並列構造解析に基づく 複合化された固有表現の曖昧性解消

澤田 悠治 [1] 寺西 裕紀 [2] 松本 裕治 [2] 渡辺 太郎 [1]
[1]奈良先端科学技術大学院大学先端科学技術研究科
[2]理化学研究所 革新知能統合研究センター
{yuya.sawada.sr7, taro}@is.naist.jp
{hiroki.teranishi, yuji.matsumoto}@riken.jp

1 はじめに

固有表現認識は，人名や地名などの専門用語 (固有表現) をテキストから自動的に抽出する自然言語処理のタスクの一つである．近年では，これまでの科学技術の発展に伴う論文アーカイブの蓄積に伴い，各分野の論文を解析して，物質や微生物の特徴・専門知識を機械的に獲得する自然言語処理技術の応用が期待されている．固有表現認識の手法は，人手による素性設計を必要としないニューラルネットワークの導入によって顕著に発展している [1, 2] が，対応が困難な固有表現が存在し，専門分野における固有表現認識の精度改善と情報検索・関係抽出などへの応用のために解決が期待される．

従来の固有表現認識手法で対応が困難な固有表現の一つとして，並列構造による省略を伴う固有表現が挙げられる．並列構造を含む固有表現には複数の固有表現が複合化されており，複数の固有表現の中で重複する単語が省略される．例えば，“Human T and B lymphocytes"では，“Human T lymphocyte"と “Human B lymphocyte"の二つの固有表現が複合されており，それぞれの固有表現で “Human"と “lymphocyte"が省略されている．このような固有表現は，生命科学分野のコーパスの一つである GENIA Term annotation で全体の 3%に含まれており，既存の手法では一つの固有表現 (“Human T and B lymphocytes") として扱うか，例外として除去するなどの処理が行われる．

本研究の目的は，既存の固有表現認識器とのパイプライン処理が可能な並列構造の解析手法を提案し，上述の固有表現を辞書や複合名詞句がアノテーションされたデータセットを用いずに抽出することである．具体的には，固有表現認識器で抽出された固有表現に内包される等位接続詞に対して並列構造の範囲を同定し，省略された単語を補完する手法を提案する．本稿では，教師情報を用いない並列構造の解析器と複合された固有表現の正規化方法について説明し，提案手法によって固有表現全体の抽出性能の向上が見られたことを示す．また，提案手法において対応が困難な事例についても示し，今後の課題について整理する．

2 関連研究

2.1 固有表現認識

固有表現認識の多くの研究は系列ラベリングによる手法に基づいており，固有表現の範囲は BIO 方式や BIOES 方式などのラベル方式によって表現される．このようなラベル方式は，連続した単語列からなる一つの固有表現しか表現できない問題があり，上述の固有表現は複合した固有表現全体や一部分の固有表現のみが抽出対象とされるか，例外として除外されるなどの処理が行われている．複合した固有表現を不連続な範囲を持つ固有表現とみなして解く手法として，複数のラベル系列を一本のグラフと表現して出力する手法 [3] や，句構造解析で頻繁に用いられる Shift-Reduce アルゴリズムに基づいた手法 [4] が提案されている．しかしながら，不連続な固有表現が出現する事例の少なさや並列構造によって複数の固有表現が入り混じりになる点で，学習が困難になる問題がある．

2.2 並列構造解析

並列構造解析では，“and"や “or"といった等位接続詞によって結びつけられた句 (並列句) の範囲を同定する．並列構造解析の多くの研究では，同一の並列

構造にある並列句同士が意味的・統語的に類似する特徴 (類似性) に着目している．黒橋ら [5] は，文字列や品詞の一致などのルールによって付与されたスコアを設けて，並列句の類似度をチャートと動的計画法で算出した．新保ら [6] は英語の並列構造において，単語や品詞，形態情報に基づいた素性の重みづき線形和を元に類似度を算出した．近年では，ニューラルネットワークによる手法とともに，可換性に着目した研究が発展している．Ficler ら [7] は外部の構文解析器で抽出された並列句の候補の中で，ベクトル同士のユークリッド距離 (類似性)，片側の候補を抜いた場合に出力されたベクトルなどを用いたスコアを算出している．寺西ら [8] は，それぞれの並列句の内部と外部の境界などをベースにしたスコアを元に，並列構造を表す木を CKY アルゴリズムによって構築し，入れ子状の並列構造や三つ以上の並列句に対応する手法を提案した．ニューラルネットワークによる手法の利点は，黒橋らと新保らのような人手による素性の設計のコストの問題が削減される点にあり，GENIA Treebank と Penn Treebank においても高い性能を示している．しかしながら，これらのモデルの学習には個々の並列構造にある並列句の範囲がアノテーションされた大量のデータが必要になることから，他分野への応用という面を考慮すると，データセットの作成にかかるコストの問題が課題として挙げられる．

