使って，品詞が付与される．
3. さらに，“A, B, C, and D” などのより大きな構造を含む core term を構築する．

MEDLINE からとった計 80 文を用いた実験の結果，正解率 94.70%, 適合率 98.84% で物質名を認識できたことが報告されている．

Krauthammer らは，文献中に含まれる遺伝子とタンパク質名を同定するのに，DNA 塩基配列を比較し，類似性 (相同性) を計算する BLAST を用いている [21]．辞書から得られた遺伝子名は，各文字ごとにあらかじめ定義された DNA 配列の組み合わせに変換され，BLAST データベースに保管される．これと対象となる文献を同じ文字-ヌレオチド列に変換したものを BLAST を用いて比較を行い，対応が取れた遺伝子，タンパク質名が文中に出現したとみなしている．また，Collier らは，隠れマルコフモデル (HMM) を用いて遺伝子と遺伝子産物名を抽出する手法を提案している [5].

単なる単語の同定ではなく，単語の意味の曖昧性を機械学習を用いて解消する手法を提案しているのが，Hatzivassiloglou らの研究 [14] である．蛋白質と遺伝子の両方の意味を持つ単語 (例. SBP2) があるときに，Naive Bayes, C4.5, PIPPER のような機械学習アルゴリズムを用いて，単語にラベル付けを行う．

3 省略語解析

ライフサイエンス分野の文献は，“gamma-aminobutytic acid (GAVA)“, Gcn5-related N-acetyltransferase (GNAT) といった多数の省略語を含む．省略語と省略前の語の組を抽出を識別することは，辞書構築や複合語の処理のための知識源として重要であり，いくつかの研究が行われている [47, 27, 34, 12].

Schwartz ら [34] は，Pustejovsky らの省略形解析問題 [30] に対して，省略形のもつ特徴的な制約 (例えば，省略形の最初の文字は元の表現の頭文字に対応しやすい) を文字列照合のアルゴリズムとして実装し，「省略形（元の表現)」および「元の表現 (省略形)」という 2 種類の対表現を高精度で抽出することに成功した．ランダムに抽出した MEDLINE 1000 文書に対し，96%の正解率と 82% の適合率を得た．

Liu らは，テキストから略語，同意語辞書の自動構築する手法を提案している [12]．括弧で囲まれた略語とその前に書かれた数単語を抽出して，その対応が尤もらしいペアを略語と正式な表記のペアとして出力する．

4 関係抽出

“遺伝子 A が病気 B を抑制する” といった関係を高い精度で抽出できれば，遺伝子間の関係解析や，未知の遺伝子の機能予測などに応用することができる．医用文献からの関係抽出（主に名詞-名詞，名詞-動詞-名詞など) は，非常にホットな研究テーマの一つである [1, 2, 41, 16, 35, 45, 24, 7, 3, 43, 8, 21, 18, 33, 38, 17, 18, 36, 29, 32, 28, 46, 26, 10, 31, 23, 4, 19].

文献中には，多くの未知語 (辞書に記載されていない語) が存在するし，ある関係を表現する言語表現も多彩である．たとえば，「A が B を抑制 (inhibit) している」という関係は，文中には以下に示した様々な表現で出現する．¡small¿

```
A inhibits B (動詞)
A is inhibited by B (動詞)
inhibition of B by A (名詞)
A inhibitor of B (名詞)
A does not inhibit B (モダリティ)
A inhibits it (it は B を示す代名詞)
```

¡/small¿ これらの表現をうまく扱うためには，パターンマッチング，形態素解析， 構文解析，照応解析などの技術が必要となる．以下は，特徴的な手法を順不同に並べたものである．

- 文の表層的な手がかりを用いるもの [3]
- 部分的な構文解析 (shallow parsing) を行うもの ([35, 32] など多数)
- 文全体に対する構文解析 (full parsing) を行うもの [46, 10, 28, 19]
- テンプレートを用いるもの [43, 18, 17, 22, 13]
- 機械学習を用いるもの [7, 23, 14, 37]
- UMLS などのシソーラスを用いるもの [32, 13]
- 抽出した遺伝子，タンパク質間の関係から反応，伝達，代謝などに関するネットワーク構造 (パスウェイ) を構築するもの [24, 18, 8, 28]

