自然言語処理　144－24
（２００１．７．１７）

帰納的学習を用いた括弧つきコーパスからの意味解析規則の自動獲得手法の性能評価

渋木 英潔　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

E-mail：{shib,araki,tochinai}@media.eng.hokudai.ac.jp

あらまし　本稿では，帰納的学習を用いて括弧つきコーパスから意味解析規則を自動的に獲得する手法を提案し，本手法を実装したシステムの評価を行う．機械翻訳や対話処理などの高度な自然言語処理を行うためには，意味的情報の活用が必要と考えられる．しかしながら，意味的情報を自動的に獲得する従来手法では，事前に多くの格パターンを用意したり，深層格やシソーラスなど意味に関係する情報を必要とする．本手法では，意味解析規則の他に深層格やシソーラスも獲得対象とし，それらの情報を括弧つきコーパスという統語的情報だけが与えられた事例から教師を必要とせずに帰納的に学習する．EDR コーパス 200 文を用いて学習した結果，オープンデータ 50 文に対して 35.3%の正解率で意味解析規則が獲得された．

キーワード　自然言語処理，帰納的学習，意味解析，深層格，シソーラス

Evaluation of a Method for Automatic Acquisition of Semantic Rules from a Bracketed Corpus using Inductive Learning

Hideyuki Shibuki　Kenji Araki　Koji Tochinai

Graduate School of Engineering, Hokkaido University
N13-W8, Kita-ku, Sapporo 060-8628, JAPAN

E-mail：{shib, araki,tochinai}@media.eng.hokudai.ac.jp

Abstract　In this paper, we propose a method for automatic acquisition of semantic rules from a bracketed corpus using inductive learning, and evaluate a system based on this method. It is necessary for advanced natural language processing to use semantic information. In previous methods for automatic acquisition of semantic information, they require many patterns for detecting and deciding information, or need other semantic information such as deep cases, thesaurus, and so on. Our method acquires semantic rules, which include deep cases and thesaurus, from only sentences analyzed syntactically by unsupervised inductive learning. Our system learned from 200 sentences in the EDR corpus. As a result, a 35.3% of accuracy rate was achieved in 50 sentences as open data using the acquired semantic rules.

key words　natural language processing, inductive learning, semantic analysis, deep case, thesaurus

1　はじめに

機械翻訳や対話処理などの高度な自然言語処理を行うためには，形態素や構文などの表層的情報だけではなく，深層格などに代表される意味的な情報の利用が必要であると考えられる．しかしながら，現在のところそのような意味的情報が誰にでも利用できる形で整備されているわけではない．意味的情報を人手で作成することは莫大な労力がかかり，さらに作成者の主観による影響を大きく受ける．それゆえ，人手を介すことなくコーパスからこれらの情報を自動的に学習することが望ましい．以上の背景から，我々は帰納的学習を用いて，単語境界，内容語と機能語のタグ，依存関係という表層的な情報だけが付与されたコーパスから意味的情報を自動的に獲得する手法を提案している [1]．本手法で獲得する意味的情報とは，深層格，深層格を決定するための規則，選択制限，分類シソーラスに相当するものである．

意味的情報を獲得する研究として，文献 [2, 3, 4] などがある．これらの手法は，事前に与えられた表現形式や格助詞などによるパターンを用いて獲得する．このような手法には，用意されたパターンだけで全てを網羅できないという問題がある．また，深層格やシソーラスなど意味に関係する深層的な情報が既に与えられており教師なし学習ではない．さらに，これらの手法は，動詞と下位範疇化される名詞間の関係，上位下位概念にある名詞間の関係と対象を限定している．本手法では，パターンを用いずヒューリスティックスと頻度を指標に，動詞，名詞，形容詞といった対象を限定せずに学習できるという利点がある．

