自然言語処理 121－6
（1997. 9. 11）

構文情報と意味情報からの構文規則の獲得

小島 丈幸　乾 伸雄　小谷 善行
(東京農工大学　工学部　電子情報工学科)

本稿では、自然言語文の構文情報と意味情報から文法規則を獲得する手法を提案する。獲得例としては、EDR コーパスのを想定している。獲得に際し、例から得られる以外の情報は用いない。

まず、構文的な構造だけでなく、意味的な構造も生成可能な文法体系を設計した。この文法体系では、規則は CFG の書換え規則に制約と意味構造生成規則を付加したものになる。次に、構文情報と意味情報の対から文法規則を獲得するアルゴリズムを考案した。最後に、実験によって規則を生成した。

Grammatical Rule Acquisition from Syntactic and Semantic Information

Takeyuki KOJIMA, Nobuo INUI, Yoshiyuki KOTANI
(Dept. of Computer Science, Tokyo University of Agric. and Tech.)

This paper proposed the method that acquire grammartical rules from examples. The examples are consist of pairs of a syntactic structure and a semantic structure. The pairs are extracted from the EDR coupus.

First, we designed the grammmar that generate not only a syntactic structure but also a semantic structure. This grammar is the context-free grammar with restrictions. Second, we desingned the grammar acquiring algorithm. Finally, we generated the grammartical rules by a examination.

1　はじめに

自然言語処理における知識獲得は、大規模なコーパスの利用と多量の計算処理が現実的なものとなった近年、ますます盛んに行われるようになっている。構文解析に必要な文法も、実用的な段階のものは人手で作成するのが困難であり、時間や労力・技能を必要とする人手による困難な作業を避け、計算機で自動的に獲得する研究が行われている。

こういったコーパスからの文法獲得では、何らかの文法体系を仮定しなければならない。この獲得する文法体系としては、文脈自由文法や係り受け文法がよく用いられている。文脈自由文法の句構造規則を獲得する研究には、品詞情報だけから統計的な手法によって獲得するもの [3] や、構文解析済みの括弧付コーパスから、類似した括弧をグルーピングして獲得するもの [2] がある。

しかし、構文解析における統語的曖昧性を減少させるためには、統語レベルの知識だけでは十分ではない。そのために、統語レベルの知識を補う形で、語彙レベル・意味レベルの制約を課す文法体系が考えられており、その文法体系における文法の獲得が行われている。これらには、統語規則の中に制約を設けるのではなく、統語的な知識と独立な語彙レベル・意味レベルの制約を用いる手法と、統語規則の中に語彙レベル・意味レベルの制約を組み込む手法とがある。前者は、下位範疇化や名詞・動詞の意味分類の情報を、統語的な情報とは独立に獲得する [1]。獲

得した制約も統語的な情報とは独立に記述する。

本稿で述べる手法は、後者に相当する。後者に相当する他の研究では、係り受け制約を組み込んだ確率文脈自由文法の枠組を対象にした研究 [6] がある。ここでは、文脈自由文法の非終端記号を意味により細分化している。そして、意味的に的確な係り受けの制約を、語の共起情報から獲得し、さらに、文法サイズの縮小のためにシソーラスを文法規則内に取り込む。

上記の研究にでは、意味情報を用いても、構文的曖昧性の解消に用いるだけである。これに対し、統語規則に組み込んだ意味的な規則により意味情報をより積極的に得る文法体系を対象にした研究もある。

構文構造と意味構造を対にして、そこから対応規則を獲得する研究 [5] では、構文構造と意味構造が単純に対応していない場合も柔軟に対応できる手法が提案されている。EDR のコーパスの利用を前提に作られているが、構文情報の中間節点には、範疇名がついており、EDR コーパスの構文情報とは異なる。また、実験規模は小さい。シソーラス情報・概念辞書を参照しつつ、自然言語と意味構造の対応規則を獲得する研究 [4] もあるが、これもやはり実験規模が小さい。

筆者らも、過去に **LFG**(語彙機能文法) のサブセットを対象にし、構文的曖昧性解消の制約となる機能スキーマを獲得する研究を行った [8]。この研究でも、意味的な情報を生成するような文法体系を対象としたが、上記二つの研究と同じく、意味情報が多量には入手できず、実験規模が小さかった。

