［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

4

バイオNLPのためのコーパスと各種リソースの現状

建石 由佳
yucca@is.s.u-tokyo.ac.jp / 科学技術振興機構

大田 朋子
okap@is.s.u-tokyo.ac.jp / 科学技術振興機構

辻井 潤一
tsujii@is.s.u-tokyo.ac.jp / 東京大学大学院情報学環

生命科学分野では，情報のデータベース化，文献のふるい分け，文献からの情報抽出，情報同士の関連付けなどを自動化あるいは補助するために，自然言語処理技術が論文やアブストラクトの解析に応用されつつある．高度に専門化された分野のテキストに自然言語処理技術を適用するには，分野特有の知識や語用法を機械可読な形に整備したオントロジ・辞書・コーパスは不可欠なリソースであり，これらを整備するための努力がなされている．本稿ではそれらのリソースについて紹介する．

背景

ポストゲノムシークエンス時代を迎えた今日，解析機器の進歩も伴って実験室では日々実験データが量産されている．そのデータを解析・評価するためには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るが，地道に文献を読んで理解するのではデータの急増に追いつくことができない．そこでまず考えられたのは情報をデータベース化し統合するということである（**図-1**）．しかしデータ登録のスピードがデータの増加スピードに追いついておらず，データベース化される情報は全体のごく一部で結局は個々の研究者が文献を多量に読まねばならないのが現状である．このような状況で，情報のデータベース化や文献のふるい分け，文献からの情報抽出，情報同士の関連付けなどを自動化あるいは補助する技術が熱望される現在，自然言語処理技術が生命科学分野の論文あるいはアブストラクトの解析に応用されつつある．

生命科学テキストに対する自然言語処理システムの構築に際し，解析ルールなどを構築するためのデータリソースとして概念体系を整備したオントロジ概念体系，用語との関連をつけるための辞書，および用語が実際のテキストでどのように表現されるかを示すコーパスが必要とされる（**図-2**）．本稿ではこれらについて述べる．

データベースとオントロジ

オントロジとはもともと哲学の用語で「存在論」と訳されるが，人工知能の分野では知識ベースシステムの知識を記述するために研究されている．オントロジは「概念の明確な体系的記述」と捉えられ，知識ベースシステムなどを構築する際のビルディングブロックとして用いられる基本概念や語彙の体系を整理し定義したものとさ

