大規模ラベルなしデータを利用した係り受け解析の性能検証

鈴木 潤　　磯崎 秀樹
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
{jun,isozaki}@cslab.kecl.ntt.co.jp

概要

係り受け解析では，正解係り受けが付与されたデータを用いた教師あり学習 (近年では主に識別学習) により解析器の学習を行うのが現在最も一般的な方法である．本稿では，正解が不明なデータも利用して係り受け解析精度を更に向上させる方法について議論する．まず，教師あり学習した解析器と比較して，係り受け解析精度を顕著に向上させたと報告のあった 3 種類の方法を取り上げ，可能な拡張を提案する．次に，それらの手法が単一の学習の枠組みで表現できることを示す．実験では，これら 3 つの従来法の単純な組み合わせがさらに解析精度を向上させることが可能であることを示す．本稿では最終的に，取り上げた 3 つの従来法は学習の枠組みは同じであるため統合が容易なこと，にも関わらず同一の情報源からそれぞれ違った質の情報を獲得し，互いの情報を補完する関係にあり更なる精度向上が可能であるという知見が得られたことを報告する．

1　はじめに

係り受け解析は，文内の語や節間の文法的な依存関係を解析する問題である．近年では，2006, 2007 年の CoNLL shared task[1, 2] で取り上げられる等，国際的にも自然言語解析技術の重要な基盤技術の一つとして研究が続けられている．

係り受け解析では，これまで，正解係り受けが付与されたデータから教師あり学習 (近年では主に識別学習) により解析器の学習を行うのが最も一般的な方法である．これは，正解データが十分な量確保できれば，教師あり学習により高い解析精度が得られることが数多くの文献で報告されているためである [3, 4, 5, 6, 7]．ただし，解析精度向上の方法として，正解データを無限に増やし続けるのはコストや時間的に非現実的なので，次のステップとして，正解データ量を固定した上で更なる解析精度向上が可能な方法が模索されている．

係り受け解析では，一般的に入力は文である．文は，web データや電子化文書から比較的容易に，かつ，大量に獲得することが可能であり，これらの多量に獲得可能なデータを利用して係り受け解析精度を向上させるという取り組みは，現実的な選択肢の一つと言える．事実，近年，大量の正解係り受けが不明なデータを利用した方法が多く提案され，いくつかの良好な結果が報告されている [8, 9, 10, 11, 12, 13, 14]．

そこで本稿では，係り受け解析タスクを対象に正解が不明なデータを利用して顕著に解析精度を向上させた 3 種類の方法 [11, 12, 13] に注目する．まず始めに，それぞれの方法の概略を述べ，個別に可能な拡張を提案する．次に，これら 3 手法は，“正解が不明なデータから識別学習用の新たな特徴表現を獲得する方法” として一般化可能であることを示す．これは，同じ学習の枠組みを用いているとみなすことができるため容易に統合が可能であることを意味する．最後に，実験において，これら 3 つの従来法の単純な組み合わせによって，さらに解析精度が向上することを示す．これらの結果から，本稿で取り上げた 3 つの従来法は学習の枠組みは同じであるにも関わらず，同一の情報源からでもそれぞれ違った質の情報を獲得していると考えられ，互いの情報を補完する関係にあることを示す．

2　正解係り受けが不明なデータの利用

係り受け解析タスクで，正解が付与されていないデータ (ラベルなしデータ) を利用することで，従来の識別学習による教師あり学習と比べて解析精度が顕著に向上した方法として以下の 3 種類が挙げられる．

2.1　クラスタリングによる方法 (手法 1)

入力を類似度などの観点でクラスタリングした後に得られるクラスを新しい特徴として識別学習で利用する方法が提案されている．特に，係り受け解析タスクにおいては，Koo ら [11] が，良好な解析精度の向上が得られることを示している．Koo ら [11] では，まず始めに，Brown アルゴリズム [15] を用いて，2 分木による単語の階層クラスタリングを行う．次に，その階層の上位 n ビットをクラス識別子として新たな特徴を構成する要

