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

2004－NL－159　(18)
2004／1／14

ライフサイエンス分野を対象とした低レベルのテキスト処理

山本 薫　小長谷明彦
理化学研究所
ゲノム総合科学センター ゲノム情報科学グループ
{kaorux,konagaya}@gsc.riken.go.jp

佐藤 賢二
北陸先端科学技術大学院大学
知識科学研究科
ken@jaist.ac.jp

ライフサイエンス分野テキストに特有な言語現象を調査し、分析結果を足掛かりに、GENIA Corpus 3.02 のわかち書き仕様と品詞体系の変更箇所を提案し、統計的自然言語処理の手法を用いて形態素解析システム「cocab」を実現した。
予備実験で、今回提案した仕様でタグ付与したコーパスから学習したモデルは、GENIA Corpus 3.02 から直接学習したモデルより、誤り率が改善されることを確認した。

キーワード：わかち書き、品詞付与、原形変換、ライフサイエンス

Low-level Text Processing for Life Science

Kaoru Yamamoto　Akihiko Konagaya
Bioinformatics Group, Genome Sciences Center
RIKEN
{kaorux,konagaya}@gsc.riken.go.jp

Kenji Satou
Graduate School of Knowledge Science
Japan Advanced Institute of Science and Technology
ken@jaist.ac.jp

We investigate language phenomena specific to PubMed abstracts, propose positive modifications to tokenization and part-of-speech tagging guidelines in GENIA Corpus 3.02, and implement morphological analyzer “cocab” using statistical natural language processing techniques.
Preliminary experiments show that the model trained from the modified corpus has improved the error rate from the original corpus.

Keywords：tokenization, part-of-speech tagging, normalization, life science

1 はじめに

ヒトゲノム塩基配列の解読宣言にも象徴されるように、ライフサイエンス分野の課題は、配列同定から機能同定にシフトしつつある。現在、遺伝子ネットワークの構築など基礎研究からオーダーメイド医療のような応用研究まで、分野を横断して研究が進められている。

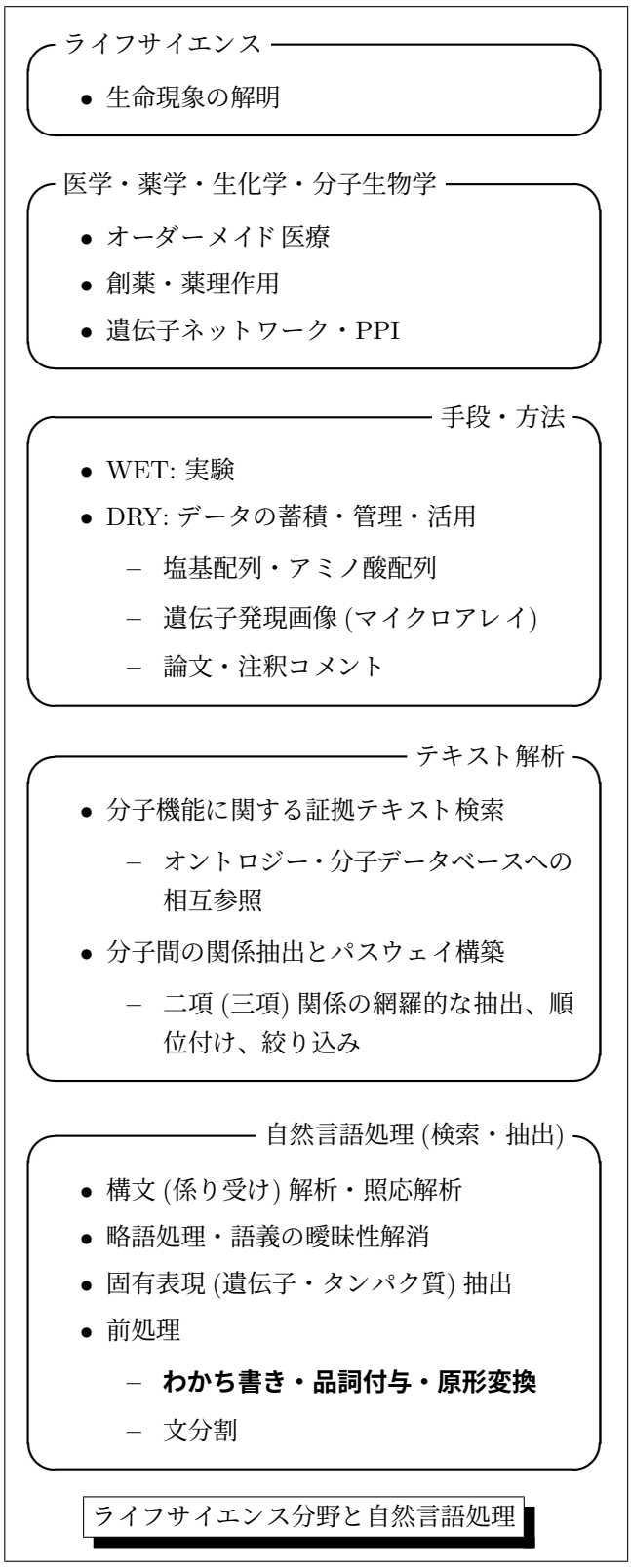
ライフサイエンス分野では、実験をしながら生命現象の説明を試みる WET 派の研究活動から、データ分析やシミュレーションしつつ生命現象の説明を試みる DRY 派の研究活動まで混在する。そして、あらゆる情報源と手段を使って、遺伝子の機能を同定する作業が進められている。

