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

2005－BIO－3（11）
2005／12／22

# 相互作用情報タグつき生命科学論文アブストラクトコーパスの作成

大田朋子[1,2]　建石由佳[1,2]　金進東[1,2]　辻井潤一[1,3]
JST/SORST[1]　東大・院・情報理工[2]　東大・院・情報学環[3]
{okap, yucca, jdkim, tsujii}@is.s.u-tokyo.ac.jp

**概要**

生命科学の分野では、分野に関連する論文のアブストラクトが網羅的にMEDLINEデータベースに収録されており、現在約1,500万件の論文アブストラクトが収録されている。この分野の論文アブストラクト中に出現する遺伝子・タンパク質の相互関係は、その名前の表記の多様性に加えて言語学的に多様な表現で記述されている。これらの多様性を解析・学習して、構文解析結果と高度な知識との関係を理解し、認識するシステムの開発を行うためには、システムの学習や評価などに用いる基礎データが必要不可欠である。そこで我々は相互作用情報タグつき生命科学論文アブストラクトコーパスの作成を開始した。

**Abstract**

Information extraction systems in biomedical field has become a widely researched application of natural language processing (NLP) technologies. Convincing results of named entity extraction have been reported and now research focus is shifting to extraction of verbal information such as interactions, locations, relations and other events between named entities such as proteins and genes. Traditionally, events and relations are extracted using patterns on surface text around a certain sets of verbs using Part-of-speech taggers and/or shallow parsers. Recently, due to limitation of the scope of verbs and expressions that pattern-based approach can handle, more strategic and systematic analysis using deeper NLP techniques are suggested. One solution to this problem is using deep parsers which can abstract the syntactic variation of a relation between a verb and its arguments represented in the text, and constructing extraction rule on the abstract predicate-argument structure.

## 1.　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、生命科学の実験室では実験データが日々量産されるようになっている。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、実験データを裏付けていく必要があるが、既に研究者が文献を読んで理解しながら対処できる限界を超えてしまった。この分野では、これまでにさまざまな情報がデータベース化されているが、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いないことや、個別にデータベース化されている情報を横断的に検索して情報を収集する技術が未熟であることなどから、個々の研究者が大量の関連文献を読まねばならないのが現状である。このような状況で、大量の文献群から効率よく関連する情報を収集する技術が切望され、自然言語処理

技術の適用が求められている。

2. 生命科学分野のコーパス

生命科学分野の論文アブストラクトは古くから MEDLINE データベース[2]として電子化され、また近年では BioMed Central[3]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。

図 1 に示すように、MEDLINE データベースに収録されているアブストラクトは、2003 年の時点で既に 1200 万件（1 件約 200 語として約 24 億語）を越えており、近年では年間約 50 万件ずつ追加収録されている。

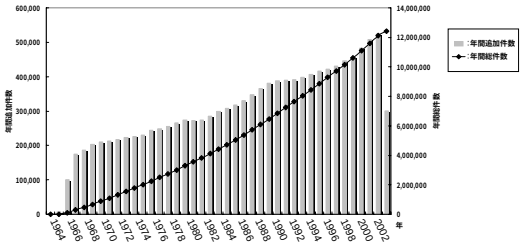


**図1 MEDLINE収録アブストラクト数の推移**

この分野の論文やそのアブストラクトは、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簡単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判定や、ルールベースのシステム構築の障害になり、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は、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表現されている文字列との対応をつけることが必要であり、したがって、生命科学のように高度に専門化された分野ではタグ付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わかってきた。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必要なのである。

3. 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設計した。前述のように、新聞などのテキストに対して既に実用可能な性能を示す品詞付けについても、この分野のテキストに対しては改良の余地があることから、現在我々は MEDLINE アブストラクトに対して、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コーパスを作成している。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うるかを観察することも目的としている。

対象とするテキストは、MEDLINE データベースから human（ヒト）blood cells（血球細胞）transcription factors（転写因子）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用いている。これは生命科学分野の中でも、かなり狭い研究領域

になっているため、辞書ベースのアプローチに利用するにはそのカバレージは十分ではない。しかし、領域を広く設定しすぎると、専門用語（特に物質名）が多岐にわたり、機械学習ベースのアプローチに有効な特徴を捉えることができなくなる。一方、生命科学分野特有の言語学的な特徴は、この絞り込んだアブストラクト群からも十分に捕らえることができると考え、この領域のアブストラクト群を用いている。

### 3.1. GENIA 専門用語コーパス

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、細胞名などの意味クラスを、生命科学の専門家が人手で付与したものである。この意味クラスを定義するために、我々はGENIA オントロジーと呼ぶ小規模な分類体系(図２)を構築した。