素とする．最後に，単語や品詞といった情報と組み合わせて，新しい特徴を導出する．最後の処理は，クラスタリング結果のクラスをそのまま特徴として利用しても効果が低いため，組み合わせて情報を混ぜ合わせることによって高い効果を得る狙いがある．また，上位 n ビットの n は，複数個を利用した方が効果的であるという知見も得られており，文献 [11] では，$n = 4, 6, 8, 10$ の 4 つが利用されている．

本稿では，基本的に文献 [11] で用いられたクラスタリング情報から導出された特徴を，そのまま用いることとする．

2.2 予測結果から特徴を自己学習的に生成する方法 (手法 2)

Chan ら [12] は，正解が不明なデータを，従来の識別学習により学習した解析器を利用して一度解析し，そこで得られた予測結果から新たな特徴を生成し利用する方法を提案している．考え方自体は古典的な自己学習 (self-training) と共通するが，Self-training のように予測結果を正解とみなして利用するわけではなく，新たな特徴を生成するために利用する点が大きな違いである．特徴として利用することにより，予測結果が曖昧なものやノイズとなりそうなものは，識別学習の観点で排除できる可能性が期待できる．

具体的には，予測結果の係り受け解析木から，係り受け単語 bigram，または，係り受け単語 trigram といった部分木を列挙し，それらを頻度順に並べたものに対し，ランキングの “上位 10%に含まれる”, “10〜20%に含まれる”, “それ以外 (20%〜100%) に含まれる” の 3 種類のラベルと，“一度も出現しなかった” の計 4 種類のラベルを識別学習時の特徴として利用する．

文献 [12] では，予測結果から単語に関する部分木のみを利用していた．しかし，実際には，単語のみではなく品詞等別の情報を用いて部分木を列挙してもよい．そこで本稿では，係り受け解析で用いられる特徴抽出テンプレートを利用して部分木の列挙を行うように拡張する．ただし，上記の順位付けと属するクラスの決定はそれぞれの特徴抽出テンプレート毎に独立でおこうこととする．つまり，本稿で用いた方法では，新しい特徴は特徴テンプレート数× 4 個追加される事になる．これにより，単語と品詞の組み合わせなど，より多様な部分木特徴も導入することができる．また，文献 [12] にあるような，係り受け対象の前後一語を利用して部分木を得るような変則的な処理も特徴抽出テンプレートとして記述することで無理なく統一的に扱うことが可能である．つまり，提案法はより汎用的な枠組みになっていると言える．

2.3 補助モデルによる方法 (手法 3)

Suzuki ら [13] は，従来の識別学習法に生成モデルによる補助モデルを導入した半教師あり学習法を提案している．具体的には，導入された生成モデルそのものが，新しい特徴として識別モデルで利用される．また生成モデルは，識別モデルとの不整合が最も小さくなるようなモデルを選択するように，正解が不明なデータから学習される．

文献 [13] では，係り受け 2 次モデル (second-order model) で用いられている特徴に関する生成モデルは利用していなかった．これは，係り受け 2 次モデルでは周辺確率の計算量が比較的大きくなるため，計算量の観点から扱われなかったためである．しかし本稿では，これら係り受け 2 次モデルに関する特徴も利用することとする．具体的な解決策として，周辺確率を直接計算せずに，ビタビ予測の結果を周辺確率の近似値として利用することで計算量を削減する．

2.4 その他

正解が不明なラベルを利用した方法は，上記 3 種以外にも様々な方法が提案されている．それらの方法は，正解が付与された訓練データが非常に小規模な場合を想定していたり，ドメイン適用を主目的としているなど，本稿で考慮したい主旨と異なるため今回は取り上げない．また，同様に，係り受け解析以外の問題で顕著な性能向上が示されている方法も存在するが，それらも，係り受け解析に適用する方法が自明ではなかったり，係り受け解析に適用しても明確な精度向上が得られなかったため，今回の考慮対象には含めないこととする．

