1A1-02

統計的日本語文解析における格フレーム辞書の利用に関する考察

Introducing case frame constraints into statistical parsing

八木豊 Yutaka YAGI　白井清昭 Kiyoaki SHIRAI　田中穂積 Hozumi TANAKA　徳永健伸 Takenobu TOKUNAGA

東京工業大学 大学院情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Linguistic knowledge plays important roles in natural language processing. Constructing linguistic knowledge manually requires much human labor and cost to give its quality. On the other hand, it is possible to construct broad-coverage knowledge automatically from a large corpora. But its quality is not always satisfactory. Combining these different kinds of knowledge is one of important issues in recent natural language processing research. This paper proposes a method to introduce case frame constraints into statistical parser. The case frame constraints are compiled by lexcographers, while the statistical parser learns parameters from annotated corpora. Experiments showed that introducing case frame constraints to filter out semantically ill-formed parses slightly improved parsing precision. The paper also discusses a relation between the improvement and strictness of selectional restriction of a case frame.

1. はじめに

自然言語を解析する過程においては，正しい解析結果の他にも多くの誤った解析結果が得られるため，複数の解析結果の中から正しい解を選択しなければならない．これは一般に曖昧性解消と呼ばれ，自然言語解析における重要な課題の一つである．曖昧性解消を行う手法は以下の 2 つに大別される．

1. 人手で作成された知識に基づく手法
2. コーパスなどの言語資源から得られる統計情報に基づく手法

人手で作成された知識に基づく手法は，利用する知識の信頼性が高いという利点を持っているが，すべての言語現象に対する知識を網羅的に作成することは困難である．一方，統計情報に基づく手法は，利用する言語資源の規模にともなって多くの言語現象をカバーできるという利点を持っているが，得られる知識の信頼性に問題が残る．このような 2 つの手法の特徴は，相反するものであるというよりも，むしろお互いの欠点を補い合う関係にあると考えられる．したがって，曖昧性解消ではどちらか一方のみを利用するのではなく，2 つの手法を組み合わせて利用することが望ましい [Klavans 96]．しかし，2 つの手法をどのように組み合わせるかについては多くの問題がある．

本論文では，統計情報に基づいた形態素構文解析に人手で作成された格フレーム辞書を利用し，その曖昧性解消に対する効果を実験的に考察することを目的とする．格フレーム辞書とは，動詞が取り得る格，およびその格要素としてどのような名詞が現れ得るかという選択制約を記述したものである．格フレーム辞書を曖昧性解消に利用した研究は過去に多く行われているが [河原 00, 松尾 98, 吉田 98]，統計情報に基づいた形態素構文解析と組み合わせて利用した研究は少ない．

形態素構文解析には MSLR パーザを用いる [白井 00]. MSLR パーザは文脈自由文法をもとに形態素解析と構文解析を同時に行うツールである．また，確率一般化 LR モデル (以下，PGLR モデル) と呼ばれる確率モデルを用いて，複数の解析結果に対して統計情報に基づいた優先順位を与えることができる [橋本 99]．しかし，MSLR パーザでは構文的な制約しか与えないので，文法的には正しくても意味的に不適格な解析木が解析結果に含まれる．このような解析結果から，格フレーム辞書の選択制約に違反する解析木を取り除くことによって曖昧性解消の精度向上を図る．

連絡先: 八木豊, 東京工業大学情報理工学研究科, 〒 152-8552 東京都目黒区大岡山 2-12-1, 電話番号: 03-5734-3016, Fax 番号: 03-5734-2915, e-mail: yutaka@cl.cs.titech.ac.jp

2. MSLR パーザによる統計的形態素構文解析

本節では，MSLR パーザで使用する文法，および PGLR モデルの学習アルゴリズムについて述べる．

2.1 文法の作成

文法は，人手で作成された文法と，その一部の文法規則を自動的に細分化することによって拡張した文法の 2 つを使用する．以下，前者をオリジナル文法，後者を拡張文法と呼ぶ．但し，オリジナル文法，拡張文法ともに文節文法である．文節文法とは，日本語における文節を 1 つの非終端記号としてまとめて，文節間の係り受け関係を表した解析木を生成する文法である．文節は，オリジナル文法では“ 文節-(B)-(C) ”，拡張文法では“ 文節- (A)-(B)-(C)-(D)-(E) ”という非終端記号を頂点とした部分木で表される．(A)〜(E) の意味と取り得る値を以下に示す．

