E-032

文全体の構造の特徴を利用した並列構造解析
Coordination Analysis with Features of String Structure

鈴木俊友 † 吉川毅 † 野中秀俊 †
Shunsuke Suzuki Takeshi Yoshikawa Hidetoshi Nonaka

1. 背景

自然言語において句や節などが並列の関係となる表現の存在は，自然言語処理の妨げのひとつである [1]．並列の表現では，主語や述部に相当する部分が二度登場するなど，通常ではありえない構造が出現するからである．並列表現は，その開始位置と終了位置，および並列として扱われている句（節）の組が曖昧になりやすく簡単に判別することができない．並列に扱われている部分がわかれば，部分ごとに処理することによって通常の文と同じように扱うことができると考えられる．そのため，並列表現の構造を正しく解析することができれば自然言語処理の精度を向上させることができる．また，長文においては複雑な形の並列表現が出現しやすいため，長文を対象とした自然言語処理には特に並列構造解析の効果が大きいと考えられる．

本論文では並列表現が持つ構造のことを並列構造とよび，ある文中に存在する並列表現がどのような構造を持つのか解析することを並列構造解析とよぶ．並列構造において並列として並べられている単語や句，文などのことを並列項目とよび，並列構造中の並列項目以外の部分を接続部と呼ぶ．

並列構造解析に関連した研究として，用いられている単語を並列項目間で比較したものがある [1-3]．これらの研究では，単語の意味的な類似度や単複の一致，ウェブ検索における共起頻度などが並列構造解析に有効であるとしている．実際に並列構造解析を行う際には，並列構造のパターンを書き出してマッチングを行う方法 [2][3] や，機械学習を利用した統計的な手法 [4][5] がある．一般に，パターンによるマッチングを用いた方法は精度が高くなりやすいが，対象となる言語の知識と多くの手間が必要となる．一方で機械学習を利用した方法は人手が余りかからないが，パターンを利用する方法に比べ精度が低いという指摘もある．

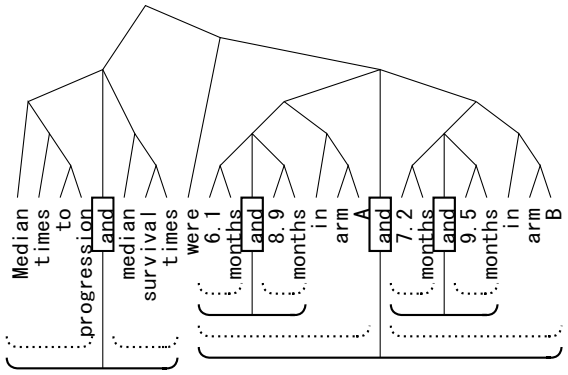
また，ほとんどの並列構造解析は並列項目の単語を比較する手法である．しかし，この手法による並列構造解析は，並列項目に含まれる単語が増えた場合に比較対象が曖昧になるなどの欠点を持つ．

本論文では，文全体の特徴を考慮に入れた並列構造解析を行うことによって，従来手法の欠点を解消することを目指す．そのために，先行研究である原らの手法 [5] に大域的素性 [7][8] を適用する手法を提案する．

2. 従来手法

提案手法を説明するために，従来手法である原らの手法 [5] を簡単に説明する．従来手法では，入力された文に対する並列構造を次の 3 段階から推定する．

1. 入力文に対する並列構造の候補を生成する．
2. 各並列構造の特徴を示す素性ベクトルを計算する．
3. 素性ベクトルを元に正しい並列構造を推測する．



破線：並列項目, 実践：並列構造，四角：接続部

図 1: ある並列構造を含んだ文を木の形で表したもの

本論文では手順 2 について考察する．手順 3 における正しい並列構造の推測は，素性ベクトルと重みベクトルの内積をとることによって並列構造の妥当性を示すスコアを計算し，そのスコアが最も高い並列構造を選択することによって正しい並列構造を推測する．重みベクトルの調整は，Collins のパーセプトロン学習の拡張 [6] を利用して行う．

3. 従来手法の問題点と改善

3.1. 従来手法の問題点

従来手法では，並列項目の単語数が少ない場合や名詞句が並列項目となっている場合には比較的精度がよい．一方で，並列項目の単語数が多い場合や文や動詞句が並列となっている場合には精度が悪くなる．この理由について考察する．

まず並列項目に含まれる単語の数によって精度が変化する原因について考える．従来手法では主に並列項目の単語を比較する．これは，並列項目同士で構成が対応しており，対になる単語が存在しているという仮定から行われる．単語を比較して意味が近かったり単複が一致していたりした場合，並列項目間における対になる単語と予想でき，並列構造解析の手がかりとすることができる．以下の例では，並列項目間で各単語の対応がわかりやすい形になっている．

> Phorbol ester-induced production of 02- was similar in cells treated [with *RNR*-60] and [without *RNR*-60] gangliosides.

