JST 科学技術用語シソーラスに基づく MeCab 用専門用語辞書

建石 由佳 † 信定知江 † 高木利久 ‡,†
† 科学技術振興機構 バイオサイエンスデータベースセンター
‡ 東京大学 大学院理学系研究科 生物科学専攻
{tateisi,nobusada}@biosciencedbc.jp,tt@bs.s.u-tokyo.ac.jp

1 はじめに

形態素解析は自然言語の基礎的な解析プロセスであり、特に分かち書きされない日本語文の解析に置いては単語同定のために不可欠である。形態素解析の結果にはそれ以後の処理がすべて依存するので、形態素解析のエラーは、情報抽出、文書分類などで思わぬ結果を生む原因となる (図 1)。

一般に形態素解析の精度は利用する辞書に大きく依存するので、専門分野に適応した辞書を利用して単語同定の精度を向上させることが広く行われている。本稿では、生命科学データベースの説明文の特徴語を抽出するために作成した、形態素解析器 MeCab[1] 用のバイオインフォマティクス分野専門用語辞書とその利用について報告する。

図 1: Google Scholar での「情報中心性」の検索結果 (2017 年 1 月 4 日)。「情報」「中心」「性」への過分割のために意図しない検索結果が上位に表示されている

2 用語辞書の概要

専門用語のセットとして、科学技術振興機構（JST）で整備している「JST 科学技術用語シソーラス（JST シソーラス）」2015 年版 [3] を利用した。用語の品詞、生起コストの割り当ては、ComeJisyo V3-1[5] の方法を参考にした。

2.1 JST 科学技術用語シソーラス

JST 科学技術用語シソーラス（JST シソーラス）は科学技術振興機構（JST）の運営する文献データベースでの索引づけに用いられる統制語彙で、1 つの技術用語は、シソーラス上で「見出し語」、「同義語」、「関係語」、「主題カテゴリーコード」のレコードを持っている。見出し語レコードは用語の標準表記を示すもので、各用語に対し、見出し語表記、語番号、ソート用のふりがなが与えられる。同義語レコードは見出し語以外の用語の表記を示すもので、語番号と見出し語以外の表記との対応が与えられる。見出し語レコード数は約 4 万、同義語レコード数は約 8 万である。関係語レコードは、用語に対して、上位、下位、関連語の関係となる用語を与える。主題カテゴリーは理工学の 14 分野が 208 に細分されたもので、主題カテゴリーレコードでは、1 つの用語に対してカテゴリーコードが最大 6 個付与される。

2.2 ComeJisyo

ComeJisyo は、医療電子記録の単語同定を目的として西南女学院大学の相良らにより開発された医療・看護分野用の MeCab ユーザー辞書で、修正 BSD ライセンスで公開されている。看護学教科書、看護師および管理栄養士の国家試験問題、Web 上に公開された看護学関連文書、電子医療記録などから専門用語を収録しており、最新版である ComeJisyo V5-1[4] は 77760 語の用語を持つ。ComeJisyo のエントリーには、MeCab の解析に必要な

> 表層形, 左文脈 ID, 右文脈 ID, コスト, 品詞, 品詞細分類 1, 品詞細分類 2, 品詞細分類 3, 活用形, 活用型, 原形, 読み, 発音

の情報 [2] と、看護教育等に利用するための属性（英語訳：出現領域：出現文書数：番号）を与えている。

品詞、コストは、IPA 辞書を利用した MeCab の結果をもとにしており、ComeJisyo V3-1[5] では、登録語の品詞を「用語を構成する中で最後に現れる単語（語根）の品詞」とし、コストは、（5000 －文字長× 2）を基準とし、登録語の中に IPA 辞書のコストの小さな単語が含まれる場合、そのコストより小さい値を登録語のコストとする、としている。IPA 辞書の情報を利用することにより、ComeJisyo は学習コーパスの少ない（または全くない）分野での専門用語辞書に対して妥当なコストを与え、専門用語の過分割を避けることができている。

2.3 作成した辞書の構造

今回作成した辞書は、JST シソーラス 2015 年提供版の見出し語と同義語のレコードに与えられた表記、見出し語のふりがな、主題カテゴリーを、MeCab のユーザー辞書として利用できるようにしたものである。用語間の関係は利用してない。