(A) 文節の分類を表す．“ 後置節 ”,“ 名詞句 ”,“ コピュラ ”,“ 動詞句 ”,“ 形容詞句 ”,“ 形容動詞句 ”,“ 他 ”のいずれかの値を取る．

(B) 文節の受け属性，すなわち，その文節が他の文節からどのような修飾を受けることが可能なのかを表す．“ 用 ”,“ 体 ”,“ 用体 ”,“ 無 ”のいずれかの値を取る．

(C) 文節の係り属性，すなわち，その文節が他の文節に対してどのような修飾を行うことが可能なのかを表す．“ 用 ”,“ 体 ”,“ 用体 ”,“ 無 ”のいずれかの値を取る．

(D) 助詞の有無および分類を表す．“ 格 ”,“ 係 ”,“ 他 ”,“ 無 ”のいずれかの値を取る．

(E) 句読点の有無を表す．“ 句 ”,“ 読 ”,“ 無 ”のいずれかの値を取る．

表 1: 文節ラベルの例

文節	オリジナル文法	拡張文法
パーザは、	文節-体-用	文節-後置節-体-用-係-読
文法を	文節-体-用	文節-後置節-体-用-格-無
使用する。	文節-用-無	文節-動詞句-用-無-無-句

表 2: 文法の概要

	オリジナル文法	拡張文法
規則数	1000	3713
非終端記号数	228	394
文節を表す非終端記号数	12	68
終端記号数	469	469

各値は文節を構成している形態素や品詞から決定される．例えば,「パーザは、 文法を 使用する。」という例文における各文節は表 1 に示す非終端記号で表される．表 1 に示すように，拡張文法では非終端記号が細分化されているので，より多くの情報を PGLR モデルに反映させることができる．しかし，文法規則数の増加にともなって学習に必要な訓練データの量も増加する．それぞれの文法の規則数を表 2 に示す.

2.2 PGLR モデルの学習アルゴリズム

PGLR モデルの学習では，使用する文法から作成された LR 表を用いて訓練データの文を解析し，その際に使用された LR 表の各アクションの使用回数を数え上げる．そして，それらを正規化して各アクションが実行される確率を推定する．学習された PGLR モデルを用いて生成された解析木は，生成時に使用した各アクションに振られた確率の積によってその生成確率が与えられる.

学習に使用する訓練データは，使用する文法に従って完全な構文木が付与された文が理想的である．しかし，このような文を大量に用意することは困難であるため，本論文では既存の構文木付きコーパスと単純なヒューリスティックスを用いて PGLR モデルの学習を行う.

学習では，まず，構文木付きコーパスから文節区切りと文節間の係り受け関係を与えた文を取り出し訓練データとする．構文木付きコーパスには，単語区切りや品詞，文節などのラベルも付与されているが，使用する文法と体系が異なるのでそのまま利用することはできない．利用するためには，何らかの形で文法とコーパスの間の整合性をとる必要がある．したがって，本論文では単語区切りや品詞，文節などのラベルの情報を利用しない.

次に，取り出した訓練データを解析する．このとき，与えられた文節区切りや文節間の係り受け関係によっては受理できない文もあるが，受理できた文からは複数の解析木が得られる．そこで，各解析木に対して形態素数最小法と文節数最小法の 2 つのヒューリスティックスによるスコアを与えて，最もスコアの高い解析木から学習を行う．スコア最大の解析木が複数ある場合は，そのすべてから学習を行い，得られた LR 表の各アクションの使用回数を学習の対象となった解析木の数で割って正規化を行う.

特に，初めは MSLR パーザ付属の未知語処理を使用せずに解析し，そこで受理できなかった文に対してのみ未知語処理を使用して解析する．初めから未知語処理を使用しないのは，未知語処理を行うことによって得られる解析木の数が増加し，未知語を含む間違った解析木から学習を行う可能性が高まるからである．一方，未知語処理を使用しない解析で受理できなかった文には，実際に未知語が含まれていると考えられる.