角括弧で囲われた部分が正しい並列項目の範囲である．この場合，“with” と “without”，各並列項目の

†北海道大学大学院情報科学研究科

Copyright © 2012 by
The Instiute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

”*RNR*-60” がそれぞれ対応していると考えられる．並列項目を構成する単語が少ない場合，対となる単語が明らかであり比較が容易に行うことができる．しかし並列項目を構成する単語が多くなると，単語と単語の対応が曖昧になる．

> NF-kappa B is activated *-58 upon cellular treatmemt by [phorbol esters] and [the cytoline tumor necrosis factor alpha(TNF alpha)].

この文では，後ろ側の並列項目の単語数が多いうえに，前の並列項目の単語数と大きな差がある．このような場合では単語の対応が明確ではない．単語の対応に曖昧さを持つ並列構造では，単純に並列項目の単語を比較しても，並列構造の特徴を正しく表現できない．この問題を解決するには，並列項目に含まれる単語の数に影響を受けないような比較の方法が必要である．

次に，文や動詞句などが並列となる場合に精度が悪くなる原因について考える．多くの並列構造では，並列項目が並列の関係として扱われていることが多い．しかし並列項目が文や動詞句などの場合は，単に二つの文を接続詞によりつなげただけの場合が多くなる．このような並列構造の場合，並列項目を構成する単語の類似度が小さくなってしまう．例えば，文が並列項目となっている次のような文を考える．

> [Receptor transcription were 4.6 kilobases(kb)], and [no variant sizes were observed *-61].

並列項目となっている文は，それぞれ動詞の形が違う上に構成している単語に似た特徴を持つものが少ない．そのため，並列項目を比較するだけでは正しい並列構造かどうかを判断することが難しい．このような場合には，並列項目を比較するだけではなく，並列項目以外の部分から手掛かりを得るような方法が必要となる．

4. 改善手法の提案

前節で述べた従来手法での問題点を改善するためには，並列項目に含まれる単語数に影響を受けづらい比較方法と，並列項目の比較では十分な手掛かりが得られない場合でも正しい並列構造を推測できる手掛かりを得る方法が必要である．

並列項目を構成する単語数に依存しない比較方法として，単語ごとではなくより広い範囲から手がかりを得る大域的素性 [7][8] を利用する方法がある．この大域的素性を利用することによって，並列項目自体の特徴を捉え並列項目を比較する．利用する大域的素性は，例えば並列項目に動詞が含まれているかといったものである．大域的素性では並列項目自体を比較するため，並列項目を構成する単語数が変化した場合でも比較に問題が出ることを抑えることができる．

並列構造解析以外の自然言語処理では，大域的素性を利用することによって精度が向上することが報告されている．Finkel らは，英文をその意図によって分類する問題において，文全体で述部が出現した回数を大域的素性として適用することにより，分類精度を向上させることができると述べている [7]．また渡邉らは，述語の語義と単語の文中の役割を同定する問題で，述語の語義と文全体の構造に対して定義される大域的素性を用いている [8]．渡邉らは，大域的素性を用いることによって複数の項が同じ意味役割を持つなどの通常ではありえない推定を抑制することなどができるとしている．

大域的素性は単語の品詞推定などにおいて有効であるとされ，ほかの自然言語処理でも有用だと考えられる．しかし素性として用いられることは希で，上記の例でも制限された使い方がされている．大域的素性の利用が少ない原因は，大域的素性を用いた場合には多くの自然言語処理で利用されているビタビアルゴリズムが利用できなくなるためである．ビタビアルゴリズムでは，近傍にある単語同士の比較で得られる素性のように，局所的な影響しか受けない素性のみでなければ利用できない．そのため，ビタビアルゴリズムと大域的素性を同時に利用する場合はさまざまな工夫や制限が必要となる．並列構造解析ではビタビアルゴリズムを利用していないため，制限なしに大域的素性を利用できる．

次に，並列項目の比較では十分な手がかりが得られない場合に対する改善手法について考える．並列項目が文となっている次のような並列構造の場合，従来手法では誤った並列構造が出力されやすい．

> [The virus with four Sp1 sites did outgrow the three Sp1 virus in 35 days of culture] and [CTG-monomer virus outcompeted the CTG-dimer virus in 42 days].

