1H1-3

機械学習と系列アラインメントを応用した日本語並列句解析

A Discriminative Learning Method for Japanese Coordinate Phrase based on Sequence Alignment

大熊 秀治 Hideharu Okuma　原 一夫 Kazuo Hara　新保 仁 Masashi Shimbo　松本 裕治 Yuji Matsumoto

奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

We explore the effective features that can be used with Shimbo and Hara's learning method for coordinations, in the task of Japanese coordinate structure analysis. Unlike their preliminary experiments which used only minimal features, we introduce more realistic features based on external resources for collocations and a thesaurus and propose efficient feature decomposition based on the distance between conjuncts. We also make modification to the label set used in the model, which is necessary for bunsetsu based Japanese processing. In an experiment with the Kyoto Text Corpus, we obtain performance exceeding that of the rule-based KNP parser, corroborating the effectiveness of the new features and the discriminative learning approach for coordinate structure analysis.

1. はじめに

自然言語処理における並列構造は，構文解析を困難にしている要因の一つである．そこでの解析誤りは，情報抽出や機械翻訳などの上位のアプリケーションでの精度低下に直接つながるため無視できない問題である．これまで構文解析の枠組みの一部として，あるいは構文解析とは独立したものとしていくつかの並列構造解析手法が提案されてきたが，未だに決定的な手法は確立されていない．

本論文では，近年新保ら [Shimbo 07] によって提案された手法に注目する．彼らは，識別学習モデルを用いて並列構造解析に取り組んでいる．彼らは提案手法の利点として，多数の素性が利用可能なことと，学習ベースのモデルであることから，新しいドメインへの適用も簡単であると主張した．しかし，彼らは英語の医学文献コーパスである GENIA を対象に，わずかな素性を用いた予備的な実験しか行っておらず，外部知識源を一切利用していない．彼らは，提案モデルは既存の構文解析器より高い性能だと主張しているが，その比較対象となった構文解析器もまた外部知識源に頼らないものであった．しかし，他の多くの並列構造解析に関連する研究では，シソーラス，あるいはウェブやコーパスから求まる統計量などの外部知識源を用いており，新保らの提案した手法がこれらの外部知識源を用いた解析モデルより有効であるかどうかは明確ではない．

新保らのモデルの有効性をより現実的な設定で評価するために，我々は彼らのモデルを日本語並列構造解析のタスクに適用する．日本語に適用するために，元のモデルを日本語解析に適したものへと修正し，またシソーラスや単語の頻度に基づく素性を導入し，モデルの有効性を Kurohashi-Nagao Parser(KNP) と比較する．KNP には黒橋らによって提案されたルールベースの並列構造解析モデル [Kurohashi 94] が組み込まれおり，また外部知識源として我々のモデルで利用するシソーラスと共通のものを利用しているため，機械学習ベースの手法の有効性を検証するのに適している．

連絡先: 奈良県生駒市高山町 8916-5
奈良先端科学技術大学院大学情報科学研究科情報処理学専攻 大熊 秀治
e-mail:hideharu-o@is.naist.jp

2. 類似度に基づく並列構造解析

並列構造は二つあるいは三つ以上の類似した構造を持つことが多い [Resnik 99]．新保らは，この特徴を利用したモデルを提案した．

新保らのモデルの基礎となるのは，上三角形型の “編集グラフ”(図 1) である．斜め方向の枝は，グラフにおいて枝の上と右に位置する語が並列関係にあることを表し，水平，垂直方向の枝は対応する語が削除あるいは挿入されていることを表す．斜め方向の枝に，枝によって対応付けられる二つの語の類似度を反映したスコアが与えられていると仮定し，編集グラフにおける最大スコアの部分経路を探索することよって並列構造解析を行う．このとき，部分経路が張られる領域の上および右に位置する一連の語が並列句を構成すると考える．詳細に関しては，新保らの論文 [Shimbo 07] を参照してほしい．