近年、機能同定のために、自然言語処理や情報検索などで、大量で多様な生テキストから埋もれている情報を抽出するという手段が注目を集めている。対象とするテキストは、PubMed[1]に集められた論文アブストラクト以外に、BioMed Central[2] から電子発行されているフルペーパー、分子データベースに注釈として記述されているコメント部分など、多岐にわたる。TREC Genomic Track[3] や BioCreAtIvE[4] などが組織され、テキスト解析の共通タスクとして、分子機能に関する証拠テキストの検索や分子の機能や関係の抽出などが提案されている。

ライフサイエンス分野の科学者にとって、テキスト解析は一手段で、自然言語処理技術は道具である。既存の自然言語処理ツールは、新聞記事を前提にしているため、分野が異なるライフサイエンス分野テキストでは、情報として重要な専門用語の周辺で解析誤りが起こりやすい。結果、テキスト解析で、当初期待していたほどの効果を得るに至っていない。

我々は、ライフサイエンス分野テキストの特徴を考慮した頑健な低レベル処理を整備すれば、状況は改善されると考えた。そこで、本研究では、ライフサイエンス分野テキスト処理で需要が高い PubMed アブストラクトを対象とした **泥くさい低レベルのテキスト処理** を課題とした。2002 年度から、既存の統計的自然言語処理の手法を使って、ライフサイエンス分野のテキストを対象とした形態素解析システム「cocab」を開発している。これまで我々が得た経験と、cocab の概要を伝えるのが本稿のねらいである。

以下、第 2 節では、ライフサイエンス分野テキスト特有の言語現象ついてまとめる。第 3 節では、前節の論点を踏まえた低レベルのテキスト処理ツール開発について報告する。第 4 節で、得られた成果と今後の課題について述べる。最後に、第 5 節でまとめる。



ライフサイエンス分野と自然言語処理

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
[2] http://www.biomedcentral.com/info/about/datamining/
[3] http://medir.ohsu.edu/~genomics/
[4] http://www.pdg.cnb.uam.es/BioLINK/

2 ライフサイエンス分野テキスト

本稿で扱うライフサイエンス分野テキストとは、PubMed に登録されている分子生物学に関連した英語の科学論文アブストラクトとする。これらのテキストでは、専門的な内容が理解可能な科学者への情報伝達しか念頭においていないため、造語が生成されやすい。そのため、正式表記からダッシュ(-) やスラッシュ(/) など特殊文字を任意に挿入・削除して生成される派生表記、略語、語でないもの (塩基配列の断片やデータベースの ID) などが多く出現する。

上記の特徴は、Penn Treebank Corpus [4] などの新聞記事では、稀れ、もしくは、全く観察されない現象である。従来のライフサイエンス分野を対象としたテキスト解析は、この事実を認識はしているが、新聞記事など異なる分野のデータを前提としたわかち書き仕様と品詞体系に基づくツールで代用している [6]。新聞記事で開発したツールは、未知現象に対しても新聞記事の特徴から推定する。そのため、分野を違えた場合、専門用語の大半が未知のため解析誤りをおこし、結果的に上位タスクに悪影響を及ぼすことが、経験的に知られている。