我々の研究の目的として，特定の言語に依存した知識を使わずに学習するということがある．これは，将来的に他言語への適用を見据えてのことである．表層的な情報である前述した知識は，以下に挙げる手法をタグなし文に用いて明示することが可能であると考えられる．我々は，文献 [5] においてタグなし文から依存関係を学習する手法をすでに提案している．内容語と機能語のタグ付けにおいては，EDR コーパスを用いた我々の実験より，出現頻度を基に 90%近い精度で分類できることを確認している．単語境界についても，中国語ではあるが文献 [6] の 91%という値から高い精度で示すことが可能であると思われる．誤りの部分が及ぼすであろう影響については，今回の研究では対象外としており今後の課題である．

本稿では，2 節と 3 節で基本的な考え方と学習の処理過程を記述した後，4 節で EDR コーパス [7] を使った実験について述べ，その結果について考察する．5 節で結論を述べる．

2　基本的な考え方

2.1　意味解析

本手法で獲得する知識を説明するために，まず，本手法で用いる意味解析について説明する．入力は，2 文節間の依存関係が明示された文を仮定する．ここでの文節とは，1 つの内容語と 0 個以上の機能語で構成された句のことを指す．内容語は意味解析の出力に出現する単語であり，機能語は出現しない単語である．1 文節内に出現する機能語全体を 1 つの記号で表現したものを機能句と呼び，文節を内容語と機能句のペアで表現したものを hamlet と定義する．従って，2 文節間の依存関係は hamlet 対（**p**air of **h**amlets, **PH**）で表現される．また，PH における両方の hamlet の機能句の組を，その PH の機能句対（**p**air of **f**unction phrases，**PF**）と呼ぶ．hamlet 間の依存関係において，中心的な役割を果たす hamlet が存在し，これを **h**ead **h**amlet（**HH**）と呼ぶ．HH 以外の hamlet は **c**omplement **h**amlet（**CH**）となり，HH に対してある深層的な役割をもって存在する．この深層的な役割を duty と呼び，格文法で用いられる深層格も duty の一種とみなされる．duty には，動詞と名詞間の関係だけではなく，名詞と名詞，名詞と形容詞などの関係も含まれる．例えば，文節と依存関係を () を用いて表現した 2 文節の文 ((太郎が)(走っている．)) において，HH は [走っている．]（内容語 “走”），CH は [太郎が]（内容語 “太郎”），PF は 〈っている.〉←〈が〉である．ここで，[ ] と ◇ は hamlet と機能句をそれぞれ示し，〈B〉←〈A〉は A が B に係っていることを示している．また，この場合の duty は “動作主格” に相当することが期待される．従って，本手法の意味解析とは，入力に対応する PH から HH となる hamlet を探し出し，その HH に対する CH の duty を決定する作業のことである．

次に意味解析を行うために必要な規則を説明す

る．深層的な役割を表す duty はシステムが入力から自動的に作成し，作成された duty は独自に割り振られた番号によって識別される．従来研究 [2] では，深層格は事前に設定されていたが，本手法では獲得対象とする．これは，設定の基準や原理が明確ではないこと [8] と，設定者の主観による影響を回避するためである．duty は，最初，学習データ中の全ての PH に 1 つずつ独自に割り当てられる．その後，類推を基にした階層的クラスタリングでボトムアップにまとめ上げていく．階層的クラスタリングの方法は，3.2 に記述する．

duty を決定するための第 1 の要因は，PF，および HH に対する CH の位置関係（依存方向）などの機能的要因である．日本語では，一般に格助詞などの機能語による影響が大きいことが知られている．しかし，同じ格助詞「が」であっても，((太郎が)(呼ぶ)) と ((太郎が)(呼ばれる)) のように係り先の機能句によって格が変化することから，本手法においては，機能句を単独では用いず，PF を用いて duty を判断する．依存方向は，英語のような分析型言語においてだけではなく，日本語においても，((太郎が)(走っている)) と ((走っている)(太郎が)) のように影響があると考えられる．

