専門用語抽出のための並列名詞句の教師なし範囲同定

澤田 悠治[1]　和田 崇史[1]　芝原 隆善[1]　近藤 修平[1,2]　松本 裕治[1,2]
[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, wada.takashi.wp7, shibahara.takayoshi.sk4, shuhei-k, matsu}@is.naist.jp

1　はじめに

科学技術論文では，物質名や種名などの専門用語が等位接続詞を用いて並列的に記述されている場合がある．並列構造を持つ専門用語では，例えば‘Amorfrutins A and B’の‘ B’のように，2 つ目の用語の前半が省略され、後半部分のみが並列になっている．従来の固有表現抽出器では，このような構造を持つ専門用語に対して認識が困難になる問題が生じる．

専門用語などの名詞句の並列構造の範囲を同定するタスクとして，並列構造解析がある．近年の並列構造解析では，Ficler ら [1] や寺西ら [2] のニューラルネットワークを用いた手法により，高い解析性能を獲得している．しかし，並列構造のアノテーションが付与されているデータセットが必要になるため，学習面でのコストがかかる．また，専門用語の抽出の際に固有表現抽出器と組み合わせる可能性を考えると，より簡便な解析手法が望まれる．自然言語処理ツールの一つである Stanford CoreNLP[1] (CoreNLP) では句構造解析から並列構造の範囲を予測できるが，名詞句の並列に対して実際よりも広めに範囲を予測する傾向がある．

本研究では，並列句の意味的な類似性に着目し，動的計画法によるマッチングを用いる手法で並列句の範囲同定を行う．黒橋ら [3] は，ルールによって付与されたスコアが最大になる範囲を並列句の範囲と予測しているが，本稿では単語分散表現による類似度を利用する簡便な手法の有効性を確認する．提案手法では，学習済みの言語モデルから等位接続詞の前後にある各単語間の類似度を算出し，動的計画法を用いて類似度の和が最大になる範囲を並列構造の範囲として出力する．筆者らのデータセットを対象とした評価実験において，提案手法が CoreNLP より高い精度を得たことを示す．

[1]https://stanfordnlp.github.io/CoreNLP/

2　関連研究

これまでの並列構造解析の研究では，並列句の類似性に基づいた手法が数多く提案されてきた．昨今ではニューラルネットワークを用いたモデルが評価データセットにおいて高い性能を示している．Ficler ら [1] の手法では，外部の構文解析器から抽出した並列句候補を分散表現として，並列句同士のユークリッド距離が近くなるように学習する．寺西ら [2] は並列句の内側と外側の境界を学習し，3 つ以上の並列句や入れ子になっている並列構造に対しても同時に解析可能な手法を提案している，これらの手法は並列構造がアノテーションされたデータセットが必要となる．専門用語抽出では，固有表現抽出器と組み合わせる可能性があるため，同じデータに対して固有表現と並列構造のアノテーションを付与する必要があり，学習面でのコストがかかる．

ニューラルネットによる手法以前では，動的計画法を用いて類似度の和の最大値を求め，並列句の範囲を予測する手法が提案されている．黒橋ら [3] は，品詞や文字タイプの一致などの文節単位のルールに基づいて経路にスコアを付与する．新保ら [4] は，系列アライメントのエッジ，接辞などの形態情報，品詞などを素性としている．原ら [5] は新保らの手法を拡張し，入れ子状になった並列構造に対して，個々の並列構造の類似度の和から範囲の予測を行っている．

本研究では，これらの動的計画法を用いた手法に基づき，学習済み言語モデルの単語分散表現を用いて並列句の類似度を求める．近年では BERT [6] や ELMo [7] などの言語モデルに，科学技術論文の単語を学習させたモデルが公開されている．これらの言語モデルから専門用語などの単語同士の類似度が計算できるため，並列句の類似度をより容易に求められる可能性がある．並列構造を持つ専門用語は名詞または名詞句であるため，本稿では名詞句における並列構造の範囲同定のみに焦点を当て，並列構造のアノテーションが付与された学習データセットを必要としない簡便な予測

モデルの構築を目指す．

3　提案手法

本研究で行う解析内容の概要を図 1 に示す．始めに，並列句の範囲の候補をルールベースで抽出し (1)，抽出した範囲で等位接続詞の前後の語系列で類似度テーブルを作成する (2)．類似度テーブルから類似度の和の最大値を動的計画法を用いて求め (3)，並列句の範囲として出力する (4)．

