生命科学分野のタグ付きコーパス：GENIA コーパスの設計と作成

大田朋子[1,2]　建石由佳[1,2]　金進東[1,2]　薬師寺あかね[2]　辻井潤一[1,3]
JST/CREST[1]　東大・院・情報理工[2]　東大・院・情報学環[3]
{okap, yucca, jdkim, akane, 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 という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。実験室で日々量産される実験データを解析・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り、自然言語処理技術の応用が切望されている。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言語リソースが必要不可欠である。そこで我々は、MEDLINE アブストラクトに対して、専門用語・品詞・構文木をタグ付けした GENIA コーパスを作成し、公開している[1]。本大会では、この GENIA コーパス作成のためのタグの設計とタグ付けの際に生じた問題点を報告する。

1. はじめに

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、

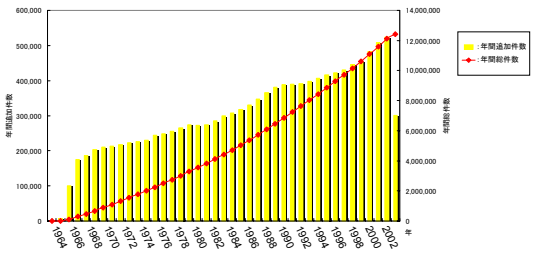

図1　MEDLINE収録アブストラクト数の推移

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。

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。

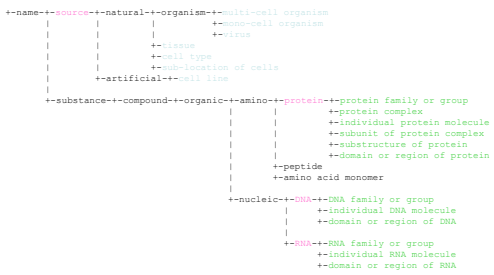

図2 GENIA オントロジー
(物質的な概念の分類)

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ている。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)。

```
<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

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。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。

4. おわりに

現在、専門用語(2000 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）、構文木コーパス（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）を公開[1]している。また Institute of Infocomm Research(シンガポール)における MedCo プロジェクト[8]との共同研究で、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。将来の拡張として、専門用語コーパスに対して新たな概念を定義してタグ付けを拡張すると共に、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。

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