［特集］ポストゲノム時代に高まるバイオ自然言語処理への期待：バイオ自然言語処理最新事情

7 企業におけるバイオNLPへの取り組み
a）ポストゲノム時代のテキストマイニングミドルウェア

浦本 直彦
URAMOTO@jp.ibm.com / 日本アイ・ビー・エム東京基礎研究所・国立情報学研究所

松澤 裕史
MATUZAWA@jp.ibm.com / 日本アイ・ビー・エム東京基礎研究所

人間の全遺伝情報（ヒトゲノム）の解読が終了し，ポストゲノム時代に突入したライフサイエンス分野において，自然言語処理技術が本質的に貢献するには，個々の技術の精度を高めるとともに，共通に使える資源，ツール，経験を整備することで，研究者が本当に難しい問題と向き合う環境を構築する必要がある．このために，単なるプロトタイプや柔軟性に欠ける完成されたアプリケーションのみならず，多くの開発者やユーザが共通に使えるミドルウェアとして定義し公開していくことに大きな意味がある．本稿では，このような観点から，IBM研究開発部門で行われているゲノム研究と自然言語処理の交点から生まれたいくつかの技術を紹介する．

ゲノム研究と自然言語処理

人間の全遺伝情報（ヒトゲノム）の解読が終了し，ゲノム研究の焦点は，塩基配列に代表される基本情報をいかに効率よく生成するかから，獲得された大量かつ多様な情報を統合・解析するかに移りつつある．ここで得られる知識は，薬や治療法を個人レベルの違いに合わせたオーダーメイド医療や，ゲノム研究を臨床研究へ応用するトランスレーショナルリサーチに代表される，人類に対し直接の恩恵を与える新しい技術として注目を集めている．

このような「知識化」の段階において，自然言語処理技術が果たす役割は大きい．従来からこの分野で培われてきた統計的手法，機械学習アルゴリズムなどが配列情報の解析に応用されてきた．加えて，この分野においては，MEDLINE（www.ncbi.nlm.nih.gov/entrez/）に代表される文献抄録情報，UMLS（www.nlm.nih.gov/research/umls/），Gene Ontology（GO: www.geneontology.org/）などの辞書情報やオントロジなど，大量の言語資源が蓄積されており，大量のテキスト情報から，有効な知識を獲得するテキストマイニング技術が大きく期待されているからである．このように多量の言語資源が公開されている分野は，他にあまり例がなく，研究対象として自然言語処理研究者の興味を引きつけている．このような背景から，医学文献からの情報抽出技術やテキストマイニングツールの研究開発などを通じ，自然言語処理技術がゲノム研究を支援する技術の柱の1つとして認知されつつある[1)]．

ライフサイエンスという新しい分野で，自然言語処理技術が本質的に貢献するには，個々の技術の精度を高めるとともに，共通に使える資源，ツール，経験を整備することで，研究者が本当に難しい問題と向き合う環境を構築する必要がある．このような観点から，単なるプロ