機能的要因の他に duty を決定する要因として選択制限の考え方がある．選択制限を用いるためには，シソーラスの存在が必要である．本手法では，学習データ中に出現した内容語（および，PF と duty）に対し，階層的クラスタリングを行うことで分類シソーラスを作成する．

選択制限と機能的要因から duty を決定するための規則を ordinance と呼ぶ．ordinance は PH から 3.3 に記述される方法で作成，一般化される．

第 3 の要因として，必須格と任意格の考え方がある．本手法では，これを必須か任意という 2 段階で表現せず，確率化された数値で duty の必須度合いを表現する．この知識を demand と定義し，全ての内容語ごとに，その内容語と共起する duty を計算し学習する．

最後に，duty を決定する要因は ordinance と demand だけによらない．例えば，((太郎は)(運んだ)) における duty を考えた場合,「太郎は次郎を運んだ」と「太郎は次郎が運んだ」では異なる．これは，[次郎が (を)] という hamlet が既に [運んだ] の中の duty を占有しているために起きる現象と考えられる．この現象を処理するために，HH を中心とした依存関係の束を用い，これを borough と呼ぶ．borough はフレーム形式に似ており，スロット名に duty，スロット値に CH がそれぞれ対応する．

2.2 派生

前節で述べた規則を用いて意味解析を行うが，規則を用いて解析するときに問題となるのは，規則の適用率，すなわち規則を適用することのできない事例をどう処理するかということである．大規模なコーパスから単純に大量の規則を獲得したとしても，そのコーパスが現実世界に起こり得る全ての事例を含んでいるという保証がない限り，規則の完全性も保証できない．また，大量の規則を獲得することは，規則間の整合性の欠如と適用時の競合を引き起こす．これらの問題は，規則を杓子定規にしか適用できないことに原因があると考えられる．

このような問題から，用例ベース [9] や統計ベース [10] の解析手法が提案されているが，それぞれ新たな問題が生じる．用例ベースにおいては，解析のために類似した大量の用例が必要になる．また，単純に用例を保持するだけであり，用例の一般化や抽象化といった学習を行うことがないため，通常の規則ベースと同様に整合性の欠如や適用時の競合が起きうる．さらに，規則化されていないため演繹的な推論を行うことができない．N-gram に代表される統計ベースでは，言語的特長を捨象し確率情報だけを保持するため，大局的な情報しか得られず，希少な事例への対処が難しい．統計の信頼性を高めるためには，用例ベースと同じく大量のデータが必要となる．また，確率的に最大の結果を選択するだけでは，どんなに確実な統計をとったとしても 100%でない限り，必ず例外となる結果が存在することになる．一方，規則ベースにおいては，規則を組み合わせて演繹的推論を行うことで未知の事例に対処でき，例外的な事例は規則に明示することで対処できる．従って，規則の適用率と整合性の問題を解決できれば，規則ベースの方が有利であると考えられる．

我々は，整合性のとれた少数の基本的な規則を保持し，直接適用できない事例に対しては，適用時に若干規則を変形させることで解決する方法をとった．事例に合わせて規則を変形させることを

派生と呼ぶ．派生には一定のコストを定め，無制限な派生を禁止する．本稿での派生は，規則を構成する要素の挿入，削除，倒置の三つである．要素の置換については，削除と挿入を行うことで可能となる．

事例に合わせて規則を派生させることにより，どのような事例に対しても必ず規則を適用して解析することができ，適用率の問題を解決できる．しかしながら，このことは逆に，ある依存関係に対して全ての規則が適用できるということであり，適用する規則を選択するための尤度関数が重要になる．規則選択のための基本的な考え方は，規則を適用することによって得られる収入 (surplus) がその規則の派生に必要な支出 (deficit) を上回るものを選択するというものである．ordinance x，demand y を使用したときの尤度 (plausibility(x, y)) は式 (1) で表される．