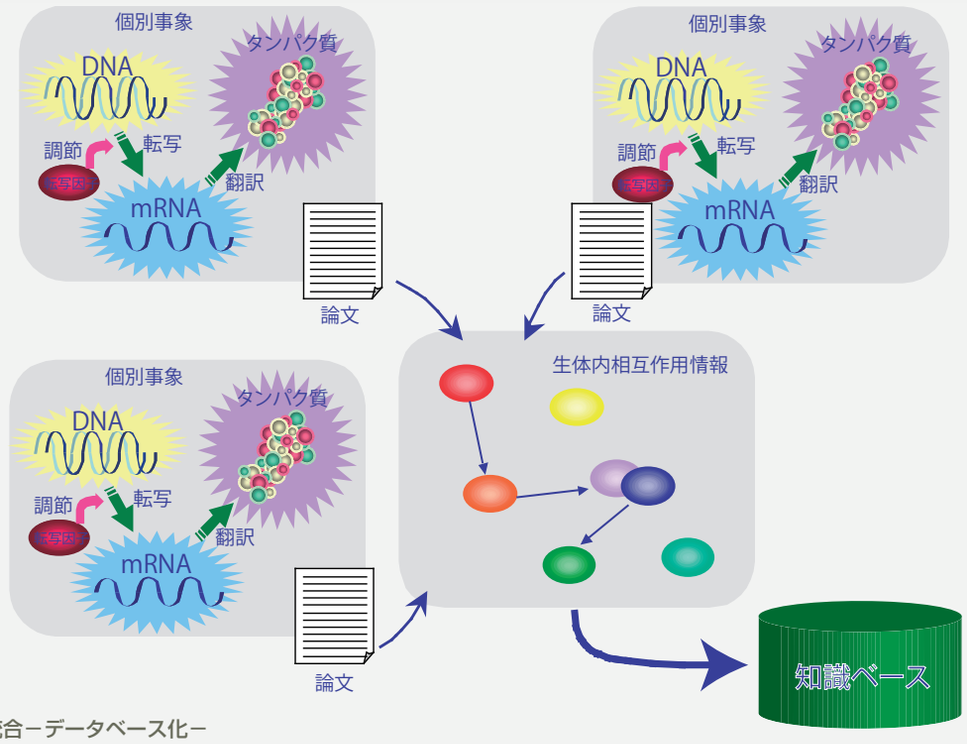

図 -1　知識の統合－データベース化－

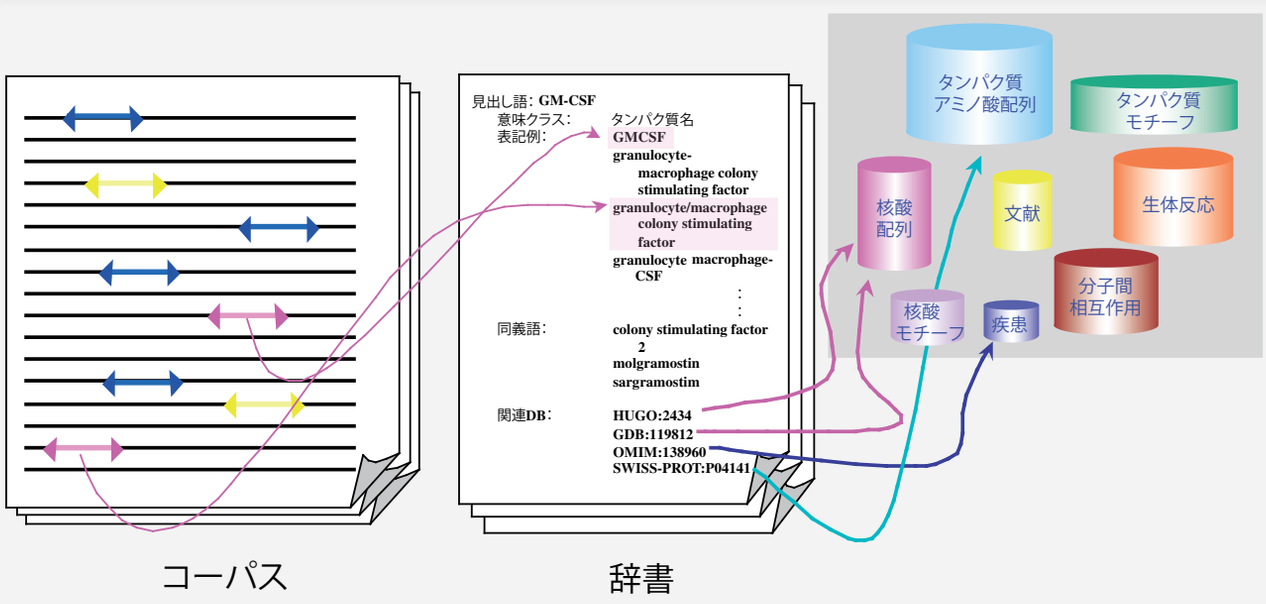

図 -2　コーパス・辞書は統合された知識と論文上の表現形との対応付けをする

れている．

生命科学のオントロジは，データベースのデータ記述のための標準語彙体系として発達してきた．同一の概念を同一の語彙で表現するためには概念そのものの整理体系化が必要であり，それはオントロジを整備することにほかならない．特に生命科学では数学・物理学と異なり日常の言語で現象を記述してきたため，言語の表現する対象の定義があいまいであるということがしばしばある．たとえば「タンパク質」も「ペプチド」もアミノ酸から構成される高分子を指すが，与えられた「アミノ酸からなる分子」をタンパク質であるとするかペプチドであるとするかについての明確な基準はない．また発見された物質の命名規則が確立していないため，**図 -3** に示すように同一物質が複数の名前を持つことも多い．これらは研究者同士のコミュニケーションにも問題を生じ得るが，特に機械化のために大きな障害となる．データベースの使用語彙の統一と体系化はこの問題を解決するために行われているものであり，最大規模の試みとして Gene Ontology がある．

見出し語：**IL-1 alpha**
表記例：**IL 1 alpha**
IL-1A
IL1A
interleukin-1 alpha
interleukin (IL) 1 alpha
：
：
同義語：**IL-1**
interleukin 1
Interleukin 1
：