この編集グラフは黒橋らによって提案されたダイナミックプログラミング用の表 [Kurohashi 94] に類似している．どちらの手法も，グラフ (あるいは表) における最大スコアの部分経路を探索しているが，決定的な違いの 1 つは，枝に与えるスコアの調整方法である．黒橋らの手法はルールベースであり，スコア関数は少数のルールに基づいてあらかじめ人手で決定されている．それと対照的に新保らの手法は，スコア関数を多数の素性の重み付き線形和として定義している．素性は一つの枝の始点のみではなく，二つ枝が連接する点にも与えられているため，素性の表現能力は高くなるが結果的にパラメータ (素性の重み) の数は膨大になってしまい，人手によるパラメータの調整は困難である．そこで新保らは機械学習の手法であるパーセプトロンアルゴリズムを利用し，素性の重みを並列句の範囲が付与された訓練セットに最適化する方法を提案した．

3. 日本語並列構造のための BIO ラベリング

新保らのモデルは，三つあるいは四つ以上の句からなる並列構造を，連続する二つの句をペアとする構造に分解して扱う．例をあげると，“(A),(B) and (C)” という 3 つの句 (語) からなる並列構造は，(A,B)，(B,C) という二つの並列構造に分解される．これらの並列構造はグラフでは鎖状に繋ぐことのできる経路 (chainable paths) として表される．英語において三つの並列句からなる並列構造は図 2(a) のようになる．図からわ

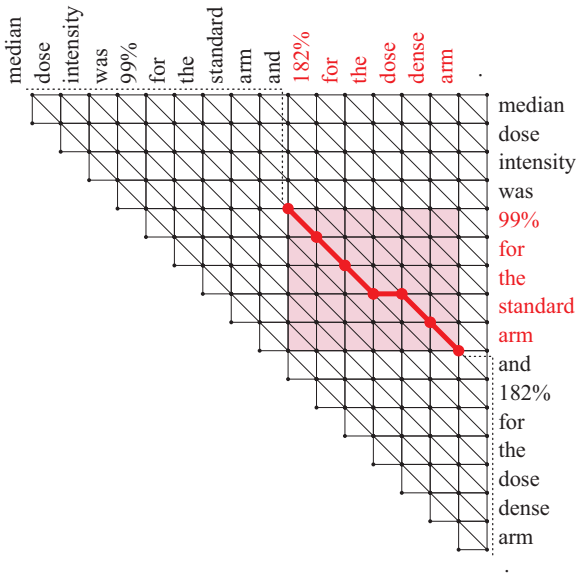
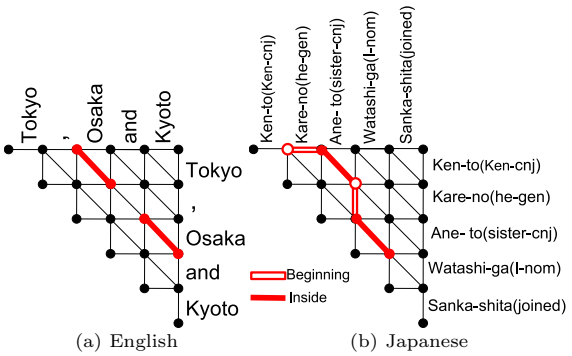


図 1: 並列構造解析ための編集グラフ [Shimbo 07].



(a) English (b) Japanese

図 2: (a) 英語，(b) 日本語における 3 つの句からなる並列構造.(a) では二つのパスは分離しているが，(b) では一つになっている．

かるように，カンマや接続詞 “and” といった並列構造の手がかりとなる表現が並列句を分離しているため，二つの並列構造を表す経路は分断された部分経路で表され，二つの構造を容易に識別できる．

日本語にはおいては，同じような部分経路の分断は起きない. 日本語の文の基本的な単位は文節である. 文節は一つあるいはそれ以上の自立語と，それに続く零個以上の機能語から成り立っている. 並列構造の手がかりとなる並列助詞の “と” や句読点は機能語であり，それらは文節内に包含されている. そのため，並列構造全体を連続する句をペアとする並列構造に分解しても，それぞれの並列構造を表す経路は分断されずに一本に繋がってしまい，並列構造の境目がが不確かになる. 図 2(b) は日本語で並列句が三つある並列構造の例である．