$$
\begin{aligned}
&\mathrm{plausibility}(x, y) \\
&\quad = \mathrm{surplus}(x, y) - \mathrm{deficit}(x) \qquad (1) \\
&\mathrm{surplus}(x, y) \\
&\quad = c_1 \times Cv_{hc}(x) \times In_{hc}(x) \\
&\quad + c_2 \times Cv_{cc}(x) \times In_{cc}(x) \\
&\quad + c_3 \times Cv_{pf}(x) \times In_{pc}(x) \\
&\quad + c_4 \times Dm(y) \\
&\mathrm{deficit}(x) \\
&\quad = c_5 \times Dir(x) \\
&\quad + c_6 \times Cv_{hc}(x) \times Out_{hc}(x) \\
&\quad + c_7 \times Cv_{cc}(x) \times Out_{cc}(x) \\
&\quad + c_8 \times Cv_{pf}(x) \times Out_{pf}(x)
\end{aligned}
$$

ここで，Cv_{hc}，Cv_{cc}，Cv_{pf} は，それぞれ，HH の内容語 (HC)，CH の内容語 (CC)，PF に対する ordinance の非適用範囲で $0 < Cv \leq 1$ の値をとり，適用できる範囲が狭い（一般化がされていない）ほど高くなる．この値は，3.2 の階層的クラスタリングで作成したシソーラスから求められる．In は適用範囲内に存在すれば 1，そうでなければ 0 を返す．Dm は HC に対する demand の値である．Dir は適用時の依存方向に必要なコストであり $0 \leq Dir \leq 1$ である．Out は，適用範囲内に存在すれば 0 を返し，そうでなければ，どのくらい外れているかを $0 < Out \leq 1$ の値で返す．この値もシソーラスを基に算出される．また，式中の c_i は係数を表す．

派生の効果は適用時にのみ及ぶものではない．学習時において，派生により生成されたと思われる規則を原形に戻すことにより，冗長な規則を削除する．また，派生時の尤度を基に規則の一般化を行い，規則間の整合性をとることができる．一般化された規則は，他の規則との比較を容易にし，学習をさらに促進させることが可能となる．これは，データスパースネス問題の解決にも繋がる．それゆえ，事例をそのまま保持し適用時にのみ類推して解決しようとする用例ベースのアプローチとは，一線を画している．規則の一般化の具体的な処理は 3.3 に記述する．

3 処理過程

3.1 全体の流れ

本システムは，下に示す 4 つの処理によって構成されている．

(1) HH の決定

(2) シソーラスの作成

(3) demand の作成

(4) 規則の一般化

上記の順番で処理を行っていく．本稿では，紙面の都合により，全ての処理を説明することはできない．そこで，文献 [1] から，特に変更のあったシソーラスの作成と規則の一般化について記述する．

3.2 シソーラスの作成

内容語，PF，duty のそれぞれについて階層的クラスタリングを行い，シソーラスを作成する．本手法のシソーラスは木構造を仮定している．基本的な考え方は，入力データと以前のシソーラスを基に類似度を算出し，その類似度から新たなシソーラスを作成するというものである．最初のシソーラスは，全ての要素が 1 つのルートの直下にまとめられている状態である．

本手法では，文献 [5] と同様のアプローチでクラスタリングを行う．まず，学習する事例に対してユニークにシンボルを割り当て，その後,「周囲の

環境が類似したシンボルは類似している」という指標に基づいてまとめ上げていく．類似度は，PHの各要素を「周囲の環境」とみなし，2つのPHを比較することで計算される．

