自然言語処理 141－1
（2001．1．25）

模擬対話音声コーパスの書き起こし文における帰納的学習を用いた構文解析手法の性能評価

正富 欣之　荒木 健治　栃内 香次

北海道大学大学院工学研究科
〒060-8628 札幌市北区北13条西8丁目
TEL: 011-706-7389　FAX: 011-709-6277
E-mail: {tome,araki,tochinai}@media.eng.hokudai.ac.jp

あらまし　自然言語処理における構文解析手法に，固定された文法規則を用いるものと用例から得られたデータを用いるものがある．これらの手法における問題点として，前者は対象に依存した言語現象への対処が困難である，後者はデータ量が膨大になる，ということが挙げられる．そこで，実例から帰納的学習を用いて構文解析規則を獲得し，得られた解析規則を適用した構文解析を行うことにより，上記の問題を解消することを目指した手法を提案する．本手法の利点としては，学習アルゴリズムは言語に依存していないので，学習データさえあれば同一のアルゴリズム（帰納的学習）でどのような言語にも適用できるということがある．すなわち，多言語対応構文解析手法としての能力を有するものと考えられる．本稿では，上述の手法に基づいたシステムについての概要を述べ，模擬対話音声コーパスの書き起こし文を用いた実験を行い，本手法の有効性について検討する．

キーワード　構文解析，帰納的学習，多言語

Evaluation of a Syntactic Analysis Method Using Inductive Learning for Transcripts of Simulated Spoken Dialogue Corpus

Yoshiyuki Masatomi, Kenji Araki, Koji Tochinai

Graduate School of Engineering, Hokkaido University
N13 W8, Kita-ku, Sapporo, 060-8628 Japan
TEL: +81-011-706-7389　FAX: +81-011-709-6277
E-mail: {tome,araki,tochinai}@media.eng.hokudai.ac.jp

Abstract　In this paper, we propose a syntactic analysis method using inductive learning from examples. In our proposed method, the system acquires the parsing rules using the examples of parsing result. And the system parses Japanese sentences using the acquired parsing rules. We consider that our proposed method can resolve problems of the Rule-based approach and the Example-based approach. Moreover, we consider that this method can be applied to the other languages. We performed the experiment using transcripts of simulated spoken dialogue corpus. We consider the result of experiment about the ability of our proposed method.

Keywords　parser, inductive learning, multi-language

1　はじめに

自然言語処理において，構文解析は文を解析する上で非常に重要である．それゆえ，これまで構文解析に関する多くの研究がなされてきた．それらの研究で，主となる構文解析手法は構文解析規則を用いたものである[1][2]．これらの構文解析規則に基づく手法では，与えられた構文解析規則が固定したものであるため，さまざまな言語現象に対処するのが困難である．また，もし新たな規則を追加しようとすれば，他の規則との整合性をとるために非常に労力のかかる作業となる．

さまざまな言語現象に対処するために機械翻訳の分野では用例に基づく手法が試みられている[3][4][5]．同様に，構文解析の分野でも用例に基づく手法を用いた構文解析システムについて提案されている[6]．一般に，用例に基づく手法では，自動的に対象に適応できるので，構文解析の精度が上昇すると考えられる．しかしながら，このような手法に基づくシステムの性能，つまりは構文解析の質を向上させるためには膨大な構文解析例を必要とする．

また，木構造を用いた構文解析手法も提案されている[7][8][9][10]．これらの手法では統計情報を用いて構文解析を行う．しかしながら，統計情報を得るために用いるコーパスをどのような基準で選択し，どれだけの量を使用するかという問題が残る．

これらの問題を解決するために，比較的少量の構文解析例を用いて精度の高い構文解析を行える手法を提案する．構文解析例からより多くの構文解析規則を自動的に抽出することができれば，少量の構文解析例でもさまざまな言語現象に適応できる精度の高い構文解析が可能となる．そこで，本稿では，構文解析例（品詞付単語列と構文解析結果の組）からの帰納的学習を用いた構文解析手法を提案する．本手法では，入力を品詞付単語列(一般的な形態素解析結果)とし，最初に文節区切処理を行う．文節区切処理では文節区切規則を用いて，品詞付単語列を文節に分ける．文節区切規則は構文解析結果から獲得する．次に，構文解析処理を構文解析規則を用いて行う．構文解析規則は帰納的学習を用いて獲得され，さらに獲得された構文解析規則から再帰的に構文解析規則を獲得する．このように，自動的に構文解析規則を獲得することによりさまざまな抽象度の構文解析規則を獲得でき，それらを用いて最適な抽象度，つまりは具体的な規則から順に構文解析規則を適用することにより解析精度を向上させる．