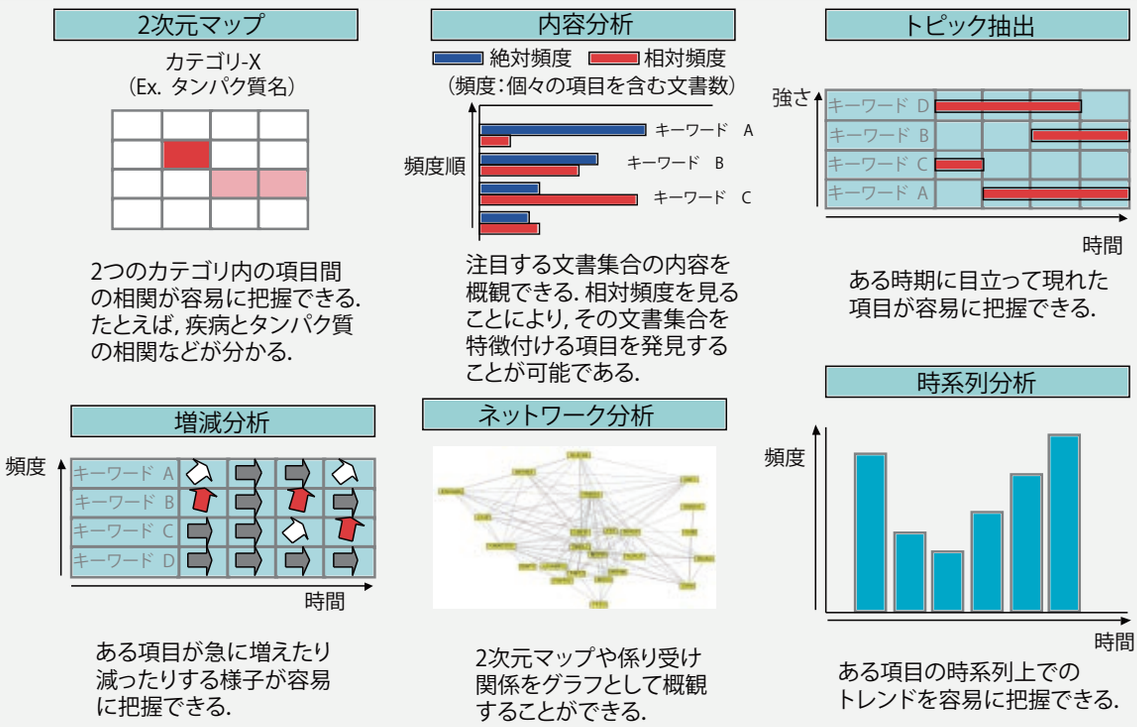

図-1　MedTAKMI が提供するマイニング機能

トタイプや柔軟性に欠ける完成されたアプリケーションのみならず，多くの開発者やユーザが共通に使えるミドルウェアとして定義し公開していくことの意義は大きい．本稿では，IBM研究開発部門で行われているゲノム研究と自然言語処理の交点から生まれた基盤および応用技術を紹介する．

テキストマイニングシステム IBM TAKMI® for Biomedical Documents

IBM TAKMI® for Biomedical Documents（以降，MedTAKMIシステムと呼ぶ）は，IBM東京基礎研究所がセレスター・レキシコ・サイエンシズ（株）と共同で開発した医学文献向けのテキストマイニングシステムであり，MEDLINE全件（開発当時約1,200万件）を処理するテキストマイニングエンジンとして実働している．MedTAKMIシステムの特徴を以下に示す．

- 約250万語の同義語辞書と組み合わせた構文解析を行うことで，名詞キーワードだけでなく，動詞（述語），形容詞，係り受け関係（主語－動詞，主語－動詞－目的語）などを抽出し，さらに，各単語に対し，遺伝子，疾病，生物種などの概念ラベルを付与している．
- テキストという非構造化情報と出版年，著者名などの構造化（定型）情報を統合的に処理することによる多角的な分析が可能である．
- キーワード間の相関関係の把握，比較，特徴の概観，分類，時系列的傾向の把握，トピック抽出など本格的な分析機能を提供する．
- 約34万語からなる階層化カテゴリ体系を使って，大まかなレベル（上位概念）から細かなレベルの概念（下位概念）まで幅広い範囲の分析が可能である．
- 集合の絞り込みと分析を柔軟に対話的に行う，いわゆる発想支援環境を提供する．
- 組織や細胞における遺伝子発現量の変化を解析するマイクロアレイ解析の実験結果（遺伝子集合）と文献情報を関連付け，遺伝子間の直接的，間接的関係を発見することができる．

MedTAKMIシステムの開発思想は，単なる検索システムやキーワードの可視化システムでは提供できない機能，人間の処理能力を超えるレベルの知識発見を支援する機能を提供することである．たとえば，MedTAKMIシステムでは，すべてのキーワードに対して，遺伝子，タンパク質，疾患といった階層化された概念体系に含まれるカテゴリコードを1つ以上割り当てることができ，複数の文献に出現する遺伝子・タンパク質間の関係や時系列に沿った話題の発見を行うことができる．また，Webサービスとして実装されているため，実験データとの統合などより大きなシステムの一部として埋め込むことも可能である．**図-1**にMedTAKMIシステムが提供する分析機能を示す．MedTAKMIシステムの詳細につ

いては，参考文献2），3）を参照されたい．

ゲノム研究における自然言語処理技術の応用が進められている中で，この分野に特有の難しい問題がいくつか見えてくるようになった．たとえば，遺伝子名の曖昧性の問題である．遺伝子名は，“IL6”や“TP53”など英数字の短い並びであるが，非常に多くの表記・同義語を持っており，2種類の曖昧性が生じる．1つ目は，遺伝子名として，“p”，“bare”，“bike”，“dark”，“db”など，一般語と見分けがつかない単語が存在することである[4]．2つ目は，同じ単語が，複数の遺伝子名の同義語として定義されており，これらを弁別することは非常に難しいことである．なぜなら，遺伝子という限定された意味カテゴリにおける曖昧性であるため，自然言語処理で用いられる選択制限などの手法がうまく適用できないためである（たとえば，文中の「リンゴ」の意味を赤リンゴと青リンゴのどちらかに同定する難しさを思い浮かべてほしい）．筆者らは，遺伝子名として曖昧性の高い遺伝子を同定する手法を開発し，現在，曖昧さ解消の手法を開発中である．