角括弧が正しい並列構造の並列項目，下線部が誤って出力されやすい並列構造の並列項目である．正しい並列構造の並列項目では，類似する特徴を持つ単語の割合が並列項目間で少ない．そのため従来手法では，似ている単語の割合が十分多くなるような，並列項目の単語数が少ない並列構造の候補を出力してしまう．このような場合には，候補の並列構造の並列項目を比較するだけでは正しい並列構造を推定することは難しい．しかし，間違った候補の場合は並列項目以外の部分の構造が崩れることになる．上記の間違った候補の場合では，“did” と “outcompeted” の動詞が離れた位置で二度出現するという通常ではあまりない形になっている．並列項目以外の部分に対して大域的素性を利用することによって，間違った候補に見られるような文の形の崩れを捉えることを目指す．

並列項目と文全体に対する大域的素性の利用は，原らの手法 [5] の候補からの推定部分にこれらの大域的素性を追加することによって行う．また，今回は並列構造を一つのみ含む文に対象を限定して考える．以下に原らの手法に大域的素性を導入する方法について述べる．まず並列項目を入力とする属性を定義する．並列項目に関する属性は

- 品詞それぞれの含まれている数

Copyright © 2012 by
The Instiute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

- 含まれている品詞の種類

の二つを利用する．並列項目 S の並列項目に関する属性 $a_{CJT}(S)$ を従来手法と同様に特徴関数を利用して定義する．並列項目に関する属性の例を次式に示す．

$$a_{CJT_4}(S) = \begin{cases} 1 & S \text{ に含まれる名詞の数が } 4 \text{ つ} \\ 0 & \text{上記以外} \end{cases}$$

同様に，並列項目以外の部分 E に対する属性を定義する．属性は，各品詞の出現回数を利用する．

これらの属性を利用し，並列項目間の素性 ϕ_{CJT} と並列項目以外の部分に対する素性 ϕ_{ELSE} を定義する．並列項目間の素性 ϕ_{CJT} の定義は，従来手法と同様に属性を二つ組み合わせることによって行う．並列項目 S_1 と S_2 を入力とし，二つの並列項目に関する属性 a_{CJT_p}, a_{CJT_q} で定義される並列項目間の素性 ϕ_{CJT_j} は以下のようになる．

$$\phi_{CJT_j}(S_1, S_2) = \begin{cases} 1 & a_{CJT_p}(S_1) = 1 \\ & \text{かつ } a_{CJT_q}(S_2) = 1 \\ 0 & \text{上記以外} \end{cases}$$

j は属性の組み合わせごとにつけられた番号である．並列項目以外の部分に対する素性 ϕ_{ELSE} も同様に，並列項目以外の部分 E を入力とする属性 a_{ELSE} を二つ組み合わせることによって定義する．

$$\phi_{ELSE_k}(E) = \begin{cases} 1 & a_{ELSE_r}(E) = 1 \\ & \text{かつ } a_{ELSE_s}(E) = 1 \\ 0 & \text{上記以外} \end{cases}$$

並列項目 S_1 と S_2 を入力とする並列項目間の素性 $\phi_{CJT_j}(S_1, S_2)$ を順に並べたベクトルを並列項目間素性ベクトル $\boldsymbol{f}_{CJT}(S_1, S_2)$ と定義する．

$$\boldsymbol{f}_{CJT}(S_1, S_2) = (\phi_{CJT_0}(S_1, S_2), \phi_{CJT_1}(S_1, S_2), \ldots,)^T$$

同様に並列項目以外の部分 E に関する素性ベクトル $\boldsymbol{f}_{ELSE}(E)$ を定義する．

$$\boldsymbol{f}_{ELSE}(E) = (\phi_{ELSE_0}(E), \phi_{ELSE_1}(E), \ldots,)^T$$

並列項目 S_1, S_2 と並列項目以外の部分 E は文と並列構造によって決まるため，$\boldsymbol{f}_{CJT}$ と $\boldsymbol{f}_{ELSE}$ の値は，入力文 x とそれに対する並列構造の候補 t によって値が決まる．

