生物学・医学文献からの情報抽出：専門用語辞書利用に関する一考察

保坂順子
理化学研究所
ゲノム科学総合研究センター
横浜市鶴見区末広町 1-7-22
jhosaka@gsc.riken.go.jp

Judice L.Y. Koh
Institute for Infocomm Research
21 Heng Mui Keng Terrace,
Singapore 119613
judice@i2r.a-star.edu.sg

小長谷明彦
理化学研究所
ゲノム科学総合研究センター
横浜市鶴見区末広町 1-7-22
konagaya@gsc.riken.go.jp

要旨

我々は、生物学・医学文献からの情報抽出を試みている。現在、その一環として、たんぱく質の相互作用表現および相互作用要素を構文情報に基づいて抽出している。生物学・医学などの文献を言語処理する場合、最も問題になるのは専門用語の認識である。そこで、一般的な辞書に専門用語辞書を加えた場合、抽出精度にどのような影響を与えるか調べるため、Medline[1]のアブストラクトから 100 文を選択して実験した。その結果、一般ドメインで学習した構文解析パーザに関しては、一般的な辞書のみを使った場合（再現率 93.0%、適合率 91.0%）と専門用語辞書を追加した場合（再現率 92.9%、適合率 89.6%）を比較すると、後者のほうが抽出精度が低下するという現象が確認されたので、これを報告する。

1 はじめに

生物学・医学分野では近年、論文数の増加が著しい。例えば、腫瘍の発現を抑えるたんぱく質に、1970年代後半に発見されたP53と呼ばれるものがある。このP53に関連する文献数を、PubMed[2]に単純な検索をかけて調べたところ、1970年代後半以降1980年から1984年までの5年間に89件であったのが、1995年から1999年までに12,699件、2000年から2003年1月現在で、すでに10,105件の論文が検索されるまでになった。1980年代と比較すると、100倍の数になっている。これらの文献すべてに目を通すのは不可能であり、必要な情報だけを取り出すなど、何らかの形での言語の自動処理が望まれる。

生物学・医学分野での情報抽出には、単語の共起情報を使ったもの (Jenssen, 2001)、フルパーザを使ったもの (Yakushiji, 2001)、抽出規則を人手で書き下したもの (Blaschke, 2001)、医学文献用に開発した自然言語処理システムを、意味カテゴリーに変更を加えるなどして分子生物学用にしたもの (Friedman, 2001) などがある。しかし、たんぱく質の相互作用などの情報を抽出するには、一文にたんぱく質名が複数出現している場合でも、これらが互いになんらかの関係を持っているとは限らないため、共起情報だけでは十分でない。また、フルパーザ用に文法規則を書き下すことは困難であり、処理には多大な計算量を要する。

言語処理をする上で、第一に問題となるのは専門用語の認識である。たんぱく質名などは、新しいものが頻繁に作り出されており、すべてのたんぱく質名を網羅したリストを用意するのは不可能である。そこで、我々は、相互作用を表す動詞を中心に、構文構造から、その動作主、被動作主を抽出している。これらが、相互作用要素の最良の候補だと思われるからである。パーザは、ニューヨーク大学で開発された Apple-Pie Parser ver.5.9[3]（再現率：77.5 パーセント、適合率：75.58 パーセント）を使っている。

[1] Medline は米国の国立医学図書館, the National Library of Medicine (NLM), の生物学・医学の文献データベースである。
[2] PubMed は、NLM が提供しているインターネット上の文献検索サービスである。
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[3] http://www.cs.nyu.edu/cs/projects/proteus/app/

2 テスト文および専門用語辞書

テスト用の文は、パーザが正しく構文解析できるものを選択した。これは、構文情報を基に抽出規則を作成しているためである。パーザには Penn Tree Bank 用の一般用語辞書が用意されているが、生物学・医学文献を扱うには不十分だと考え、オンラインの専門用語辞典を用い、生物学・医学ライブラリー辞書, Medical Library Dictionary (MLD), を作成した。

2.1 テスト文の選択

生物学の専門家から、相互作用を表す動詞のリストと、PubMed を使って Medline から検索した論文の抄録 1000 件を受け、これを基にテスト文を選択した。抄録は白血球間相互作用にかかわるインターロイキン 6 に関するものである。

相互作用を表す単語のリストから、2 要素の相互作用をよく表すと思われる "activate" を選択し、この動詞から始めることにした。

テスト文の選択のために、まず、"activat*[4]" という文字列を含む文に構文解析パーザをかけ、動詞の "activat*" を含む文を選択した。これは、約 1000 文あった。次に、言語の専門家二名が "activat*" とそれにかかわる動作主、被動作主を含むフレーズのマーキングと、フレーズに関係する構文解析の評価をしたもののうち、評価者の判断が一致した文から、ランダムに 100 文選択した。