本節では、Penn Treebank とほぼ同様のわかち書き仕様と品詞体系でタグ付与された GENIA Corpus 3.02 [1] を参照しながら、ライフサイエンス分野テキストに特有な言語現象を分類し、問題点を整理する。

2.1 GENIA Corpus 3.02

GENIA Corpus 3.02 は、独自のオントロジーに基づく意味体系のタグが付与されたコーパス (意味コーパス; `GENIAcorpus3.02.xml`) と、Penn Treebank とほぼ同様の仕様でわかち書きや品詞情報が付与されたコーパス (品詞コーパス; `GENIAcorpus3.02.pos.xml`; 品詞体系は付録 1 を参照) と、両者を統合したコーパスの (統合コーパスと呼ぶ; `GENIAcorpus3.02.merged.xml`) の 3 種類から構成されている。Human, Blood Cells, Transcription Factors という MeSH[5] カテゴリが付与されている 約 2000 アブストラクトが収録されており、ライフサイエンス分野テキストの言語処理用タグ付きコーパスとしては世界最大級である。

統合コーパスでは、品詞コーパスのトークンが意味コーパスの境界と一致しない場合、残片が * でタグ付与されている。例を表 1, 2 に示す。**ライフサイエンス分野テキストのタームは境界判定が困難であり、必ずしも空白文字と隣り合っているわけではない。また、特殊文字を含む語の品詞は自明でない。**表 3 に、本稿で提案するわかち書きと品詞付与を示す。以下に、この提案に至った動機について述べる。

[5] `http://www.nlm.nih.gov/mesh/meshhome.html`

2.2 1 トークン $\neq$ 1 見かけ語

本稿で使う用語の整理をする。トークンとは、わかち書き処理で同定された最小単位を指すことにする。一方、見かけ語 (graphic word[6]) とは、空白文字で区切られた単位とする。

英語は、中国語や日本語と異なり、空白文字が境界として機能し、punctuation などの一部の例外を除いて、「1 トークン = 1 見かけ語」とされてきた。自然言語処理では、ペンシルベニア大学が配布する sed tokenizer[7]が多く使われ、一部の例外を除き、空白文字でトークンを同定している。しかし、上の仮定は、次のような理由で、危険である。

2.2.1 1 トークンの長さ < 1 見かけ語の長さ

ライフサイエンス分野のタームは、必ずしも空白文字で区切られているとは限らない。例えば、表 1 で、 ER は G#protein_family_or_group とタグ付けがあるが、見かけ語の部分文字列である。GENIA Corpus 3.02 において、専門家が意味付与しているタームの定義も自明ではないが、少なくとも現時点で意味コーパスと品詞コーパスで境界が一致していない事例が散見された。そこで、集計したところ、全体で約 8 %の不一致度が観察された。内訳を表 4 にまとめる。この結果は、Penn Treebank とほぼ同様のわかち書き仕様で定義した品詞コーパスのトークン単位から、まとめあげ (チャンキング) による固有表現抽出を行なっても、再現率の上限は 92 % に留まることを示している[8]。

自然言語処理では、固有表現抽出をチャンキング問題とみなし、チャンクタグを最大エントロピー法やサポートベクタマシーンなどで学習する手法が一定の成功を納めている。BioCreAtIve でも、固有表現抽出は、検索エンジンのインデックス作成や二項関係抽出などの重要な要素技術と認定されており、高い精度の実現が求められている。テキストに出現する遺伝子名やタンパク質名を抽出するタスクに、固有表現抽出で培った知見を応用するのは自然な流れだと考えるが、その際、分割不足なトークンの存在は、不必要に精度を下げる原因となる。

2.2.2 1 トークンの長さ > 1 見かけ語の長さ

ライフサイエンス分野に頻出するラテン語由来の学名 (E. coli; Escherichia coli) や学術熟語 (in vivo;

[6] Foundations of Statistical Natural Language Processing では 125 ページに、*Kučera and Francis* (1967) による graphic word の定義が引用されている。"a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks." [2]

[7] `http://www.cis.upenn.edu/~treebank/tokenizer.sed`

[8] 境界が一致しない現象は、意味コーパスのターム内部のタグに出現していることが多い。内部タグは見落されている可能性もあり、表 4 の不一致度は下限であると見積るべきであろう。

表 1: “ER-mediated repression” に対する、意味コーパス、品詞コーパスそれぞれのタグ付与

生テキスト	*E*	*R*	-	m	e	d	i	a	t	e	d		r	e	p	r	e	s	s	i	o	n
品詞コーパス	**JJ**												**NN**									
意味コーパス	G#protein																					
	G#other_name																					

