生命科学系日本語学術文書の係り受け解析

建石 由佳† 薬師寺あかね‡ 飯田啓介§ 山本泰智§
†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
‡フリーランス
§情報・システム研究機構ライフサイエンス統合データベースセンター
tateisi@biosciencedbc.jp, akane.yakushiji@gmail.com, iida@med.tohoku.ac.jp, yy@dbcls.rois.ac.jp

1 はじめに

生命科学分野の文書からの情報抽出における重要なターゲットの一つとして、文中に出現する生命科学用語の意味的関係の抽出がある。意味的関係は構文構造に反映されるので、構文解析はその関係を抽出するための手掛かりとなる。英語論文については、PubMed上の論文アブストラクトを対象として1990年代おわりごろから言語処理の研究がおこなわれており、その成果がPubMed自身、あるいは、その他の検索システムに実装されて近年の論文出版の増加による読者の負担を軽減する助けとなっている。しかし、生命科学日本語文献に対する言語処理の研究は少ない。

情報・システム研究機構ライフサイエンス統合データベースセンター（DBCLS）で公開されている「ライフサイエンス 新着論文レビュー」[1]（以下「新着論文レビュー」ともいう）は、Nature、Science、Cellなど生命科学分野のトップジャーナルに日本人を著者として発表された論文の著者自身によるレビューである。レビュー記事は、そのトピックを専門としない生命科学研究者に理解しやすくするため、専門の編集者である新着論文レビュー編集人（飯田）による綿密な編集作業を経たもので、内容の信頼性が高いだけでなく日本語の文章としての品質も高い。2010年9月1日のサービス開始より2018年11月30日[1]までに1279本の記事が作成され、「クリエイティブコモンズライセンス（CC）表示2.1日本」ライセンスの下で公開されている[2]。

我々はこの新着論文レビューに自然言語処理に有用な種々の情報を付加したコーパスの作成に取り組んでいる。このコーパスは、生命科学の日本語学術文書を対象とした自然言語処理ツールの分野適応に利用され、それを用いて文書の構造化、さらに構造化文書を利用した効率の良い文書検索・情報抽出に貢献できるものとすることを目指している。

本稿では新着論文レビューの文に係り受け関係をアノテートしたコーパスとコーパスを利用した係り受け解析システムCaboCha[2]の学習実験について述べる。

2 コーパスの作成

まず、新着論文レビューからランダムに選んだ500文について、以下のような手順で係り受けをアノテートし、結果をDBCLSのGitHubリポジトリ[3]より「クリエイティブコモンズライセンス（CC）表示4.0日本」ライセンスの下で公開した。

1. 500文をMeCab[3]（IPA辞書）で形態素解析した。このとき、Entrez Gene[4]の遺伝子シンボルリストとライフサイエンス辞書[4]をもとに生成したユーザー辞書を利用した。
2. その結果に対しCaboChaで係り受けをアノテートした。
3. 自動アノテーションの結果に対して、生命科学の英語論文からのテキストマイニングの経験を持つ二名のアノテータ（薬師寺、建石）が修正を行った。修正にはChaki[5]を利用し、係り受けのスキーマは京都大学テキストコーパスの作業基準[6]に準拠した。
4. 二名の結果の不一致の部分について新着論文レビュー編集人の判断を仰ぎ不一致を解決した。

500文に対しアノテータ間の一致率は係り受け箇所を単位とする単純一致率で90.5%となった。不一致の内訳は

[1]以後更新が停止されている
[2]http://first.lifesciencedb.jp
[3]https://github.com/dbcls/FA_corpus
[4]https://www.ncbi.nlm.nih.gov/gene

係り先不一致 7.3%

係りタイプ（並列、同格）不一致 2.8%

係り元不一致（文節区切りのずれ）1.4%

となっていた。かなり良い一致率が得られたと判断したので、さらに 1000 文に対して同様のアノテーションを行った。

3 CaboCha 学習実験

作成したコーパスを用いて CaboCha の学習実験を行った。CaboCha の学習は、CaboCha 公式ページ[5]のデフォルトの設定を用いて、作成したコーパスのみで学習したモデルと、コーパスを追加してオリジナルモデルを再学習したモデルを作成した。

評価用に、モデルを作成した 1500 文とは別の 100 文についてアノテーションを行った。この 100 文については 1500 文をアノテートしたうちの一名（建石）が新着論文レビュー編集人の判断を仰ぎながらアノテーションを行った。