MeCab のユーザー辞書のエントリに必要とされる情報のうち、「表層形」はシソーラス中の見出し語あるいは同義語の表記、「原形」は、見出し語については表層形と同じ、同義語については対応する見出し語とした。同義語の原形として見出し語の表記を持たせたのは、同義/異表記を標準化する目的に利用できるようにするためである。

「読み」は、見出し語についてはふりがなを用いたが、同義語にはふりがなが与えられていなかったので、対応する見出し語（原形）のふりがなを用いた。当面の目的である用語抽出には読みが必要ないこと、読みを正確に付与するには専門知識が必要であることから、表層形の正確な読みの付与は行わないこととした。同義語をデフォルトの辞書で解析した結果を利用することも考えられたが、修正の手間が膨大になりそうであったため、今回は見送った。「発音」は「読み」を正規表現 を用いて変換したものとした。

さらに、「シソーラス区分」（エントリーの表層形が見出し語であるか同義語であるかに応じて C または V)、「シソーラス ID」（エントリーに対応する見出し語の JST シソーラス語番号)、「主題カテゴリー」（JST シソーラス上の主題カテゴリー。複数ある場合は/で区切り 1 つの文字列とする)、JST の運営する科学技術統合検索サービス J-GLOBAL[1]の ID（エントリーに対応する見出し語の J-GLOBAL ID）を付与した。

品詞とコストの推定のためには、IPA 辞書（IPAdic-2.7.0-20070801) を用いた MeCab0.996 で用語を単独で解析した結果を用いた。品詞は、用語が MeCab で分割された結果最後に現れる形態素の品詞細分類に「サ変」を含めば「名詞, サ変接続」、そうでなければ「名詞, 一般」とした。用語の品詞として最後に現れる形態素の品詞そのものを用いなかったのは、末尾に数字を含む用語（例：ＩＧＦ 2、ＬＤ 50、ヒト乳頭腫ウイルス 16）が全体で数扱いされてしまうことと、最後の形態素が誤って名詞以外とされるケース（表 1）が散見されたことによる。

コストは、ComeJisyoV3-1 に習い、(5000-文字長× 2）と（MeCab で分割された結果に現れる形態素のコストの最小値-1) の小さいほうの値にした。

表層形	形態素列
うきぶくろ	うき（形容詞, 自立）/ぶ（動詞, 自立）/くろ（形容詞, 自立）
おしべ	おし（形容詞, 自立）/べ（助詞, 終助詞）
加味逍遥散	加味（名詞・サ変接続）/逍遥（名詞, サ変接続）/散（動詞, 自立）
特発性壊そ	特発（名詞, サ変接続）/性（名詞, 接尾）/壊そ（動詞, 自立）

表 1: 用語に対する誤った品詞付与の例：形態素列を/で区切り（）内に品詞と品詞細分類 1 を示す

3 用語辞書の作成とシステム辞書の修正

前節で述べた方法を用いて、JST シソーラスの主題カテゴリーに LS01～LS74（ライフサイエンス分野)、EG01（電気分野:電子計算機)、BA01（管理･システム技術分野 : ドキュメンテーション)、CA01～CA24（基礎化学分野)、CC03～CC23(工業化学分野)、IA01～ID01（共通分野）のいずれかを持つ見出し語と、それらの語に対する同義語より、67589 語（見出し語 20338 語、同義語 47251 語）からなる辞書を作成した。作成した用語辞書をユーザー辞書として MeCab 0.996 で Integbio データベースカタログ[2]のデータベース説明文をいくつか解析したところ、データベース名である MEDALS を MED/ALS（筋萎縮性側索硬化症） とするなど、英文字からなる略語によって単語を分割してしまう問題が発生した。そこで、以下のように英文字列に関する MeCab のシステム辞書エントリーを変更し、英文字のみからなる短い用語によって英単語が分割されるのを防ぐようにした。

[1] http://jglobal.jst.go.jp/

[2] http://integbio.jp/dbcatalog/?lang=ja

まず、システム辞書で英文字のみからなる表層形を持つエントリー (以下、英文字エントリー）の品詞および品詞細分類を抜き出した。IPAdic-2.7.0-20070801では、「感動詞」、「接頭詞, 数接続」、「名詞, サ変接続」、「名詞, 一般」、「名詞, 形容動詞語幹」、「名詞, 固有名詞, 一般」、「名詞, 固有名詞, 人名, 一般」、「名詞, 固有名詞, 人名, 姓」、「名詞, 固有名詞, 人名, 名」、「名詞, 固有名詞, 組織」、「名詞, 固有名詞, 地域, 一般」、「名詞, 固有名詞, 地域, 国」、「名詞, 接続詞的」、「名詞, 接尾, 一般」「名詞, 接尾, 助数詞」の１５品詞（細分類）が英文字エントリーを持っていた。