表 2: “ER-mediated repression” に対する統合コーパスのタグ付与; 境界が一致せず残片 (*) を含む例

生テキスト	*E*	*R*	-	m	e	d	i	a	t	e	d		r	e	p	r	e	s	s	i	o	n
統合コーパス	*****		**JJ**										**NN**									
	G#protein																					
	G#other_name																					

表 3: “ER-mediated repression” に対する本稿で提案するわかち書きと品詞付与; 意味タグは変更せず、ターム境界だけわかち書きに反映させる

生テキスト	*E*	*R*	-	m	e	d	i	a	t	e	d		r	e	p	r	e	s	s	i	o	n
提案コーパス	**NN**		-	**VBN**									**NN**									
	G#protein																					
	G#other_name																					

表 4: (a) 意味タグと品詞タグの境界が一致しない箇所の頻度; (b) 意味タグ本来の頻度; (c) それぞれの意味タグにおける不一致度 [(a) の数を (b) の数で除じたもの]

	(a)	(b)	(c)
全体	6686	83890	0.08
G#protein	2963	32303	0.09
G#other_name	1490	17885	0.08
G#organic	706	6618	0.11
G#cell	524	10530	0.05
G#(DNA|RNA)	523	10492	0.05

表 5: 特殊文字 (-/) と表記の揺らぎ

(1)	T cell receptor-CD3 complex
(2)	T cell receptor/CD3 complex
(3)	TcR-CD3 complex
(4)	TCR/CD3
(5)	CD3-T-cell receptor complex

in vitro; in silico) は、空白文字を含んでいる。しかし、これらは、1 トークンとして扱うほうが、E. が文末かどうか、in が前置詞かどうかなど、余計な曖昧性を排除する効果があると考える。トークンを空白文字で過分割してしまうと、逆に曖昧になる場合があり、解析精度が下げる原因になりうる。

2.3 特殊文字を含む見かけ語の品詞

ライフサイエンス分野テキストでは、特殊文字が頻出するが、その機能は、文脈を理解できる専門家でない限り、わからない場合がよくある。特に、ハイフン (-) と スラッシュ (/) は、前後と容易に連結できるため、複合語が作られやすくなる。そして、表 5 の (1) と (5) のように - の前後の反転現象も可能になり、表記の揺らぎの原因となっている。

付録 1 に、本稿で議論する Penn Treebank と GENIA Corpus 3.02 の品詞体系の要点をまとめた。Penn Treebank も GENIA Corpus 3.02 も、ハイフンで連結された見かけ語は、JJ とタグ付けされてることが多いが、NN と NN で囲まれた場合は NN とタグ付けする。GENIA Corpus 3.02 の品詞体系では、Penn Treebank と違い、記号に対して SYM を廃止し、そのトークンが暗示している品詞を付与する仕様になっている。例えば、+ は positive を意図しているのであれば JJ 、and なら CC といった具合である。

一般的に、特殊文字の用法は、文法制限を受けにくいため、多機能で曖昧である。分野の背景知識がない限り、特殊文字を含む見かけ語の品詞は自明ではなく、品詞タグ付与が揺れる可能性が高い。例えば、表 1 では、ER-mediated が JJ になっているが、“repression has been mediated through ER” と書き換え可能で、mediated の品詞は VBN である。 表 2 のように、ER-mediated が JJ とするのは理解できるが、意味タグと不整合だった残りの部分 -mediated を JJ とするのに、疑問が残る。今度は、ハイフンが任意の位置に挿入できるため、見かけ語の品詞の一貫性が乱れる例を示す。表 6 に、タンパク質 GM-CSF と同義語を載せた。ハイフンの有無によってトークン単位とそれに伴う品詞が異なる。参考までに、タンパク質 G-CSF の品詞は、ハイフン連結にもかかわらず NN とタグ付けされている。

このように、「1 トークン = 見かけ語」とした場合、品詞が自明でないため、モデル学習に不可欠なコーパスの一貫性が保たれにくくなる。しかも、表記の揺らぎのために辞書エントリが不必要に増加し、学習時に深刻なデータスパースネス現象に陥る可能性がある。