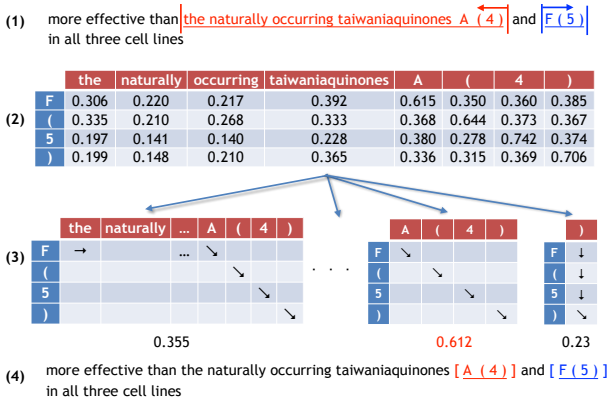

図 1: 提案手法の過程

3.1　並列句の候補の抽出

並列構造の範囲の上限をルールベースで抽出する．以下の 3 つのルールに従い，いずれかのルールを満たす前後の語系列をそれぞれ等位接続詞から後ろ向き，前向きに見て分割を行う．なお，3 つ目のルールは，‘(HT-29 and T84)’ などの丸括弧内にある並列構造を抽出するために適用する．実際に図 1 の (1) では前置詞の出現に従い，‘the naturally occurring taiwaniaquinones A (4)’ と ‘ F (5)’ が上限の範囲として抽出されている．

- 前置詞が 1 回出現
- 動詞が 2 回出現
- 前の系列で ‘ (’，後の系列で ‘)’ が先に出現

3.2　類似度テーブルの作成

次に，前後の単語系列に含まれる各単語同士の類似度を求め，類似度の和が最大となる範囲を並列構造の範囲として出力する．本手法では，学習済み言語モデルから得られる分散表現を用いて，コサイン類似度から単語間の類似度を算出する．前後の全単語の組み合わせで類似度を計算し，類似度テーブルを作成する．

3.3　動的計画法による範囲同定

この類似度テーブルから Viterbi アルゴリズムを用いて，類似度の和の最大値を得る経路を求める．ここで，単語の対応関係を示す経路を選択する際，横方向および縦方向にはペナルティとして任意の負の値を付与する．これは抽出された前後の語系列から，同じ長さの並列句を取りやすくするためである．また，抽出された語系列から最適な類似度の和を求めるため，前の語系列 $\mathrm{w}_{i:N}(1 \leq i \leq N)$ の N 通りの組み合わせで類似度テーブルを作成し，類似度の和の最大値をそれぞれ求める．そのため，これらの和を対して各テーブルの前の語系列の長さ N^{α} (α は $0 \leq \alpha \leq 1$ の任意のパラメータ) で正規化し，最大となったテーブルの範囲を並列句として出力する．

また，名詞句の並列構造の他にも，文や動詞句などの並列句が出現するため，名詞句以外の並列構造に対してルールで誤って出力される可能性がある．このような並列句はルールによって部分的に抽出されるため，語系列の類似度は低くなると考えられる．そこで，これらの語系列は類似度の和が低くなると仮定し，0 から 1 の任意の閾値を上回る類似度を持つ範囲のみを名詞句の並列構造として出力する．

4　実験

4.1　評価方法

本稿では，天然物化学に関する研究の論文誌である Journal of Natural Products(JNP) を評価データセットとして用いる．951 件のアブストラクト全 3,398 文から，等位接続詞 ‘ and’ を含む文をランダムに 100 件サンプリングし，人手でアノテーションを行い，評価データセットを作成した．100 文のうち並列句は 158 件，そのうち 122 件が名詞句による並列句となる．なお，入れ子になっている並列構造も評価対象とし，3 つ以上の名詞句からなる並列構造に関しては，等位接続詞に隣接する名詞句のみを正解としてアノテーションを付与している．

専門用語の抽出を目的としているため，名詞句における並列構造を適合率・再現率・F 値で評価する．名詞句として出力され，前後の並列句の両方の範囲が一致した事例を正解とし，いずれかの並列句が誤った範囲で予測されていれば不正解とする．また，36 件の名