3 提案手法

本研究で用いる手法の概要を図 1 に示す．提案手法は，固有表現認識器と並列構造解析器の二つのモジュールで構成されている．固有表現認識器では複合した固有表現を一つの固有表現として抽出し，並列構造解析器では文中に含まれる名詞句と形容詞句の並列構造の範囲を同定する．これら二つのモジュールの解析結果に基づいて，複合した固有表現を個々の固有表現へと分解する．

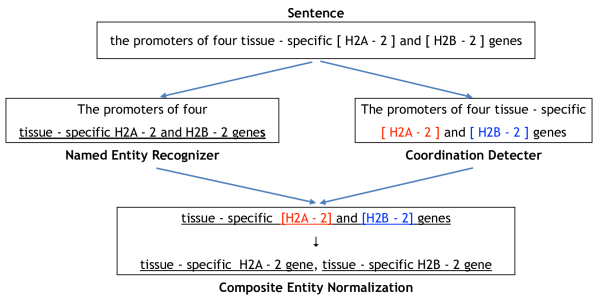

図 1 提案手法の概要図

3.1 並列構造解析器

Algorithm 1: 並列構造解析器

Input: $\{w_1, w_2, ..., w_N\}$ and k
Output: best_span
i, j = Preprocess($w_{1:N}$, k);
best_score = $-\infty$;
best_span = ϕ;
for $b = i$ *to* k - *1* **do**
  score, span = Alignment($w_{b:k-1}, w_{k+1:j}$);
  if *score* > *best_score* **then**
    best_score = score;
    best_span = span;
  end
end

固有表現のアノテーションのみが利用できる状況を想定し，並列構造の教師情報なしで並列構造の範囲を同定する解析器を使用する [9]．本研究で構築する並列構造解析器の実行過程を Algorithm 1 に示す．語数 N で構成される文 $w_{1:N} = \{w_1, w_2, ..., w_N\}$ と並列キー w_k に対して，全ての可能な並列構造範囲の組み合わせの中でスコアを計算し，最もスコアの高い組み合わせを並列句の範囲として同定する．複合した固有表現に含まれる並列構造は，“T and B" や “myeloid and lymphoid"のような並列する名詞句や形容詞句からなるため，本手法は名詞句と形容詞句による並列構造を同定の対象とする．また，“-2, -4, -5 and -13"のような三つ以上の並列句からなる並列構造の場合は，等位接続詞の前後にある並列句の範囲 (“-5 and -13") のみを同定する．

まず，文や節などの名詞句と形容詞句以外からなる並列構造を対象外とするための前処理として，品詞タグを用いたルールに従って並列句の可能性がある範囲を抽出する．本手法では，等位接続詞の前後で動詞，前置詞，カンマ (“,”)，コロン (“:”)，セミコロン (“;”)．三点リーダ (“…”) を含まない最長の範囲を抽出する．例えば，“Antigen complexed with major histocompatibility complex class I or II molecules on the surface of antigen presenting cells ...”という文が入力された場合は，“major histocompatibility complex class I or II molecules”が並列句の候補の最長範囲として抽出される．また，“-M1 , and -M2"のような等位接続詞と先行する並列句の間に出現するカンマは，事前に除去してから前処理を行う．

次に，系列アライメントから並列構造の範囲の候補に対してスコア計算を行い，最大スコアを持つ範囲を並列構造の範囲として同定する．系列アライメントでは，一方の系列を置換・挿入・削除の三つの編集操作を通して他方の系列へ変換し，これらの変換にかかる一連の操作の中で系列間の対応関係が求められる．それぞれの編集操作に対してはスコアが割り当てられ，編集グラフでスコアの総和が最も高い操作系列が最適なアライメントとして決定される．本手法は，並列構造のアノテーションを用いずに範囲を同定するため，最適な並列構造をもつ範囲のスコアが最大になるように，学習済みの言語モデルを用いてスコアの総和の最大値を算出する．具体的には，挿入・削除をスキップ，置換をマッチの操作として定義し，二つの編集操作に対して学習済みの言語モデルから得られる単語分散表現を元にしたスコアを付与する．マッチの操作に対するスコアは，対応する単語同士のコサイン類似度に基づいて計算し，スキップの操作に対しては，開発データ [10] を使用した事前実験によって決定された特定の値を割り当てる．本稿では，スコア計算に用いる学習済み言語モデルとして，筆者ら [9] の評価実験にて最も高い性能を示した ELMo[11] を採用し，PubMed で入手可能な論文を用いて訓練したモデルを使用する．これらのスコアの算出を，前処理で抽出された範囲の中での全ての組み合わせで実行し，最大のスコアを持つ組み合わせを並列構造の範囲として同定する．編集グラフのサイズが大きくなるほどスコアも他の候補に比べて大きくなるため，各編集グラフで算出された最大スコアに対して，操作系列の長さで正規化したスコアを使用する．