本手法の利点としては，規則を人手により予め与えていないので，会話文などの非文法的な文や省略の多い文にも柔軟に対処できることが考えられる．また，一般的には構文解析を行うのに困難である対象に依存した文の解析も，システムが動的に適応することにより解決できることが多い．これは，このような限定された対象にのみ有効な構文解析規則を学習により自動的に生成することができるからである．さらには，構文解析例から学習できるので，適応能力が高く，さまざまな文章に対応できる．特に，本手法の学習アルゴリズムは言語に依存していないので，学習データさえあれば同一のアルゴリズム（帰納的学習）でどのような言語にも適用できる．すなわち，多言語対応構文解析手法としての能力を有するものと考えられる．

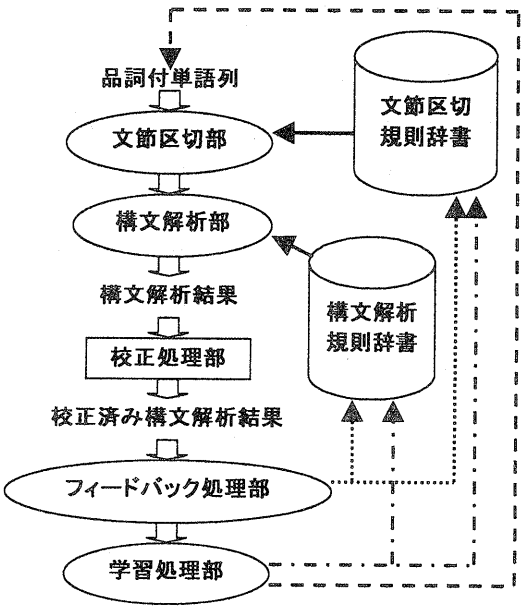


図1 処理過程

2　処理過程

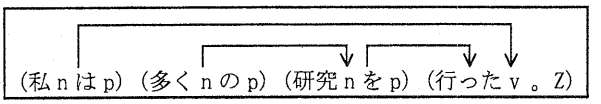
図1に本手法の概要を示す．本手法に基づく実験システムは日本語文の構文解析を対象として

表 1: 品詞付単語列および文節区切例

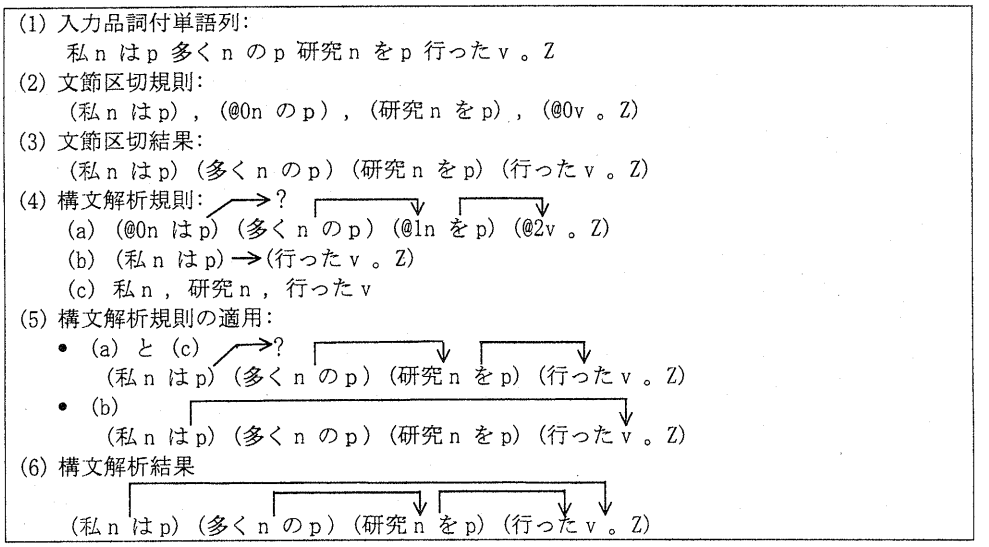
文	私は多くの研究を行った。							
品詞付単語列	私 n	はp	多く n	のp	研究 n	をp	行った v	。Z
文節区切	私 n はp		多く n のp		研究 n をp		行った v 。Z	

n:名詞, p:助詞, v:動詞, Z:特殊

表 2: 構文解析結果例



( ):文節, n:名詞, p:助詞, v:動詞, Z:特殊, →:係り受け関係



( ):, n:名詞, p:助詞, v:動詞, Z:特殊, →:係り受け関係

図 2: 構文解析処理例

作成されている. 各処理の詳細についてはすでに文献[11]で述べているので, 本稿では, 本手法の概要と新しく追加した処理である文節区切部について詳細に記述する.