見出し語：**GM-CSF**
表記例：**GMCSF**
granulocyte-macrophage colony stimulating factor
granulocyte/macrophage colony stimulating factor
granulocyte macrophage-CSF
：
同義語：**colony stimulating factor 2**
molgramostin
sargramostin
：

図 -3　専門用語の異表記・同義語

■ Gene Ontology と Open Biology Ontology

Gene Ontology は FlyBase（ショウジョウバエゲノムデータベース），SGD（酵母ゲノムデータベース），MGD（マウスゲノムデータベース）の作成者たちが遺伝子の配列の持つ機能の記述（アノテーション）をするための共通の用語体系を作る目的で 1998 年に作成を始めた．データベースで使用する用語を標準化することによって複数のデータベースにわたる検索を可能とするためである．現在では Wormbase（線虫ゲノムデータベース）など多種の生物のゲノムデータベースや EBI（タンパク質データベース SwissProt の作成元）が参加して Gene Ontology Consortium[1] という組織をつくりオントロジの整備拡張を行っている．

Gene Ontology は Molecular Function（分子機能），Biological Process（生体内反応），Cellular Component（細胞構成要素）の 3 つのコンポーネントからなる．各コンポーネントは上位下位の階層に整理された用語のリストで，概念の定義は自然言語の文で記述される（**図 -4**）．Molecular Function は遺伝子の転写により生成された分子（生成物）の機能，たとえば transcription factor activity（転写因子としての機能）や signal transducer activity（シグナル伝達機能）などを指す．Biological Process は生体レベルの視点の反応，たとえば mitosis（有糸分裂）や purine metabolism（プリン代謝）などが含まれる．Cellular Component は細胞の部分構造や高分子複合体，たとえば nucleus（核）や telomere（テロメア）などを含む．

この 3 つのコンポーネントを用いて，各遺伝子には「その遺伝子の生成物が果たす 1 つ以上の Molecular Function」，「生成物がかかわる 1 つ以上の Biological Process」，「生成物が生成される生体内での場所としての Cellular Component」が記述される．

しかし，Gene Ontology で記述できる遺伝子の機能は遺伝子にかかわる情報の一部に過ぎない．たとえば遺伝子の異常にかかわる疾患などは Gene Ontology の体系には含まれない．これらについては，Gene Ontology を拡張するのではなく既存のオントロジ体系を取り込むことによって対応しようとしており，そのための枠組みとして Open Biology Ontology プロジェクト（OBO）[2] がある．

OBO は，生命科学の諸分野のオントロジを，書式を共通化しオープンリソース化することによって共有する試みである．このプロジェクトは，すでにプロジェクトに参加しているオントロジに記述される概念体系とは独立した体系を記述すること，他のオントロジと同じツールで扱える書式であること，（自然言語による）用語の定義を含むこと，などの条件を満たし，かつ使用に制限を設けないという条件を備えたオントロジをまとめるものである．現在は Gene Ontology をはじめ 7 つの組織が参加している．

用語辞書

専門的な分野のテキストの解析では専門用語の処理が重要な課題であり，専門用語に対して構文的意味的情報を付与した網羅的な辞書はこの処理を容易にする．生命科学分野においても用語辞書の作成更新が自動・手動を問わず広く行われている．ここでは例として特に医学分野で広く使われている Unified Medical Language System [3] を紹介する．

■ Unified Medical Language System（UMLS）

UMLS は米国 National Library of Medicine（NLM）が構築している医学一般の機械可読辞書である．NLM は文献データベース MEDLINE を管理する組織で，MEDLINE には検索用のキーワードとして統制語彙 MeSH Term が定義されている．この MeSH Term 自身，用語の階層化と定義関連語などの注釈を持つ語彙体系であるが，UMLS はこれを拡張したものであり，統制語彙集（Meta Thesaurus），意味ネットワーク（UMLS Semantic Network），医学生物学用語辞書（SPECIALIST Lexicon）からなる．Meta Thesaurus は 80