Unstructured Information Management Architecture（UIMA）

MedTAKMIシステムのようなテキストマイニングシステムを構築するためには，テキスト情報を解析し，そこからさまざまな情報を抽出する必要がある．ライフサイエンス分野においても，MEDLINEに代表される医学文献に対する形態素解析（単語の同定），構文解析（単語間の係り受け関係の同定），概念抽出（遺伝子名の抽出など），関係抽出（タンパク質の相互作用など）が研究の中心となっている．これらの処理では，さまざまなアルゴリズムや実装手法が考えられるが，テキスト文字列が入力で，テキスト中の部分文字列（単語，文，段落，テキスト全体など）に対する付加情報（アノテーション）の集合が出力であると考えることができる．つまり，構文解析器も情報抽出器も，共通のインタフェース（API）とそれを呼ぶミドルウェアを用意してやれば，異なる実装のプログラムを自由に組み合わせることが可能であることを意味している．Unstructured Information Management Architecture（UIMA）は，異なるテキスト解析ツールを協調して動かすためのミドルウェアであり，IBM米ワトソン研究所を中心にIBMの各基礎研究所との共同で，現在活発に研究開発が進められている[5]．Javaで記述されたUIMA ライブラリ（UIMA SDK）は，IBMのalphaWorksサイトで公開されており（http://www.alphaworks.ibm.com/tech/uima），ダウンロードして使用することができる．

自然言語文に対する高精度での意味的解析や機械翻訳技術は，いまだ発展途上であり，そもそも，単一の手法だけでの完全な解決は困難である．そこで，複数の異なるアルゴリズム，たとえば，統計，論理，経験的手法などを組み合わせることで，単体手法の単なる足し算以上の成果を上げる結合仮説（Combination Hypothesis）[6]がUIMAの根底をなす思想である．また，従来から，さまざまなテキスト解析ツールが提案されているが，現実的な問題として，それらを大きなコミュニティの中で共有することは，なかなか難しい．ライセンス上の問題は別にしても，同じようなテキストに対して，同じような処理を行うツールを集めてきたとしても，関数の形式が異なったり，テキストを特別な形式に変換しないといけなかったりすることが多いからである．UIMAの目的の1つは，異なる開発者が作ったツールを自由に組みあわせることで，本質的に難しい問題を解くためのコストを最小限にすることにある．

UIMAのアーキテクチャを**図-2**に示し，個々のコンポーネントを簡単に説明する．

- **Text Analysis Engine（TAE）**：入力テキストに対して，何らかの処理を行い，結果をアノテーションとして出力するコンポーネント．情報抽出器，構文解析器などさまざまなツールがTAEとして表現される．複数のTAEを結合し1つのTAEとして定義することもできる（Aggregate TAE）．各TAEに関する情報（出力アノテーションの構造や起動プログラムの指定など）は，TAE Descriptorと呼ばれるXML文書で記述される．
- **Common Analysis（CAS）**：テキストとTAEが出力するアノテーションを保持する構造を持ったオブジェクト．アノテーションのデータ構造として素性構造（feature structure）を用いており，複雑な構造を持ったアノテーションを定義できる．文中で交差するアノテーションを表現するため，文中でのアノテーションの開始位置と終了位置を保持している．CASオブジェクトは，CAS Initializerによって生成され，TAE群に順次渡されていく（図-2では，TAEごとに別のTAEが表現されているが，実際は，1つのCASオブジェクトがTAEに渡されていき，アノテーションが追加されていく）．
- **CAS Consumer：**（複数の）TAEによって生成されたアノテーション構造（CAS）から，個々のアプリケーションが必要とするデータ構造を生成するコンポーネント．
- **Collection Processing Manager（CPM）**：処理対象となるテキストは，ファイルシステム，Webデータなど集合（コレクション）として存在することが多い．CPMは，コレクションを扱うためのフレームワーク

