BioIE に向けて – 形態素解析編

山本 薫† 工藤 拓‡ 小長谷明彦† 松本裕治‡

†理化学研究所 ゲノム科学総合研究センター (GSC) ゲノム情報科学グループ

〒230-0045 神奈川県横浜市鶴見区末広町 1-7-22 E209

{kaorux,konagaya}@gsc.riken.go.jp

‡奈良先端科学技術大学院大学 情報科学研究科

〒630-0192 奈良県生駒市高山町 8916-5

{taku-ku,matsu}@is.aist-nara.ac.jp

1 はじめに

現在、PubMed には 1200 万件もの生物医学関連の論文の論文概要が集積されている。遺伝子の機能推定には、大量の論文概要を読まなければならない。研究者は、個々の遺伝子に着目するのではなく、システムレベルの理解が必要である。このような背景から、高効率の情報抽出が望まれている。

本研究は、生物医学分野における情報抽出-BioIE-を実現するための基礎技術の確立を目指している。その第一歩は、テキスト中の遺伝子名やタンパク質名を認識することである。この課題は、自然言語処理分野でもバイオインフォマティクス分野でも、重要性が認識され、多くの先行研究が存在する。それらを大別すると、(1) 機械学習による手法 (2) 規則主導型手法 (3) 辞書検索手法 の 3 つに分類できる。

自然言語処理分野では、(1) が主流である。つまり、固有表現抽出とみなし、分類問題に帰着させ、SVM などを応用している。一方、バイオインフォマティクス分野では、(2) と (3) が提案されている。配列データベース、オントロジー、辞書などの外部資源の活用方法を知っていて、分野の専門家の意見を採り入れやすい環境にあるためと考えられる。(2) は、各々に異なる言葉を使っているが、正規表現などを使って規則を記述し、基本名詞句を認識する問題としている。(3) は、外部資源を検索し類似文字列が存在するかを判定する問題としている。

(1) と (2) は、両方とも、入力文字列に対して規則でトークン列に区切り、品詞付与を行ない、そのトークン列をチャンキングする処理である。チャンキングする手法こそ異なるが、一般的に自然言語処理で利用されている前処理を行なう。一方、(3) は、辞書構築を工夫して、BLAST や独自の類似文字列検索アルゴリズムを使う。文字種などは使うが、言語処理を前提としない。

実験設定が異なるが、報告された精度だけを比較すると、(3) が優勢である。これは、外部資源の活用が有効であること、表記の揺らぎにも対応できていることを示している。ただし、問題点もある。部分的に照合した位置はわかっても、言語処理を行なっていないために、名前の区切りが確定できないという問題がある。現状では、適当に閾値を設けて、それ以内の文字列を名前と認識させている。

本研究で、生物医学英語用の形態素解析を開発した動機は 2 つある。

1 つめは、(3) の問題点で指摘したように、わかち書きが困難であるということである。実は、(1) や (2) でも、わかち書きの困難さは存在する。前処理で、品詞付与をするために、空白文字を基に規則的にトークンに区切る。次節で述べるように、分子名区切りと空白文字は一致せず、見かけ上の一単語に [-] や [/] で複数の分子名が出現する場合がある。そのため、日本語の固有表現抽出で問題になる現象、トークンより細かい単位を抽出できない現象、が起こる。

2 つめは、(3) のように、豊富な外部資源を言語処理にも活用して、分子の名前認識問題に必要な属性を一括して付与する前処理を実現したい、ということである。従来から自然言語処理で研究された品詞などの言語情報だけでなく、外部資源から得られる属性、例えば、分子の名前と対応する実体へのリンク (配列 ID) なども同時に付与すれば、より付加価値が高くなると考えた。

従来、英語は、わかち書きされていると仮定されていた。しかし、本稿では、わかち書きされていないとし、日本語形態素解析の枠組を応用した。外部資源を活用して、配列 ID などの言語外属性も辞書エントリに付記した。そして、辞書エントリに基づく柔軟なわかち書きをし、言語属性と外部資源属性を統合的に付与できる、生物医学英語用の形態素解析の開発した。以下に、その概要について述べる。