この不確定性を解消するために，我々は新しいラベル，*Beginning* を追加して，並列構造を表す部分経路の最初の枝を他の枝と区別する. 従って，用いられるラベルは全部で三つ，*Beginning*,*Inside*,*Outside* となる. 最後の二つのラベルは新保らのモデルで元々使われているものである. このラベルの集合を “BIO ラベリング” 法と呼ぶ. 図 2 において白抜きで表されている枝が *Beginning* ラベルの枝であり，部分経路の先頭を表している. この新しいラベルにより，二つの並列構造を表す経路が繋がってしまっていても，並列構造を区別することが可能になる．

4. 素性

新保らのモデルで用いられている素性に加えて，我々は新しいクラスの素性を導入する. これらの素性は，シソーラスや単語の頻度に基づいて計算される. このクラスの素性を外部素性と呼ぶ. これに対して，元のモデルで使われている素性は，外部知識を用いない素性であるため内部素性と呼ぶ. さらに，素性を区別する効果的な方法も新たに提案する. 簡潔に言えば，素性を枝で対応付けられる二つの文節の距離に基づいて二つに分解しそれぞれ独立に重みを学習する．

4.1 内部素性

内部素性は，文節内の情報のみに基づく素性である. 日本語で実験をする際に用いる素性は，基本的には新保らが英語に適用した際に用いた素性 [Shimbo 07] に従ったが，日本語に適用するために若干修正した．

“**素性**” を定義するためにまず，文内の各文節を “**属性**” の集合として表す. 属性は表層形や，品詞や品詞のサブカテゴリ，動詞や形容詞の活用形，助詞，句読点などである. 属性は 1 つの文節について定義される値だが，これに対して，“ **素性**” は一つあるいは二つの文節について定義される. なぜなら，素性は編集グラフの枝の始点，あるいは連続する枝の接点で定義されるからである. 枝で対応付けられる二つの文節が並列構造を構成しているかは，これらの素性の重み付き線形和の値によって推定される. これらの素性は基本的にバイナリの表示関数 (indicator function) で，一つあるいはそれ以上の文節の属性の特定の値が，枝の始点，あるいは連接する枝の接点の周辺で出現しているかでその値が決まる. 素性の例として，始点がグラフの i 行 j 列であるような斜め方向の枝を考えると，

$$f = \begin{cases} 1 & \text{if } POS_i = \text{名詞}, POS_j = \text{形容詞}, \text{and} \\ & Label = Beginning, \\ 0 & \text{otherwise.} \end{cases}$$

といったものが考えられる．

予備実験において，自立語の表層形を素性として用いると素性空間がスパースになりパーセプトロンの学習が安定しなかったため今回は用いていない. ただし，並列構造の内部と外部の境，つまりラベルが *Inside*(または *Beginning*) と *Outside*，あるいは *Outside* と *Beginning* の枝が連接する点については表層を用いる. これは “どちらも” や “両方” といった並列構造の周辺に表れやすい語を素性として利用するためである．

品詞の情報は，文節内の主辞の品詞を用いる. 主辞とは，文節内で一番後方に位置する自立語を指す．

文節内に名詞があるときは，主辞としての品詞情報の他にも，文節内にサブカテゴリが “固有名詞”，“地名”，“人名”，“組織名”，“数詞” となる名詞が含まれているかの属性値も素性に用いる. 助詞の情報は，同一の表層でも異なるサブカテゴリの助詞となるものがあるため，サブカテゴリまで含めて一つの助詞として扱う. もし文節内に複数の助詞が存在する場合は，それらをまとめて一つの助詞として扱う. 例えば，“～ にも” という文節においては，“に” と “も” の二つの助詞が文節に存在することになるが，“にも” を一つの助詞として扱う. また内部素性として，枝で対応付けられる文節間の文字の一致の数も利用し，文字の一致数それぞれについてバイナリ値の素性を利用する．

4.2 外部素性

外部知識源として，分類語彙表 [国立 04] を用いた. 分類語彙表では各語に五桁のコードが与えられており，各桁の値はシソーラス木における語の位置 (語が所属する意味的なクラス) を表している. 上位の桁ほど根に近い位置を表し，同じ値を持つ語の類似性も粗くなっていく.