以下，代表的な手法を説明する．

Sekimizu ら [35] は，MEDLINE 文書から，タンパク質の結合 (protein binding) に関する関係を抽出

する手法を提案している．抽出対象の関係は，“activate”, “interact”, “bind” などの文中に頻出する動詞と主語，目的語の組である．対象となる 34,343 文に対し，「浅い」構文解析器 (シャローパーザ，shallow parser) である EngCC を使って構文解析を行い名詞句を認識する (noun phase braketing)．注目する動詞と名詞句中のヘッドとなる名詞と組み合わせるために，(1) 動詞を同定する (2) 動詞の前方にある名詞句のヘッドを主語とする (動詞が受身形の場合は目的語) (3) 動詞の後方にある名詞を目的語とする，という手順で主語-動詞-目的語の組を抽出する．実験の結果，72.9%の正解率で，正しく主語と目的語の組を抽出できたと報告している．

Blaschke らは，文の表層的な手がかりをルールとして用い，タンパク質間の相互作用 (protein-protein interaction) を抽出する実験を行っている [3]．処理の手順は以下の通り．

- タンパク質をひとつ選び，人手で関連するアブストラクトを収集する．
- 反応を表現する 14 個の動詞 (例．“acetylate”, “activate”, “bind” など．活用形，名詞形も含む) を定める．
- アブストラクト集合からターゲットとなるタンパク質と動詞を含む文を見つける．
- 表層的な手がかりを元に (“.”/“,”/“;”)，文をさらに断片に分割した後で，2 つのタンパク質名と上にあげた動詞を含む断片を抽出する．

ある頻度以上で出現する関係を元に，反応のネットワークを構築する．論文では，構築されたネットワークが本当に意味があるかの検討を行っている．

Craven らは，機械学習を用いて MEDLINE からタンパク質，組織，疾病，細胞下構造間の関係を抽出する 2 つの手法について述べている [7]．例えば，subcelluar-localization(Protein, Subcellular-Structure) という関係 (あるタンパク質が細胞内の構造のどこかで局在化する) を学習したいとする．タンパク質と，細胞内構造カテゴリに属する 2 つの単語を含む文集合を用意し，この文に含まれる関係が subcelluar-localization を満たすか否かを判定した正解および不正解集合を作成する (タンパク質や細胞内構造に関する辞書が存在することを仮定している)．これを用いて，タンパク質と，細胞内構造カテゴリに属する 2 つの単語を含む文が，正解，不正解のどちらに分類されるかを，Naive Bayes を使って判別する．もう一つの手法では，シャローパーザの一種である Sundance を用いて，文を解析し句構造 (名詞句，動詞句，前置詞句など) に分解する．さらに，subcelluar-localization 関係を正しく表現している訓練文を用意し，句間の隣接関係，包含関係，係り受け関係などの組み合わせ (規則) を記録する．これらの正解となる組み合わせ集合から，subcelluar-localization 関係を抽出する規則を学習する．例えば，“ある文が subcelluar-localization 関係を表現するならば，protein を含む句が，句 A の左にあり，句 A の右には Location を示す句が来る ” といった規則 (規則) を学習する．

NLM の Unified Medical Language System (UMLS) メタシソーラス (metathesaurus) や Gene Ontology(GO) などの分野依存の情報をドメイン知識として解析の精度を上げる試みも多くなされている．

Rindflesh らの研究 [32] の特徴は，UMLS Metathesaurus をドメイン知識として，関係 (bind(X, Y)) を抽出する点にある．彼らが開発した ARBITER システムは，Xerox タガーと UMLS の語彙をベースとする構文解析器 SPECIALIST を用いて解析された単語に対し，結合 (bind) され得る単語の意味クラスに属す (例えば，アミノ酸は結合され得る，言い換えると，動詞”bind” の目的語となり得る) かどうかを判別することで，より正確に関係を抽出することが可能である．