あるPHxとPHyを比較する場合を仮定する．この場合，xにおける依存方向，duty，HC，CC，PFを，それぞれ，dir_x，dt_x，hc_x，cc_x，pf_xとすると，xとyを比較したときのdutyの類似度$Sim\alpha_{dt}$，HCの類似度$Sim\alpha_{hc}$，CCの類似度$Sim\alpha_{cc}$，PFの類似度$Sim\alpha_{pf}$は以下の式で表される．

$$
\begin{aligned}
Sim\alpha_{dt}(dt_x, dt_y) &= A + C + D + E && (2)\\
Sim\alpha_{hc}(hc_x, hc_y) &= A + B + D + E && (3)\\
Sim\alpha_{cc}(cc_x, cc_y) &= A + B + C + E && (4)\\
Sim\alpha_{pf}(pf_x, pf_y) &= A + B + C + D && (5)\\
A &= c_9 \times Al_{dir}(dir_x, dir_y)\\
B &= c_{10} \times Cs_{dt}(dt_x, dt_y)\\
C &= c_{11} \times Cs_{hc}(hc_x, hc_y)\\
D &= c_{12} \times Cs_{cc}(cc_x, cc_y)\\
E &= c_{13} \times Cs_{pf}(pf_x, pf_y)
\end{aligned}
$$

ここで，Alは依存方向が同じなら1，そうでなければ0となる．Csは，シソーラスにおける距離の近さを$0 \leq Cs \leq 1$で表した値であり，近いほど大きくなる．全てのPHの組に対して上記の計算を行い，平均をとったものが基本類似度$Sim\beta$となる．

本手法では，知識を決定的に学習しない．そのため，学習される知識が常に変動し，収束しない可能性がある．変動する可能性がある中で常に同じ結果が導き出されるとしたら，その結果は非常に安定しており信頼できると考えられる．逆に絶えず異なる結果であった場合，その結果はあまり信頼できない．以上の考えから，信頼度Rlを安定度Sbを基に式(6)で計算する．

$$
Rl(x,y) = \frac{20 + Gd(x,y) \times (Sb(x) + Sb(y))}{40} \quad (6)
$$

ここで，$Gd(x,y)$はxとyの上位概念が同一であれば1，そうでなければ-1を返す．Sbは，$0 \leq Sb \leq 10$の値をとり，直前の結果と同じ結果になった場合+1，そうでなければ0となる．

次に，クラスタリングの目的について述べる．クラスタリングの目的は，類似した事例をまとめることで，出現頻度の低い事例に対処することである．従って，出現頻度が十分に高いクラスタは既に1つの独立したクラスタであり，それ以上他のクラスタと統合される必要がなく，逆に，類似度がさほど高くなくとも，その事例の頻度が低ければ統合するべきと考えられる．このことから，独立度Idを出現頻度Fqから式(7)を用いて計算する．

$$
Id(x,y) = Fq(x) + Fq(y) \quad (7)
$$

最終的な類似度Simは，$Sim\beta$，Rl，Idを用いて，式(8)に従って計算される．

$$
Sim(x,y) = \frac{Sim\beta(x,y) \times Rl(x,y)}{Id(x,y)} \quad (8)
$$

このSimを用いて，階層的クラスタリングを行う．最初にシンボル一つだけからなるクラスタの初期集合を求める．初期集合のクラスタ間の類似度はシンボル間の類似度と同じである．次に，最も類似度の高い組から順に，その二つのクラスタを包含するクラスタを作成し，ボトムアップにまとめ上げていく．本手法は二分木であることを仮定していないので，作成されたクラスタに三つ以上の要素が含まれる可能性がある．そこで，クラスタを作成したときの類似度と同じ類似度を持つクラスタも，そのクラスタに包含させる．新しく作成されたクラスタには新たなシンボルを割り当て，そのシンボルとの類似度を計算する．このシンボルはPH中に出現しないため，$Sim\beta$は包含したときの$Sim\beta$を使用し，RlとIdを新たに算出してからSimを再計算する．以上の作業を一つのクラスタになるまで繰返す．