本稿では、構文情報とともに意味情報が得られる文法体系を対象とする。シソーラスや概念辞書といった、例以外の情報は用いずに規則を獲得する。獲得の際の意味情報には、意味情報が多量に得るため EDR コーパスを利用する。まず、上記の文法を設計する。これは、文脈自由文法に制約を加えた文法である。次に、解析済み構文情報と属性構造で与えられる意味情報の対から帰納的に文法規則を獲得する手法を提案する。

2 例の形式

本稿では、大量の学習例を得るために、EDR コーパスの構文情報と意味情報を用いることを想定している。しかし、EDR コーパスの情報そのままでは複雑なため、一部制限された構造を利用する。

2.1 構文情報

構文情報は、依存構造を基本とした木構造で表現されている。中間節点に範疇名という情報はなく、その節点が子の節点をどのようにまとめているかという情報、すなわち、合成関係が記されている。さらに、同じ節点を親の節点としてもつ子の節点のうち、どれか一つは必ず主節点であり、その情報も記載されている。

EDR コーパスには修飾合成・統合合成・数合成・複合合成の四つがあり、統合合成は、さらに従属合成と等位合成に分類できる [7]。そのうち、本稿では次の二つの合成関係だけを利用する。

修飾合成 子ノードの依存関係を記した合成関係

従属合成 概念を持つ自立語と概念を持たない付属語の合成

図 1 に EDR 構文情報の例を示す。これは、「車/の/中/に/潮/の/香り/が/し/て/き/まし/た/。」という形態素列に対応した構文情報である。主節点は太い四角で囲ってある。末端以外の節点には、合成関係が付加している。

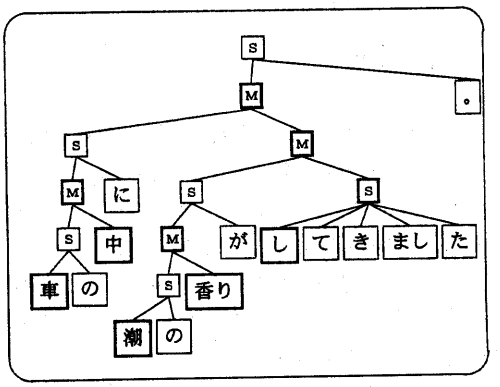

図 1: 構文情報の例

2.2 意味情報

意味情報は、文の述語となっている概念と、文を構成する他の要素との関係を、属性構造の形で表現したものである。

EDR で用意されているすべてのスロット名は使わない。たとえば、attribute スロットはテンス情報やアスペクト情報を持っているが、これは扱わない。また、EDR の意味情報は概念へのリンク情報が付加しているがこれも利用しない。

たとえば、「車/の/中/に/潮/の/香り/が/し/て/き/まし/た/。」という形態素列に対応する意味情報を属性構造で表現すると、図 2 のようになる。

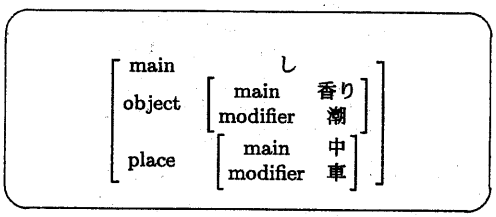

図 2: 意味情報の例

2.3 構成要素情報

EDR の構成要素情報は、文を構成する形態素ごとに記載された、表記・概念へのリンクなどの情報である。本稿で利用するのは、そのうち品詞情報だけである。構文情報に品詞の情報が無いので、構成要素情報からの品詞情報を付け加える。

3 獲得する文法体系

3.1 文法体系への要請

本節では、2 節で説明した例を解析可能な文法体系を設計する。基本的には、文脈自由文法に構文的・意味的な制約を課した文法体系を用いる。さらに、構文解析の際に、構文情報だけでなく意味情報も生成するような文法体系を用いる。つまり、ここでは想定する文法体系には、次のことが要請される。

(1) 文脈自由文法に制約を課した文法であること

(2) 構文解析の結果として意味情報も得られること

3.2 各節点に付加する情報

LFG は 3.1節の要請を満たすが、解析結果として生成される機能構造には、意味情報のほかに統語的な情報も含まれてしまう。そこで本稿で設計した文法では、各節点に対し、範疇のほかに接続情報と意味情報の二つの情報を付加する。これによって、明示的に意味情報と統語情報を分けることができる。接続情報と意味情報はそれぞれ属性構造で表現する。

解析は、与えられた形態素列に対して規則 $R \in \boldsymbol{R}$ を繰返し適用することで行われる。各規則がその子節点に適用できるかどうかは、規則に付加した制約で決まる。制約が満たされて規則が適用されれば、親節点の意味情報や接続情報が決定する。