3 ラベルなしデータからの特徴表現の獲得

ここで，$\boldsymbol{x}$ を入力，$\boldsymbol{y} \in \mathcal{Y}$ を出力とする．係り受け解析では，$\boldsymbol{x}$ は文，$\boldsymbol{y}$ は係り受け構造である．また，$\mathcal{Y}(\boldsymbol{x}) \subseteq \mathcal{Y}$ を入力 $\boldsymbol{x}$ が与えられた時の解候補の集合とする．次に，$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$ を $\boldsymbol{x}$ と $\boldsymbol{y}$ から抽出される特徴ベクトル，$\boldsymbol{w}$ を $\boldsymbol{f}$ に対応するパラメタベクトルとする．簡単のため，識別学習後の最尤候補 $\hat{\boldsymbol{y}}$ を推定する識別関数が以下の式で与えられるとする．

$$\hat{\boldsymbol{y}} = \arg \max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{w} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

これは，文献 [4, 5, 6] 等で用いらている大域的最適解を求めるアルゴリズムを用いた際の識別関数と一致する．

このとき，上記のような識別学習に 3 つの手法のいずれかを適用して学習した後の識別関数は，一般化すると以下のように書き表すことができる．

$$\hat{\boldsymbol{y}} = \arg \max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{w} \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{v} \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}), \tag{2}$$

このとき，$\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})$ を正解が不明なデータから新たに導出した特徴の集合 (ベクトル表現) を表すとし，$\boldsymbol{v}$ を $\boldsymbol{w}$ と同様に新たに導入された特徴 $\boldsymbol{u}$ に対応するパラメタベクトルであるとする．つまり手法 1 では，クラスタリング結果のクラスを組

データセット	文数	単語数
学習 (Sec. 02–21)	39,832	950,028
開発 (Sec. 22)	1,700	40,117
評価 (Sec. 23)	2,012	47,377
ラベルなし	1,796,379	43,380,315

表 1: 実験データ (学習・開発・評価セット:Penn Treebank III, ラベルなしデータ: BLLIP コーパス)

み合わせて新たに生成した特徴の集合，手法 2 では，頻度により順序付けした部分木が属する量子化したクラスの集合，手法 3 では学習後の生成モデルの出力が $\boldsymbol{u}$ となる．また，$\boldsymbol{v}$ の値は，$\boldsymbol{w}$ と同様に識別学習により決定される．細かい差としては，手法 1,2 は追加される特徴は，自然言語処理でよく用いられるバイナリの特徴であるのに対して，手法 3 で追加される特徴は実数値であるという点，或は，手法 1 は $\boldsymbol{y}$ の情報は実質的に利用していないが，手法 2,3 は $\boldsymbol{y}$ の情報を利用して特徴を抽出しているという点が挙げられる．とはいえ，本稿で取り上げた 3 手法は，いずれも “正解が不明なデータから識別学習用の新たな特徴表現を獲得する方法” と捉えることができる．

このように上記の 3 手法は，追加の特徴 $\boldsymbol{u}$ の導出法が違うだけで，学習法としては同じ枠組みであると言える．故に，学習の枠組みとしてこれらの手法を組み合わせるのは，非常に容易である．つまり単純に 3 種類の $\boldsymbol{u}$ を追加すればよい．

4 実験

本稿の実験では，英語係り受け解析の標準的な評価テストセットと設定を用いて実験を行う．具体的には，文献 [11, 12, 13] で用いられている実験データと設定をそのまま用いる．表 1 に，学習，開発，評価データ，および，正解が不明なデータのデータ量を示す．

実験では，教師あり識別学習法として，文献 [13] で用いられている拡張 MIRA を用いた．繰り返し数は最大 20 とし，開発セットで最も良い解析精度が得られた学習結果を用いて，評価セットの評価を行った．

本実験では，前節で取り上げた 3 つの従来法の単純な組み合わせにより解析精度にどのように変化するか検証することを目的とする．

4.1 実験結果および考察

表 2 は，第 2 節で取り上げた各手法の結果と，それらを組み合わせた場合の結果を示している．前述の通り，開発セットの結果は，最も良い結果が得られた際の結果を示しており，評価データの結果は，開発データでチューニングして得られたハイパーパラメタの値を用いた際の結果となっている．