$\boldsymbol{f}_{CJT}$ と $\boldsymbol{f}_{ELSE}$ を，従来手法で生成する素性ベクトル $\boldsymbol{f}$ に連結して利用する．このときの入力文 x に対する並列構造の候補 t'_x の推測は次のような式になる．

$$\begin{aligned} t'_x &= \arg\max_{t \in T_x} \boldsymbol{w}^T(\boldsymbol{f}, \boldsymbol{f}_{CJT}, \boldsymbol{f}_{ELSE}) \\ &= \arg\max_{t \in T_x} \boldsymbol{w}^T \hat{\boldsymbol{f}}(x, t) \end{aligned}$$

T_x は入力文 x に対して生成される並列構造の候補の集合，$\boldsymbol{w}$ は重みベクトルである．提案した素性ベクトルを追加した新しい素性ベクトルを $\hat{\boldsymbol{f}}$ とする．

Algorithm 1 重みベクトルの学習アルゴリズム

Input: 訓練データ $((x_1, t_1^*), (x_2, t_2^*), \ldots, (x_L, t_L^*))$
Initialization: $\boldsymbol{w} \leftarrow \boldsymbol{0}$
repeat
  for $i = 1 \ldots L$
    $t' = \arg\max_{t \in T_{x_i}} \boldsymbol{w}^T \hat{\boldsymbol{f}}(x_i, t)$
    if $t' \neq t_i^*$ **then**
      $\boldsymbol{w} \leftarrow \boldsymbol{w} + \mu\hat{\boldsymbol{f}}(x_i, t_i^*) - \mu\hat{\boldsymbol{f}}(x_i, t')$
    endif
  endfor
until 重みベクトル $\boldsymbol{w}$ の更新があった場合
Output $\boldsymbol{w}$

重みベクトル w の調整は Collins のパーセプトロン学習の拡張 [6] を利用する．Collins の手法を原らの手法に適用した重みベクトル $\boldsymbol{w}$ の学習アルゴリズムを **Algorithm** 1 に示す．入力データとして与えられるのは，並列構造をひとつだけ含んだ英文 x_i と，x_i に対する正しい並列構造 t_i^* の組み合わせ $(x_i, t_i^*)_{i=1,2,\ldots,L}$ の集合である．重みベクトル $\boldsymbol{w}$ はゼロベクトルで初期化される．英文に対する正しい構文木の推測は構文木のスコアを利用することによって行われる．アルゴリズム中の T_{x_i} は 文 x_i から生成される並列構造の候補全ての集合である．今回は接続部の位置を教師データから与え，接続部の位置と入力文 x_i を元にとりうる全ての並列構造を T_{x_i} として与えている．

5. 実験

以上の考察をもとに，従来手法と提案手法の並列構造解析における精度を検証した．実験は，従来手法，並列項目間素性を利用した手法，並列項目間素性と並列項目以外の部分に関する素性の両方を利用した手法の三種類に関して行った．実験の対象となるデータには Genia Treebank Beta [9] を利用した．実験の結果，提案手法は従来手法よりよい結果を示した．

5.1.Genia Treebank Beta

今回の評価実験では Genia Treebank Beta [9] を用いて評価した．Genia Treebank Beta は，アメリカ国立医学図書館のオンライン医学文献検索サービスである MEDLINE のアブストラクトの文に対して品詞情報や構文構造などの情報を付与したものである．Genia Treebank Beta には並列項目の範囲が明示的に示されているほか，並列項目が名詞句であるなどの並列構造の種類も示されている．Genia Treebank Beta は医学論文をもとにして作られたコーパスのため，並列表現がよく出現するという特徴がある．今回の実験では Genia Treebank Beta に含まれる文のうち，並列構造を一つだけ含む文のみを対象とした．

5.2. 結果

今回の実験では，対象となる Genia Treebank Beta の 623 文に対し 5 分割交差検定を行った．その結果，すべての文の中で提案手法が並列構造を正しく推測した文の割合は，従来手法が正しく推測した文の割合を，並列項目間素性のみを利用した場合 4.81%，両方を利用

Copyright © 2012 by
The Instiute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 1: 実験結果

並列項目の種類	(f, f_{CJT})	(f, f_{CJT}, f_{ELSE})
NP	1.12	1.03
VP	1.2	1.11
ADJP	1.12	1.05
S	0.27	1.09
PP	1.00	0.8

NP：名詞，VP：動詞，S：文
ADJP：形容詞，PP：前置詞

した場合は 2.21 % 上回った．実験の結果を並列構造の種類ごとにまとめたものを 表 1 に示す．表 1 の値は，各項目での従来手法の正答率の値を 1 とした時の提案手法での正答率の値である．