規則は次の三種類とした。

語彙的規則 語彙と品詞を結び付ける規則

修飾合成規則 EDR 構文情報の修飾合成に対応する規則

従属合成規則 EDR 構文情報の従属合成に対応する規則

次節から、それぞれの規則について述べる。

3.3 語彙的規則

3.3.1 語彙的規則に対する要請

語彙的規則は、形態素とその品詞を結び付ける規則である。品詞に対応する節点、すなわち、親節点に構文的・意味的な情報を与えなければならない。それ自身が意味を持つ自立語の場合には意味的な情報の記述が、付属語の場合には構文的な情報の記述が不可欠である。

3.3.2 語彙的規則の形式

語彙的規則 R_{L} の定義は、次のように書ける。

$$
\begin{aligned}
R_{\mathrm{L}} &= C \rightarrow \langle m, \boldsymbol{u} \rangle,\ C \in \boldsymbol{P}, m \in V_{\mathrm{T}}\ , \\
\boldsymbol{u} &= \{u_1, u_2, \cdots, u_p\}\ , \\
u_k &= \langle a_k, v_k \rangle,\ a_k \in A, v_k \in V_b\ .
\end{aligned}
$$

ここで、$\boldsymbol{P} \subseteq V_{\mathrm{N}}$ は品詞の集合である。$\boldsymbol{u}$ は、接続情報を決定する。

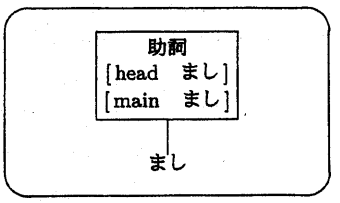

図 3: 語彙的規則の適用例

3.4 修飾合成規則

3.4.1 修飾合成規則に対する要請

修飾合成規則は、意味を持つ二つの節点を合成する。一方が被修飾節点になり、他方が修飾節点になる。たとえば、名詞句とそれに係る形容詞の二つの節点から名詞句を生成する規則、用言句とそれに係る連用句の二つの節点から用言句を生成する規則などがこの修飾合成規則に相当する。

親節点の接続情報には、被修飾節点の接続情報がそのまま伝わればよい。意味情報に関しては、親節点の意味情報は被修飾節点の意味情報で、それに修飾節点の意味構造がある意味的な関係でつながっていればよい。したがって、修飾合成規則には、修飾側の意味情報が被修飾側の意味情報とどのような関係にあるのか記述する必要がある。

3.4.2 修飾合成規則の形式

修飾合成規則 $R_{\mathrm{M}} \in \boldsymbol{R}_{\mathrm{M}}$ は、

$$
\begin{aligned}
R_{\mathrm{M}} &= C \rightarrow X_1\, X_2 \\
X_i &= \langle C_i, \phi_i, \boldsymbol{\lambda}_i \rangle \\
C_i &\in V_{\mathrm{N}} \\
\phi_i &\in \{\text{主}\} \cup \{\langle a, b\rangle \mid a, b \in \Pi\} \\
\boldsymbol{\lambda}_i &= \{\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{ir_i}\} \\
\lambda_{ij} &= \langle a_{ij}, v_{ij}\rangle
\end{aligned}
$$

と書ける。

ϕ_i は、その子節点が主節点かどうか、もし、主節点でなければ、親節点と意味的にどのような関係にあるかを示している。すなわち、ϕ_i = 主 であればその子節点は主節点である。そうでなければ $\phi_i = \langle a, b\rangle$ は、その子節点が親節点に関係 a で係り、逆にその子節点に親節点が関係 b で係ることを表す。ただし、関係 $\diamondsuit \in \Pi$ は、意味的に係らないという関係を表す。

$\boldsymbol{\lambda}_i$ は、その規則が適用可能な子節点を、接続情報の持つべき値で制限している。つまり、a_{ij}-v_{ij} という属性名-属性値対が、i 番目の子節点の接続情報に含まれていなければ、この規則は適用されない。