次に、それらの品詞に対して新たに英文字エントリー専用の別の品詞を設け、それぞれ「感動詞 A」などとした。品詞の細分類はそのままにして新たに 15 品詞（細分類）を設け、品詞 ID と接続 ID を割り当てた。新しく割り当てた接続 ID に対する接続コストは、新しい接続 ID 同士、および、新しい接続 ID と ID 4（記号, アルファベット）間では「デフォルトの matrix.def での接続コストの最大値の 2 倍」という大きな値とし、それ以外は対応する既存 ID（例 :「名詞 A, 一般」の場合は「名詞, 一般」の ID である 1285) のものを引き継いだ。

さらに、システム辞書、用語辞書の英文字エントリーの品詞、接続 ID を新しいものに変更した。コストなどは変更しなかった。また、未知語辞書（unk.def）で英文字列に割り当てる品詞を、元のエントリーに対応する新しい品詞とした。

4　用語辞書を利用した形態素解析

修正したシステム辞書と用語辞書を利用して Integbio カタログより 20 個のデータベース[3]の説明文を解析し、データベース説明文を特徴づける語として、用語辞書のエントリーとシステム辞書で固有名詞とされた語を抽出した。テキストはあらかじめ半角文字を全角に変換してから形態素解析をした。

その結果、形態素総数 1569 のうち、用語辞書のエントリーがのべ 382 語（異なり 185 語)、固有名詞がのべ 4 語（異なり 3 語）抽出された。用語をシソーラスの主題カテゴリーに基づいて分野分けすると表 2 のようになった。

目視による結果、おおむね抽出したい用語が抽出できていたが、下のような問題も見つかった。

[3]データベース ID : NBDC00005, NBDC00006, NBDC00015, NBDC00016, ... , NBDC00095,NBDC00096

分野	語数 (のべ)	語数 (異なり)	用語例
共通	186	81	情報、配列、データ、機能、検索
生命科学	130	69	遺伝子、ゲノム、タンパク質、ヒト、遺伝子発現
化学	25	17	リガンド、化学的構造、アミノ酸、複合体、塩基
計算機	24	6	データベース、ソフトウェアツール、ダウンロード、データバンク、ブラウザー
ドキュメンテーション	17	12	アノテーション、アブストラクト、公開

表 2: Integbio データベースカタログ 説明文より抽出した用語

抽出されていない語

機関名、データベース名等の固有名　EBI、NCBI、 DDBJ、 EMBL、FASTA、GenBank、Gene Ontology、PubMed、PDB、RefSeq、Ensembl