Hahn ら [13] らは，MEDSYNDIKATE と呼ばれるテキストマイニング・システムを開発中で，従来の情報抽出システム以上に深い知識を獲得するために，照応解析を含む (依存文法に基づく) 構文解析と，テンプレート型の情報抽出および UMLS などのシソーラスを利用している．最近では，UMLS のオントロジー部分に記述論理 (description logic) による推論を適用し，これに手作業による修正を加えて，is-a/part-of の関係を含んだ知識ベースの構築を行っている．

Pustejovsky ら [31] は，語彙意味論 (lexical semantics) をベースに構成された意味オートマトン (semantic automata) を用いて，より柔軟に関係を抽出する手法について述べている．中心となる処理は，以下の通りである．

1. Brill のタガーを用いて品詞付けを行う．
2. 意味オートマトンを用いた浅い構文解析．以下の 5 つの処理に対応するオートマトンをそれぞれ適用する．

   Level I: 名詞のチャンキング．固有名詞，一般名詞のグルーピング．2 重前置詞, 複合的な関係語のグルーピング．

   Level II: 名詞句のチャンキング (前置詞句は含まず)

   Level III: 並列句（名詞，動詞) のチャンキング

   Level IV: of でつながる前置詞句を含む名詞句のチャンキング

Level V: 従属節の識別

3. 解析された構文構造を用いて関係を抽出する．

このように，より精密な文法規則 (オートマトン) を用いることで，多様な関係表現に対応することが可能であり，MEDLINE からとった 59 文献 95 文中に含まれる “inhibit”(名詞形も含む) に対し，正解率 90.4%，適合率 58.9%という従来研究を上回る結果を得ることができた．

4.1 テンプレートの使用

情報抽出規則を定義したテンプレートに基づいてタンパク質間の相互作用 (protein-protein interaction) に関する関係を抽出するのが，Thomas らのシステムである [43]．この研究では，新聞記事からの情報抽出を行うツールであった Highlight システム[2]を，ライフサイエンス文献にカスタマイズして用いている．動詞 “interact”, “associate”, “bind” に注目し，構文的な情報をメインとするテンプレートを適用することで，それらの動詞と (係り受け関係がある) タンパク質名の関係を抽出している．

また，Humphreys らは，酵素の相互作用とタンパク質に関する情報抽出プロジェクトである EMPathIE(Enzyme Metabolic Pathway IE) と PASTA (Protein Active Site Template Acquisition) を通じて，テンプレートを用いた情報抽出の手法を提案している [18, 17] ．EMPathIE は代謝パスウェイのための酵素反応に関する情報抽出で，20,000 以上のレコードを持つ酵素反応データベース EMP が訓練データ，評価データとして使われている．PASTA はタンパク質のアミノ酸，構造的特徴，アクティブサイト，相互作用などに関する情報を抽出することができる．Thomas らの手法と比較して，より意味的な情報を考慮したテンプレートを用いている．

4.2 フルパーザの使用

上で挙げた論文の多くは，比較的軽い自然言語処理を用いている．例えば，文の構文解析 (parsing) においては，文全体ではなく，単語，名詞句，動詞句レベルの認識を行う解析 (shallow parsing) が多い．これは，対象とする文章が専門用語や未知語，複合語を多数含んでいるため，文全体に対する解析 (full parsing) を行うと全体的に精度が落ちてしまうことが予想されるからである．

しかし，“activate”, “inhibit” のような限られた動詞とそれらを含むパターンを構築するだけでは，十分な知識が得られない可能性がある．また前述したように，“inhibit” は “inhibited”, “inhibitor”, “inhibition” など様々な形式で出現するので，動詞ごとに対応する細かなパターンを記述するのは困難である [46]．そのため，できるだけ精度を落とすことなく積極的に深い解析を使いながらより深い知識を得るための研究がいくつか行われている．

Friedman らは，NLP の技術を積極的に適用したパスウェイ構築システム GENIES を提案している．GENIES では，MEDLINE のアブストラクトではなく，論文全体を処理対象とする．また，文法的および意味意味情報を元にするフルパーザを用いて文を解析し[3]，より深いレベルでの関係抽出を目指している．例えば，“Raf-1 activates Mek-1” という文から，

```
[action, activate, [protain, Raf-1],
                   [protein, Mek-1]]
```