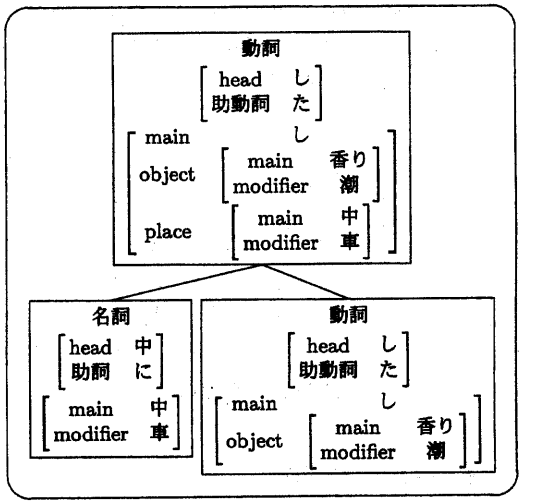

図 4: 修飾合成規則の適用例

3.5 従属合成規則

3.5.1 等位合成規則に対する要請

従属合成規則は、意味を持つ自立語と意味を持たない付属語をまとめ上げる規則である。たとえば、名詞句と格助詞の合成や動詞と助動詞の合成などがこれにあたる。子節点の数は複数だが、一般に 2 とは限らない。

親の意味情報は、主節点、すなわち、自立語の意味情報である。主節点以外の節点、すなわち付属語は、意味情報には影響を与えず、接続情報に影響を与えていると考えられる。したがって、親節点の接続情報は子節点の接続情報をまとめあげたものとする。

3.5.2 従属合成規則の定義

従属合成規則 $R_S \in \boldsymbol{R}_S$ は、

$$\begin{aligned} R_S &= C \rightarrow X_1\, X_2 \cdots X_n \\ X_i &= \{C_i, \theta_i, \boldsymbol{\lambda}_i\} \\ \theta_i &\in \{\text{主}, \gamma\} \\ \gamma_i &= \{\gamma_{i1}, \gamma_{i2}, \cdots, \gamma_{ip_i}\} \\ \gamma_{ij} &\in A \end{aligned}$$

と書ける。ただし、θ_i は、その子節点が親節点とどう関わるかを表している。もし、$\theta_i =$ 主 であれば、その子節点が主節点、すなわち自立語であることを示す。$\theta_i = \gamma$ であれば、その子節点は付属語の節点であることを示す。親に伝える接続情報を γ で示している。

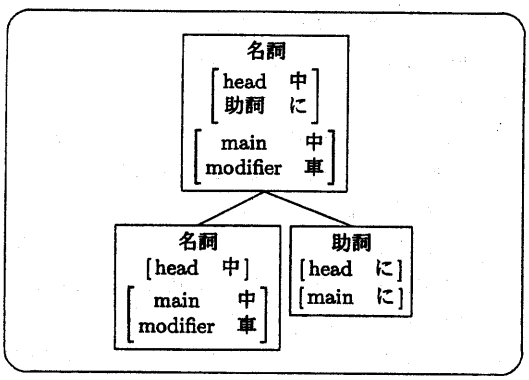

図 5: 従属合成規則の適用例

4 獲得手法

獲得の過程は、生成フェーズと精錬フェーズの二つのフェーズからなる。生成フェーズは、後段の精錬フェーズで必要となる情報を抽出する役割を果たす。精錬フェーズは、統計的な手法を用いて生成された規則を精錬する。

4.1 生成フェーズ

このフェーズでは、非常に簡単な文法を生成する。まず、各節点の接続情報は、次の三つとする。

(1) 主節点の形態素 (head)

(2) 末尾の節点の品詞 (pos)

(3) 末尾の節点の形態素 (morph)

つまり、語彙的規則における $\boldsymbol{u}$ は、常に

$$\boldsymbol{u} = \{\langle \text{head}, m\rangle, \langle \text{pos}, C\rangle, \langle \text{morph}, m\rangle\}$$

となる。

このフェーズでは、中間節点の範疇名は、その節点の主節点の範疇名とする。つまり、通常は名詞句とラベルが付けられる範疇は "名詞" となる。名詞句と格助詞の連接による連用句も "名詞" となる。

与えられた構文情報から、使われる規則の種類がわかる。意味情報は修飾合成規則だけで構成されるので、意味関係が変数であるような意味情報が一意に作れる。そしてこの意味情報と与えられた意味情報とのマッチングを取る。マッチングか成功した部分では、変数だった意味関係がインスタンシエートされ、その修飾合成規則における意味関係が決まる。

接続情報に対する制約は、与えられた例だけを生成できるように最もきつい制限を生成する。つまり、例において子節点に付加していた属性は、すべて必要であったものとして、$\boldsymbol{\lambda}_i$ に含める。

4.2 精錬フェーズ

前段の生成フェーズで生成される規則は、語彙に依存しているために膨大な量になる。したがって、一般化して文法のサイズを減らしていく。一般化の戦略としては、次のようなことが考えられる。