2 わかち書き

英語におけるわかち書きにも2つ問題がある、と山下らは述べている [4]。第一の問題は、空白文字を含む複数単語が1つの辞書エントリに対応する場合 (many-to-one) である。第二の問題は、1 単語が区切られ複数の辞書エントリに対応する場合 (one-to-many) である。これらの問題を生物医学英語における例で説明する。

many-to-one は、複合語の扱いに現れる。例えば、文字列 "phospholipase C-gamma1" や "PLC gamma 1" は、一つの単位として認識される方が好ましい。なぜなら、タンパク質配列データベースにおいて一エントリになっているからである。[PLC] は [phospholipase C] の略語であり、空白文字だけで区切った場合、[phospholipase][C-gamma1] と [PLC][gamma][1] は、単語同士が対応しない状況にあり、それぞれを対応させることが困難である。生物医学分野には、このような造語が数多くみられる。

one-to-many は、[-] や [/] などの記号文字を含む文字列で多く見られる。例えば、文字列 "SLP-76-associated substrate" では、[SLP-76] が分子名である。空白文字をもとに区切った場合、[SLP-76-associated][substrate] となり、[SLP-76] が切り出せない。規則ベースのわかち書きの場合、文脈を考慮して [-] を区切るか区切らないかという処理ができない。

生物医学英語において、誰もが納得するわかち書き単位は存在しない。上位アプリケーションの用途に依存している。GENIA コーパス [1] で指定したわかち書きの定義も有力な候補の一つである。しかし、我々のグループでは、タンパク質相互作用に絞った情報抽出を目指している。この場合、GENIA コーパスで出現する [X-dependent](X は分子名) は、[X][-dependent] とし、X を切り出してほしい。なぜなら、かたまりのままでは、固有表現抽出での段階で X を切り出せないという問題がある。一方、すべて細かい単位に区切って、多段チャンキングでまとめあげることも可能である。しかし、分子生物学では、部位の特定 (domain)、や機能の特定 (inhibitor) など、分子に対する修飾が多い。まとめあげの範囲を決定するのが困難である。階層構造になっているため、複合語解析を要するが、現段階では、実用的でないと判断した。

本稿では、このような言語学的な追求をせず、わかち書きを lexeme 単位の列に区切る処理と割り切った。ここでいう lexeme とは、辞書エントリを指す。生物医学分野では、SwissProt などの配列データベース、UMLS や GO などの辞書やオントロジー整備が進んでいる。我々は、辞書エントリに基づく柔軟なわかち書きの実現した。これにより、名前認識など上位アプリケーションで必要なコンポーネント処理の精度向上につながると期待している。

3 生物医学分野テキストを対象とした形態素解析

前節での問題点を解消するために、山下らは、特定の言語に依存した部分を明らかにし、どの言語にも共通した部分と切り分けた「言語に依存しない形態素解析」を提案した。詳細については、文献 [4] に譲るが、ポイントは、(1) 形態素片の辞書検索を 共通接頭辞検索 (common prefix search) にする、(2) マルコフモデルによるコスト最小法を採用し、トレリスから Viterbi-like な動的計画法による最適解選択をする、の2点を共通部分としてコンポーネント化したことにある。一方、わかち書き (形態素片認識) と未知語処理は、言語依存部分として残されている。

本研究で開発した英語形態素解析器は、山下らの基本モデルを念頭において、高速日本語形態素解析器 MeCab[3] を、生物医学英語を効果的に処理できるように拡張したものである。実装にあたり次の点を考慮する必要がある。

- コスト最小法に必要なパラメータの学習
  - 品詞連接コスト
  - 単語生起コスト
- わかち書き処理における単語相当単位の認識
- (未知語処理)
- 生物医学分野における言語外属性

3.1 コスト最小法に必要なパラメータ学習

マルコフモデルによる形態素解析では、品詞タグ列をマルコフ連鎖とみなす。わかち書きされた 形態素片列 $W = w^1, \ldots, w^n$ が与えられたときに、品詞タグ列 $T = t^1, \ldots, t^n$ を決定する処理である。以下の式による確率を最大にするような品詞タグ列 T と形態素片列 W を求める問題に形式化できる。

$$\begin{aligned} T &= arg\max_T P(T|W) \\ &= arg\max_T \frac{P(W|T)P(T)}{P(W)} \\ &= arg\max_T P(W|T)P(T). \end{aligned}$$

ベイズの定理を利用して式展開すると、品詞連接確率 $P(T)$ と単語生起確率 $P(W|T)$ の積に帰着する。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値の逆数の対数に適当な係数をかける整数値 (コスト) を用いている。コストの和演算で最適な解を選択する方法をコスト最小法と呼ばれている。

表 1: fnTBL で PTB(02-21) を学習した結果

評価データ	PTB(23)	GENIA(5/5)
適合率	95.574	93.064
再現率	95.592	93.205

表 2: fnTBL で GENIA(4/5) を学習した結果

評価データ	GENIA(1/5)
適合率	97.310
再現率	97.329

3.1.1 品詞連接コスト

英語の品詞タグ付けにおいて、先行研究の多くは、トライグラムモデルで良好な精度を報告している。MeCab では、連接表を 3 次元配列で実現している。このような都合から、品詞連接コストは トライグラムモデルを基本に算出した。