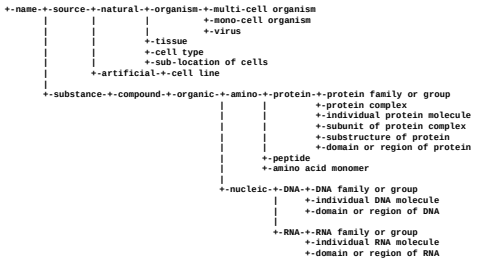


**図2 GENIA オントロジー**
**(物質的な概念の分類)**

GENIA オントロジーは Substance(物質名)、Source（所在）、および Other(その他)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を意味クラスと付与している。コーパスの設計段階では、葉ノードだけではなく、意味クラスとしていずれのノードを選択してもよく、また、複数の意味クラスを付与することも考慮に入れていた。しかし実際にタグ付け作業を進めてみると、作業者に十分な背景知識があれば、葉ノードを割り当てることが可能であることがわかった。さらに、専門用語には当然多義語も存在し、同一の表現で複数の意味クラスに属する可能性があるが、それぞれの出現箇所では、その文脈に依存して一つの意味クラスを割り当てることが可能であった。ただし、現在の GENIA オントロジーでは、物質の化学的な性質のみに基づいて分類しており、機能に関する分類を行っていない。今後、生体内での役割などの物質の機能に関する概念を定義し、現在タグ付けされている用語に対してさらに属性を追加していく予定である。

### 3.2. GENIA 品詞コーパス

MEDLINE アブストラクトに対して、汎用の品詞タガーである Junk tagger[5]を適用すると、Wall Street Journal に対して 96.8%であった精度が 83.5%まで落ちる。そこで我々は、専門用語コーパスと同じアブストラクトセットに対して、品詞タグを付与した GENIA 品詞コーパスを作成した。

```
<abstract>
…
<sentence><cons lex="IL-2-mediated_T_cell_proliferation"
sem="G#other_name"><cons lex="IL-2"
sem="G#protein_molecule">
<w c="*">IL-2</w></cons><w c="JJ">-mediated</w> <cons
lex="T_cell" sem="G#cell_type"><w c="NN">T</w> <w
c="NN">cell</w></cons> <w
c="NN">proliferation</w></cons>…
…
</abstract>
```

**図3 GENIA品詞コーパスの例**

GENIA 品詞コーパスでは、専門用語コーパスの対象アブストラクトの各単語に、Penn Treebank[3]のセットに基づく品詞を割り当てている。まず前処理として、Penn Treebank に付属のトークナイザーを用い

て機械的にトークンに分割し、文脈に依存せずに品詞が確定する語を機械的に処理した後、Junk tagger で初期タグ付けを行い、それを人手で修正することによって品詞タグを付与した。品詞タグは基本的に、Penn Treebank の指針に従って付与しているが、この分野のテキストでは、人名などのいわゆる固有名詞ではない物質名や細胞株名などが大文字で始まる特性を持ち、固有名詞と普通名詞の区別がつきにくい。しかし、固有名詞と普通名詞の区別は構文解析などの次のプロセスにはあまり重要ではなく、細かい基準を作って混乱するよりも安定して判断できることの方が重要であると考え、論文の著者などの人名および研究機関名等を除き、普通名詞として扱うこととした。また、論文アブストラクトは限られた文字数の中で論文の概要を説明しようとすることから、新聞記事などと比較して明らかに等位接続の使用頻度が高い。特に物質名等の専門用語については、トークンよりもさらに狭い単位で共通する接頭語などを共有する語同士の等位接続がある。このような場合には共有する部分を省略して表現されることもあり、トークンが分割されることがある。このように分割されたトークンに対しては、品詞の代わりに"*"を割り当てている。

### 3.3. GENIA 構文木コーパス

構文木コーパスは、専門用語コーパスの対象アブストラクトのサブセットに対して、GDA-DTD[7]を拡張した DTD を用い、Penn Treebank[4]形式の構造を XML 形式で付与している(図 4)。タグ付けの基準は Penn Treebank に基づいているが、作業者間の一致度を目安として検討した結果、スキーマが不徹底であったり解釈が難しい箇所があること、生命科学分野特有の表現の取り扱いなどが不一致の原因であることがわかった。

```
<gda>
…
<S><PP>In <NP>the present paper </NP></PP>, <NP-
SBJ id="i55"><NP>the binding </NP><PP>of <NP>a
[125I]-labeled aldosterone derivative
</NP></PP><PP>to <NP><NP>plasma membrane rich
fractions </NP><PP>of <NP>HML
</NP></PP></NP></PP></NP-SBJ><VP>was
<VP>studied <NP NULL="NONE"
ref="i55"/></VP></VP>.</S>
…
</gda>
```

**図4 GENIA構文木コーパスの例**

特に、生命科学分野特有の表現では、時には専門用語中に前置詞句を含む場合もあるなど、用語が非常に長く、用語内の構造をどこまで分析するのか判断が揺れていた。そこで GENIA 構文木コーパスでは、専門用語内の構造については分析しないこととし、その他の分野特有な表現について、例を示したガイドラインを作成した。