3.2 複合化した固有表現の正規化

上述の並列構造解析器と従来法による固有表現認識器の出力を利用して，複合した固有表現を個別の固有表現に分解する．具体的には，並列構造を含んでいる固有表現の中で省略されている単語を識別し，それぞれの固有表現に結合させる．省略されている単語は，固有表現認識器で抽出された範囲から並列構造の範囲を取り除くことで識別できる．例えば，固有表現認識器で “Human T and B lymphocytes” が抽出され，並列構造解析器で “T and B” が同定された場合，“Human” と “lymphocytes” が省略されている単語であることが分かる．これらの単語をそれぞれの並列句と組み合わせることで，二つの固有表現 (“Human T lymphocyte”，“Human B lymphocyte”) が抽出される．

固有表現認識器には，従来法の使用を想定して BioBERT [12] を固有表現認識タスク向けに fine-tune したモデルを使用する．BioBERT は，PubMed で保管されている生命科学分野の論文のアブストラクトと生物医学とライフサイエンス分野の論文のフルテキストを訓練コーパスとして BERT [13] を事前学習した，バイオ分野向けの学習済み言語モデルの一つである．固有表現認識モデルへの fine-tune を行う際は，Lee ら [12] と同様に，BioBERT に Softmax 層を加え，各サブワードの固有表現ラベルを予測する．複合した固有表現については分解せず，一つの固有表現の事例として抽出されるよう，データセットの fine-tune を行った．

4 評価実験

本稿では，複合化した固有表現について個々の固有表現がアノテーションされた GENIA Term annotation [14] を用いて評価実験を行う．GENIA Term annotation では，11 タイプ[1)] の等位接続詞のタグで複合名詞が分類されており，全体の 97%が AND と OR のラベルで占められる．本実験では AND と OR を実験対象の等位接続詞として性能の比較を行い，並列構造解析器の前処理で用いる品詞は gold の品詞を使用した．

4.1 評価方法

本実験では，Muis ら [3] の評価方法に基づいて，コーパスの最初の 80%と 10%の文を固有表現認識器の訓練及び開発用，残りの 10 %を評価用のデータセットとして用いる．固有表現は DNA, RNA, Protein, cell_line，cell_type の 5 タイプを対象とし，DNA と RNA，Protein にあるサブカテゴリは親カテゴリに統合する．また，GENIA Term annotation には “EBV - transformed human B cell line” の “human B cell line” のような入れ子になった固有表現が全体の 10%に含まれており，固有表現認識器の学習のため，訓練データと開発データには入れ子状の固有表現を除去した．

本稿では，全ての固有表現と並列構造を含んだ固有表現に限定した設定で精度・再現率及び F 値で性能を比較する．全ての固有表現には BioBERT を

1) AND, BUT_NOT, AS_WELL_AS, AND/OR, AND_NOT, TO, NEITHER_NOR, THAN, VERSUS, NOT_ONLY_BUT_ALSO

ベースラインとして使用し，並列構造を含んだ固有表現を対象にした評価では，固有表現認識器の代わりに全ての連続した範囲にある固有表現を Oracle として組み合わせたモデルを提案手法の upper bound として使用した．また，複合する固有表現を不連続な固有表現として抽出する手法の一つである Dai ら [4] の手法で実験した結果についても示す．Dai らの手法は，複合した固有表現の中で連続した範囲にある固有表現に対して偽陽性を判定できないため，並列構造を含む固有表現のみの評価については再現率のみを示す．

4.2 実験結果

	精度	再現率	F1
	全固有表現		
BioBERT	0.756	0.636	0.691
BioBERT+ours	0.764	0.652	0.703
Dai et al [4]	**0.791**	**0.762**	**0.776**
	複合する固有表現		
BioBERT+ours	0.325	0.441	0.374
Oracle+ours	0.625	**0.678**	0.650
Dai et al [4]	-	0.644	-