学習用コーパスには、品詞情報が必要である。候補として、Penn Treebank と PubMed アブストラクトを抜粋した GENIA 2.1 がある。後者は、一部を除き、Penn Treebank に準拠した品詞体系が採用されている。Penn Treebank は、トークン数で GENIA 2.1 の約 8 倍の量であるが、Penn Treebank で学習したモデルが PubMed アブストラクトを処理するのに適しているか未知数である。予備実験として、Penn Treebank Sections 02-21 を Brill Tagger と同じ Translformation-based error-driven learning である fnTBL [2] で学習したモデルで、Penn Treebank Section 23 と GENIA 2.1 (670 アブストラクト) を評価した。結果を表 1 に示す。GENIA 2.1 をアブストラクト単位で 5 分割して、4/5 で学習し、1/5 で評価した結果を表 2 に示す。

Penn Treebank で学習したモデルは PubMed アブストラクトに対してあまり有効でないことが実験結果から読みとれる。今回は、GENIA 2.1 のみを利用し、 4/5 (536 アブストラクト) をトライグラム学習し、5431 パターンの品詞連接コストを計算した。

3.1.2 単語生起コスト

単語生起コストも、GENIA 2.1 を基にした。4/5 の学習データから、11026 の辞書エントリを獲得した。

本研究では、分子名などを精度よく認識したいという要望がある。GENIA 2.1 のみの辞書エントリでは、明らかに絶対数が少ない。ここで、2 つの対処方法が考えられる。一つは、分子名などの規則性を考慮した未知語処理を工夫する方法である。もう一つは、外部資源から辞書エントリを拡充する方法である。今回は、抽出したい遺伝子名やタンパク質名をできるかぎり辞書に書きつくすことを努力し、その代わり、未知語処理は簡単な方法ですませることにした。

表 3: デリミタ辞書と形態素片辞書

	トークン	指定した文字列
デリミタ辞書	なれない	空白文字
形態素片辞書	なれる	.,:;"'%/[]{}!?%$&-()

バイオインフォマティクス分野では、発見された核酸配列データベース (GenBank) やアミノ酸配列データベース (SwissProt) が整備されている。これらのデータベースには、それぞれの配列に対する名前と同義語 (バイオインフォマティクスでは alias と呼ばれている) が、定義されている。さらに特徴的なのは、異なるデータベースが同じ配列に対するエントリを所有していた場合、相互参照リンクが張られている。さらに、ヒトに関しては、各種データベースの相互参照リンクをデータベース化したもの (HUGO) まで存在する。

そこで、HUGO の連結テーブルからヒトに関係あるタンパク質配列 ID を洗い出し、それに対応している名前記述をすべて列挙した。こうして、163431 エントリを獲得した。これらの辞書エントリは、すべて一般名詞 (NN) とし、コストは、GENIA コーパスにおける NN の最大コスト (最小単語生起確率) を付与した。

3.2 わかち書き処理

わかち書きは、言語に依存した部分である。ここでは、英語のわかち書きについて述べる。山下らのモデルでは、文中での「辞書検索を始めて良い位置・終えて良い位置」を言語ごとに明確に定義する。その定義をもとに、「辞書検索を始めて良い位置」から始まるすべての lexeme を common prefix search で一括して取り出し、「辞書検索を終えて良い位置」で終っている lexeme すべてをトレリス構造に追加する。MeCab では、common prefix search を効率的なトライ構造であるダブル配列で実現して、高速化に成功した。

入力文字列から、辞書検索を始めて良い位置と終えて良い位置を計算するために、デリミタ辞書と形態素片辞書を用意する。前者は、トークンの境界として働くが、それ自体は独立したトークンにはならない文字列の辞書である。lexeme の開始と終了位置にはデリミタは現れない。一方、形態素片辞書は、トークンとなる特殊な文字・文字列の辞書を指す。記号文字のようにトークンの境界として働き、それ自体もトークンとして扱われる文字列を格納する。表 3 のように、デリミタ辞書と形態素片辞書を設定した。

入力文字列中の辞書検索を始めて良い位置・終えて良い位置を求めた例を図 1 に示す。D はデリミタ文字を、M は形態素片文字を示す。↑が辞書検索を始めて良い位置を ↓が辞書検索を終えて良い位置を表す。図 1 では、4 文字目の S の位置から common

