自然言語処理 133－13
（1999. 9. 10）

医学・生物学論文からのタグ付きコーパスの作成

大田朋子，建石由佳，Nigel Collier，野畑周，辻井潤一

東京大学大学院理学系研究科

概要

情報抽出プログラムのテストデータおよび学習データとして使用するため，医学生物学分野の論文アブストラクトに物質名などをマークアップしたタグ付きコーパスを作成した。タグ付けにあったって,物質の概念モデル(オントロジー)を作成し，それにもとづいたタグセットをSGMLで定義した。
MEDLINE データベースに登録されたアブストラクト 100 件をタグ付けして初期コーパスとするとともに，われわれとは独立な研究者にもタグ付けを依頼し,タグの信頼度を検証した。検証の結果，タグ付け作業者間で揺れの大きかった点についてタグおよび作業指示書の改定を行った。

Building Annotated Corpus From Biomedical Research Papers

Tomoko Ohta, Yuka Tateisi, Nigel Collier, Chikashi Nobata, Jun'ichi Tsujii

Graduate School of Science, University of Tokyo

Abstract

As a part of a project on information extraction from the research papers in genome domain, we are creating an expert-tagged corpus of MEDLINE abstracts which will be used for training and testing the information extraction systems. The markup scheme is based on a conceptual domain model (ontology) and implemented in SGML. We created a preliminary corpus of 100 MEDLINE abstracts, and also conducted a cross-validation experiment with independent biologists.

1 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。そこでわれわれは，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および統計学習プログラムの学習データとするため，われわれは医学・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にタグ付けした文書（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。

2　タグセットの設計

タグ付けは,作業者に理解しやすく,明確な基準に基づいて行わなければならない。われわれは，まず医学生物学分野の概念モデル（オントロジー）を構築し,それに基づいたタグセットを設計することにした。タグの書式はSGML[4]を採用した。

2.1　オントロジー

近年,物質の分子構造や性質についての情報が膨大なデータベースとして整備されていくにともなって,それらに対する検索を半自動化する必要が認識された。このため，現在は研究者個人が持っている分子生物学分野の広範な知識を機械可読な形でオントロジーとして構築する研究が行われ[5,6]，独立に作成された種々のデータベースを統合したビューの作成に応用されつつある[7]。

われわれは図 1 に示すような概念の階層構造を仮定し，これに基づいてタグセットを設定することとした。現時点では単純な is-a 関係に過ぎないが，これを拡張してオントロジーを構築していく予定である。図 1 の階層構造を作る際に，われわれは物質を生体内での役割でなく化学的（構造的）性質により分類することにした。これは，化学的性質による分類のほうが生体内の環境（したがって論文中の文脈）に依存することが少ないと考えたからである。たとえば，あるタンパク質の化学的性質は構造式などの形で定義できるが，そのタンパク質は生物種によって酵素として働いたり毒物となったりしうるからである。

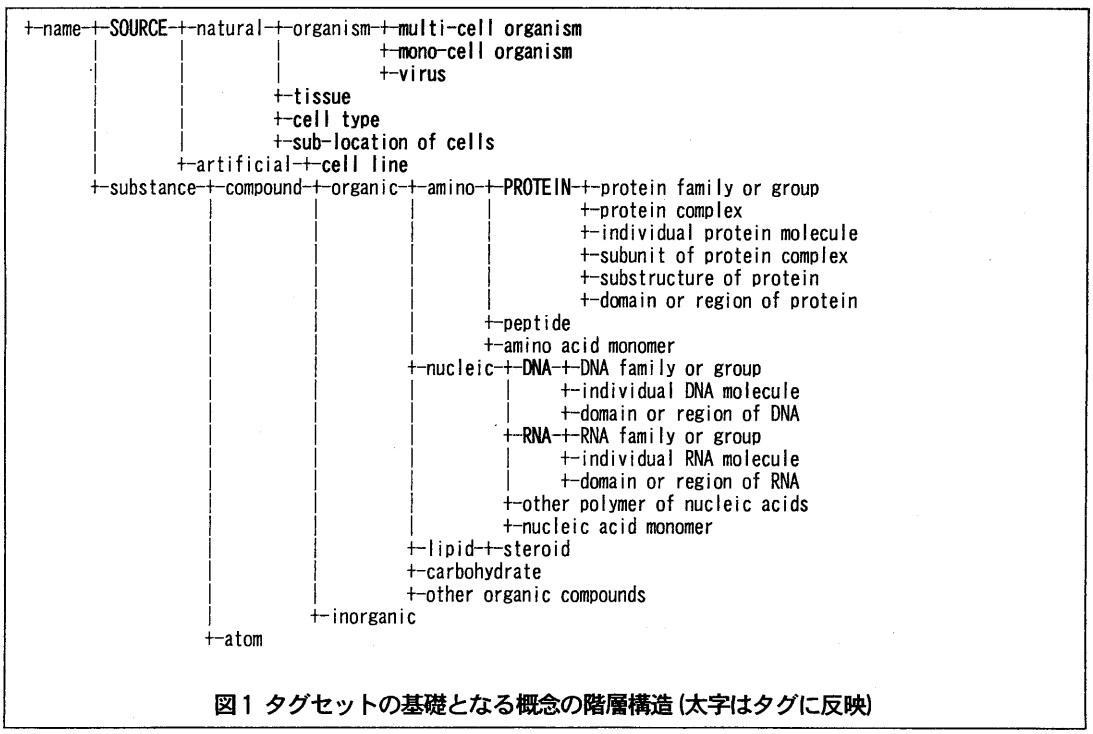

図1 タグセットの基礎となる概念の階層構造 (太字はタグに反映)