表 6: ハイフンの有無によって見かけ語の境界が変わり、付与された品詞が異なる例

	GM-CSF NN								
97362232	granulocyte-macrophage JJ			colony-stimulating JJ			factor NN		
99172186	granulocyte-macrophage JJ			CSF NN					
99048705	granulocyte NN		macrophage-colony NN			stimulating NN			factor NN

表 7: macrophage がない例; granulocyte-macrophage の品詞は JJ だが、granulocyte-colony の品詞は NN; macrophage も colony も 1 見かけ語で出現するときの品詞は NN

	G-CSF NN					
97467112	granulocyte-colony NN		stimulating NN		factor NN	

3 低レベルのテキスト処理

我々は、前節の問題点を踏まえて、GENIA Corpus 3.02 のわかち書き仕様と品詞体系を変更した。そして、統計的自然言語処理の手法を使ってモデル学習を行ない、形態素解析システム「cocab」を実現した。本節では、cocab の基本的な考え方を述べ、特徴について説明する。

3.1 基本的な考え方

3.1.1 「1 トークン $\neq$ 1 見かけ語」

GENIA Corpus 3.02 のわかち書き仕様と異なり、見かけ語に捕らわれないトークンを定義する。そのために、cocab では、2.2.1 節の「1 トークンの長さ $<$ 1 見かけ語の長さ」(以下、短い単位) も 2.2.2 節の「1 トークンの長さ $>$ 1 見かけ語の長さ」(以下、長い単位) も扱えるような仕組みを用意する。今回は、山下らの方法 [8] を採用し、ライフサイエンス分野における英語を日本語のように見かけ語を持たない言語とみなして辞書検索を行ない、トークン境界の曖昧性解消と品詞の曖昧性解消を同時に行なう。

短い単位も長い単位も許容するので、コーパスに様々な単位が混在する問題が発生する。我々は、できる限り、トークンと最小単位が一致するのが望ましいと考え、わかち書き仕様は、長い単位 (トークンが分割不足な状態) より短い単位 (トークンが過分割である状態) の方を優先するようにした。理由は、固有表現抽出などの上流のタスクを考えた場合、過分割の方が後処理しやすいからである。

3.1.2 語彙化品詞の導入

英語の to には、前置詞と不定詞の用法があるが、Penn Treebank の仕様では、品詞の曖昧性解消をせず、TO という語彙化品詞を採用している。

2.3 節で見たように、特殊文字を含む見かけ語の品詞は曖昧である。cocab では、人間が判断に迷う特殊文字に関しては、品詞の曖昧性解消を行なわず、特殊文字だけをトークンとして切り出して、解析処理を簡単化する。GENIA Corpus 3.02 の品詞体系と、語彙化した品詞を導入する点が異なる。

何を語彙化品詞にすべきかは、議論の余地がある。今回は、固有表現抽出で綴り素性として有効だといわれている特殊文字 (表 8 参照) に限定する。語彙化品詞の導入により、表 6 と表 7 に例示したようなタグの齟齬を、ある程度、抑制できる。

3.1.3 付加情報が追加可能

ライフサイエンス分野におけるテキスト解析では、わかち書きや品詞以外の情報も有用な手がかりになりうる。辞書エントリに付加情報が簡単に追加できる仕組みを用意し、使いやすさを意識する。

cocab のわかち書きと品詞付与のモデル学習には直接関係ないが、以下にあげる付加情報を追加する[9]。

- GENIA Corpus 3.02 の意味タグ
  階層構造で最下位に位置する意味タグ (innermost tag) のみを付与する。
- 活用語の原形
  動詞など活用する語の原形を登録する。例えば、activates, activating, activated のエントリに対して、activate を登録する。英辞郎、UMLS、Sussane Corpus を利用した。
- 派生語の動詞語幹
  GENIA Corpus で頻出する動詞から派生した名詞に関して、元になった動詞を登録する。例えば、activator, activation のエントリに対して activate を登録する。

遺伝子やタンパク質を主格・目的格にとる動詞の活用原形や派生語の情報を付与する目的は、ライフサイエンス分野のテキスト解析でステミングと類似効果を提供するためである。

[9]その他、略語と正式名称のマッピング、分子データベースの ID などが有用だと思われる。

表 8: cocab で語彙化品詞にした特殊文字

特殊文字	.	,	;	:	“	”	’	(	)	[	]	{	}	-	+	/
語彙化品詞	.	,	:	:	”	”	’	(	)	(	)	(	)	-	+	/

3.2 形態素解析システム「cocab」

cocab は、マルコフモデルによる形態素解析システムで、トークン候補から構成されるトレリスを構築し、ビタビアルゴリズムで最適解を選択する、という古典的な統計的自然言語処理の手法を用いる。