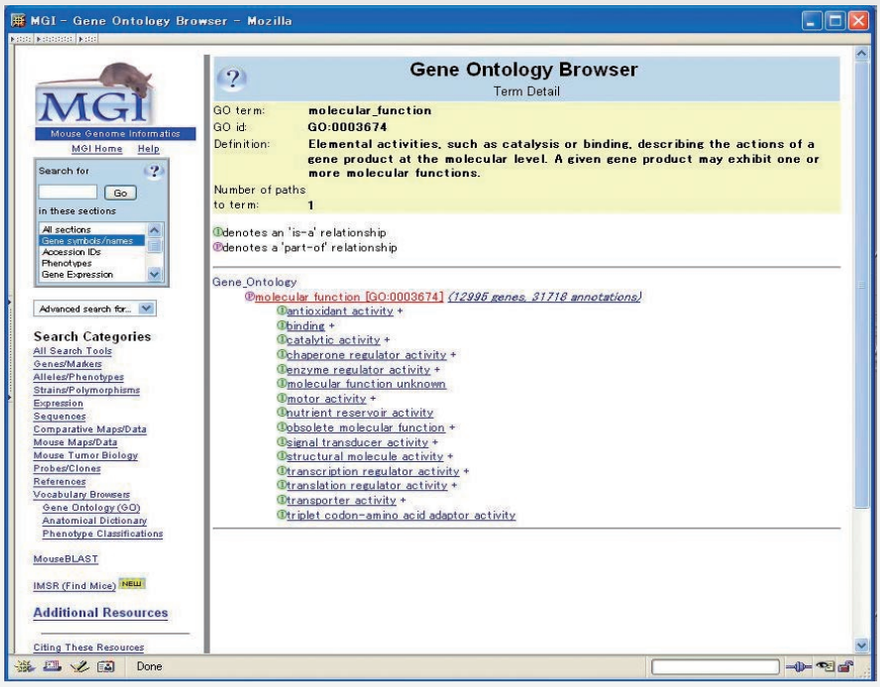

図 -4　Gene Ontolgy（Mouse Genome Informatics のブラウザでの表示）

万の階層化された概念とそのラベルとしての1,900万語の語彙を持つ大規模な体系となっている．UMLS Semantic Networkはカテゴリ間の上位－下位関係を示すis-a関係のほかに「is temporally related to」「is physically related to」など54種の関係でシソーラス中の134種の概念が結びつけられたネットワークで，Meta Thesaurusの概念を概念間の関係によって定義していることになる．SPECIALIST lexiconは一般辞書にない専門用語の辞書で，語の意味のほかに統語情報も含む．

しかしながら，これでも実際のテキストにあたってみると登録されていない用語が多く，現実には各データベースのエントリーのキーと内容からおのおのの研究グループが辞書を作成して使用している場合が多い．このときSwissProt などのデータベースを用いると，各エントリーの関連項目がGene Ontologyに関連付けられているためGene Ontologyを共通の意味体系とした辞書を得ることができる．

また，文献から用語を自動抽出した結果から辞書を作成する試みも行われている．詳しくは本号の「生命科学文献からの知識抽出と辞書構築」(小池･高木)を参照してほしい．

文献の解析とコーパス

コーパスとは，言語学研究のために収集された大量のテキストの集合，特に機械処理できるように電子化されたテキスト資料を指す．最も単純なものは，テキストをそのまま集めたもので，「生コーパス」（raw corpus）と呼ばれる．品詞，構文木，対訳など各種の情報を付加したものは「タグ付きコーパス」（tagged corpus annotated corpus）と呼ばれる．コーパスは古くから実際の用例を分析して言語理論を組み立てるために不可欠のリソースで，英語ではLOB Corpus，British National Corpus，Brown Corpus，University of Pennsylvania Corpusなどが有名である．

自然言語処理においては，共通のコーパスに対するカバレージや精度を測定することによって異なるシステムの客観的な比較を行うことが可能になる．また解析ルールの構築のためのリソースとしても利用され，近年機械学習の応用が盛んになるにつれて学習のためにコーパスの必要性が増すとともに，より大規模なコーパスが求められている．

生命科学分野のコーパス

生命科学分野では情報の共有が重視されており，論文アブストラクトは古くからMEDLINEとして電子化されてきた．また近年ではBioMed Central[4]，PubMed Central[5] のように論文自体をフリーアクセスにする動きも出てきている．このためこれらを生コーパスとして利用することができ，大規模な生コーパスは容易に手に入るといえる．たとえばMEDLINE収録アブストラクトは2003年の時点で1,200万件（1件約200語として約24億語）を超えている（**図-5**）．