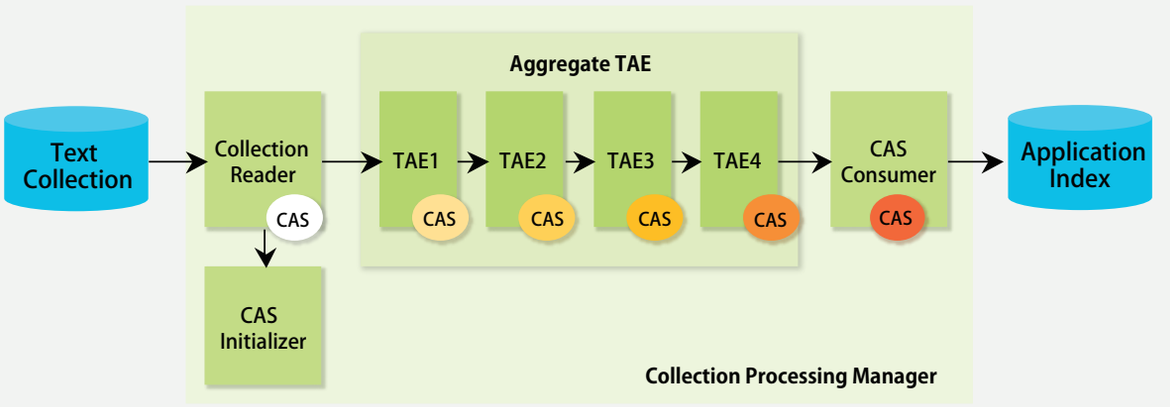

図 -2　UIMA の構成要素

であり，Collection Readerが，コレクションから一文書ずつ取り出し，対応するCASを生成して，TAEに渡す．

たとえば，形態素解析器，構文解析器，辞書を用いた化合物名抽出器，統計的手法を用いた遺伝子名抽出器，構文情報を用いたタンパク質間の関係抽出器，などの異なる手法・実装に基づく複数のTAEを組み合わせることで，高度な処理を行うことができる．もちろん，UIMAそのものは分野に依存しない基盤である．

ライフサイエンス分野向けのテキストマイニングミドルウェアBioTeKS

BioTeKS[7]は，UIMAをベースに，ライフサイエンス分野向けのテキストマイニングミドルウェアを構築するプロジェクトであり，IBM米ワトソン研を中心に，筆者のグループを含むIBMの各基礎研究所が参加している．BioTeKSでは，ライフサイエンス向けのTAEとして，遺伝子，タンパク質，疾患名，化合物といった概念の抽出や，それらの間の関係（相互作用）などを抽出するコンポーネントが開発されている．上で挙げたMedTAKMIシステムで用いられるインデックスファイルも，複数のTAEを組み合わせることにより生成することが可能である．UIMAアーキテクチャに基づいてツールを開発することで，他の研究所で作られたTAEとその結果を簡単にMedTAKMIシステムに取り込むことが可能である．また，ライフサイエンス分野では，マイクロアレイを用いた遺伝子発現解析などの実験データ（構造化情報）とテキスト情報（非構造化情報）をいかに統合するかが鍵であり，筆者らのグループでは，マイクロアレイ解析から得られる遺伝子集合と文献を関連付け，遺伝子間の関係をレポートするツールを開発している．

さらなる発展へ向けて

本稿では，テキストマイニングミドルウェアの観点から，現在IBMで進められているいくつかのプロジェクトを紹介した．このような共通基盤を構築することで，この分野でのブレイクスルーが起きることに，筆者らは大いに期待している．

謝辞　本稿で紹介した技術は，日本IBM東京基礎研究所の竹内広宜，長野徹，吉田一星，猪口明博，村上明子，武田浩一の各氏の協力によるものである．

参考文献

1) 浦本，松澤，猪口，武田：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，情報処理学会情報学基礎研究会報告，No.130-004(2003)．
2) Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. and Takeda, K.：A Text-mining System for Knowledge Discovery from Biomedical Documents, IBM Systems Journal, Vol.43, No.3 (2004) (http://www.research.ibm.com/journal/sj/433/uramoto.html).
3) 松澤，長野，村上，浦本，武田：ライフサイエンス向けテキストマイニングツール MedTAKMI，情報処理学会データベースシステム研究会報告，No.130-005(2003)．
4) 大井，大田，今一，丹羽，久光：一般語との曖昧性を持つタンパク質名の自動検出，情報処理学会情報学基礎研究会報告，No.76-04(2004)．
5) Ferrucci, D. and Lally, A.：Building an Example Application with the Unstructured Information Management Architecture, IBM Systems Journal, Vol.43, No.3 (2004) (http://www.research.ibm.com/journal/sj/433/ferrucci.html).
6) Spector, A.：Architecting Knowledge Middleware, Invited Talk at The WWW Conference 2002 (http://www2002.org/spector.pdf).
7) Mack, R., Mukherjea, S., Soffer, A., Uramoto, N., Brown, E., Coden, A., Cooper, J., Inokuchi, A., Iyer, B., Mass, Y., Matsuzawa, H. and Subramaniam, L. V.：Text Analytics for Life Science Using the Unstructured Information Management Architecture, IBM Systems Journal, Vol.43, No.3 (2004) (http://www.research.ibm.com/journal/sj/433/mack.html).

（平成17年1月4日受付）