- 類似した規則を見つけ、独立と見なせる制約を消す。
- 類似した規則を見つけ、同じ環境で使われている名詞や動詞をまとめる。

また、同じ範疇名でも、接続情報を元に細分化することが考えられる。どちらも、統計的な手法を用いることで、他の情報を用いること無く一般化が行えると考えられる。

5 実験

EDR コーパスから、2節で定義した形式の例を生成し、規則を獲得する実験を行った。今回は主として、生成フェーズに関する実験を行った。規則の種

類ごとの生成数を図6に、修飾合成規則の生成数を図7に示す。

図6では、修飾合成規則で意味関係規則が獲得できたものを、獲得成功、失敗したものを獲得失敗としている。

図7における強・弱は制約を一部消すことによる一般化を行ったものである。おそらく、構文的な曖昧性をあまり解消していないものと考えられる。図7における成功修飾とは、修飾合成規則で意味関係規則が獲得できたものの、範疇が同じものを一つとして計数したものである。

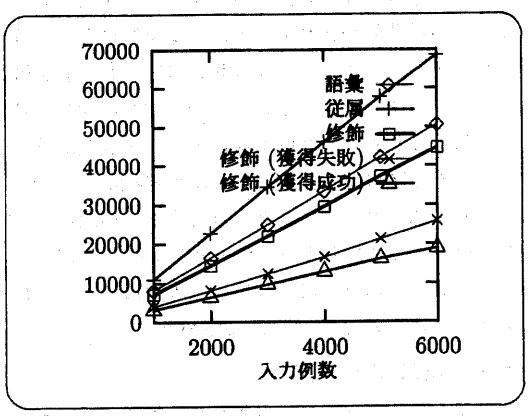

図 6: 規則の生成数

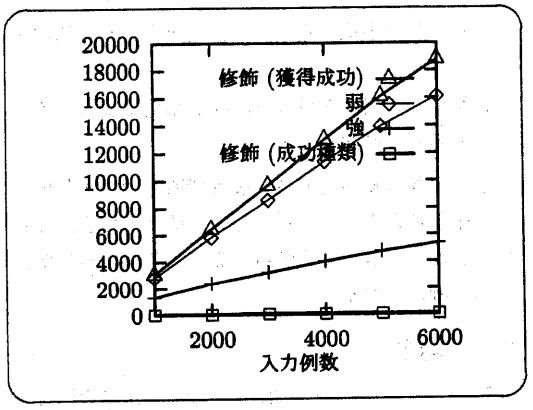

図 7: 修飾合成規則の生成数

6 おわりに

本稿では、EDR の意味情報と構文情報を解析する文法を設計した。さらに、それらを例として文法を獲得する手法を提案した。今後は、より多くの意味情報・構文情報に対応するよう文法を拡張し、洗練フェーズを設計・実現する必要がある。

参考文献

[1] 宇津呂 武仁, 宮田 高志, 松本 裕治 : 最大エントロピー法による下位範疇化の確率モデル学習および統語的曖昧性解消による評価, 情報処理学会自然言語処理研究会, NL-119-11, 1997.

[2] THEERAMUNKONG Thanaruk, OKUMURA Manabu : Learnig a Grammar from a Bractketed Corpus 情報処理学会自然言語処理研究会, NL-116-13, 1996.

[3] 森 信介, 長尾 眞 : タグ付きコーパスからの統語規則の獲得, 情報処理学会論文誌, Vol. 37. No. 9, 1996.

[4] 荒木 健治, 宮永 喜一, 栃内 香次 : 自然言語-意味構造対応ルールの獲得と適用, 電子情報通信学会技術報告論文集, NLC95-42, 1995.

[5] 大倉 清司 : タグ付コーパスからの統語・意味的知識の自動獲得, 電子情報通信学会技術報告論文集, NLC95-75, 1995.

[6] 田辺 利文, 冨浦 洋一, 日高 達 : 係り受け制約を含む文脈自由文法, 電子情報通信学会技術報告論文集, NLC95-21, 1995.

[7] EDR 電子化辞書 1.5 版 使用説明書, 日本電子化辞書研究所, 1996.

[8] 小島 丈幸, 山口 昌也, 乾 伸雄, 小谷 善行, 西村 恕彦 : LFG 機能スキーマの帰納的獲得, 情報処理学会自然言語処理研究会, NL-??, 1996.