注意点としては，この結果は，各手法が提案された各文献で報告されている解析精度とは大幅に違う結果となっている点である．具体的には，元文献よりも解析精度が高い結果が得られている．これは，識別学習の学習アルゴリズムがより高精度な結果が得られるものを使用している点，第 2 節で述べたようにそれぞれの方法を改良している点等が作用しているためである．

実際に各文献では，細かい実験設定が完全に一致しているわけではないので，厳密な意味で性能比較するのは難しい状況であった．表 2 に示されている結果は，正解が不明なデータから獲得した特徴以外の実験設定は全く同じ設定で行った結果であるため，一つの公平な評価結果であるといえる．ただし，各手法固有の事前知識的な設定やパラメタの設定などは，各手法を紹介した各文献に基づいて設定を行っている．ゆえに，各手法固有のハイパーパラメタのチューニングや固有の性質を利用した改良を施すことによって，各手法の解析精度は大幅に変化する可能性があり，ここに示した結果の優劣が，絶対的な手法の優劣を決定するものではない．ここでの目的は各手法の優劣を決めることではない．ここで注目してもらいたいのは，各手法とも同じように解析精度を大幅に向上できることを示しているにも関わらず，それぞれの手法を組み合わせることによって，更に解析精度を向上させることができることが示されている点である．一般的に，各手法は，それぞれの方法で情報を獲得しているが，同一の情報源から情報を獲得しているという点で，統合することでさらに性能が向上することは自明な事柄ではない．この結果は，同一の情報源から各手法がそれぞれ違った質の情報を獲得していることを示唆していると考えられる．

このように，これらの手法は，同じ学習の枠組みとみなせるため容易かつ自然に各手法の統合が可能であり，識別学習した解析器をシステム統合するような煩雑な処理やチューニングなどは不要であることや，精度向上に関して，いずれかの方法を排他的に選択しなくてはいけない関係ではないという非常に理想的な関係であることが示された．つまり実際に高精度な係り受け解析器を構築したい場合には，特に気にせずに全ての方法を同時に利用すればよい．

4.2 既存のトップシステムとの性能比較

表 3 に，過去文献で示された係り受け解析結果と本稿で得られた最も良い結果を比較を示す．ここから，3 種類の手法を統合することで，これまでに報告されている最も良い結果を更に上回る結果が得られたことがわかる．つまり本稿で得られた結果は，用いた英語係り受け解析の標準データセットにおいて現在までに報告された結果の中で最も良い結果であった．

5 まとめ

本稿では，係り受け解析タスクにおいて，正解が付与されたデータから教師あり学習 (識別学習)

評価対象	開発セット		評価セット	
評価尺度	係り受け正解率	文完全正解率	係り受け正解率	文完全正解率
拡張 MIRA (教師あり学習)	93.02 -	48.00 -	92.76 -	46.90 -
+ 手法 1 (クラスタリング)	93.69 (+0.67)	51.35 (+3.35)	93.34 (+0.58)	48.72 (+1.82)
+ 手法 2 (予測結果からの特徴獲得)	93.46 (+0.44)	49.71 (+1.71)	93.10 (+0.34)	47.52 (+0.62)
+ 手法 3 (補助モデル)	93.50 (+0.48)	49.59 (+1.59)	93.43 (+0.67)	48.59 (+1.69)
+ 手法 1,2	94.23 (+1.21)	53.00 (+5.00)	93.82 (+1.06)	50.70 (+3.80)
+ 手法 1,3	94.29 (+1.27)	53.47 (+5.47)	93.91 (+1.15)	51.08 (+4.18)
+ 手法 2,3	93.76 (+0.74)	50.24 (+2.24)	93.52 (+0.76)	49.01 (+2.11)
+ 手法 1,2,3	94.59 (+1.57)	55.06 (+7.06)	94.09 (+1.33)	51.86 (+4.96)