2.1 概要

最初に, 日本語文の品詞付単語列を入力する. この品詞付単語列は帰納的学習による単語の認識[12]によって得られる単語列に品詞を付属させたものである. また, 品詞付単語列は一般的な形態素解析結果といえる.

次に, 後述する学習処理部で獲得された文節区切規則を用いて文節区切処理を行う. 表 1 に入力の品詞付単語列と文節区切の例を示す.

さらに, 構文解析部では学習処理部で獲得された構文解析規則を用いて構文解析を行う. 表 2 に本稿における構文解析結果例を示す. ここで, 本稿で用いた構文解析結果表現の各記号の意味を説明する. “( )” で囲まれた部分が文節を表す. 単語の後ろに付属しているアルファベットは, その単語の品詞を表している. さらに, 矢印はそれぞれ係り元の文節から係り先の文節に到達し, 係り受け関係を表す. 構文解析結果が出力されるまでの例を図 2 に示す. 図 2 で, 矢印の先にある “?” は係り先が未定であることを示している.

さらに, 校正処理部では構文解析結果に誤りが含まれている場合に人手により校正が施され, 校正済み構文解析結果が得られる.

フィードバック処理部では, 文節区切部および構文解析部より出力された文節区切および構文

解析結果に誤りが含まれている場合に，構文解析結果と校正済み構文解析結果の差分より誤りの原因となった文節区切規則および構文解析規則を推定することによりフィードバック処理を行う．

学習処理部では，得られた校正済み構文解析結果から文節区切規則および構文解析規則を獲得する．さらに，獲得された規則より帰納的学習を行い，新たな抽象化された文節区切規則および構文解析規則を抽出する．この操作は獲得された規則に対して新たな規則が抽出されなくなるまで繰り返し行う．

以上のような処理を繰り返し実行することにより，文節区切規則辞書および構文解析規則辞書が充実し，これらの辞書を用いて行う構文解析の精度が上昇する．

2.2 文節区切部

文節区切部では，システムが文節区切辞書を用いて入力された品詞付単語列を文節に区切る．文節区切規則は，具体的なもの（変数が少ない）から抽象度の高いもの（変数が多い）の順で最長一致する規則が適用される．複数の規則が適用可能な場合は（A）式の尤度評価関数が最大になる規則を選択する．最大関数値が等しい場合は最も新しく獲得された規則を適用する．図2の（1）～（3）に文節区切部の処理例を示す．（2）の文節区切規則は学習処理部で獲得されたものである．学習処理部では校正済みの構文解析結果から正しい文節区切を得て，それを文節区切規則として獲得する．さらに，文節区切規則の組から帰納的学習を用いることによって，新たに抽象化された規則を抽出する．なお，「@」は変数であり，直後の数字は規則中の変数番号を表す．

$$\mathbf{CEF} = \mathbf{AF} + \alpha \times \mathbf{CF} - \beta \times \mathbf{EF} \cdots (\mathbf{A})$$

AF：出現頻度，**CF**：正適用度数，**EF**：誤適用度数

α, β：係数

3 実験

本手法の性能を評価する実験を行った．実験では模擬対話音声コーパス[13]の書き起こし文1,800文を用いた．コーパスの内容はスケジューリングを対象としたものとし，電気通信大学で作成されたスケジューリング会話とスケジュール管理，大阪大学で作成されたスケジュール調整タスク，および京都大学で作成された秘書システムである．このコーパスを用いたのは，本手法の会話文などの非文法的な文や省略の多い文への対処能力を調べるためである．実験結果を評価する際には，係り受け関係にのみ着目した．

3.1 実験手順

初期状態では文節区切規則辞書および構文解析規則辞書は空である．以下の手順により実験を行った．

① 模擬対話音声コーパスの書き起こし文を品詞付単語列に変換し，システムに入力．
② 入力された品詞付単語列について文節区切処理，構文解析処理を行い，構文解析結果を得る．
③ 構文解析結果に誤りが含まれている場合は校正処理を行い，校正済み構文解析結果を得る．
④ 校正済み構文解析結果から文節区切規則と構文解析規則を獲得し，さらに帰納的学習を用いて規則を抽出する．
⑤ ④を新しい規則が抽出されなくなるまで再帰的に繰り返す．
⑥ 以上の操作を1文入力毎に繰り返し行う．