表 1 GENIA Term annotation での実験結果

実験結果を表 1 に示す．全ての固有表現を対象にした評価について，提案手法はベースラインの性能を上回ったが，並列構造を含む固有表現のみを対象にした設定では，Oracle の固有表現と組み合わせたモデルと比べ，0.28 ポイントの改善の余地が見られた．Dai らの手法と比較すると，全ての固有表現において提案手法が近い性能を示し，複合した固有表現に対しては oracle の固有表現を代わりに使用した手法が同等の再現率を示した．この結果から，複合化された固有表現の正規化において並列構造解析器が有効に働いていると考えられる．また，提案手法と Dai らの手法にある性能の差の原因として，Dai らの手法が入れ子の固有表現も対象にしている点や後述の固有表現に対する誤りによる点が挙げられる．

5 今後の課題

提案手法では，三点の特徴をもつ固有表現に対して抽出が困難になる．一つ目は，並列構造を含んだ長い範囲の固有表現である．評価データセットで生じた固有表現認識器の誤りの例を図 2 に示す．この例では，正解が “Oct-1"と “Oct-2A"と独立した固有表現であるものの，認識器では “recombinant Oct-1 and Oct-2A protein"と一つの固有表現として抽出されている．このような抽出の誤りは，複合された固有表現の出現パターンが類似している点が原因として考えられる．実際に “PRDII and tetrahexamer binding proteins"では，“PRDII binding protein", “tetrahexamer binding protein"と別々の固有表現として扱われており，このような一部のタグで頻出する単語が周辺にあることで，複合した固有表現と誤って認識されていると考えられる．

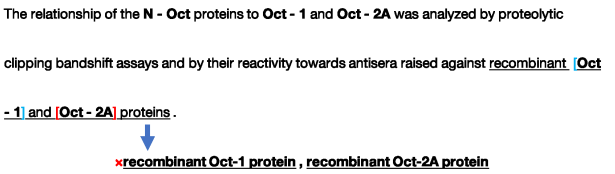

図 2 BioBERT におけるエラー例

二つ目は，等位接続詞を含んだ単一の固有表現である．等位接続詞を含む固有表現には，“Signal transducer and activator of transcription protein"のような，“Signal transducer"と “activator of transcription"が並列構造を持っているにも関わらず一つの固有表現として扱われる事例が存在する．本手法は，二つのモジュールで出力される範囲に重複がある場合は必ず複合した固有表現であると仮定しているため，固有表現認識器で正しく予測されても並列構造解析器が別の固有表現として正規化される．また，このような固有表現には略語 (“STAT") が範囲に加えられる場合もあり，正規化がより困難になる．抽出された固有表現が複合されたものか分類する機構との組み合わせとともに，コーパスの整備が今後必要になる．

三つ目は，三つ以上の並列句からなる並列構造を含む固有表現である．複合された固有表現の中には “human interleukin -2, -4, -5 and -13"のような三つ以上の固有表現 (“human interleukin -2", “human interleukin -4", “human interleukin -5", “human interleukin -13") が複合されたものも存在し，これらの固有表現の正規化には三つ以上の並列句それぞれの範囲を同定する必要がある．本手法の並列構造解析器は等位接続詞の前後にある並列句しか同定できないため，三つ以上の並列句からなる並列構造への対応も課題として挙げられる．

参考文献

[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[2] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1064–1074, Berlin, Germany, August 2016. Association for Computational Linguistics.

[3] Aldrian Obaja Muis and Wei Lu. Learning to recognize discontiguous entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 75–84, Austin, Texas, November 2016. Association for Computational Linguistics.

[4] Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5860–5870, Online, July 2020. Association for Computational Linguistics.

[5] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, Vol. 20, No. 4, pp. 507–534, 1994.

[6] Masashi Shimbo and Kazuo Hara. A discriminative learning model for coordinate conjunctions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 610–619, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[7] Jessica Ficler and Yoav Goldberg. A neural network for coordination boundary prediction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 23–32, Austin, Texas, November 2016. Association for Computational Linguistics.

[8] Hiroki Teranishi, Hiroyuki Shindo, and Yuji Matsumoto. Decomposed local models for coordinate structure parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3394–3403, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Yuya Sawada, Takashi Wada, Takayoshi Shibahara, Hiroki Teranishi, Shuhei Kondo, Hiroyuki Shindo, Taro Watanabe, and Yuji Matsumoto. Coordination boundary identification without labeled data for compound terms disambiguation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3043–3049, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[10] Jessica Ficler and Yoav Goldberg. Coordination annotation extension in the Penn Tree Bank. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 834–842, Berlin, Germany, August 2016. Association for Computational Linguistics.

[11] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2019.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 2003.