というフレームを用いた意味表現を出力することができる．このフレーム表現は入れ子を許すので，より複雑な表現も記述することができる．125 個の動詞を 14 個のグループに分けて，前述の意味表現を抽出し，分子間のネットワーク (パスウェイ) を構築し表示することができる．実験では，7,790 個の単語からなる文献から，2 項の関係を 96%の正解率で取り出すことができた．

Yakushiji らは，full parser を使うことの長所 (前述) と短所 (処理速度，曖昧性 (多数の候補を出力してしまうこと)，カバレッジ) を論じつつ，HPSG ベースの full parser(XHPSG) を用いて MEDLINE 文書を構文解析し，argment structure と呼ばれる関係表現とそれから導かれるフレームに基づく意味表現を抽出する手法を提案している [46]．full parser の欠点を補うために，複合語をまとめ上げる term recognizer と，局所的な制約を用いて解析結果の候補の数を除去する shallow parser を，解析の前処理ツールとして用いている．MEDLINE から取った 97 文から抽出できた 133 個の argment structure のうち，31 個 (23%) が一意に，32 個 (24%) が複数の構造と共に (曖昧性あり) 正解であり，残りの 70 個 (53%) が正しく関係構造を抽出できなかったと報告している．

Park らも，同様の問題意識のもと，文脈自由文法より生成力が大きい Combinatory Categorial Grammar (CCG) をベースにした構文解析を用いて MEDLINE 文書を構文解析し，自動的にパスウェイを構築する手法について述べている [28]．

Hosaka らは，生物学医学分野の専門用語辞書とフルパーザを用いて，人手で作成した抽出規則を元に特定の動詞に対する動作主・非動作主・動詞句の

[2]情報抽出のコンテストである Message Understanding Conference (MUC) のために開発された．

[3]解析精度を上げるために，フルパーザで扱えない部分をシャローパーザで補う仕組みを用いている

抽出を行っている [19]．抽出対象の３つの中で動作主の抽出が一番困難であり，これは専門用語辞書を用いることにより精度が向上した．しかし残りの動詞句・非動作主の抽出に関しては専門用語辞書を用いると反って精度が下がっている．これは専門用語辞書のサイズが単語認識率に十分貢献できるほどのサイズではないこと，単語境界や品詞が正しく与えられても異なるドメインで学習を行ったフルパーザを用いることにより句の組み立てに失敗していることなどが原因ではないかと報告している．

4.3 パスウェイ構築と可視化

遺伝子やタンパク質間の関係を抽出し，可視化を行う研究も多数なされている．遺伝子やタンパク質間を含む物質間の伝達，反応，代謝などの関係をネットワーク化したものはパスウェイ (pathway) と呼ばれている．実際の使用に耐えるパスウェイを文献情報から自動的に構築するのはほぼ不可能であるが，以下に示すようないくつかのアプローチがある．

Stapley らは，2 つの遺伝子の類似性を，文献内での共起の度合いに置き換えて計算し，グラフとして可視化する手法を提案している [36]．類似度の計算には相互情報量を用いる．計算結果を基に，遺伝子-遺伝子行列を計算し，遺伝子をノード，類似度をノードの長さに対応させグラフを構築し，可視化を行っている．

Ng らは，MEDLINE 文書からタンパク質名と，“inhibit”や “activate”などの動詞を含むパターンを使って，タンパク質間の関係を抽出し，グラフとして可視化する [24]．主要なコンポーネントは，検索を行う BioKleisli，情報抽出を行う BioNLP, 可視化を行う BioJAKE からなる．関係抽出で用いるのは，表層的な情報を用いたパターンであり (例 ” タンパク質 関係語 of タンパク質”)．構文解析などの処理は行われない．

Dickerson らは，MEDLINE から代謝ネットワークに関する関係抽出結果とマイクロアレイデータの検索結果を，代謝ネットワークとして可視化するためのツールキット Gene Expression Toolkit (GET) を提案している [8]．処理は，(1) パスウェイに関する分子名を入力，(2) 同義語辞書を使って MEDLINE を検索，(3) 検索結果から文章ととりだす，という操作を，異なる分子名に対して繰り返すことで行う．マイクロアレイデータの検索は各タイムスタンプにおける変化を指定した検索などが可能である．