ここで，手順④で帰納的学習を行い，新たな文規則を抽出する対象となる文規則は直前に入力された100文である．

3.2 実験結果

図4に実験結果を示す．図4は係り受け関係200個ごとの正解析率と誤解析率の変化を表している．誤解析率は文節区切が正しく，構文解析部で係り受け関係が誤っていたものである．

また，図5に文節区切の結果を示す．図5は文節数200個ごとの正解率の変化を表したものである．

4 考察

図4の実験結果のグラフから本手法に基づく構文解析結果の精度が徐々に向上しているのが認められる．初期状態では文節区切辞書が空の状態であるので，入力文数が少ない段階では入力品詞付単語列に対して適用できる規則が辞書中に無いことが多い．そのため，文節区切部で文節を

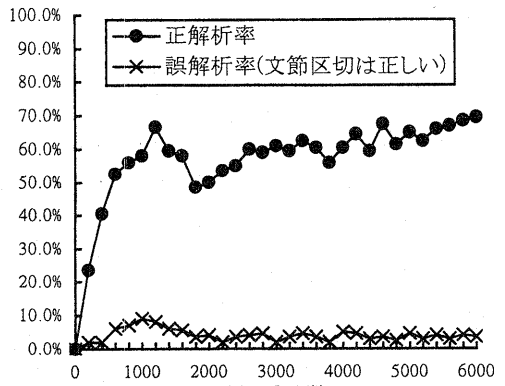


図4:係り受け関係正・誤解析率の変化

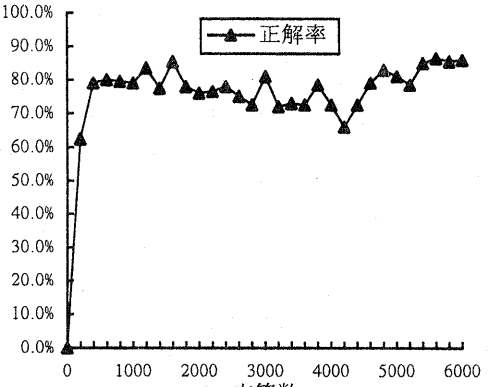


図5:文節区切正解率の変化

(1) 入力品詞付単語列:
修士n 課程n のp 人n はp , Z 火曜日n のp 午前n 中s はp 講義n がp ありv ますs 。Z
(2) 文節区切結果:
(修士n 課程n のp) (人n はp , Z) (火曜日n のp) (午前n 中s はp)
(講義n がp) (ありv ますs 。Z)
(3) 構文解析結果
(修士n 課程n のp) ⟶ (人n はp , Z) ⟶
(火曜日n のp) ⟶ (午前n 中s はp) ⟶
(講義n がp) ⟶ (ありv ますs 。Z)

( ):, n:名詞, p:助詞, v:動詞, Z:特殊, →:係り受け関係

図 6: 正解析例

(1) 入力品詞付単語列:
5n 時s 半s にp 私n のp 部屋n とp いうv ことn でp いかがでしょうj かp 。Z
(2) 文節区切結果:
(5n 時s 半s にp) (私n のp) (部屋n とp) (いうv)
(ことn でp) (いかがでしょうj かp 。Z)
(3) 構文解析結果
(5n 時s 半s にp) ⟶
(私n のp) ⟶ (部屋n とp) ⟶ (いうv) ⟶ (ことn でp) ⟶
(いかがでしょうj かp 。Z)

( ):, n:名詞, p:助詞, v:動詞, Z:特殊, →:係り受け関係

図 7: 誤解析例

正しく決定することができないので文節が確定せず,構文解析を行うことができない未解析のものが多く見られた.入力文が増加するにつれ,文節が正しく区切られるようになり未解析のものが減少した.ただし,構文解析規則が不充分である状態では,正しく文節が区切られた入力に対して,構文解析部で係り受け関係を誤る誤解析率が増加した.しかし,入力文が増加し,文節区切辞書および構文解析辞書の精度が上がると,正解析が増加し,誤解析率は減少した.また,係り受け数が1,800および3,800近辺で正解析率の一時的減少が見られるが,これは入力文の種類が変化したためである.その後は,多少の変動があるものの精度は上昇した.したがって,本手法には係り受け関係について徐々に精度を向上させる能力があり,その有効性を見出すことができた.