タグ付きコーパスは1990年代の終わりごろから作成され始めた．初期のコーパスは情報抽出システムが抽出すべき情報の「正解」をタグ付けしていた．たとえばテキスト中に存在するタンパク質の名前をすべてタグ付けするなどである．論文・アブストラクトは高度に専門化された内容を表現しているため，門外漢である自然言語処理研究者にとって内容を容易には理解できないテキストを処理することになる．このことは情報抽出の正解が判定しづらい，ルールベースのシステムを構築する際ルールが書きにくい，という点で情報抽出システムの構築を困難にしている．これらの問題点を解消するために専門家の知識を何らかの形でエンコードしてテキストに

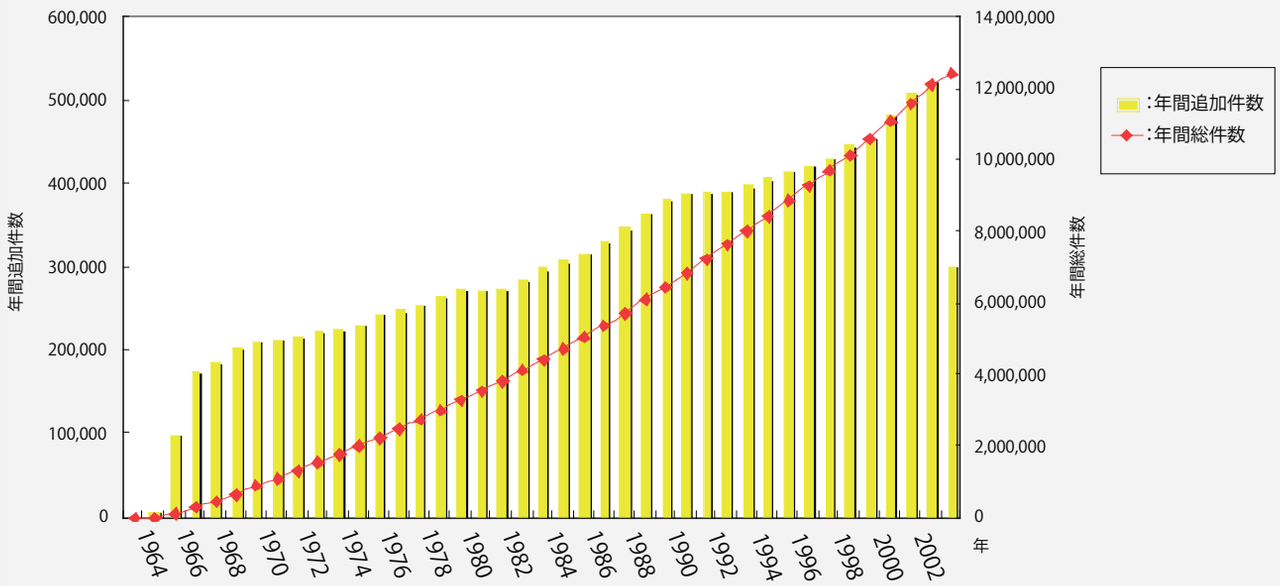

図 -5　MEDLINE 収録アブストラクト数の推移

表現されている文字列との対応をつけることは有用な手段であり，生命科学のように高度に専門化された分野ではタグ付きコーパスがより重要である．

また，研究が発展するにつれ，新聞などのテキスト処理に使われた言語解析システムを論文・アブストラクトにそのまま使用したのでは精度やカバレージが不十分で，分野に特化した改良や新しい手法の開発が必要であることが分かってきた．意味が十分に把握できないと品詞付け・構文解析などの浅いレベルの処理に対しても精度の評価がしづらい．そのため，専門家の知識を利用して品詞・構文・照応などのあいまいさを解消したコーパス，すなわち言語学的情報をタグ付けしたコーパスも作成されるようになった．

タグ付きコーパスの例

実際の処理システムの開発においては，いまだに多くの場合個々の開発者が，独自の，たいていは小規模なコーパスを使用してシステムの評価開発を行っている．公開された大規模なコーパスの例として我々のグループで開発しているGENIAコーパスを挙げる．