5 その他

その他，MEDLINE 文書を用いたテキストマイニング技術に関連する論文をいくつか挙げる．

Marcotte らは，蛋白質の相互作用に関して書かれている MEDLINE アブストラクトを，Poisson 分布とベイズ学習法を用いて取り出す手法を提案している [23]．関連文書の抽出は，文書を区別するのに重要な単語を計算することで行われる．全文書での全単語数 N，ある単語 w の頻度 f とすると，訓練データで単語 w が n 回出現する確率は $P(n|N,F) = e^{-Nf}\frac{(Nf)^n}{n!}$ である．この値の対数値が小さい単語ほど，重要な単語であると定義することができる．

Ding[9] らは，情報抽出を行うテキスト情報の単位 (抄録全体，隣接文の対，単文，および句) とその特徴について論じている．人手で解析された約 300 件の MEDLINE 文書に対し，タイトルを含む抄録全体に現れる 2 種類の対象の相互作用を記述した文書件数，与えられた単位のテキスト情報に現れる相互作用の件数，与えらたた単位に 2 種類の対象が共起する件数を計算した．一般に大きな単位ほど相互作用のカバレージ (recall) が高く，小さな単位ほど精度 (precision) が高くなるが，総合的にみて，単文を単位とすることがが効率的であると論じている．

東大辻井研究室を中心に行われている GENIA (Genome Information Acquisition) プロジェクト [6, 25] では，タンパク質や遺伝子といった意味的なクラスの固有表現を抽出するために，Swissprot 等から得られた単語リストから統計的学習による単語分類や決定木を用いた手法を開発している．さらにオントロジー構築・管理のためのシステムや，シソーラスの自動構築手法が研究されている [42]．同プロジェクトのホームページ (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/) からは，研究用のコーパスがダウンロード可能である．

情報検索やマイクロアレイによる遺伝子クラスタリングへのテキストマイニング技術の応用については，以下のようなシステムが知られている．

Tanabe ら [41] の MedMiner は，MEDLINE の検索結果に対するフィルタリングシステムとして有名である．ユーザは興味のある遺伝子名（あるいはその遺伝子名や遺伝子名病名の対) についての検索条件を指定し，MedMiner はこれを GeneCard という遺伝子情報 DB への検索式として処理し，特定のマイクロアレイに関連する遺伝子名や同義語に展開する．ユーザはこの結果を編集可能であり，これが PubMed への検索式に変換され MEDLINE 文献集合を得ることができる．最後に得られた文献集合に MedMiner で事前に定義された関連度フィルターを適用し，関連情報を含む文が出現する文献が表示される．MedMiner は PubMed と同様に，米国 NIH(National Institutes of Health) のサ

イト (http://discover.nci.nih.gov/textmining/) で利用可能である．

Kankar ら [20] は，MedMeSH Summarizer という，マイクロアレイ出力から得られた遺伝子クラスタに対して，MEDLINE 文献情報から得られた MeSH タームを対応づけるシステムを開発した．各クラスタに含まれる遺伝子名をもとに対応する文献集合を求め，各クラスタに固有の MeSH タームの集合を統計的に計算し，これをサマリー情報として表示する．

6 技術動向と今後の展開

前節まで，特にライフサイエンス文献に対して適用されているテキストマイニング技術，特に情報抽出の技法を概観した．ここまでの議論からもわかるように，自然言語処理，情報検索の分野で研究が進められてきた情報抽出の代表的な手法はほぼ試みられ，それなりの成果が上がっている．これらの結果を元に，今後は，より分析的な手法が提案され，文中に直接現れないような知識を得られるようになるだろう．しかし，いくつかの問題もある．例えば，マイニングによって獲得したい知識の代表的なものは，遺伝子やタンパク質の機能および反応関係であると考えられるが，本論文で紹介した手法を使うことで，使用者が知らなかった知識を得られるかは不明である (多くの関係はすでに知られているものであろう．これはテキストおよびデータマイニングに共通の課題でもある)．また，ライフサイエンス分野が，従来の分野と何が違うのか，どのような新しい手法が提案され，マイニングコミュニティに還元されていくのかは，今後も注目すべき問題である．