我々はこのシソーラスの情報を二つの方法で素性として利用する. 一つは，斜め方向の枝で対応付けられる二つの文節について，主辞のコードが上位桁から比較していきどの桁まで一致しているか，つまり二つの主辞がシソーラス木のどの深さで同じクラスに属しているかを表す素性である. 分類語彙表は，桁の上位から下位に向かうにつれて “類”，“部門”，“中項目”，“分類項目” と分類が細かくなっていくが，“類” までの一致は実質品詞レベルでの一致であるため，コードが “類” までしか一致していない場合はこの素性は使わない. 黒橋らの並列構造解析モデルでも同様に，シソーラス木において二つの語が一致する位置に依存する素性が用いられている.

もう一つは，主辞のコードそのものである. 主辞の表層形を使うと素性空間がスパースになってしまうので，表層語の代わりに分類語彙表のコードを素性として用いる. コードの一致に依存する素性は，二つの語がどのクラスで一致しているかには依存せず，一致する深さが同じであれば素性は単一の素性として扱われるため，均一な重みが素性に与えられる. しかし語のコードそのものを使うことにより，どのグループの語の組み合わせが並列になりやすいのか，なりにくいのかといった特徴を，各グループごとに独立に重みとして学習することができる.

共起表現に関連する素性も利用する.“A の B と C が” というフレーズがあったときに，“A の B”，“A の C” の両方がコーパスに表れやすいならば，並列構造として ((A の B と)(C が)) という構造よりも (A の ((B と)(C が))) という構造を他に手がかりがなければ選ぶべきである. この特徴を素性として組み込むためにまず共起表現を収集する. 京都コーパス (version4.0) 内に出現する全ての依存関係にある文節ペアを集め，Dunning による尤度比検定の手法 [Dunning 93] を用いて抽出される語のペアを共起表現とする. ただし，語そのものの頻度を数えるのではなく，語に対応する分類語彙表のコードについて頻度を数えた. これは，語そのものについて共起表現を収集するには京都コーパスは規模が小さく，極めて少数の表現しか抽出できなかったためである. 共起表現はバイナリ値の素性として用い，編集グラフで並列構造外部から内部へ，あるいは内部から外部へ変化する点において，対応する文節の主辞のペアが収集された共起表現にあるか否かでその値を決める.

4.3 距離による素性の分解

本節では新しい素性を導入するのではなく，素性を編集グラフで出現する位置によって区別する効果的な方法を提案する. 簡潔に言うと，前節までに導入した素性を，編集グラフで素性が出現する位置と編集グラフの対角線との距離が閾値 θ[*1] を越えるかに否かに基づいて二つに分ける. この距離は，素性が与えられている枝に対応する文節間の距離を表している. これにより，元々は一つの素性だったものが距離によって二つの素性に分解され，それぞれ独立の重みが学習される. ある素性が条件 X が成り立つかどうかを表す素性だったとすると，その素性は次の二つの素性 f_1 と f_2 に分解される.

$$f_1 = \begin{cases} 1 & \text{if } X \text{ and (distance from the diagonal)} \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$$f_2 = \begin{cases} 1 & \text{if } X \text{ and (distance from the diagonal)} > \theta, \\ 0 & \text{otherwise.} \end{cases}$$

これにより，素性は距離の値によって互いに疎になるように分解されるが，距離の値が重なるように分解することも可能である. つまり，距離に依存しない素性と，距離が閾値以下 (あるいは以上) の素性の二つに分解される. しかし，予備実験の結果，今回の実験では互いに疎になるように分解する. 名詞句並列構造では並列ペアとなる文節間の距離が小さくなりやすいため，距離と組み合わせることによりこの特徴を素性として利用することができる.