■ GENIAコーパス

GENIAコーパスは機械学習ベースの自然言語処理の手法を応用するための学習および検証データとして開発された．MEDLINEデータベースからhuman（ヒト），transcription factors（転写因子），blood cells（血球細胞）の3つをキーワードとして検索された結果のテキストをベースとし，そのテキストから考えられる限りの情報を陽にマークアップしていこうという方針で作成されている．同一のテキストに各種の情報をタグ付けすることによって情報相互間の関係，たとえば用語抽出のためにどんな言語知識が手がかりとなり得るか，を観察することも目的としている．

現在，専門用語（2,000アブストラクト約40万語），品詞（専門用語コーパスと同一セット），構文木（専門用語コーパスのサブセット200アブストラクト約4万語）をタグ付けしたコーパスを公開している．またInstitute for Infocomm Research（シンガポール）におけるMedCoプロジェクトではGENIAコーパスの一部（228アブストラクト）に照応関係（代名詞などの参照関係）をタグ付けしている．

GENIA専門用語コーパスは，物質とその所在（ソース）の名の位置を同定するとともに各々の用語についてタンパク質名細胞名などのクラス分けをする．このクラス分けのために，我々はGENIAオントロジと呼ぶ小規模な分類体系（**図 -6**）を構築した．GENIAオントロジはSubstance（物質名），Source（所在）およびOther（そのほか）というサブコンポーネントからなる概念の階層関係の木構造で，専門用語にはこの葉ノードのいずれかが意味クラスとして割り当てられる．また品詞コーパスではアブストラクトの各単語にPenn Treebank（ペンシルバニア大学で作成され広く使われている品詞・構文

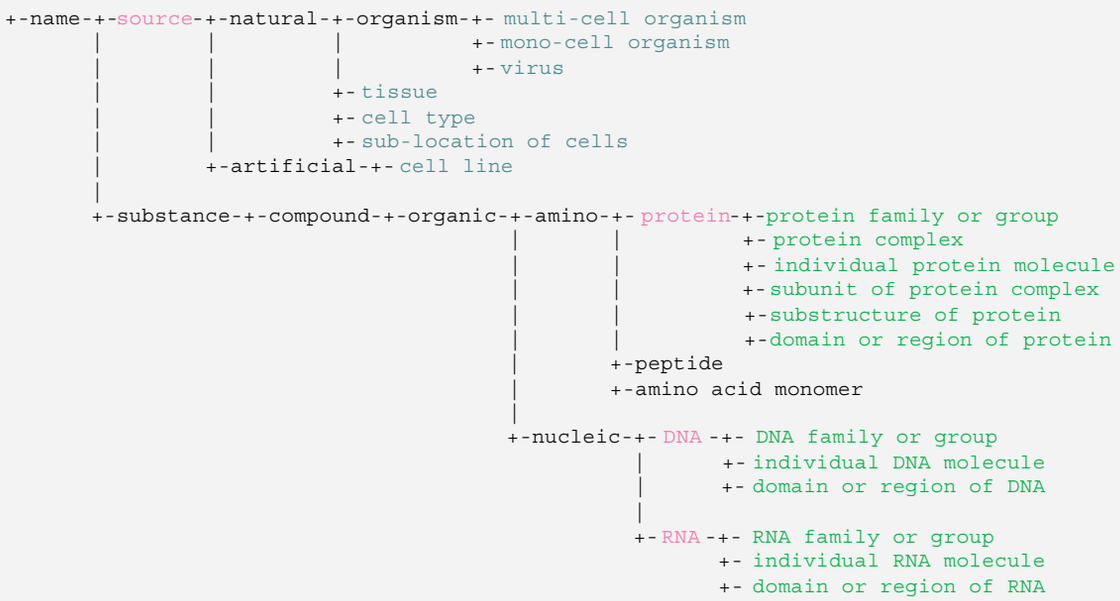

図 -6　GENIA オントロジ

タグ付きコーパス）のセットに基づく品詞を割り当てており，構文木コーパスはPennTreebankと同様の構造をXML形式で付与したものである．将来の拡張として，動詞とその主語・目的語などとの関係（述語項構造）のタグ付け方式を研究中である．