詞句以外の並列構造が正しく除去されているかについて，全ての並列句での一致率から評価する．この評価では，名詞句の範囲の一致と共に，名詞句以外の並列構造に対して何も出力されなかった場合を正解，いずれかの単語が出力された場合を不正解とする．これを全並列句 158 件で行い，正解数の割合を計算する．

4.2 ベースライン

比較するモデルに関しては CoreNLP をベースラインとし，類似度算出で使用する言語モデルは SciBERT [8]，ELMo [7] を用いて精度の検証を行う．まずベースラインの CoreNLP では，等位接続詞の前後にある名詞または名詞句，数詞，代名詞を抽出し，名詞句の並列構造として出力する．SciBERT は Semantic Scholar の約 114 万件の論文のフルテキストで BERT [6] を学習したモデルで，ELMo は PubMed を用いて学習したモデルを使用する．また，SciBERT では最終層で出力される分散表現 (N=12) に加え，Embedding の層にあるサブワード の分散表現 (N=0) も比較対象とする．物質名など複数のサブワードで構成される単語は，サブワードベクトルを平均して単語ベクトルを生成する．

上記のモデルに加え，ルールによる系列分割の有効性を検証するため，CoreNLP と ELMo を組み合わせたモデルを作成する．このモデルではルールの代わりに CoreNLP からベースラインと同様の方法で並列構造の範囲上限を定め，その後 ELMo で範囲を絞り込む．各モデルのペナルティ，正規化のパラメータ，閾値を表 1 に示す．ベースライン以外の各モデルは，評価データセットで最も良い結果を得た設定を用いて評価する．

	Penalty	Norm	Threshold
SciBERT(N=0)	-0.4	0.7	0.1
SciBERT(N=12)	-0.2	0.9	0.6
ELMo	-0.1	0.9	0.45
CoreNLP+ELMo	0.0	0.3	0.4

表 1: 各モデルのペナルティ，正規化のパラメータおよび閾値の設定

4.3 実験結果

各モデルの評価を表 2 に示す．NP は名詞句，All は全並列句による一致の割合を表す．名詞句の一致の評価においては，ELMo を使用したモデルが最も高い性能を示しており，ELMo の双方向 LSTM を用いて文脈を考慮させた単語ベクトルが，同一文内の単語マッチングにおいて有用であると考えられる．また，SciBERT を用いたモデルでは Embedding の層を用いたモデルがベースラインより高い性能を示したのに対し，最終層はベースラインを下回る性能を示している．最終層の出力では，入力文中の単語同士で比較した時，本来類似する単語同士が位置によって類似度が低くなり，他の単語との類似度より下回る傾向があった．そのため，SciBERT を用いたモデルでは，文脈を考慮しないサブワードの分散表現を用いた方が単語のマッチングでは有用であると考えられる．

	NP			All
	P	R	F	Acc
CoreNLP	0.521	0.496	0.508	0.576
SciBERT(N=0)	0.584	0.664	0.621	0.608
SciBERT(N=12)	0.506	0.301	0.421	0.399
ELMo	**0.680**	**0.713**	**0.696**	0.639
CoreNLP+ELMo	0.675	0.631	0.652	**0.703**

表 2: JNP での比較

全ての並列句を対象とした場合，CoreNLP と ELMo を組み合わせたモデルがルールを用いた ELMo のモデルより 0.06 ポイント高い一致率を示している．これは，CoreNLP では名詞句による並列句のみを抽出しているのに対して，形容詞句，動詞句などの名詞句以外の並列構造がルールによって抽出されたためだと考えられる．実際に (‘hyperglycemic’, ‘antihyperglycemic’) など，類似性の高い形容詞の並列構造が，名詞句として予測されているなどの誤りが生じている．そのため，全体の一致率に関しては現状 CoreNLP と組み合わせた方が高い性能を示しているが，ルールベースでは CoreNLP より 0.08 ポイント高い再現率を示していることから，今後提案モデルで予測した範囲に対して句の識別を行うことで，全並列句の一致率，名詞句の適合率が向上する可能性がある．

次に，ルールを用いた ELMo のモデルにおける，各閾値での一致率の推移を図 2 に示す．0.0 から 0.45 への閾値の増加にかけて，一致率も増加していることが分かる．閾値を設けない場合の一致率は 0.56 であり，0.08 ポイントの性能の向上が確認できる．実際に，(‘ elucidated on the basis of spectroscopic data analysis’, ‘ compared with the literature’) のような文や一部の動詞句による並列構造は閾値によって除外されている．