### 3.4. GENIA 相互作用情報コーパス

相互作用情報コーパスは、構文木コーパスの対象アブストラクトのサブセット(500 件)に対して、現在共同研究を行っている Caderige プロジェクト(フランス)の仕様に準じた DTD を用いて、XML 形式で付与した(図 5)。

```
<sentence id="U92176657:4" title="no">

    <genic-interaction regulation="directionless"><gaf1>A T
    cell-specific protein <ga1 direct="yes" type="protein">
    Neg-1</ga1></gaf1> and a ubiquitous protein Neg-2
    <if><i>binding</i> to</if> <gtf1>
    <gt1 type="gene">NRE-I</gt1></gtf1> and NRE-II,
    <if>respectively</if>, were identified.</genic-interaction>

    <genic-interaction regulation="directionless">A T cell-specific
    protein Neg-1 and <gaf1>a ubiquitous protein <ga1 direct="yes"
     type="protein">Neg-2</ga1></gaf1> <if><i>binding</i> to</if>
    NRE-I and <gtf1><gt1 type="gene">NRE-II</gt1></gtf1>,
    <if>respectively</if>, were identified.</genic-interaction>

</sentence>
```

**図5 GENIA相互作用情報コーパスの例**

このコーパス中では、アブストラクト

を文単位に区切って<sentence>エレメントとし、各文が相互作用情報を含むか否かによってまず 3 種類に分類している。エレメント名は、遺伝子性分子同士の相互作用は<genic-interaction>、非遺伝子性分子を含む相互作用は<non-genic-agent-interaction>、<non-genic-target-interaction>のように相互作用の種類を使用している。また、相互作用情報を含まない文は、実験に関連する場合は<experiment>、それ以外の場合は<conteins-no-interaction>としている。さらに、相互作用に関連する要素をエレメント化して、相互作用の種類に応じたタグを付与した。具体的には、肯定文か否定文か(Assertion)、相互作用の性質(Regulation)、相互作用の種類(Type)、表現の確実度(Uncertainty)、情報がその文章で完結しているか否か(Self-contained)、作業者の自身度(Confidence)などである。これらのほかに、文中で相互作用の作用者(Agent)と被作用者(Target)を示す表現や、具体的な相互作用(Interaction)を示す表現、時期や時間、場所、実験手法に関する表現などをタグとして付与した。

500 件のアブストラクトに対してタグ付けを行った結果、総文章数 4,572 中で何らかの相互作用に関する記述のあった文章は 2,942 文(64.4%)であった。また、それぞれの文中に記述されていた相互作用の総数は 5,630 で、平均すると一文中に 1.91 個の相互作用が記述されていたことになる。記述されていた相互作用の内訳は、遺伝子性分子同士の相互作用が 3,180 個(56.5%)、作用者のみ非遺伝子性分子の相互作用が 1,585 個(28.2%)、被作用者のみ非遺伝子性分子の相互作用が 865 個(15.4%)であった。さらに、それぞれの相互作用中での作用者・被作用者の分子種は表１に示すとおりである。

| 遺伝子性分子同士の相互作用 | | | |
|---|---|---|---|
| | タンパク質 | 遺伝子 | RNA |
| 作用者 | 85.91% | 12.82% | 1.27% |
| 被作用者 | 90.37% | 8.75% | 0.88% |
| **作用者のみ非遺伝子性分子の相互作用** | | | |
| | タンパク質 | 遺伝子 | RNA |
| 被作用者 | 53.35% | 39.27% | 7.32% |
| **被作用者のみ非遺伝子性分子の相互作用** | | | |
| | タンパク質 | 遺伝子 | RNA |
| 作用者 | 81.23% | 15.82% | 2.41% |

**表1 作用者･被作用者の分子種**

今回用いた Caderige プロジェクトの仕様に準じた DTD は、微生物に関する論文のアブストラクトを対象に設計されたものであり、GENIA コーパスが対象としているヒトのような多細胞生物に関する論文のアブストラクト中に記述されている相互作用をタグ付けするにはいくつか不都合な点が認められた。そこで、今回のタグ付け結果を参考にして、新しく文中に記述される生物学的なイベント情報をタグ付けするための仕様とイベントオントロジーを設計し、タグ付け作業を開始した。

## 4. まとめ

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ためには、その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ー･辞書･コーパスは必要不可欠なリソースであり、現在世界でもいくつかのグループによ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ひとつとなっている。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ため、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。今回我々は、これまでに作成してきた GENIA 専門用語・品詞・構文木に加え、相互作用情報タグを付与したコーパスを作成した。

**参考文献**


1. GENIA Project. 2005. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
2. MEDLINE. http://www.ncbi.nlm.nih.gov/PubMed/
3. Bio Med Central. http://www.biomedcentral.com/
4. Beatrice Santrini. 1991. Part-of-speech tagging guidelines for the Penn Treebank project. *Penn Treebank II CD-ROM*
5. Jun'ichi Kazama, et al. 2001. A Maximum Entropy Tagger with Unsupervised Hidden Markov Models. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium* (NLPRS2001), pp. 333--340
6. Mitchell P. Marcus, et al. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, Vol.19.*
7. Koichi Hashida. Global Document Annotation. *Proceedings of the Second Natural Language Processing Pacific Rim Symposium* (NLPRS1997)
8. MEDCo Project. http://nlp.i2r.a-star.edu.sg/medco.html
9. Alphonse, E. et al. 2004. Event-based Information Extraction for the biomedical domain: the Caderige project, Workshop BioNLP at Coling 2004.