■ そのほかのコーパス

米国ペンシルバニア大学のBioIEプロジェクトでは腫瘍（oncology）およびタンパク質CYP450に関する情報抽出を行うためのデータとして，関連する論文アブストラクトに物質名（遺伝子，タンパク質など），Penn Treebank形式の品詞，および構文のタグ付けを行っている．2004年11月現在，物質名および品詞については2,257件のアブストラクト，構文については643件のアブストラクトについてタグ付けが完了している．また述語項構造タグ付けの計画もある（2005年1月公開予定）．

米国テキサス大学ではMEDLINEから “human” をキーワードとして検索したアブストラクト750件について遺伝子・タンパク質名をタグ付けしたコーパス，およびタンパク質反応データベースDIPに参照される文献のアブストラクト200件にタンパク質名とDIPデータベースに登録されている反応をタグ付けしたコーパスを作成し，これらを，遺伝子名に言及があっても遺伝子間の反応に関する記述のないアブストラクト30件とともに公開している．

より小規模なコーパスとして公開されているものではSwedish Institute of Computer Science（スウェーデン）のYAPEXコーパス，Brandies大学（米国）のMedstractコーパスがある．前者はGENIAコーパスのサブセットを含む200件のアブストラクト（約4万語）にタンパク質名をタグ付けしたものである．後者は略語と正式名称の関連付けをタグ付けしたもの（288アブストラクト，約6万語）と照応関係をタグ付けしたもの（32アブストラクト，約1,000語）である．

さらに，共通の課題を解くことによってシステムの相互比較を行うタスクのためのコーパスがある．2004年にはISMBに併設するBioLinkワークショップでBioCreAtIvEとよばれる情報抽出の共通タスクが，Colingに併設するBioNLPワークショップで専門用語抽出の共通タスクが，それぞれ設定された．BioCreAtIvEで使用されたコーパスは，公開はされていないが主催者へリクエストすれば手に入れることができる．BioNLPの専門用語抽出はGENIAコーパスを簡略化したコーパス（トレーニング用2,000アブストラクト評価用400アブストラクト）を用いた．このコーパスはタスクのWebページから入手可能である．

主なコーパスの入手先などを**図-7**に示す．

まとめと将来の課題

高度に専門化された分野のテキストに自然言語処理技術を適用するには，分野特有の知識や語用法を機械可読な形に整備したオントロジ，辞書，コーパスは不可欠

コーパス名	作成元	ベース	大きさ	付けられた情報	関連URL
GENIA	東京大学	アブストラクト	2,000件（約 40万語）	専門用語(物質名, 所在, その他)	http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
				品詞	
			200件 (約4万語)	構文木	
GENIA-MedCo	Institute of Infocomm Research (シンガポール)	アブストラクト（GENIA コーパスのサブセット）	228件 (約4万5千語)	照応	http://nlp.i2r.a-star.edu.sg/medco.html
Yapex	Swedish Institute of Computer Science (スウェーデン)	アブストラクト	200件 (約4万語)	タンパク質名	http://www.sics.se/humle/projects/prothalt/
MedStract	Brandeis 大学 (米国)	アブストラクト	288件 (約6万語)	略称と正式名称の対応	http://www.medstract.org/gold-standards.html
			32件 (約800語)	照応	
AIMED	Texas大学（米国）	アブストラクト	750件	タンパク質名・遺伝子名	http://www.cs.utexas.edu/users/ml/
			200件+反例30件	タンパク質の反応	
BioNLP/NLPBA　Shared Task		アブストラクト	トレーニング用 2,000件＋テスト用 4,000件（計約　5万語）		http://research.nii.ac.jp/~collier/workshops/JNLPBA04st.htm
BioCreAtIvE		アブストラクトから抜き出した文	トレーニング用 7,500文＋テスト用 2,500文 (計約20万語)	遺伝子名, タンパク質名	http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/results/

図 -7　公開されているタグ付きコーパス

なリソースであり，これらを整備するための努力がなされている．生命科学の分野では知識の増加スピードが速いので，リソースの整備による知識の統合整理のみならず，その統合整理を効率化する手法の開発も課題の1つである．また自然言語処理のためのリソースの開発には専門分野の知識のみならず言語学の知識も要求されるので，生命科学者と自然言語研究者の協力体制の確立も求められている．

参考文献
1) http://www.geneontology.org/
2) http://obo.sourceforge.net/
3) http://www.nlm.nih.gov/research/umls/
4) http://www.biomedcentral.com/
5) http://www.pubmedcentral.nih.gov/

（平成16年12月28日受付）