2.2　タグセット

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，およ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。<SOURCE>タグには subtype 属性を持たせ，生物種・組織・器官・細胞株・細胞の区別に用いる（表 1）。

表 1：subtype 属性の値とタグ付けの対象

値	対象
org-multi	多細胞生物
org-mono	単細胞生物
Virus	ウイルス
Tissue	組織・器官
cell-type	細胞
cell-line	細胞株
sub-location	細胞内の小器官

物質名・生物種名には同一物質・同一種をさす別名が存在することがある（たとえば生物種名の human と homo sapiens など）。また，正式名と略称が並行して使われることもある。これらの別名，略称を明示するために，すべてのタグには id 属性を持たせ，同一物質(種)をさす名前は同じ id 値をとるようにした。

物質名の中には，他の物質，種，細胞などの名を含むものがある。たとえば，RNA の名である T helper cell mRNA には，細胞の名である T helper cell が含まれる。このようなときは，タグを入れ子にせず，一番広い範囲にのみタグを付けることとした。

タグ付けしたテキストの例を図 1 に示す。

3 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>.
4 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2：タグ付けテキスト例

3 タグ付け実験

3.1 実験 1：タグ付けの作業量

MEDLINE から，``human''，``blood cell''，``transcription factor''のすべてを MeSH Header[8]に持つものを選び,そのうち 100 件について著者の 1 人（分子生物学分野の博士号を持つ）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。

3.2 実験 2：作業者間の一致

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る。

実験 2 は複数の作業者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[9]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（医学部病理学科講師，医学部小児科医員，基礎工学部助手）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およ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)[10]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。結果は表 2 に示す。

表 2：作業者間の一致率（F-スコア）

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	平均
A0-A1	1.000	0.691	0.382	0.828	0.698	0.839	0.741	0.833	0.883	0.917	0.773
A0-A2	1.000	0.610	0.661	0.677	0.805	0.723	0.727	0.901	0.842	0.714	0.768
A0-A3	0.952	0.591	0.576	0.966	0.861	0.838	0.696	0.796	0.857	0.849	0.784
A1-A2	1.000	0.838	0.412	0.606	0.620	0.672	0.830	0.777	0.808	0.787	0.726
A1-A3	0.952	0.525	0.667	0.786	0.615	0.766	0.778	0.707	0.797	0.765	0.728
A2-A3	0.952	0.513	0.473	0.636	0.850	0.761	0.855	0.820	0.836	0.654	0.739
平均	0.976	0.628	0.528	0.750	0.742	0.767	0.772	0.806	0.837	0.781	0.759

注）T1-T10 はアブストラクト，A0-A3 は作業者

次に，目視で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の原因についての分析を試みた。これらは以下のように特徴づけられた[3]。

タグで囲む範囲がずれている（87 ケース）

① 後ろに続く名詞をタグで囲む範囲に入れるか入れないかで揺れる。この場合，後ろに続く名詞をとるかとらないかによってタグ付けされた名前のさす物質が変わる（タグの種類が変わることもある）場合（例 1）と同じ場合（例 2）がある（48 ケース）。

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

例 2　<PROTEIN>Stat 91 protein</PROTEIN>
　　<PROTEIN>Stat 91</PROTEIN> protein

② 正式名称と略称が続いて現れる場合，それらを 1 つにまとめるか分けるかで揺れる（13 ケース）。

例 3　<PROTEIN id=1>interleukin-2(IL-2)</PROTEIN>
　　<PROTEIN interleukin-2</SOURCE> (<PROTEIN id=1>IL-2</PROTEIN>)

③ 同格句がある場合，まとめてタグ付けする・同格句と別々にタグ付けする・同格句にはタグ付けしない，で揺れる（8 ケース）。

例 4　<SOURCE id=1 subtype=cl>U937</SOURCE>, a human monocytoid cell line

---

[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 人の作業者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