表 1: 並列句の終端のペア (“P”) の精度 (%)，再現率 (%)，F 値 (%)

Method	P	R	F
KNP	87.2	83.6	85.3
提案手法 (内部素性のみ)	86.8	84.2	85.5
提案手法 (全素性)	**87.9**	**85.5**	**86.7**

5. 実験

評価実験には，京都コーパス (version4.0) を用いた. コーパスは 38,400 文からなり，新聞記事を元に作成されている. 全ての文は文節単位に区切られ，形態素，依存関係の情報が付与されている.

京都コーパスでは，並列構造は文節間の依存関係の 1 つとして扱われ，“P” で表される. 京都コーパスには並列句の先頭位置を表す情報は付与されておらず，並列構造を構成する並列句の終端の文節間に “P” というタグが与えられている.

今回の実験では，名詞句並列構造のみを含む文を使用した. 理由は，名詞句並列構造がコーパス内に出現する並列構造の中で最も頻出する構造だからである. ただしかぎ括弧文を含むものや，ネスト並列構造を含む文は除外した. かぎ括弧内における “P” は通常の “P” とは異なる意味でも用いられることと，ネスト並列構造は我々のモデルでは扱えないことが理由である. 名詞句並列構造は，“P” で関係づけられた二つの文節の主辞がともに名詞である構造を指す.

我々のモデルは学習に並列句の先頭位置と終端位置を必要とするが，先述のとおり京都コーパスには並列構造の先頭位置が付与されていない. このため，二人の作業者が京都コーパス中の名詞並列句の先頭位置を付与し，また同時に明らかに並列構造ではないものやネスト構造になっている文を除外した. その結果 4,154 文が残り，評価実験にはこれらの文を使った.

6. 結果と考察

比較対象として，ルールベースの並列構造解析器を内蔵した Kurohashi-Nagao Parser(KNP version2.0b) を用いた. KNP への入力は JUMAN(version 5.1)[*2] の出力を与えた. JUMAN の出力は助詞や品詞などの形態素情報，あるいは文節区切りの情報が京都コーパスとは一致しないこともあり，この誤った情報が KNP の性能に影響を及ぼす可能性がある. 公平な比較をするために，評価用のセットとして JUMAN の出力が京都コーパスの文節区切り，形態素情報と一致する文，2,067 文を用い，残りの JUMAN の出力が京都コーパスと一致しない

∗1 実験では $\theta = 6$ を用いた.

∗2 KNP: http://nlp.kuee.kyoto-u.ac.jp/nl-resource/
JUMAN: http://nlp.kuee.kyoto-u.ac.jp/nl-resource/

表 2: 各素性を用いたときの精度 (%)，再現率 (%)，F 値 (%)

素性	P	R	F
内部素性	86.9	84.8	85.9
内部素性 + シソーラス (コードの一致)	86.9	84.8	85.9
内部素性 + シソーラス (コードの一致) + 距離	88.0	86.0	87.0
内部素性 + シソーラス (コードの一致，コード) + 距離	88.6	86.4	87.5
内部素性 + シソーラス (コードの一致，コード) + 距離 + 共起	**88.9**	**86.9**	**87.9**
KNP	86.2	82.1	84.1

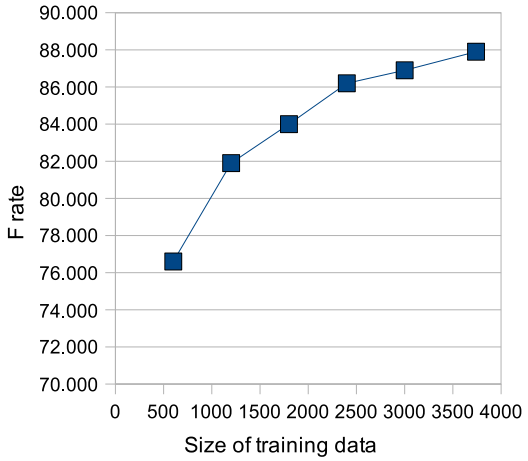


図 3: 訓練セットのサイズに対する F 値の変化

2087(= 4154 − 2067) 文は，訓練用にのみ用いる. 分割交差検定を行うために評価用の 2067 文をランダムに五つのセットに分割する．つまり実験は五回行い，それぞれの回では異なるセットをテストセットとして用いる. 訓練セットとしては，残りの 4 セットに，JUMAN の出力が京都コーパスと一致しない 2087 文を加えたセットを用いる。