この 100 文に対して、モデルによる解析精度の比較を行った。また、テキストから直接係り受けまで解析した結果と、アノテーションから作成した形態素結果を用いて係り受けのみを解析した結果の比較も行った。

100 文の文節単位の係り受け精度を表 1 に示す。なお、ここでは文節内の形態素の区切りと品詞については考慮せず、文節の区切りと係り受け関係のみの精度を評価している。

表 1 からわかるとおり、コーパスを追加することにより精度はわずかながら向上する。また、1500 文のみからの学習によってもオリジナルのモデルとほぼ同程度かそれ以上の精度が達成できる。しかしながら、表 1 の結果は、モデルの学習よりもむしろ形態素解析を理想的なものにする効果のほうが大きいことも示唆している。これは、新着論文レビューの文章が専門の編集者によって推敲を重ねられ日本語の構文としては整ったものになっている一方でオリジナルのモデルで想定されている新聞記事の文章とは分野の違いにより語彙が大きく異なっていることを反映している結果といえるであろう。コーパスをさらに増やすことでさらに良いモデルを作成することも考えられるが、学習コーパス作成の負担を考慮すると、形態素解析の整備を先に考えるほうが効率がよいであろうということが表 1 から示されている。

テキストから直接

設定	精度（Accuracy）
オリジナル	68.0%
1500 文のみ学習	68.7%
再学習	68.8%

形態素解析アノテーション利用

設定	精度（Accuracy）
オリジナル	83.6%
1500 文のみ学習	83.8%
再学習	84.4%

表 1: 再学習した CaboCha の解析精度: 「テキストから直接」は評価用テキストをオプションなしの CaboCha で解析した結果。「形態素解析アノテーション利用」は評価用テキストの人手でのアノテーション結果から形態素解析情報を抽出したものを-I1 オプション（形態素解析レイヤを入力とする）つきの CaboCha で解析した結果。

4 形態素解析の係り受け解析に対する効果

前節の結果を受けて MeCab の設定を変え、形態素解析ならびに係り受け結果の改善を試みた。まず、目視により解析誤りの個所を確認したところ、遺伝子名など「英数字の列」を不適切に区切り、その結果が文節区切りの結果に影響しているケースが目立った。たとえば、//を文節の区切り、/を形態素の区切り、( ) 内を人手によるアノテーションとして

- ヒストン/H/3/の/4/番目/の （ヒストン H3/の//4/番目/の）
- tgrB//1/-/tgrC//1/遺 伝 子/座/が （tgrB1/-/tgrC1/遺伝子座/が）
- SphK//2/ノックアウト/マウス/の（SphK2/ノックアウトマウス/の)
- TET//2/遺伝子/, （TET2/遺伝子/,）

のようなものである。ただし、英数字以外でも「組/換え」「遠/位」「摂/食」 のような区切り誤りによる文節区切り誤りもあった。

[5] https://taku910.github.io/cabocha/

この観察を受けて、MeCab の解析に利用する辞書や英数字列の区切りかたの設定を変えて形態素解析と係り受け解析の精度を比較した。具体的には、

ユーザー辞書: ライフサイエンス辞書、Entrez Gene の遺伝子シンボルリスト、「科学技術用語形態素解析辞書」[6]、および、IPA 辞書の英数字と記号を半角文字としたものをユーザー辞書として利用する。

制約つき解析: 池上によるコード [7] を、英数字に加えてギリシャ文字および+-などの記号からなる文字列の中を区切らないようにし、かつ、そのような列とそれ以外の文字列との境界が形態素境界になるかどうかは問わないように改変したものを利用する。

モデル学習: 係り受けアノテーションつき 1500 文をオリジナルモデルに追加して再学習したモデルを利用する。

のそれぞれと、それらの組み合わせで形態素解析の精度を測定し、さらに、各設定での形態素解析結果を利用した係り受けの精度を測定した。結果は表 2 のようになった。

設定	区切りのみ	区切り＋品詞（細分類含む）	係り受け（オリジナル）	係り受け（再学習）
d d d	80.7	79.0	68.0	68.7
u d d	92.9	91.2	76.8	77.9
d c d	83.4	80.3	78.8	79.1
u c d	95.2	92.8	82.7	83.5
d d m	80.9	78.4	63.4	64.4
u d m	91.4	86.8	71.1	72.0
d c m	83.7	79.9	75.4	76.3
u c m	92.8	87.6	76.7	77.7