ここで示した学習アルゴリズムでは，文節区切りと文節間の係り受け関係だけを与えた文から完全な解析木を生成して学習を行っているため，文節の内部構造については正しく学習されていない．文節の内部構造の学習については本論文では扱わない.

3. 格フレーム辞書を用いた曖昧性解消

前述したように，MSLR パーザでは構文的な制約しか与えないので，文法的には正しくても意味的に不適格な解析木が解析結果に含まれる．本論文では，解析結果からこのような解析木を取り除くために格フレーム辞書をフィルタとして利用する.

3.1 格フレーム辞書

格フレーム辞書には日本語語彙大系第 5 巻構文体系を使用する [池原 97]．構文体系では，6,118 個の用言に対して 14,819 個の格フレームが記述されており，格要素の選択制約は，同じく日本語語彙大系第 1 巻意味体系のシソーラスにおける普通名詞の意味属性 2,710 個によって記述されている．構文体系を使用するにあたって以下の点を考慮した.

- 用言には「齎す (もらたす)」のように，通常は漢字による表記をしないものも含まれているので，用言の読みも格フレーム辞書の見出しとして利用した.
- 用言の直前に記述されている格要素が意味属性ではなく名詞の事例で表されている場合には，名詞，助詞，用言の 3 つを組み合わせたものを定型表現であると考え，格フレーム辞書の見出しとして利用した.
- 構文体系では，用言ではなく格要素となっている名詞に係っている要素まで記述されていることがある．このような要素は，文型をより特定するものではあるが，用言の格要素ではないので格フレームからは取り除いた.

その結果，16,498 個の見出しに対して 28,632 個の格フレームが得られた．1 つの見出しが持つ格フレーム数の平均は 1.74 個，1 つの格フレームに含まれる格要素数の平均は 1.95 個である.

3.2 フィルタリング手法

フィルタリングでは，MSLR パーザによって出力された上位 N 個の解析木それぞれについて格フレーム辞書との対応付けを行い，選択制約に違反しない解析木のみを改めて出力する．各解析木と格フレーム辞書との対応付けの手順を以下に示す.

1. 解析木に含まれる格構造を取り出す．格構造とは，解析木に含まれる用言を中心とした係り受け関係のことである．具体的には，用言と，用言に係っている名詞句および用言が連体修飾している名詞の組のことをいう.
2. 取り出した格構造に含まれている格要素が格フレーム中のどの格と一致し得るのかを決定する．このとき，表層格が明示されているか否かによって処理が異なる.

(a) 表層格が明示されている場合，つまり取り出した助詞が格助詞である場合は，その表層格をそのまま格の候補とした．
(b) 表層格が明示されていない場合，つまり取り出した助詞が係助詞である場合，または被連体修飾句である場合は，ガ格，ヲ格，ニ格を格の候補とした．但し，被連体修飾句の場合には，連体修飾している用言と格関係にないことがあるので [阿辺川 01, Baldwin 98]，格の候補にそのような場合も加えた．
(c) 以下に示すような場合を特殊な例として扱った．ニ格の格要素として時間を表す名詞が現れている場合は，時格を格の候補に加えた．同様に，ニ格またはデ格の格要素として場所を表す名詞が現れている場合は，場所格を格の候補に加えた．助詞が「の」の場合は，ガ格を格の候補とした．

3. すべての格要素に対する格の候補が決まったら，二重格のがないかを調べる．二重格を含む解析木は意味的に間違っていると判断し，解析結果から取り除くことにした．実際には二重格が可能な場合もあるが，今回の実験では考慮しなかった．
4. さらに，格要素となっている名詞をシソーラスを用いて抽象化し，その意味属性が格フレーム辞書の選択制約に違反していないかを調べる．複合名詞を抽象化する際には，まず複合名詞全体の抽象化を試み，抽象化できない場合には，複合名詞の一番先頭の名詞から順に取り除いていき，その都度抽象化を試みる．抽象化した名詞が複数の意味属性を持っている場合は，そのうちの 1 つでも選択制約を満たしていれば良く，すべての意味属性が選択制約に違反するときのみ，その解析木を解析結果から取り除く．
名詞の抽象化ができない，あるいは格の候補と一致する助詞の記述が格フレームにないならば，対応付けは行わない．人手で作成された知識はすべての情報を網羅しているわけではないので，記述されている部分のみを信頼するためである．