次に，作成されたシソーラスからクラスタの個数を求める方法について述べる．階層的な木構造が求められていれば，要素間の距離を計算するのは難しいことではなく，必要に応じてクラスタの個数を変更することができる．出現頻度の低い事例に対処するというクラスタリングの目的に従って，各クラスタ内の要素の総出現頻度が等しくなるように分類する．作成時とは逆に，総出現頻度が等しくなるようなクラスタ集合をルートクラスタからトップダウンに探し出す．そのような集合は何通りか考えられるが，閾値以上の個数でクラスタの総数が最も少ないものを選択する．閾値を設けたのは，純粋に工学的な理由，あまりに少なすぎるdutyでは解析時の分類が上手く行かない

という理由からである．また，閾値は 30 クラスタであり，各クラスタに含まれる最低要素数は 2 である．

作成されたシソーラスが直前のシソーラスと同じになるか，一定の回数を繰返すまで，シソーラスの再構築を行う．今回，再構築を打ち切る回数は 3 回とした．

3.3 規則の一般化

同じ duty を導く複数の規則の存在は，解析時に競合を起こし精度の低下を招く．それゆえ，同じ duty を導く規則は 1 つの規則に一般化されることが望ましい．しかしながら，同じ duty を導く規則が必ずしも 1 つの規則に一般化されるとは限らない．さらに，学習事例の中には不自然な表現が含まれている可能性があり，それらの事例から獲得される規則は例外として扱った方が良い場合もある．以上の理由から，以下の三通りの方法で規則の保持を行う．

- 獲得された規則を個別に保持する．
- それらの規則全てを包括する一般化された一つの規則を保持する．
- 不自然な事例から獲得された例外的な規則であり保持しない．

どの方法で保持するかは，学習事例を最も効果的に解析できる方法で保持する．

規則を一般化する方法は，シソーラスの作成と同様に，具体的な規則から優先度を基にしてボトムアップに一般化していく．まず，ある学習事例だけを説明する規則の集合から，同じ duty を導く規則のペアを作成する．ペアごとに，両者を独立した規則とみなした場合，両者を一般化する規則を作成した場合，一方の規則を他方の規則の例外的な規則とみなした場合を仮定し，それぞれの場合において両者が説明する事例を最も効果的に解析できる場合を選択し，その場合の尤度をそのペアの優先度とする．尤度は，式 (1) から求められた値に，その規則が説明する事例の数をかけたものである．一般化された規則を保持する場合と独立した規則を保持する場合を比較したとき，どちらの場合においても deficit は 0 であり，surplus は一般化されていない分だけ独立した規則の方が高くなる．我々は，文献 [11] で主張するように，規則を記憶することにもコスト（記憶コスト）が必要であると考え，可能な限り規則数を減少させるように学習を行う．従って，独立した規則を保持する場合の優先度は，上記の尤度から記憶コストを引いた値とした．ペアの優先度が高い順に，選択された方法で保持，一般化を行う．このとき，一般化する規則が作成されれば，その規則と同じ duty を導く規則から新たなペアを作成し，同様にして優先度を決定する．すべての規則が独立した規則と判断されるまで，この作業を続ける．

以上で述べた本手法の一般化は，文献 [12] の便宜的選言記述を実現していると考えられる．しかしながら，現段階で本手法の一般化は，否定属性を表現できないため，便宜的連言記述による完全性を表現することができず，今後の課題である．

表 1: Results from the experiment

Sentence	PH	AR
50	391	34.2%
100	831	34.4%
200	1,676	35.3%

4 実験

4.1 実験手順

我々は，本手法を実装するシステムを作成し実験を行った．実験には EDR コーパスを用いた．本手法の入力条件に合致する 250 文をランダムに選択し，その内の 50 文をテストデータとし，残りの 200 文をトレーニングデータとした．200 文のトレーニングデータ 1 文あたりの平均 PH 数は 8.4 である．トレーニングデータを用いて学習を行い，獲得された規則を用いて，テストデータの解析を行った．また，トレーニングデータ数と精度の関係を調べるため，トレーニングデータの 50 文と 100 文のサブセットを作成し，同様の実験を行った．今回の実験では，式中の係数 c_i は全て 1 として実験し，処理効率のために，予備実験の結果から，尤度評価が 11 位以下の結果を枝狩りしながら解析を行った．