データの信頼性を調べるため、カッパ値（K）を求めた。これは、偶然の一致を考慮した判定者間の一致性の指標で、主観が入る判定が複数の観察者の間で、または同一観察者では複数回の判定間で、どの程度一致するかを知るものである。今回実施した構文解析に関する判断の一致度は、0.54（保坂、梅津, 2002）であり、信頼性は中程度であるとみなされる (Carletta, 1997)。

2.2 生物学・医学ライブラリー辞書

生物学・医学のドメインでは生物学、化学および医学の専門用語が使われていると仮定して、この分野のオンライン用語辞典を探した。MLD をつくるのに 4 つの用語辞典を使った：Biochemical Glossary[5] (BG), Cancernet Dictionary[6](CD), Medical Chemistry Dictionary[7] (MCD), Life Science Dictionary[8] (LSD)。MLD に加えてさらに Medical Subject Headings (MeSH[9]) の C chapter（疾病）も使った。MeSH は、NLM が提供する制限された語彙のシソーラスである。

辞書サイズを表 1 に示す。MeSH の用語数はユニークなもので、同義語と化学名を含む：

辞書	ソース辞典	用語数
MLD	BG	723
	CD	2,414
	MCD	122
	LSD	32,405
MeSH	MeSH	300,263

表 1. 用語辞書サイズ

MLD に含まれるユニークな用語数は 32,698 である。一般用語辞書, general-purpose dictionary (GPD), は 88,707 用語からなる。そこから MLD 用語を取り除いたところ MLD の用語数は 25,772 に減少した (uniMLD)。さらに、MeSH と MLD の間に 401 の重複があったが、この場合は MLD の用語を活かした。そのため MeSH 用語は 300,263 (uniMeSH) になった。実験には uniMLD と uniMeSH (MLD-M) を組み合わせて使った。ここでは、GPD と MLD-M の組み合わせを MLD+ と呼ぶことにする。

表 2 に、実験に使った辞書の用語数を示す：

辞書			用語数
MLD+	GPD		88,707
	MLD-M	uniMLD	25,772
		uniMeSH	119,599

表 2. 実験に使用した用語辞書サイズ

4 種類の用語辞典の中で LSD には part of speech (POS) があるが、他の 3 辞典にはない。これは、LSD が双方向から引ける日本語と英語の 2 言語辞書であるためであろう。MeSH に関しては、名詞のみを含む。我々が使ったパーザは未知語処理を行うが、適切な POS が辞書に書かれて

[4] "*" は任意の文字列を表す。

[5] http://www.fhsu.edu/chemistry/twiese/glossary/biochemglossary.htm
[6] http://www.cancer.gov/dictionary/
[7] http://www.chem.qmw.ac.uk/iupac/medchem/
[8] http://lsd.pharm.kyoto-u.ac.jp/index.html
[9] http://www.nlm.nih.gov/mesh/meshhome.html

いるほうが精度が向上すると考え、3 辞典に半自動的に POS を付与した。

3 抽出規則

能動態と受動態のための抽出規則を人手で作成した。これは、例えば、相互作用を表す "activat*" を含む動詞句とそれに対応する動作主および被動作主を抽出する、というものである。パージング結果は XML 形式に変換し、この構文解析結果に抽出規則を適用した。

4 実験

100 テスト文の構文解析結果に抽出規則を適用した。構文解析は GPD と MLD+ を使い二度実施した。

抽出精度を計算するための正解コーパスとして、生物学者が動詞の "activat*" とそれに対応する相互作用要素にマーキングしたものを使った。正解コーパスとしたものの中にも不整合が見受けられたため、システムの抽出はマーキングされた句を含めば正解とみなした。また、"by" のような前置詞が、正解コーパスではマーキングされていて、システムの抽出にはない場合も正解とした。

正解コーパスとシステムの抽出結果の組み合わせを A, B, C として、表 3に示す：

	正解コーパス	システム
A	抽出あり	抽出あり
B	抽出あり	抽出なし
C	抽出なし	抽出あり

表 3. 評価用マトリックス

再現率と適合率は以下の式により求めた：

再現率： A/ (A+B)

適合率： A/ (A+C)

抽出精度を、表 4に示す：

	再現率%		適合率%	
	GPD	MLD+	GPD	MLD+
VP	98.9	97.9	94.9	93.9
Agent	83.3	86.4	80.6	88.4
Recipient	96.6	94.2	87.6	86.2
合計	93.0	92.9	91.0	89.6

表 4. 抽出精度

表 4から、動作主の抽出が最も難しいことがわかる。この抽出には MLD+ が効果を示しているが、他の抽出では GDP のみを使った場合のほうが良い結果になっている。

5 専門用語辞書使用の影響

100 テスト文は約 2,500 単語からなっており、パーザは MLD-M の 236 用語 (uniMeSH 48, uniMLD 188) を認識した。このことから、専門用語は、単語認識に約 9 パーセント貢献したことになる。 表 2からわかるように、uniMLD は GPD の約 3 分の 1 のサイズだということを考えると、ヒット率はそれほど高くないことになる。