同分野でのこれまでの研究発表は，アカデミックな応用，特に辞書・用語・オントロジー構築，データベースへのアノテーション支援といった共有リソースの充実に重点がおかれてきた観があるが，ポストゲノム時代の産業的応用 (特に創薬やパーソナル医療) をリードするような技術面での研究開発も極めて重要である．特に重要文献のフィルタリングや要約，マイクロアレイ出力の分析支援，より高度な情報検索といった分野での需要が高いと予想される．この点では，KDD CUP 2002 (http://www.biostat.wisc.edu/ craven/kddcup/) におけるデータベースへのアノテーション付与のための関連文献のフィルタリング・タスクや，今年開催が予定されている TREC (Text Retrieval Contest) 2003 の Genome Track によるライフサイエンス文献検索タスクによって，機械学習や情報検索の更なる技術レベル向上が期待できよう．

参考文献

[1] M. Andrade and A. Valencia. Automatic annotation for biological sequences by extraction of keywords from medline abstracts. development of a prototype system. In *Proc. of ISMB '97*, pages 25–32, 1997.

[2] M. Andrade and A. Valencia. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *BioTechniques*, 14:600–607, 1998.

[3] C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Proc. of ISMB '99*, pages 60–77, 1999.

[4] B. Bruijn and J. Martin. Literature mining in molecular biology. In *Proc. of the EFMI workshop on NLP in Biomedical Applications*, pages 1–5, 2002.

[5] N. Collier, C. Nobota, and J. Tsujii. Extracting the names of genes and gene products with a hidden markov model. In *Proc. of COLING 2000*, pages 201–207, 2000.

[6] N. Collier, H. Park, N. Ogata, Y. Tateisi, C. Nobata, T. Sekimizu, H. Imai, and J. Tsujii. The genia project: corpus-based knowledge acquisition and information extraction from genome research papers. In *Proc. of EACL '99*, 1999.

[7] M. Craven and J. Kumlien. Constructing biological knowledge-bases by extracting information from text sources. In *Proc. of ISMB '99*, pages 77–86, 1999.

[8] J. Dickerson, D. Bedeant, Z. Cox, W. Qi, D. Ashlock, and E. Wurtele. Creating metabolic network models using text mining and expert knowledge. In *CBGIST '00*, pages 26–30, 2000.

[9] J. Ding and D. Berleant. Mining medline: Abstracts, sentences, or phrases? In *Proc. of PSB '02*, pages 326–337, 2002.

[10] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl 1):S74–82, 2001.

[11] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Toward information extraction: Identifying protein names from biological papers. In *Proc. of ISMB '98*, 1998.

[12] C. Friedman H. Liu. Mining terminological knowledge in large biomedical corpora. In *Proc. of PSB '03*, pages 415–426, 2003.

[13] U. Hahn, M. Romacker, and S. Schulz. Creating knowledge repositories from biomedical reports: The medsyndikate text mining system. In *Proc. of PSB '02*, pages 338–349, 2002.

[14] V. Hatzivassiloglou, P. Duboue, and A. Rzhetsky. Disambiguating proteins, genes, and rna in text: a machine learning approach. *Bioinformatics*, 17(Suppl 1):S97–106, 2001.

[15] M. Hearst. Untangling text data mining. In *Proc. of ACL'99*, 1999.

[16] T. Hishiki, N. Collier, C. Nobata, T. Ohta, N. Ogata, T. Sekimizu, R. Steiner, H. Park, and J. Tsujii. Developing nlp tools for genome informatics: An information extraction perspective. In *Proc. of GIW '98*, pages 81–90, 1998.

[17] K. Humphreys, G. Demetriou, and R. Gaizauskas. Automatically extracting enzyme interaction and protein structure information from biological science journal articles. In *Proc. of AISB '00*, 2000.

[18] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. In *PSB '00*, pages 502–513, 2000.

[19] A. Konagaya J. Hosaka, J. Koh. Effect of utilizing terminology on extraction of protein-protein interaction information from biomedical literature. In *Proc. of EACL '03*, 2003.