表 2: 形態素解析の設定と精度、および、係り受けの精度：「設定」欄は 1 文字目がユーザー辞書の有無、2 文字目が制約つき解析の有無、3 文字目がモデル再学習の有無をあらわし、「d」はデフォルト、「u」はユーザー辞書使用、「c」は制約付き解析、「m」は再学習モデルの使用をあらわす。また、「係り受け」の「オリジナル」「再学習」は前節で利用したオリジナルと再学習済みのモデルである。

表 2 から、形態素解析精度の向上にはユーザー辞書の利用（表 2 udd および udm の行）が効果が高いことがわかる。ユーザー辞書と制約つき解析を組み合わせれば（ucd、ucm）さらに精度が向上し、ユーザー辞書を利用せず制約つき解析のみ（dcd、dcm）でもある程度の精度向上が見込めることがわかる。モデルの再学習は単独では形態素精度向上に寄与しているように見える（ddd と ddm の比較）が、制約やユーザー辞書を使った場合はオリジナルのモデルよりも精度が下がっている。

一方、係り受けの解析では、制約付き解析のみでも大幅に精度が向上し、ユーザー辞書のみを利用した場合よりも精度が高い。両者を組み合わせると精度がさらに向上することは形態素解析の場合と同様である。形態素解析の再学習モデルでは係り受けの精度はかえって下がってしまう結果となっている。

経験的に、MeCab のユーザー辞書に半角英数字の文字列（たとえば、IL-1）を登録すると、それが他の単語（たとえば、IL-16）の一部になっている場合に辞書登録された語を優先するせいで元の単語が辞書にない場合その単語を過剰に区切ってしまう（今の例でいえば IL-1 と 6 に切る）ことがわかっている。ユーザー辞書単独では制約つき解析を組み合わせたほどの効果が出ないことはこのことを裏付けているといえ、辞書の整備とともに未知語処理の改善をおこなうことが解析精度の向上に寄与すると期待される。

5 おわりに

「ライフサイエンス新着論文レビュー」の 1500 文に係り受け関係をアノテートしたコーパスを作成した。このコーパスを利用して、CaboCha と MeCab の学習実験を行った。今回の結果では、モデルを再学習するよりも、形態素解析精度の向上をはかるほうが最終的に係り受け解析結果も改善されることがわかった。形態素解析の設定を様々に変えた係り受け解析の結果から、辞書の整備だけでは不十分で、英数字部分（一般的な日本語ではない部分）の解析精度を別の手段で向上させることが必要であることわかった。英数字部分を切らないように解析する制約つき解析はその手段の一つとして有効であったが、さらに吟味することが必要であろう。

今回作成したコーパスは DBCLS の GitHub リポジトリより公開する。なお、新着論文レビュー全文は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の生命科学系データベースアーカイブ[7]から一括ダウンロード可能である。

[6] https://dbarchive.biosciencedbc.jp/jp/mecab/desc.html

[7] https://dbarchive.biosciencedbc.jp/jp/first-authors/desc.html

謝辞

「ライフサイエンス辞書」の使用をご許可くださった京都大学大学院薬学研究科の金子周司先生に感謝いたします。

参考文献

[1] 飯田啓介. 新しい日本語 Web コンテンツ,「新着論文レビュー」と「領域融合レビュー」. 情報管理, Vol. 56, No. 3, pp. 148–155, 2013 年 6 月.

[2] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. Vol. 43, No. 6, pp. 1834–1842, 2002 年.

[3] 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.

[4] 金子周司. 無料ライフサイエンス辞書の活用と効能. ファルマシア, Vol. 42, No. 5, pp. 463–468, 2006 年 5 月.

[5] Yuji Matsumoto, Masayuki Asahara, Kiyota Hashimoto, Yukio Tono, Akira Ohtani, and Toshio Morita. An annotated corpus management tool: ChaKi. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, pp. 1418–1421, 2006.

[6] 黒橋禎夫, 居蔵由衣子, 坂口昌子. 形態素・構文タグ付きコーパス作成の作業基準 version 1.8. `http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/corpus/KyotoCorpus4.0/doc/syn_guideline.pdf`, 平成 12 年 4 月. 最終閲覧 2019 年 1 月 15 日.

[7] 池上有希乃. Python で MeCab の制約付き解析を使う. `https://qiita.com/yukinoi/items/4e7afb5e72b3a46da0f2`, 2015 年 1 月 10 日. 最終閲覧 2019 年 1 月 15 日.