形態素解析は、以下の式による $\langle W^*, T^* \rangle$ の確率が最大になるようなトークン列 $W^* = w_1^*, \ldots, w_n^*$ と品詞タグ列 $T^* = t_1^*, \ldots, t_n^*$ を求める問題と定義できる。

$$
\begin{aligned}
\langle W^*, T^* \rangle &= \arg\max_{\langle W,T \rangle} P(T|W) \\
&= \arg\max_{\langle W,T \rangle} \frac{P(W|T)P(T)}{P(W)} \\
&= \arg\max_{\langle W,T \rangle} P(W|T)P(T) \\
&\simeq \arg\max_{\langle W,T \rangle} \prod_i p(w_i|t_i)p(t_i|t_{i-2}, t_{i-1})
\end{aligned}
$$

モデルに必要な確率は、品詞連接確率 $p(t_i|t_{i-2}, t_{i-1})$ と単語生起確率 $p(w_i|t_i)$ である。実際の実装は、積演算より和演算の方が効率的に処理できるという理由から確率の逆数の対数に適当な係数をかける整数値 (コスト) を用いる。コスト最小法と呼ばれるコストの和演算で最適な解を選択する方法で実現している。コスト最小法に必要なパラメータは、品詞タグ付きコーパスを用いて学習する。

3.3 実験

3.1 節で述べた基本的な考え方に基づき、ランダムに選んだ 100 アブストラクトに対して、人手で、コーパス仕様変更を反映した。ライフサイエンス分野テキストの特徴を前提としたわかち書き仕様と品詞体系に変更することで、どの程度、わかち書きと品詞付与のタグの齟齬が抑制できるかを観察する。

3.3.1 データ

モデル学習に利用したコーパス仕様の変更点を列挙する。

- 見かけ語に表 8 の特殊文字が含まれている場合 (統合コーパスで * となっている箇所も該当) は分割し、特殊文字は語彙化品詞を、残片については次の基準で品詞を付与する。
  - 数字 (Roman numeral) なら CD
  - アルファベット (Greek alphabet) なら NN
  - 意味タグ (innermost) もあるなら NN
  - 単位 (kDa など) は NN
  - 一般名詞なら NN
  - 形容詞なら JJ
  - 過去分詞なら VBD
  - 現在分詞なら VBG
  - 上記以外は、元の品詞を尊重
- ラテン語由来の学名、学術熟語、遺伝子の端を示す 3’ や 5’ は、分割しない。
- 意味タグの境界を尊重し、分割不足を除去する。
- 意味タグが付与している箇所で、品詞タグが NN 以外の場合、品詞を見直す。

表 9: わかち書きと品詞体系の変更前と変更後; 変更前は、GENIA Corpus 3.02 統合コーパス; 変更後は、提案した変更点を反映したコーパス

	変更前		変更後	
トークン数	21061		23923	
辞書エントリ数	3436		2968	
わかち書き誤り	79	0.37 %	33	0.14 %
品詞付与誤り	444	2.12 %	387	1.62 %
合計誤り	523	2.48 %	420	1.75 %

3.3.2 結果と考察

実験の目的は、コーパス仕様の変更によりタグ付与のくい違いがどの程度改善されるかを観察することである。今回は、修正した 100 アブストラクトをすべてを学習データにし、学習時の誤りについて実験をした。

結果を表 9 に示す。変更前は、GENIA Corpus 3.02 の統合コーパス (表 2)、変更後は、3.3.1 節にまとめた変更点を反映したコーパスを指す。トークン数とは、100 アブストラクトをわかち書き仕様に基づいた単位の延べ数である。辞書エントリ数とは、モデル学習後の辞書エントリの総数である。テストの結果を、わかち書き誤り、品詞付与誤り、合計誤りに示す。わかち書き誤りは、わかち書きの境界が不正解になった数で、品詞付与誤りは、わかち書きの境界が正解であったトークンの内、品詞が不正解になった数である。合計誤りとは、総トークンの内、わかち書きもしくは品詞が不正解になった数である。

実験に用いたコーパスは少量 (GENIA Corpus 3.02 の 1/20) ではあるが、わかち書き及び品詞付与の誤り率が下がることを確認した[10]。これは、わかち書きをできるだけ短い単位にそろえ、語彙化品詞を導入したことにより、学習するコーパスの一貫性が向上したためと考えられる。