このようにしてすべての対応付けがうまくいった解析木のみを優先順位を変えずに出力する．

4. 評価実験

4.1 PGLR モデルの学習

毎日新聞 1995 年 1 月 1 日から 1 月 17 日までの全記事 19,669 文と，1 月から 12 月までの社説記事 18,714 文，計 38,383 文に構文木を付与した京都大学テキストコーパス Version3.0(以下，京大コーパス) から [黒橋 00]，文節区切りと文節間の係り受け関係を与えた文を取り出した．このうち，用意した 2 つの文法で受理できた文数は，未知語処理を用いて受理できた文を含めても全体の約 60%，22,739 文に過ぎなかった．受理できなかった原因は，文法の不備や，文法と京大コーパスにおける文節区切りの相違によるものである．実際の学習には，受理できた文のうちヒューリスティックスによるスコア最大の解析木の数が 100 以下だった 18,987 文から行った．評価データには，正解の構文木を人手で付与した 242 文を用いた．評価データの概要を表 3 に示す．

4.2 実験結果と考察

表 4 は，学習した PGLR モデルを用いて評価データに対する文節の係り受け解析を行ったときの上位 1 位の解析木における文節の正解率と文の正解率である．文節の係り受け解析とは，入力として文節ごとに区切られた文が与えられたときに，文節間の係り受け関係を決定するものである．文節の係り受け解析による評価では文節の内部構造を考慮しないので，本論文で示した文節の内部構造を正しく学習できない学習アルゴリズムに対する評価としても適当といえる．文節の正解率は，文節の総数に対する係り先の正しい文節の数の割合，文の正解率は，文の総数に対するすべての文節の係り先が正しい文の数の割合である．文節の正解率は，文の末尾にある 2 文節は除いて計算してある．ベースラインは，すべての文節が次の文節に係るとする手法を表す．オリジナル文法では文節の正解率は約 70%，文の正解率は約 25%となり，ベースラインを大きく上回った．拡張文法では，さらに，文節の正解率で 10%，文の正解率で 15%ほど良くなっている．

表 3: 評価データの概要

文数	242 文
平均文節数	8.51(最大 15, 最小 4)
平均単語数	23.98(最大 45, 最小 13)
平均文字数	75.11(最大 120, 最小 38)

表 4: 文法による正解率の変化

	文節の正解率	文の正解率
ベースライン	56.95%	0.00%
オリジナル文法	70.11%	24.79%
拡張文法	81.40%	41.74%

表 5 は，拡張文法を用いた解析結果から，それぞれ，二重格を含む解析木を取り除いたとき，格フレーム辞書の選択制約に違反する解析木を取り除いたときの上位 1 位の解析木の文節の正解率と文の正解率である．文節の正解率，文の正解率ともに，解析木を取り除く制約を厳しくしていってもわずかずつしか良くならなかった．これは，格フレーム辞書の利用法が原因であると考えられる．格フレーム辞書をフィルタとして利用した場合，MSLR パーザによる解析結果の上位 1 位の解析木が格フレーム辞書の選択制約に違反していなければ，他にもっと良く格フレーム辞書との対応がとれている解析木があってもパーザの出力した上位 1 位の解析木を覆すことはないからである．実際に，上位 1 位の解析木に変更があった文数は 242 文中 29 文で，文節の係り受けが正しくなったのは，そのうちの 20 文と非常に少なかった．残りの 9 文は，フィルタリングをすることによって係り受けが誤りになってしまった．格フレーム辞書やシソーラスの記述が不足していたからである．

次に，格フレームの選択制約の厳しさと文節の係り受け精度との関係を調べた．格フレーム辞書によるフィルタリングをする際には，格要素となっている名詞の持つ意味属性が格フレーム辞書の選択制約に違反していないかを調べる．このとき，名詞が複数の意味属性を持っていると，そのうちのいくつかのみが選択制約を満たすことがある．すなわち，文中で使われている意味属性を絞り込むことができる．意味属性の絞り込みが行われるのは，用言が比較的厳しい選択制約を持っているときであり，格要素となっている名詞がその厳しい制約を満たすならば，その用言と係り受け関係にある可能性が高いと考え調査した．その結果，意味属性の絞り込みが行われる文節に限