[20] P. Kankar, S. Adak, A. Sarkar, K. Murari, and G. Sharma. Medmesh summarizer: Text mining for gene clusters. In *Proc. of the 2nd SIAM Int. Conf. on Data Mining*, 2002.

[21] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman. Using blast for identifying gene and protein names in journal articles. *Gene*, 259(1-2):245–252, 2000.

[22] G. Leroy and H. Chen. Filling preposition-based templates to capture information from medical abstracts. In *PSB '02*, pages 350–361, 2002.

[23] E. Marcotte, I. Xenarios, and D. Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363, 2001.

[24] S. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. In *GIW '99*, pages 104–112, 1999.

[25] T. Ohta, Y. Tateisi, H. Mima, and J. Tsujii. Genia corpus: an annotated research abstract corpus in molecular biology domain. In *Proc. of HLT '02*, 2002.

[26] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.

[27] S. Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proc. of ACL '03*, pages 160–167, 2003.

[28] J. Park, H. Kim, and J. Kim. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *Proc. of PSB '01*, pages 396–407, 2001.

[29] D. Proux., F. Rechenmann, and J. Laurent. A pragmatic information extraction strategy for gathering data on genetic interactions. In *Proc. of ISMB '00*, pages 279–285, 2000.

[30] J. Pustejovsky, J. Castano, B. Cochran, M. Kotecki, and M. Morrell. Automation extraction of acronym-meaning pairs from medline databases. *Medinfo*, 10:371–375, 2001.

[31] J. Pustejovsky, J. Castano, and J. Zhang. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proc. of PSB '02*, pages 362–373, 2002.

[32] T. Rindflesch J. Rajah and L. Hunter. Extracting molecular binding relationships from biomedical text. In *Proc. of ANLP-NAACL '00*, pages 188–195, 2000.

[33] T. Rindflesch, L. Tanabe, J. Weinstein, and L. Hunter. Edgar: extraction of drugs, genes and relations from the biomedical literature. In *Proc. of PSB '00*, pages 514–525, 2000.

[34] A. Schwartz and M. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *PSB '03*, pages 451–462, 2003.

[35] T. Sekimizu, H. Park, and J. Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome Informatics*, pages 62–71, 1998.

[36] B. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline asbtracts. In *Proc. of PSB '00*, pages 529–540, 2000.

[37] B. Stapley, L. Kelley, and M. Sternberg. Predicting the sub-cellular location of proteins from text using support vector machines. In *Proc. of PSB '02*, pages 374–385, 2002.

[38] M. Stephens, M. Palakal, S. Mukhopadhyay, and J. Mostafa R. Raje. Detecting gene relations from medline abstracts. In *Proc. of PSB '01*, 2001.

[39] D. Swanson. Medical literature as a potential source of new knowledge. *Bullten of the Medical Library Association*, 78(1):29–37, 1990.

[40] D. Swanson and N. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Iatelligeace*, 91(2):183–203, 1997.

[41] L. Tanabe, U. Scherf, L. Smith, J. Lee, L. Hunter, and J. Weinstein. Medminer: An internet tool for filtering and organizing gene expression and pharmacological information. *BioTechniques*, 27:1210–1217, 1999.

[42] Y. Tateisi, T. Ohta, T. Takai, and J. Tsujii. An ontology for biological reaction events. In *Proc. of GIW '99*, 1999.

[43] J. Thomas, D. Milward, C. Ouzounis, and S. Pulman. Automatic extraction of protein interactions from scientific abstracts. In *Proc. of PSB '00*, pages 538–549, 2000.

[44] J. Tsujii and S. Ananiadou. An introduction to information extraction. In *Tutorial at PSB'01*, 2001.

[45] M. Weeber and R. Vos. Extracting expert knowledge from medical texts. In *Proc. of Intelligent Data Analysis in Medicine and Pharmacology Workshop (IDAMAP)*, pages 23–28, 1998.

[46] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. Event extraction from biomedical papers using a full parser. In *PSB '01*, pages 408–419, 2001.

[47] M. Yoshida, K. Fukuda, and T. Takagi. Automatic construction of biological abbreviation dictionary from abstracts of biomedical papers. In *GIW '98 (poster session)*, 1998.