4 成果と今後の課題

4.1 成果

本研究では、ライフサイエンス分野テキストに特有な言語現象を調査し、分析結果を足掛かりにわかち書き仕様と品詞体系の変更箇所を提案し、統計的自然言語処理の手法を用いて形態素解析システム「cocab」を実現した。

実験で、今回提案した仕様でタグ付与したコーパスから学習したモデルは、GENIA Corpus 3.02 から直接学習したモデルより、誤り率が改善されることを確認した。解析対象としているライフサイエンス分野テキストの特徴を吟味することが重要であることを示した。

4.1.1 固有表現抽出への応用

見かけ語に左右されないトークンの定義や特殊文字の語彙化品詞の導入は、テキスト中に出現する遺伝子やタンパク質の固有表現抽出など、上流タスクからの利用を意識している。

ライフサイエンス分野のテキスト解析では、遺伝子やタンパク質に絞った固有表現抽出の研究が盛んである。特殊記号などの綴りが効果的な手がかりだが、品詞は期待したほどでない、という研究報告が多い [5]。 これらの手法では、前処理として、新聞記事とほぼ同じわかち書き仕様と品詞体系を仮定しており、本稿で議論した問題点を無視している。

表 10 に示すように、特殊文字 - は、省略できたり、等位で接続して複合体や作用を修飾したり、接頭辞のために利用したり、負の意味を持たせたりと、様々な用法がある。

今回開発した cocab で前処理を行なえば、見かけ語の非境界部分に位置している - を切り出すことができ、かつ、- の前後トークンの品詞付与により、陰に - の前後の特徴を捉えることが可能になる。2.2.1 節で指摘した分割不足なトークンが排除され、2.3 節で指摘した品詞タグ付与のくい違いが抑えられていれば、- の用法が明白でなくても、十分有効であると考える。実際、我々は、cocab を使って、固有表現抽出タスクを行ない、F 値が 75 ポイントという肯定的な結果を得て、従来の手法より固有表現の境界をより正確に認識できることを確認した [7]。

[10]品詞コーパスのわかち書き仕様は、第 2.2.1 節で議論したように、コーパスが上流のタスクで必要としているわかち書き単位になっていないので、比較実験の対象にしなかった。統合コーパスの結果集計では、残片の品詞は * とみなした。

表 10: 特殊文字 (-) を含む文字列とその用法; 強調するために該当文字を — で示す

用法	文字列
省略可	NF—kappa B-dependent
省略可	EBNA—2-PU.1 interaction
等位接続 (and)	EBNA-2—PU.1 interaction
等位接続 (and)	T cell receptor—CD3 complex
接頭辞	anti—CD3 mAb
接頭辞	non—B cells
負	CD8—
負	—0.8
修飾 (形容詞)	NF-kappa B—dependent
修飾 (現在分詞)	NF-kappaB—binding site
修飾 (過去分詞)	anti-CD3—induced c-Rel expression

4.2 今後の課題

cocab は、マルコフモデルによる形態素解析という古典的な手法で実現したが、最大エントロピー法による方法など自然言語処理の手法については、改善できる部分が多く残されている。しかし、ここでは、ライフサイエンス分野テキストの特徴をもっと的確に捉えるための課題について述べる。

4.2.1 外部辞書エントリの統合基準

ライフサイエンス分野では、オントロジーやデータベースが数多く存在し、辞書エントリを拡充するための外部資源が豊富である。これらのデータを考慮することにより、被覆率が向上すると予想される。しかし、現状では、明確な外部辞書エントリの統合基準がない。以下に、基準を設計する上で、検討すべき事項を述べる。

- **トークン単位**: オントロジーやデータベースはほとんどが複合語 (長い単位) である。長い単位の場合、どこまでをトークンとして認めるべきかは自明ではない。
- **品詞**: 品詞情報は無いので推定する必要がある。簡単には、すべて名詞とみなすことができる。もしくは、短い単位で区切ったときの主辞トークンの品詞に揃えることも考えられる。
- **単語生起コスト**: 品詞と同様、単語生起コストも算出する必要がある。PubMed アブストラクトだけで 1200 万件ほどあるが、このような大量のラベル無しデータを利用した推定方法を考案する必要がある。