次に,実験結果の例を挙げる.図6に本実験での正解析例を示した.入力品詞付単語列に対して獲得された文節区切辞書を適用し,正しい文節区切結果が得られた.係り受け関係についても構文

解析規則を適用することで，各文節の係り受け関係が正しく決定され，正しい構文解析結果が得られた．図7には誤解析例を挙げた．入力品詞付単語列に対して，文節区切規則を用いて行った文節区切は正しい結果を導いた．しかし，構文解析処理を行ったところ，“(５ｎ 時ｓ 半ｓ にｐ)” という文節が “(いかかでしょうｊ かｐ ． Z)” という文節に係ったが，正しくは文節 “(いう ｖ)” に係るべきである．この誤りの原因は，抽象化された規則の中に，誤解析例のような係り受け関係が成り立つ規則を獲得していて，その規則を適用したためである．このように，抽象化された規則を適用して誤る，過学習が起きていると考えられる．これは，少量のコーパスからできる限り多くの構文解析規則を抽出し，可能な限り入力に対して規則を適用したために起こったものである．しかしながら，この誤りについてはフィードバック処理を繰り返し，構文解析規則辞書が充実することによって，次第に減少していくと考えられる．また，規則の抽出数および適用率が減少することが考えられるが，帰納的学習により規則を抽出する時の条件を厳しくすれば過学習を減少させることができるであろう．

5 おわりに

人手により与えられた構文解析規則を用いた構文解析手法と用例に基づく構文解析手法の問題点を解消するために，本稿では，実例からの帰納的学習を用いた構文解析手法を提案した．さらに，提案した構文解析手法に基づくシステムを作成し，本手法の性能を評価する実験を模擬対話音声コーパスの書き起こし文を用いて行った．実験結果より，係り受け関係の正解析率が徐々に上昇することを確認した．また，本手法が会話文などの非文法的な文や省略の多い文への対処能力を有することが確認された．したがって，本手法には，構文解析システムの精度を徐々に向上させる能力があり，その有効性を見出した．

今後は，さらに精度の向上を図るために，規則適用アルゴリズムの調整を行う．また，本手法が多言語対応の能力があることを実証する予定である．

参考文献

[1] E. Brill and P. Resnik: “A rule-based approach to prepositional phrase attachment disambiguation”, Proceedings of the 15th COLING, pp. 1198-1204, 1994.

[2] U. Germann: “A deterministic dependency parser for japanese”, Asia-Pacific Association for Machine Translation, pp. 547-555, 1999.

[3] V. Sadler and Vendelmans: “Pilot implementation of a bilingual knowledge bank, Proceedings of the 13th COLING, pp. 449-451, 1990.

[4] C. Stanfill and D. Waltz: “Toward memory-based reasoning”, Communications of the ACM, Vol. 29, No. 12, pp. 1213-1228, 1986.

[5] S. Sato and M. Nagao: “Toward memory-based translation”, Proceedings of the 13th COLING, pp. 247-252, 1990.

[6] M. H. Al-Adhaileh and T. E. Kong: “A flexible example-based parser based on the SSTC”, Proceeding of the ACL, pp. 687-693, 1998.

[7] 春野雅彦，白井諭，大山芳史：“決定木を用いた日本語係受け解析”，情報処理学会論文誌，Vol. 39, No. 12, pp. 3177-3186, 1998.

[8] R. Bod: “A Computational Model of Language Performance Data Oriented Parsing”, Proceedings of the 14th COLING, pp. 855-859, 1992.

[9] 森信介，長尾真：“係り受けを用いた確率的言語モデル”，情処学NL研報，Vol. 122, No. 6, 1997.

[10] 白井清昭，徳永建伸，田中穂積：“括弧付きコーパスからの日本語確率文脈自由文法の自動抽出”，自然言語処理，Vol. 4, No. 1, pp. 125-146, 1997.

[11] 正富欣之，荒木建治，栃内香次：“実例からの帰納的学習を用いた構文解析手法”，信学技報，TL99-44, 2000.

[12] 荒木健治，栃内香次：“帰納的学習による語の獲得および確実性を用いた語の認識”，電子通信学会論文誌 D-II, Vol. J75, No. 7, pp. 1213-1221, 1992.

[13] (株)メディアドライブ研究所：“模擬対話音声コーパス”，1995.