4P－5

医学・生物学文献からのタグ付きコーパスの作成

建石由佳，大田朋子，Nigel Collier，野畑周，辻井潤一
東京大学

はじめに

分子生物学の分野において，近年分子構造についての非常に幅広い情報はデータベースとして整備されてきているが，分子間の相互作用などの高次情報は未だ論文としてしか提供されていない。そこで我々は，医学・生物学分野の論文からの高次情報の抽出を目的とした自然言語処理システム[1]の研究開発を進めており，現時点では，MEDLINE データベース[2]上の論文アブストラクトから統計学習的手法を用いて反応に関わる物質名を自動的に抽出することに重点をおいている[3]。

情報抽出プログラムの評価データ及び統計学習プログラムの学習データとするため，我々は医学・生物学の論文アブストラクトに対して人手でタンパク質名や遺伝子名などをタグ付けした文書（タグ付きコーパス）を作成した。本論文では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について報告する。

タグセットの設計

タグの書式は SGML[4]を採用した。タグ付けの対象は反応に関わる主な物質のうちタンパク質，DNA，RNA の名前，及びそれらの物質の由来を示す生物種・組織・器官・細胞株・細胞などの名前とする。

タンパク質名には PROTEIN タグ，DNA 名には DNA タグ，RNA 名には RNA タグ，物質の由来に関する名前には SOURCE タグを付ける。すべてのタグには id 属性をもたせ，同一物質をさす別名には同じ id 値をもたせることとする。さらに，SOURCE タグには subtype 属性を持たせ，生物種・組織・器官・細胞株・細胞の区別に用いた。

タグ付け実験（実験 1）

MEDLINE から，``human''，``blood cell''，``transcription factor''のすべてを MeSH Header に持つものを選び,そのうち 100 件について著者の 1 人（大田）がタグ付けを行い，タグ付けの作業量を確認した。タグ付けは Mule エディタ上のマクロを用いて行った。

100 件のアブストラクトについてタグ付けするのに約 40 時間を要し，PROTEIN タグ 2125 個所，DNA タグ 358 個所，RNA タグ 30 個所，SOURCE タグ 801 個所が付けられた。タグ付けされたテキスト例を図 1 に挙げる。

UI　- 91092267
TI　- The actions of cyclosporin A and FK506 suggest a novel step in the activation of <SOURCE id=1 subtype=cell-type>T lymphocytes</SOURCE>.
AB　- Cyclosporin A and FK506 are immunosuppressive compounds that have similar inhibitory effects on the expression of several lymphokines produced by <SOURCE id=1 subtype=cell-type>T lymphocytes</SOURCE>.
Despite their similar effects the drugs bind to two different <SOURCE id=2 subtype=sub-location>cytosolic</SOURCE> protein, <PROTEIN id=3>cyclophilin</PROTEIN> and <PROTEIN id=4>FKBP</PROTEIN> respectively, which raises the possibility that they have different modes of action.

図 1：タグ付けテキスト例

アノテータ間の一致に関する実験（実験 2）

タグ付きコーパスの質は，タグが矛盾なく付けられていること，すなわち，ある語句に対してどのようなタグを付けるか（あるいは付けないか）がコーパス全体を通して揺れていないことにかかってい

Building Annotated Corpus from Biology Research Papers
Yuka Tateisi, Tomoko, Ohta, Nigel Collier, Chikashi Nobata, Jun'ichi Tsujii,
University of Tokyo

る。実験 2 は複数のアノテータがタグ付けをする際にも矛盾がないようにできるかどうかを検証するために行った。

まず，タグ付けの基準を記したタグ付けマニュアル[5]を作成した。100 件のタグ付けに際して迷った点については判断基準を決めるとともにマニュアルに実例を記した。

次に，実験 1 で用いたアブストラクト 100 件の中からランダムに選んだ 10 件[1]について，3 人の研究者にボランティアでタグ付けを依頼した。タグ付け作業は紙の上でラインマーカーを用いて行ってもらうこととし，作業用紙とマニュアルを渡すとともに作業及びマニュアルの内容についての簡単な説明を口頭で行った。

タグがどの程度一致しているかを定量的に測るため，2 人ずつペアにした一致率を Message Understanding Conference(MUC)[6]で用いられている方法に基づき，F-スコア[2]により測定した。4 人を 2 人ずつペアにした 6 組のペアの平均の F-スコアは 75.9%であった。これは MUC の結果と比較して十分ではない。

タグ付けされたテキストを比較することによりタグの不一致について分類したところ，タグで囲む範囲がずれているケース(図 2 に例を挙げる)が多く(87 ケース)，同じ個所に別のタグを付けているケースは相対的に少なかった(18 ケース)。このことから，タグ付けの範囲についての基準を明確化する必要があることがわかった。

例 1：
<PROTEIN>IRF-2</PROTEIN> repressor
<PROTEIN>IRF-2 repressor</PROTEIN>

例 2：
<PROTEIN>interleukin-2 (IL-2)</PROTEIN>
<PROTEIN>interleukin-2 </PROTEIN> (<PROTEIN>IL-2</PROTEIN>)

図 2：タグ付けの範囲がずれている例(属性は省略)

[1]この 10 件については実験 1 でタグを付けたものと合せて 4 人のアノテータによりタグ付けされることになる。

[2] 適合率と再現率の調和平均

現状

実験 2 の結果を受け，アノテータ間での不一致の個所，特にタグ付けの範囲が一致しない例を見直し，個々のケースについてどう判断すべきかを決めたタグ付け基準を作成し，マニュアルを改訂した[7]。現在，新しい基準に基づいて実験 1 の 100 件のアブストラクトをタグ付けし直すとともに，新しいテキストをタグ付けし，コーパスを拡張している。

おわりに

情報抽出プログラムの学習，テストのために用いるため，医学・生物学の論文アブストラクトに対して人手でタンパク質名や遺伝子名などをタグ付けした文書（タグ付きコーパス）を作成した。

将来は他の物質（脂質，炭水化物など）にタグ付けの範囲を広げる，属性（生物学的役割など）を増やす，など，タグセットの拡張も行う予定である。

参考文献

1. N. Collier et al., ``The GENIA Project: Corpus-based Knowledge Acquisition and Information Extraction from Genome Research Papers", Proc. EACL'99, 1999.
2. http://www.ncbi.nlm.nih.gov/PubMed
3. C. Nobata et al., ``Automatic Term Identification and Classification in Biology Texts", to appear in Proc. NLPRS 99, 1999.
4. ISO/IEC 8879, ``Standard Generalized Markup Language", 1986
5. http://www.is.s.u-tokyo.ac.jp/~okap/annotate-bio.html
6. Proceedings of Sixth Message Understanding Conference (MUC-6), 1995.
7. http://www.is.s.u-tokyo.ac.jp/~okap/annotate-bio-new.html