表 2: 各手法とそれらの組み合わせによる解析精度比較

係り受け解析器	D. 正解率	S. 正解率
MIRA(1 次係り受けモデル)[4]	90.9	37.5
MIRA(2 次係り受けモデル)[6]	91.5	42.1
VP (+ 手法 1 相当)[11]	93.16	–
システム統合 (+ 手法 1,2 相当)[12]	93.16	47.15
拡張 MIRA(+ 手法 1, 3 相当)[13]	93.79	–
本稿での最高解析精度	**94.09**	**51.86**

表 3: これまでのトップシステムとの解析精度比較 (D. 正解率: 係り受け正解率, S. 正解率:文完全正解率)

した解析器の解析精度を基準とし，正解が不明なデータを利用して更に解析精度を向上させる方法について検討を行った．具体的には，これまでに係り受け解析精度を顕著に向上させた報告がある 3 種類の方法を取り上げ，これら従来法の単純な組み合わせによってさらに解析精度が向上することを示した．これは，用いた英語係り受け解析の標準データセットで，これまで報告されている中で最も良い解析精度であった．本稿では，最終的に，本稿で取り上げた 3 つの従来法は，学習の枠組みは同じであるため統合が容易なこと，にも関わらず同一の情報源からそれぞれ違った質の情報を獲得し，互いの情報を補完する関係にあるという知見が得られた．

参考文献

[1] S. Buchholz and E. Marsi. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proc. of CoNLL-X*, pages 149–164, 2006.

[2] J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. The CoNLL 2007 Shared Task on Dependency Parsing. In *Proc. of EMNLP-CoNLL*, pages 915–932, 2007.

[3] H. Yamada and Y. Matsumoto. Statistical Dependency Analysis with Support Vector Machines. In *Proc. of IWPT*, 2003.

[4] R. McDonald, K. Crammer, and F. Pereira. Online Large-margin Training of Dependency Parsers. In *Proc. of ACL*, pages 91–98, 2005.

[5] R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. Non-projective Dependency Parsing using Spanning Tree Algorithms. In *Proc. of HLT-EMNLP*, pages 523–530, 2005.

[6] R. McDonald and F. Pereira. Online Learning of Approximate Dependency Parsing Algorithms. In *Proc. of EACL*, pages 81–88, 2006.

[7] X. Carreras. Experiments with a Higher-Order Projective Dependency Parser. In *Proc. of EMNLP-CoNLL*, pages 957–961, 2007.

[8] J. Blitzer, R. McDonald, and F. Pereira. Domain Adaptation with Structural Correspondence Learning. In *Proc. of EMNLP-2006*, pages 120–128, 2006.

[9] D. A. Smith and J. Eisner. Bootstrapping Feature-Rich Dependency Parsers with Entropic Priors. In *Proc. of EMNLP-CoNLL*, pages 667–677, 2007.

[10] Q. I. Wang, D. Schuurmans, and D. Lin. Semi-supervised Convex Training for Dependency Parsing. In *Proc. of ACL-08: HLT*, pages 532–540, 2008.

[11] T. Koo, X. Carreras, and M. Collins. Simple Semi-supervised Dependency Parsing. In *Proc. of ACL-08: HLT*, pages 595–603, 2008.

[12] W. Chen, J. Kazama, K. Uchimoto, and K. Torisawa. Improving dependency parsing with subtrees from auto-parsed data. In *Proc. of EMNLP*, pages 570–579, 2009.

[13] J. Suzuki, H. Isozaki, X. Carreras, and M. Collins. An empirical study of semi-supervised structured conditional models for dependency parsing. In *Proc. of EMNLP*, pages 551–560, 2009.

[14] G. Druck, G. Mann, and A. McCallum. Semi-supervised learning of dependency parsers using generalized expectation criteria. In *Proc. of ACL/IJCNLP*, pages 360–368, 2009.

[15] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai. Class-based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479, 1992.