関連する事項として、略語の追加がある。略語は複合語を 1 見かけ語にしているため、上記と同様の問題が発生する。特に、正式名称の品詞と略語の品詞をど

こまで一致させるかが難しい。例えば、monoclonal antibody は mAb と略される。mAb の品詞は NN なので m(onoclonal) も A(nti) も b(ody) も NN とも解釈できる。monoclonal の品詞は、通常 JJ であり、NN にするかは、迷うところである。

5 おわりに

本稿では、ライフサイエンス分野テキスト特有の言語現象を分析し、わかち書き、品詞付与、原形変換などの低レベルのテキスト処理について報告した。

今後、ライフサイエンス分野を対象とした情報抽出は、量も増加し、質も多様化するであろう。例えば、情報源をテキストデータに限らず、BLAST などを利用した類似配列データやマイクロアレイによる実験データなど異種データを統合したデータベース作成とそれに伴うマイニングが盛んになると考えられる [3]。

こうした状況の下、本研究で開発した低レベルのテキスト処理ツールは、想定される高レベルなタスクで有機的なつながりが実現できるように配慮した。我々が得た経験と成果を公にすることにより、分野の発展に少しでも貢献できれば、これに代わる喜びはない。

付録 1: 品詞体系

Penn Treebank の品詞体系は、タグ付けマニュアル [4] に記述されている。GENIA Corpus 3.02 の品詞体系に関する情報は、プロジェクトページ [1] にある。表 11 に、本稿で参照した品詞の記述を抜粋する。

表 11: Penn Treebank [PTB] と GENIA Corpus 3.02 [GENIA] の品詞体系

PTB	GENIA	説明
CC	CC	接続詞
CD	CD	数詞
JJ	JJ	形容詞
NN	NN	名詞、単数、不可算
SYM	n/a	記号
TO	TO	to
VBG	VBG	動詞、現在分詞
VBN	VBN	動詞、過去分詞

本稿での議論に関係あるタグ付与の現象:

- ハイフンで二つのモノが連結され、名詞が後続する (見かけ) 語の品詞ならば、JJ。(e.g. NFkappaB-DNA)
- ハイフンで一つのモノと数字やアルファベットなど型番が連結されている (見かけ) 語の品詞は、NN。(e.g. NF-kappa)
- ハイフンで一つのモノと形容詞、現在分詞、過去分詞が連結されている (見かけ) 語の品詞は、JJ。(e.g. NFkappaB-dependent)
- NN と NN に囲まれた (見かけ) 語の品詞は、NN。

謝辞 本研究の実施にあたり、コーパスを公開していただいている東大辻井研究室の皆様には、数々の疑問にご回答いただき、大変お世話になりました。特に、建石由佳氏から頂いた品詞コーパス仕様書 (非公開) からは、多くを学びました。ツールの実装に関しては、cocab の土台になった MeCab を開発した工藤拓氏、及び、奈良先端大松本研究室の関係者各位との日頃の議論が、必要不可欠でした。ここに感謝の意を表します。本研究は、理研 GSC ゲノム情報科学グループと奈良先端大松本研究室の共同研究の成果の一部です。

参考文献

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/ genia/topics/Corpus/3.0/GENIA3.0p.intro.html. *GENIA Project*. 2003.

[2] C. Manning and H. Schütze. *Foundation of Statistical Natural Language Processing*. MIT Press, 1999.

[3] S. Raychaudhuri, H. Schütze, and R.B. Altman. Inclusion of Textual Documentation in the Analysis of Multidimentional Data Sets: Application to Gene Expression Data. In *Machine Learning*, Vol. 52, pp. 119–145, 2003.

[4] B. Santorini. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. Linguistic Data Consortium, Penn Treebank II CD-ROM, 1991.

[5] K. Takeuchi and N. Collier. Bio-Medical Entity Extraction using Support Vector Machines. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pp. 57–64, 2003.

[6] L. Tanabe and W. J. Wilbur. Tagging gene and protein names in biomedical text. In *Bioinformatics*, Vol. 18(8), pp. 1124–1132, 2002.

[7] K. Yamamoto, T. Kudo, A. Konagaya, and Y. Matsumoto. Protein Name Tagging for Biomedical Annotation in Text. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pp. 65–72, 2003.

[8] T. Yamashita and Y. Matsumoto. Language Independent Morphological Analysis. In *6th Applied Natural Language Processing Conference*, pp. 232–238, 2000.