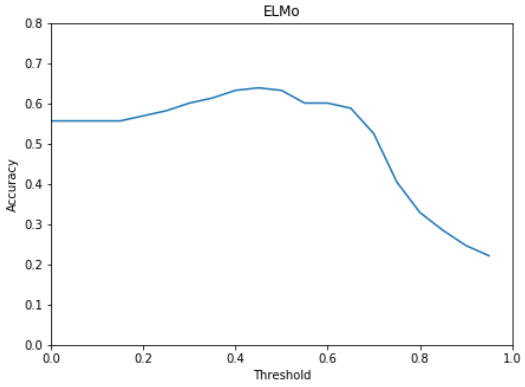

図 2: ELMo の一致率の推移．横軸は閾値，縦軸は全並列句の一致率を示す

4.4 今後の課題

Viterbi アルゴリズムによる単語マッチングの手法では，並列句の長さが前後で異なる場合に予測が困難になる．基本的に前後で語系列が同じ長さになる場合では，並列構造の範囲を比較的容易に予測できる．しかし，これらの範囲が異なる場合では，ペナルティや正規化のパラメータによって前後の範囲が同じ範囲として予測される傾向があった．特に前の範囲が 2 単語以上，後ろの範囲が 1 単語の場合はペナルティ，逆の場合は正規化のパラメータの値に影響を受けやすくなる．評価データセットで生じた実際の誤りの例を図 3 に示す．

Lupeol is a lupane-type triterpene isolated from Sorbus commixta , an oriental medicine used to treat [arthritis]correct and [inflammatory diseases]correct

図 3: ELMo モデルでの失敗例. 鉤括弧が正解の範囲，下線部がモデルで予測された範囲を表す

この例では，(‘ arthritis ’ , ‘ inflammatory diseases ’) の並列句が (‘ arthritis ’ , ‘ inflammatory’) と誤って出力されている．各単語同士のコサイン類似度は ‘arthritis’ と ‘inflammatory’ で 0.52，‘arthritis’ と ‘diseases’ で 0.54 と後者の方が高いが，類似度の差がペナルティよりも小さいため，1 対 1 の並列句として予測されている．これらの単語は意味的には類似しているが，それぞれ名詞，形容詞と品詞が異なる．そのため，単語の分散表現の他に品詞の分散表現を加えることで，類似度を下げられる可能性がある．

また，並列構造が広範囲に及ぶ場合に，現行のルールでは実際の範囲よりも狭く抽出されやすい．これらの点を踏まえ，動的計画法を用いず，広範囲の並列構造に対しても頑健なマッチングの手法が必要と考えられる．

5 おわりに

本稿では，並列構造を持つ専門用語の抽出のための簡便な名詞句の並列構造解析の手法を提案した．学習済みの言語モデルのみを用いて動的計画法から類似度の和の最大値を求め，対象となる前後の系列を名詞句の範囲として出力する．JNP による実験の結果，ベースラインである CoreNLP を上回る精度を示し，名詞句に限定すれば容易に並列句の範囲同定が可能であることを示した．今後は，提案手法の性能向上を目指すと共に，‘ and ’ 以外の等位接続詞，3 つ組の並列構造も対象にした解析方法の提案を目指す．また固有表現抽出器と提案手法を組み合わせ，並列構造を持つ専門用語抽出の応用に取り組む．

参考文献

[1] Jessica Ficler and Yoav Goldberg. A neural network for coordination boundary prediction. In *EMNLP*, pp. 23–32, 2016.

[2] Hiroki Teranishi, Hiroyuki Shindo, and Yuji Matsumoto. Decomposed local models for coordinate structure parsing. In *NAACL*, pp. 3394–3403, 2019.

[3] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, pp. 507–534, 1994.

[4] Masashi Shimbo and Kazuo Hara. A discriminative learning model for coordinate conjunctions. In *EMNLP-CoNLL*, pp. 610–619, 2007.

[5] Kazuo Hara, Masashi Shimbo, Hideharu Okuma, and Yuji Matsumoto. Coordinate structure analysis with global structural constraints and alignment-based local features. In *ACL and AFNLP*, pp. 967–975, 2009.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, pp. 2217–2237, 2018.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, pp. 3613–3618, 2019.