今回の実験では、専門用語辞書を使っても抽出精度は必ずしも向上したわけではなかった。GPD を使った場合には正しく情報抽出ができたが、MLD-M に拡張した場合には誤った抽出がされた文の分析を行う。このような文は 6 文あり、9 句が該当した。情報抽出に悪影響を与えた理由として、以下の 3 点が考えられる：

1. POS が正しく付与されなかった (3 句)
2. 個々の用語は正しく認識されたが、複合用語を組み立てるのに失敗した (2 句)
3. 用語は正しく認識されたが、句の組み立てに失敗した (4 句)

次に誤り例を示す。カテゴリー名は Penn Tree Bank[10]に準拠している。 左が GPD のみを使った場合で、右が MLD+ を使った場合の解析結果である：

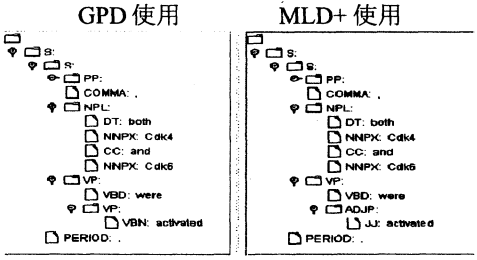

図 1. POS 付与誤り

図 1では、この文脈では "activated" という文字列は動詞であるべきだが、誤って形容詞と認識

[10] NNPX はパーザ独自のもので、NNP または NNPS を表す。

されたことを示す。LSD では “activated” は、動詞と形容詞として登録されているので、MLD にはこの両者を取り入れた。この誤りは、一般的なドメインでは、このような構文構造の文脈で使われる “activated” は、形容詞であることが多いということを示唆している。抽出規則では形容詞が相互作用を表すことは考慮していないため、何も抽出されなかった。

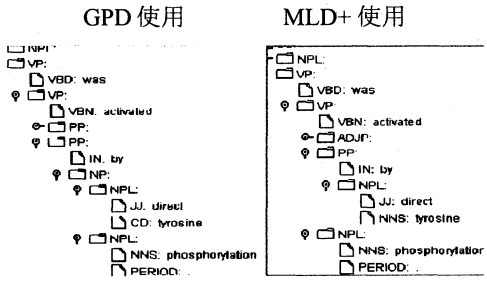

図 2. 複合用語構築誤り

図 2からわかるように、“tyrosine” の POS は正しく付与されている。しかし、システムは “phosphorylation” と複合用語を形成することに失敗している。

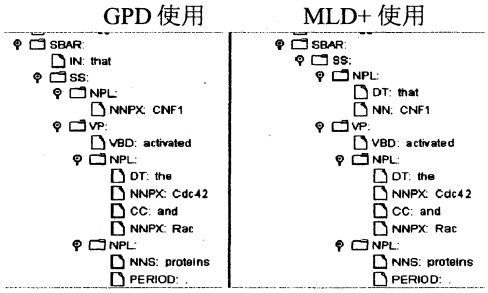

図 3. 句構築誤り

図 3では、“CNF1” の POS は正しく付与されている。しかし、前接の関係代名詞の “that” は誤って限定詞と認識された。これは、一般ドメインでは、普通名詞は冠詞を伴うことが多いからだと推測される。

我々は、POS の付与誤りや句構築誤りは、一般ドメインと生物学・医学分野で使う構文構造の分布が異なることに起因していると考える。

6 おわりに

今回の実験では、一般用語のみを使った場合のほうが、専門用語を加えた場合より情報抽出精度が若干良いという結果になった。

専門用語辞書を使うことにより POS が正しく付与されたとしても、パーザが生物学・医学分野とは異なるドメインで学習を行った場合、句を組み立てるのに失敗することがある。新しいドメインで学習するにはコーパスが必要となり、その構築には多大な時間がかかる。現状では、ドメイン特有な構文構造を集め、人手で規則に手直しをいれるのが良いと考える。

また、個々の用語が認識されているにもかかわらず、複合用語の認識に失敗するケースがあった。このような問題に対処するには、Unified Medical Language System[11] などを取り入れて、さらに専門用語辞書を充実させるのが有効だと考える。

今回の実験は小規模で、かつパーザが構文構造を正しく解析できる文のみを使った。この結果の有効性を調べるためにも、実験規模を拡大する予定である。

文献

Blaschke, Christian and Valencia, Alfonso. 2001. The potential use of SUISEKI as a protein interaction discovery tool. *Genome Informatics*, 12: 123-134.

Carletta, Jean, et al. 1997. The reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23(1): 13-31.

Friedman, Carol, et.al. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Proc. of ISMB*, 17(Suppl.1): S74-S82.

保坂順子、梅津亮. 2002. 免疫学文献からのたんぱく質相互作用情報抽出に向けて. 自然言語処理,150: 15-20.

Jenssen, Tor-Kristian, et al. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28: 21-28.

Yakushiji, Akane, et al. 2001. Event extraction from biomedical papers using a full parser. *Proc. of PSB*, 6: 408-419.

[11] http://www.nlm.nih.gov/research/umls/