手法名オリゴキャッピング法、CpG アイランド、ゲノムビューワ

物質の構造を表す語　領域、小分子、低分子、完全長、結晶

物質の動作、実験操作に関する語　共起、伝達、標的、アライメント

データの製作者、利用者にかかわる語　コンソーシアム、機関、アカウント、ユーザ、提供

その他　文献、画像、糖、癌、日本人

用語が分割されてしまうもの

機関名　かずさ DNA 研究所

物質名　β グロビン

手法名　ゲノムブラウザー、マイクロアレイ、2 次元 ゲル電気泳動、テキストベース、ゲノム解析、ネットワーク解析

物質の構造を表す語　クラスター、立体構造、連鎖不均衡構造、転写開始点、糖分子 (糖分/子 と分割された）

その他　創薬、らん藻、生理機能

分割したほうがよいもの

画像データベース、位置情報、化学的情報、ソフトウエアツール

過剰に検出されているもの

もと （「情報 をもとに」という文脈で「酒母」の意味で抽出）

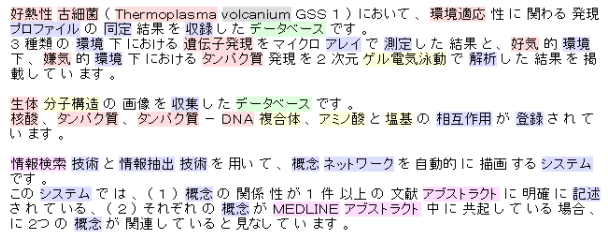
B、C（「B 型、C 型」という文脈でそれぞれ「ホ

ウ素」「炭素」の意味で抽出）

また、問題にはならなかったが、テキストを全角化したために数字列が 1 文字ずつ分割されていた。

これら抽出に失敗した語のうち、英文字からなる固有名は、未知語として品詞「名詞 A」で切り出されるので、後のプロセスでそのような語も用語候補のセットに含めることができる。また、辞書のエントリーを増やして抽出できる用語を増やす方針として次のようなものが考えられる。 1 ）機関名、データベース名はカタログに記載されるデータベースのメタデータ自体から辞書を作成して対応する。 2 ） いくつかの用語は辞書に含める主題カテゴリーを増やすことにより抽出を可能にできる。たとえば、小分子、標的、アライメント、低分子、結晶は物理分野、立体構造は土建分野、画像は電気分野、など今回利用しなかった分野の主題カテゴリーを持っていたために辞書に含まれていなかった。 3 ）JST の We b サービス J-GLOBAL ではシソーラス用語の他に「準シソーラス用語」「化学物質名」が検索できる。今回抽出に失敗した語のうち、「文献」「癌[4]」など一般によく知られている語は「準シソーラス用語」として検索することができたことから、将来、これらの用語も辞書化することを検討したい。

現在の辞書を使って、用語、固有名詞、英文字からなる未知語を抽出した例を図 2 に示す。


好熱性 古細菌 （ Thermoplasma volcanium GSS １ ）において 、 環境適応 性 に 関わる 発現 プロファイル の 同定 結果 を 収録 した データベース です 。
３ 種類 の 環境 下 における 遺伝子発現 を マイクロ アレイ で 測定 した 結果 と、 好気 的 環境 下 、 嫌気 的 環境 下 における タンパク質 発現 を ２ 次元 ゲル電気泳動 で 解析 した 結果 を 掲載 して い ます 。

生体 分子構造 の 画像 を 収集 した データベース です 。
核酸 、 タンパク質 、 タンパク質 － DNA 複合体 、 アミノ酸 と 塩基 の 相互作用 が 登録 され て い ます 。

情報検索 技術 と 情報抽出 技術 を 用い て 、 概念 ネットワーク を 自動的 に 描画 する システム です 。
この システム で は 、（１）概念 の 関係 性 が １ 件 以上 の 文献 アブストラクト に 明確 に 記述 され て いる 、（２）それぞれ の 概念 が MEDLINE アブストラクト 中 に 共起 して いる 場合 、 に 2つ の 概念 が 関連 して いる と 見なし て い ます 。

図 2: Integbio データベースカタログ 説明文の形態素解析結果の一部：用語は分野別に色分けして示す（赤:ライフサイエンス、ピンク：ドキュメンテーション、黄：化学、緑：電子計算機、青：共通、グレー：固有名詞と英字の未知語）

5 おわりに

JST シソーラスをベースにした MeCab 用のユーザー辞書を作成し、英字列周りのシステム辞書を改変することにより、バイオインフォマティクス分野の文書を形態素解析し、用語抽出を行った。この結果を利用して、説明文を利用した類似データベース検索、データベースの分類を可能にし、データベースカタログの利便性の向上を図る。

また、作成した辞書は上記目的以外にもバイオインフォマティクス関連分野の文書の形態素解析用の辞書として利用できるものであり、CSV 形式で Creative Commons CC-BY 4.0 ライセンス[5]のもとで公開する予定である。

参考文献

[1] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[2] 工藤拓. Mecab: Yet another part-of-speech and morphological analyzer. `http://taku910.github.io/mecab/`.

[3] 国立研究開発法人科学技術振興機構情報企画部. ＪＳＴ科学技術用語シソーラス　シソーラスファイル使用の手引き（ＳＪＩＳ漢字コード版）, 平成 27 年 11 月.

[4] 相良かおる. Comejisyo. `https://ja.osdn.net/projects/comedic/`.

[5] 相良かおる, 小野正子, 小木曽智信, 小作浩美. 電子医療記録の分ち書き用ユーザ辞書 ComeJisyo の紹介と単語生起コスト. 言語処理学会第 18 回年次大会発表論文集, pp. 621–624, 広島市立大学, 2012 年 3 月. 言語処理学会.

[4] シソーラスには「がん」として登録

[5] `https://creativecommons.org/licenses/by/4.0/legalcode.ja`