客観的な評価を行うために，EDR コーパスの

表 2: Details of the results

Sentence	direction	duty
50	82.5%	41.5%
100	77.6%	44.3%
200	79.2%	44.6%

格フレームを正解として正誤を判断した．評価は，尤度評価が 1 位である結果のそれぞれの PH に対して行い，CH に対する HH が一致し，かつ，その HH に対する duty が EDR の深層格と一致する PH を正解とした．深層格との一致の判定は，duty のクラスタに対応する深層格の集合中で最も多数を占める深層格と一致しているかどうかで行った．ただし，EDR コーパス中には，表層文中に明示されていながら格フレーム中では省略されているもの，あるいは，その逆に格フレーム中だけに出現するものがあり，それらの依存関係は無効なものとして評価の対象から除外した．正解率 (**A**ccuracy **R**ate, **AR**) は，式 (9) により計算される．

$$\mathrm{AR} = \frac{NAR}{NPH} \times 100[\%] \tag{9}$$

NAR は正解である PH の数，NPH は無効な PH を除いた全ての PH の数である．ここで我々が主張したいのは，EDR コーパスの結果を完全な正解としたいわけではないということである．意味という主観的な知識が，万人に全て共通しているとは考えにくく，将来的には，本手法により自動的に構築された知識を，直接アプリケーションに利用することを意図している．しかしながら，基本的な大部分の意味に対しては共通しているのではないかと考えられるため，今回の基準を採用した．

表 1 に実験結果を示す．左から順に文数，依存関係数，正解率である．また，規則の派生を行うことにより，規則の適用率は 100%である．

4.2 考察

表 1 からトレーニングデータ数が精度に大きな影響を与えていないことが分かる．これは，派生を基にした規則の一般化による効果であると考えられる．一般化された ordinance の数は，50 文から順に 44，41，56 であり，トレーニングデータ数の増加に比例して，ordinance が無用に増加することを防いでいる．このことから，規則の一般化が有効に作用していることが確認できた．また，50 文から 200 文という少ない文数で一定の成果を収めたことから，データスパースネス問題に対しても効果的であると考えられる．

正解率の内訳を表 2 に示す．左から順に，文数，CH に対する HH の一致率，その HH に対する duty の一致率である．HH を同定する段階で既に精度が 8 割程度に低下している．これを防ぐ最も単純な方法は，日本語の依存方向は前から後ろに係るというヒューリスティクスを与えることである．しかしながら，特定の言語に依存しないという我々の目的から，このような方法は採らない．HH の決定過程における結果を調べた場合，後ろから前に係ると判断された依存関係の割合は 6.6%であった．この値は，改善の余地が残されてはいるが，HH の決定が妥当に行われたことを示している．また，コストを必要とせずに後ろから前に適用できる ordinance の割合も 3.6%と低く，この誤りの原因の殆どが適用時の倒置による派生が原因であると推測される．しかしながら，この問題は，倒置に要するコストの係数を高めに設定することで，解決できると考えられる．

次に duty の精度について考察する．文献 [2] では，事前に用意された多くの格パターンとシソーラスを用いて，既に設定された深層格を割り当てる実験を行い，EDR 日本語共起辞書に対して 80.3%の精度を達成している．文献 [2] とは条件が大きく異なるが，我々の結果は，彼らの半分程度の精度しか達成されていない．200 文の学習データから作成された duty のクラスタ数は 32 個であり，1 クラスタあたりの平均 duty 数は 52.1 個だった．従って，ランダムに duty を割り振った場合に比べて，表 2 の 40%以上という結果は，明らかに正しくクラスタリングされていることが分かる．しかしながら，実用的な水準には達していない．直感的にも，duty のクラスタリングには，PF の影響が大きくなるように係数を設定した方が良いと考えられる．これは，HH の同定と同じくパラメータの問題であり，最適化により，どこまで精度を向上させることができるかが今後の課題である．