表 1 に実験結果をまとめた. 五つのセットについて実験を行った結果を平均した結果である．KNP は並列句の範囲の情報を出力しないため，直接我々のモデルの出力と比較することができない. そこで，評価は並列構造を構成する並列句の終端文節について行った．これは京都コーパスにおいて “P” で関係づけられた文節について評価することと同じである．“提案法 (内部素性のみ)” の行は，内部素性のみを用いて，素性を距離と組み合わせない場合の BIO ラベリングによる新保らのモデルの結果である．“提案法 (全素性)” の行は，内部素性に加えて外部素性を用い，さらに 4.3 節で導入した距離による素性分割を行ったときの結果である. 表からわかるように，外部素性や距離による分割を用いなくても KNP とほぼ同等の結果が出ており，また外部素性や距離と組み合わせることにより KNP よりも高い結果を出せている.

表 2 は各素性の効果を比較したものである. 時間の都合上この比較実験は五つのセットの中の一つのセットについてのみ行った．コードの一致に関する素性はほとんど有効ではないが，各素性を距離と組み合わせることにより性能が改善されている．また分類語彙表のコード自身や，共起素性を用いることでさらに性能が上がっている.

解析誤りの例を見てみると，“～ 情報，知識に関する産業が～” という文において，“情報，知識に” の部分を正しく並列構造と認識できていない例がある. 分類語彙表を見ると，“情報” と “知識” は “相” の階層までしか一致しておらず，これはかなり粗い類似度になるが，直感的にはこの二つの語の類似度は高い. 他の誤り例として “～ 実現か中止かの ～” といった例があるが，“実現” と “中止” は意味的に対立する語ではあるが，対立している語は逆にペアになりやすい. これらの例から，語の類似度を計るのに分類語彙表が必ずしも適しているとは言えない．解析精度を向上させるためには，分類語彙表の意味的なクラスを単純に利用するのではなく，コーパスから何らかの単語の分布に関する統計量を計算し，それに基づいた語の類似度やクラスを別に定義する必要があるだろう.

図 3 は，一つのテストセットについて，横軸に訓練セットの文の数，縦軸に F 値をプロットしたものである. 訓練セットのサイズが大きくなるにつれて F 値の改善幅は小さくなっているとはいえ，訓練セットが増えればさらなる性能の改善が期待できる.

最後に，外部素性や距離との組み合わせといった素性を導入できたのは，我々の手法が機械学習に基づく手法だからであることを強調しておく. 新しい素性の導入により調整する重みの数は膨大となり，人手で重みを最適な値に設定することは困難である. 特に，分類語彙表のコードによって語は 1000 程度の次元に落としこまれるが，それでもそれらのバイグラムの数は 10^6 にもなり，これらの素性の重みを人手で調整するのは不可能であろう.

今後の予定としては，日本語における述語並列構造へのモデルの適用や，3 節の BIO ラベリングの英語ドメインへの適用などを予定している.

参考文献

[Dunning 93] Dunning, T.: Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, Vol. 19, pp. 61–74 (1993)

[Kurohashi 94] Kurohashi, S. and Nagao, M.: A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures, *Computational Linguistics*, Vol. 20, pp. 507–534 (1994)

[Resnik 99] Resnik, P.: Semantic similarity in a taxonomy, *Journal of Artificial Intelligence Research*, Vol. 11, pp. 95–130 (1999)

[Shimbo 07] Shimbo, M. and Hara, K.: A discriminative learning model for coordinate conjunctions, in *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 610–619 (2007)

[国立 04] 国立国語研究所：分類語彙表, 大日本図書 (2004)