5.3. 考察

提案手法は全体を通して従来手法より高い精度となった．並列構造の種類で見ると，並列項目間素性のみを利用した場合は名詞句並列や動詞句並列，形容詞句並列の場合に精度が向上した．これは，並列項目間素性の効果として期待した並列項目の単語の数に左右されない性質が有効に働いたためと考えられる．またこのこと以外にも，従来手法では捉えることのできない単語間の境目の手がかりを得ることができたことが理由として考えられる．一般に，並列項目の最後は名詞が来る場合が多く，品詞全体の中でも名詞の割合が多い．このため，名詞句並列では並列構造の直後に名詞が存在することが多くなり，並列構造とそれ以外の部分との境目で品詞が変化しないことが多くなる．並列構造の境目で品詞の変化がない場合は，従来手法では境目の判断が難しい．一方提案手法は並列項目に注目して特徴を得るため，並列構造の境目に変化がなくても従来手法より有効に手掛かりを得ていると考えられる．

並列項目が文の場合（表 1 中の S）に関しては，並列項目間素性のみを利用した場合では従来手法より精度が悪くなった．これは，並列項目間素性を導入することによって従来手法に比べ並列項目を比較することをより重視したためと考えられる．並列項目の比較では正しい推定が難しい場合が多い文並列の場合では，並列項目を比較することを重視することが悪い影響となる．しかし，並列項目以外の部分の素性を同時に利用した場合では，この部分が改善され従来手法よりもよい結果となっている．文並列のように並列項目の比較だけでは推定が難しい場合に，並列項目以外の部分から得られる手がかりが有効であることを示している．

すべての文に対する結果では，並列項目間素性のみを利用したほうが提案した二つの素性両方を利用した場合よりよい結果となっているのは，それぞれの得意な並列構造の種類とそれが全体に占める割合の違いが原因である．Genia Tree Beta に含まれる並列構造のうち，半分以上の並列構造が名詞句並列であり，これは並列項目間素性のみを利用した場合に良い結果が出る並列構造である．一方で，両方の素性を利用した場合に良い結果が出る文並列は，Genia Tree Beta に含まれる並列構造の一割程度しかない．このような割合の差は一般の文でも同様であると考えられる．

6. まとめ

本論文では，より高い精度で並列構造解析を行うことを目標とし，原らの手法 [5] に二種類の大域的素性を用いる方法を考察した．従来手法と比較実験を行った結果，従来手法よりもよい結果を得た．

今後の課題として，今回提案した素性よりもより正確に並列項目やそれ以外の部分の特徴をとらえられるような方法を考えたい．理想的には，それぞれの部分の構造を捉え，その構造に沿った比較を行うことが望ましい．それらの特徴の効果的な組み合わせ方も重要である．また，今回は教師データから与えるとした候補集合の生成も大きな問題である．候補集合の生成を含めた並列構造解析全体の改善を目指す．

参考文献

[1] P. Resnik: “Semantic Similarity in a Taxonomy”, J. of Artificial Intelligence Research, pp.95-130,(1999)

[2] R. Agarwal *et al.*: “A SIMPLE BUT USEFUL APPROACH TO CONJUNCT IDENTIFICATION”, Proc. of the 30th annual meeting on Association for Computational Linguistics, pp.15-21, (1992)

[3] A. Okumura *et al.*: “Symmetric Pattern Matching Analysis for English Coordinate Structures”, Proceedings of the fourth conference on Applied natural language processing, pp.41-46, (1994)

[4] 原 一夫ほか: “アラインメントと機械学習を応用した並列句解析”, 人工知能学会論文誌, Vol. 22, No. 3, pp.248-255, (2007)

[5] 原 一夫ほか: “文法制約と系列アラインメントによる並列構造の解析”, 人工知能学会論文誌, Vol. 25, No. 5, pp.560-569, (2010)

[6] M. Collins: “Discriminative Training Methods for Hidden Markov Models: Theory and Experiment with Perceptron Algorithms”, Proc. of the Conference on Empirical Methods in Natural Language Processing, pp.1-8, (2002)

[7] J. R. Finkel *et al.*: “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling”, Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, pp.363-370, (2005)

[8] 渡邉 陽一郎ほか: “述語語義と意味役割の結合学習のための構造予測モデル”, 人工知能学会論文誌, Vol. 25, No. 2, pp.252-261, (2010)

[9] J.-D.Kim *et al.*: “GENIA corpus: a semantically annotated corpus for bio-textmining”, Bioinformatics, 19, suppl1, pp.i180-i182, (2003)

Copyright © 2012 by
The Instiute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.