5 まとめ

本稿では，深層格やシソーラスを含めた意味解析に必要な規則を，括弧つきコーパスから自動的に獲得する手法を提案した．本手法は，従来手法と異なり，事前に多くの格パターンや，深層格やシソーラスなどの意味に関係する情報を必要とせずに，教師なしで帰納的に学習する．規則の派生という概念を用いることで，規則ベースの手法の問題点を解決し，100%の適用率と，整合性のとれた規則の獲得，そして，データスパースネス問題に対処できるようになった．

本手法を実装したシステムは，EDR コーパスを用いた実験において，200 文という少数の事例から獲得した規則を用いて，オープンデータである 50 文を 35.3%の正解率で解析できた．正解率では，未だ実用的な水準に達していないが，今回の実験では，パラメータの最適化を行っておらず，最適なパラメータの設定により向上を図ることが期待できると考えられる．

今後の課題として，パラメータの最適化の他に，規則の一般化において便宜的連言記述を可能とすることで，より適切な規則を学習し精度の向上を図ることが挙げられる．また，今回括弧つきコーパスからの獲得であったが，入力として必要な情報（単語境界，内容語と機能語のタグ，依存関係）は，先行研究から，タグなしコーパスに対して教師なしで自動的に付与することができると考えられ，これらの手法を組み合わせることで，タグなしコーパスから意味解析規則を自動的に獲得したいと考えている．

参考文献

[1] 渋木英潔, 荒木健治, 栃内香次, “帰納的学習を用いたタグなしコーパスからの統語規則の自動獲得手法,” 言語処理学会第 7 回年次大会発表論文集, pp.472–475, March 2001.

[2] 大石亨, 松本裕治, “格パターン分析に基づく動詞の語彙知識獲得,” 情報処理学会論文誌, vol.35, no.11, pp.2597–2610, Nov. 1995.

[3] 宇津呂武仁, 松本裕治, 長尾眞, “二言語対訳コーパスからの動詞の格フレーム獲得,” 情報処理学会論文誌, vol.34, no.5, pp.913–924, May 1993.

[4] S.A.Caraballo, “Automatic construction of a hypernym-labeled noun hierarchy from text,” Proc. of 37th Annual Meeting of the ACL, pp.120–126, June 1999.

[5] 渋木英潔, 荒木健治, 栃内香次, “帰納的学習を用いたタグなしコーパスからの統語規則の自動獲得手法,” 電子情報通信学会論文誌 D-II, vol.J83-D-II, no.10, pp.2003–2016, Oct. 2000.

[6] S.Maosong, S.Dayang, B.K.Tsou, “Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data,” Proc. of 17th International Conference on Computational Linguistics(COLING ’98), pp.1265-1271, Aug. 1998.

[7] 日本電子化辞書研究所, “EDR 電子化辞書使用説明書,” 1995.

[8] 田中穂積監修，“自然言語処理—基礎と応用—，” 東京，電子情報通信学会，1999.

[9] M.H.Al-Adhaileh, T.E.Kong, “A flexible example-based parser based on the SSTC,” Proc. of 36th Annual Meeting of the ACL, pp. 687–693, 1998.

[10] M.Collins, “Three Generative, Lexicalised Models for Statistical Parsing,” Proc. of 35th Anuual Meeting of the ACL and 8th Conference of the ECACL, pp.16–23, 1997.

[11] 渋木英潔, 荒木健治, 栃内香次, “帰納的学習により獲得された統語規則の記憶コストと曖昧性減少に基づく最適化手法,” 平成 12 年度電気関係学会北海道支部連合大会講演論文集, pp.239–240, Oct. 2000.

[12] 仁木和久, 石崎俊, “概念の帰納的学習,” 人工知能学会誌, vol.3, no.6, pp.696–703, Nov. 1988.