表 5: フィルタリングによる正解率の変化

	文節の正解率	文の正解率
拡張文法	81.40%	41.74%
+二重格のチェック	82.10%	42.56%
+格フレーム辞書	82.35%	42.98%

ると，文節の正解率は 96.49%ですべての文節を対象としたときよりもかなり高い数値になっていることがわかった．今後，格フレーム辞書を利用して解析結果に優先順位を与えるときには，このように意味属性の絞り込みがどの程度行われているかによって重みを付けることが考えられる．

最後に，拡張文法における学習曲線を図 1 に示す．訓練データを 0 文から 1000 文刻みで増やしながら学習を行い，文節の正解率と文の正解率を調べた．図 1 を見る限りでは，文節の正解率，文の正解率ともに飽和傾向にあり，訓練データは拡張文法に対しても十分足りているといえる．したがって，本論文における文節の正解率，文の正解率は，先行研究と比較するとかなり低い数値になっているが，それが訓練データの不足から来るものではないことがわかる．反対に，訓練データの量を 1000 文としたあたりから正解率がすでに飽和傾向にあるので，文法を一般化し過ぎて統計情報がうまく反映されていないと考えられる．

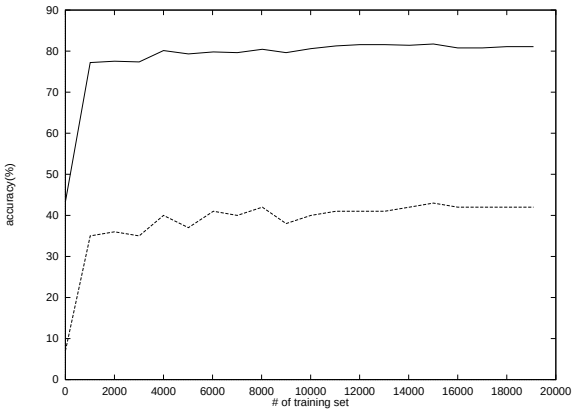

図 1: 拡張文法の学習曲線

5. おわりに

本論文では，MSLR パーザによる解析結果に対して格フレーム辞書をフィルタとして利用する手法を提案した．文節の係り受け解析実験では，わずかながら正解率は良くなったが，人手による知識と統計情報をどのように組み合わせるかについての問題は残されたままである．今後は，格フレーム辞書の情報と PGLR モデルにおける統計情報を同時的に用いる枠組みについて検討していきたい．現在その一環として，筆者らは格フレーム辞書の情報を取り入れた新たな文法を作成し，調整途中である．

参考文献

[阿辺川 01] 阿辺川武：統計情報を利用した日本語連体修飾節の解析, 言語処理学会第 7 回年次大会発表論文集, pp. 269–272 (2001).

[Baldwin 98] Baldwin, T. J.: The analysis of Japanese relative clauses, 修士論文, 東京工業大学 大学院情報理工学研究科 (1998).

[橋本 99] 橋本泰一：一般化 LR 法による構造付きコーパスからの統語的知識の自動獲得とその精密化, 修士論文, 東京工業大学 大学院情報理工学研究科 (1999).

[池原 97] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林：日本語語彙大系 — 全 5 巻 —, 岩波書店 (1997).

[河原 00] 河原, 鍛冶, 黒橋：大規模コーパスからの格フレーム辞書構築とそれを用いた格解析, 言語処理学会第 6 回年次大会発表論文集, pp. 24–27 (2000).

[Klavans 96] Klavans, J. L. and Resnik, P.: *The Balancing Act*, The MIT Press (1996).

[松尾 98] 松尾, 白井：格フレーム解析を統合した日本語係り受け解析, 言語処理学会第 4 回年次大会発表論文集, pp. 89–92 (1998).

[黒橋 00] 黒橋, 居蔵, 坂口：コーパス作成の作業基準, 京都大学, 第 1.8 版 (2000).

[白井 00] 白井, 植木, 橋本, 徳永, 田中：自然言語処理のための MSLR パーザ・ツールキット, 自然言語処理, Vol. 7, No. 5, pp. 93–112 (2000).

[吉田 98] 吉田, 峯, 雨宮：既存の電子化辞書から獲得した格フレームによる構文的曖昧さ解消, 言語処理学会第 4 回年次大会発表論文集, pp. 77–80 (1998).