[3] 例では属性は省略する。

<SOURCE id=1 subtype=cl>U937, a human monocytoid cell line</SOURCE>

例 5 <PROTEIN id=1>transcription factor AP-2</PROTEIN>

transcription factor <PROTEIN id=1>AP-2</PROTEIN>

④「生物種,細胞などの名＋物質名」という形の名前で，全体をタグで囲むか分けるかで揺れる（6 ケース)。

例 6 <PROTEIN>Human erythroid 5-aminolevulinate synthase</PROTEIN>

<SOURCE>Human erythroid</SOURCE> <PROTEIN>5-aminolevulinate synthase</PROTEIN>

⑤ 直前の(限定)修飾句を含めるか含めないかで揺れる（6 ケース)。

例 7 <RNA>housekeeping ALAS mRNA</RNA>

housekeeping <RNA>ALAS mRNA</RNA>

⑥ 接続詞 and または or でつながれた名詞句をまとめるか分けるかで揺れる（3 ケース)。

例 8 <RNA>ferritin or transferritin receptor mRNAs</RNA>

<PROTEIN>ferritin</PROTEIN> or <RNA>transferritin receptor mRNAs</RNA>

⑦ タンパク質複合体の名前を 1 つにまとめるか構成成分に分けるかで揺れる（2 ケース)。

例 9 <PROTEIN>p50-p65</PROTEIN>

<PROTEIN>p50</PROTEIN>-<PROTEIN>p65</PROTEIN>

⑧冠詞をタグで囲む範囲に含めるかどうかで揺れる（1 ケース)。

作業者によってタグが付けられたり付けられなかったりする（25 ケース）

⑨ 物質の部分構造をタグ付けするかしないかで揺れる（14 ケース)。

例 10 <DNA>21 bp repeats</DNA>

⑩ 物質のファミリー（サブクラスに当たる）をタグ付けするかしないかで揺れる（7 ケース)。

例 11 <PROTEIN>cytokine</PROTEIN>

⑪ その他（4 ケース)。

同じ個所に別のタグを付けている（19 ケース）

⑫同一の名前が別の意味で使われうるときに判断が揺れる（13 ケース)。

例 12 <DNA>TCF-1</DNA>

<PROTEIN>TCF-1</PROTEIN>

3.3 考察

実験 2 における不一致の分類のうち②，⑧，および①のうち例 2 のようにタグ名と同じ語があとに続くものについては，単純にマニュアルの記述の不備によるものと考えられる。また，⑥は省略をともない，入れ子を許さない現在のタグ書式では対応がむずかしいものである。また，以下に挙げるように，この分野特有の困難さがあると推測されるケースがある。

物質の概念が研究者によって揺れている 作業者によってタグが付けられるかどうかが別れたケース（⑨，⑩）は，物質のファミリーや部分構造が研究者によって物質であると認識されているかどうかが揺れること，⑦は複合体を 1 つの物質と認定するかしないかが研究者によって揺れることを示す。

物質の性質の記述と名前の区別があいまいになっている 新たに発見・同定された物質の名前を付ける際，その性質の記述そのものを名前とすることが良く行われる。たとえば，"B-cell specific transcription factor"などはタンパク質の名前であるが，特に語頭を大文字であらわすこともなく書かれるので，物質名なのか，（この例の場合では「B 細胞に特有の転写因子」という）クラスをあらわしているのか一見しただけでは判断が難しい。このため，「物質名＋一般名詞」の形全体を名前とすべきか物質名を含

むクラスととらえるべきか（①の例１），同格句を物質の性質の説明ととらえるべきか物質名ととらえるべきか（③），限定修飾句を含んだ全体を物質名とすべきか限定句を除いたものを物質名とすべきか（⑤，また④もこの特殊なケースと考えられる）などの判断に非常に高度な専門知識が要求され，分野の研究者であっても迷いや判断ミスが生じると考えられる。

タンパク質を転写する遺伝子をタンパク質名で参照する たとえば，タンパク質 NF-kappaB を転写する DNA を NF-kappa B gene と呼ぶなどが慣習的に行われているが，論文の著者によっては，文脈から推測可能と判断するときはこの gene を省略してしまう。これが付けるべきタグの種類に対する判断ミスの原因になる（⑫）と考えられる。

不一致を減らす対策として，物質のファミリー・部分構造・複合体などの区別を属性として定義し，ファミリー，部分構造などについてもこれらの区別を付けた上でタグを付けるように改めた。また，「生物種,細胞などの名＋物質名」，「物質名＋一般名詞」の形についてはタグ付けの時点では統一的に全体を 1 つの名としてタグ付けすることにし，個々のケースについて後からデータベースとの照合などによってチェックし直すことにした。

4 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。現在，実験 2 を受けてタグ付けの基準およびマニュアルの記述を改め[11]，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。将来はオントロジーの拡張と並行して，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。

5 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. S. Sculze-Kremer, ``Ontologies for Molecular Biology", Proc. PSB98, 1998
6. P. G. Baker et al., ``An Ontology for Bioinformatics Applications", Bioinformatics, Vol 15 No 6, pp 510-520, 1999
7. P. G. Baker et al., ``TAMBIS:Transparent Access to Multiple Bioinformatics Information Sources", Proc ISMB98, 1998.
8. http://www.nlm.nih.gov/mesh/meshhome.html
9. http://www.is.s.u-tokyo.ac.jp/~okap/annotate-bio.html
10. Proceedings of Sixth Message Understanding Conference (MUC-6), 1995
11. http://www.is.s.u-tokyo.ac.jp/~okap/annotate-bio-new.html