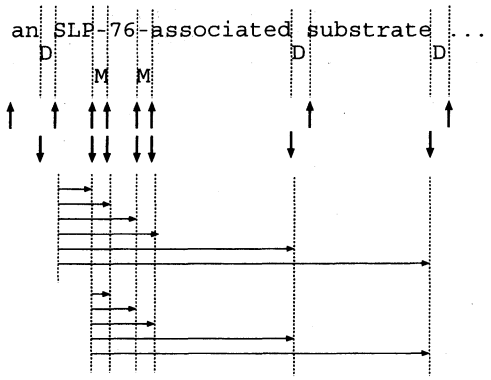

図 1: わかち書き処理

prefix search で辞書検索されるトークンの内トレリスに格納可能なものと 7 文字目の - の位置からトークンになりうるものを → で書いた。[-76] は、↑と↓に囲まれた文字列なので、辞書に存在すれば、トークンとして格納されるが、[-7] は、↓位置で終了していないので、トークンになれない。

3.3 言語外属性

分子名を認識する先行研究は、主に、文字列特性、品詞、構文構造などのテキストから抽出できる言語的な属性を手がかりとしている。しかし、分子名を認識するのに有効な属性は、他にもあると考える。例えば、alias や配列 ID などが考えられる。ChaSen や MeCab などの辞書に基づく形態素解析では、付加情報を辞書に追加でき、対象分野のニーズに応じた拡張が容易に実現できる。

3.1.2 節で述べたように、辞書エントリに追加した分子名に対応する配列 ID は既知なので、付加情報に追加した。このような工夫することにより、アブストラクトに書かれている既知の分子名を網羅的に検索し、かつ、その実体へのリンクを提供できるような前処理ツールになった。従来のわかち書きをして品詞を付与する以上の付加価値をもたらしている。

3.4 実験結果

本稿では、言語属性と外部資源属性を統合的に付与する生物医学英語用の形態素解析を目指している。言語属性の評価として、形態素解析の結果を表 4 に示す。G は、GENIA 2.1 のみを使う、GH は GENIA 2.1 にヒト辞書を追加したという意味である。我々が目指したわかち書きの単位と GENIA 2.1 の単位は異なるので精度はよくない。形態素解析がわかち書き誤ったとされる 10 % のうち、複合語が辞書エントリとしてヒト辞書に存在したもの (many-to-one) や、

表 4: 形態素解析の精度 (Precision/Recall)

連接	単語	わかち書き	品詞付与
G	G	90.722/95.661	86.374/91.094
G	GH	90.538/94.314	86.179/89.791

表 5: 分子名認識の精度 (Precision/Recall)

データ	延べ	異なり
GENIA 3.0	43.877/13.605	35.374/8.678
Yapex	64.889/45.158	44.318/37.209

部分文字列が辞書エントリとしてヒト辞書に存在したもの (one-to-may) があり、見かけの精度を悪くしている。

外部資源属性の評価として、分子名認識の結果を表 5 に示す。これは、開発した形態素解析器が GENIA 2.1 のわかち書き単位と合致していなくても、拡張した辞書エントリに基づくわかち書きした場合、どの程度、分子名が認識されるかを調査する目的で実施した。実験では、GENIA 3.0 から G#protein_molecule を抜き出しデータとタンパク質だけが注釈付けられた Yapex コーパス [5] を使った。実験データにも依存するが、形態素解析で付与した外部資源属性のみで、タンパク質名を 40 % 以上の適合率で認識できた。分子名認識問題の属性として、有効に働くのではないかと期待できる結果が得られた。

4 おわりに

本稿では、辞書エントリに基づく柔軟なわかち書きをし、言語属性と外部資源属性を統合的に付与できる、生物医学英語用の形態素解析の開発について述べた。まだ、表記の揺らぎに十分対応できていないので、今後は、未知語処理と併せて検討する予定である。

謝辞 GENIA コーパスに関して数々の疑問に答えてくださった建石由佳氏と大田朋子氏、バイオインフォマティクス分野におけるデータベースの扱い方を解説してくださった長嶋剛史氏に感謝の意を表する。

参考文献

[1] GENIA Corpus, 東大辻井研究室. http://www-tsujii.is.s.u-tokyo.ac.jp/genia/. 2002.

[2] fnTBL, Radu Florian, and Grai Ngai. http://nlp.cs.jhu.edu/~rflorian/fntbl/. 2001.

[3] MeCab, 工藤 拓. http://cl.aist-nara.ac.jp/~taku-ku/software/mecab/. 2002.

[4] T. Yamashita and Y. Matsumoto. Language independent morphological analysis. In *6th Applied Natural Language Processing Conference*, pp. 232–238, 2000.

[5] Yapex. http://www.sics.se/humle/projects/prothalt/. 2002.