単語の共起データを用いた構文的な統計情報の学習に関する研究

白井 清昭　徳永 健伸　田中 穂積
東京工業大学 大学院情報理工学研究科

1 はじめに

構文的な統計情報を用いた構文解析に関する研究は近年盛んに行われている [2, 11]．構文的な統計情報とは，ここでは文の構文的な構造や文節間の係り受け関係などについて，どのようなパタンが現われやすいのかといった傾向を示す統計的な情報を指す．一般に，構文的な統計情報は，その訓練データとして，構文木や文節間の係り受け関係などの構文的な情報が付与されたコーパスが必要となる．日本語のテキストコーパスについても，EDR コーパス [7] や京大コーパス [6] など，構文的な情報が付与されたコーパスの整備が進んだことが，統計的構文解析に関する研究が行われるようになった要因のひとつとして挙げられる．

しかしながら，これらのコーパスを構文的な統計情報の学習に直接用いることができない場合も多い．なぜなら，構文的な統計情報を利用した構文解析システムの基盤となる文法や品詞体系が，コーパスに付加された構文情報の基盤となる文法や品詞体系と必ずしも一致しないためである．例えば，句構造文法に基づく構文解析システムがあり，その解の優先順位付けのために構文的な統計情報を学習する場合には，文節間の係り受け関係が付与されたコーパスや，異なる句構造文法や品詞体系に基づく構文木が付与されたコーパスからは直接学習することはできない．統計情報を用いて文節の係り受け解析を行うシステムについても，京大コーパスのような文節の係り受け関係が付与されたコーパスを学習に利用することが考えられるが，その場合，文節切りの基準がシステムと学習用コーパスとで一致していなければならない．文節の単位としては“1 個の自立語と 0 個以上の付属語を 1 つの文節とする”という基準が一般的ではあるが，例えば「三十日夜､」という文字列について，これを 1 つの文節とするかそれとも「三十日」と「夜､」の 2 つの文節に分けるかといった複合名詞の切り方などで，文節切りの基準が一致しない場合もある．同様に，品詞体系の違いも構文的な統計情報の学習においてしばしば問題となる．品詞体系の違いとは，具体的には単語分割の基準の違いや，品詞をどれだけ細かく分類するかといった品詞の粒度の違いなどがある．これに対し，異なる 2 つの品詞体系の品詞間の対応を取る試みもいくつか行われている [4, 10]．

このような背景から，構文的な統計情報を学習する際に，構文的な知識が付与されていないコーパスを利用する，いわゆる教師なし学習を行うことを考える必要がある．しかしながら，教師なし学習によって獲得される知識は一般に品質が悪い．正解事例に相当する知識を全く使わないのではなく，構文的な統計情報の学習の手がかりとなるような知識は積極的に利用するべきである．そのような知識のひとつとして，単語の共起関係や単語の出現頻度といった語彙的な統計情報が挙げられる．語彙的な統計情報は構文的曖昧性解消の精度向上に有効であると考えられ，曖昧性解消のためのモデルにこのような語彙的な統計情報を組み込む試みは数多く行われている [1, 2, 11]．また，語彙的な統計情報は，品詞付きコーパスなど，構文的な知識が付与されていないコーパスからでも十分学習することが可能である．本研究では，品詞付きコーパスから事前に学習された単語の共起関係や出現頻度を構文的な統計情報の学習に利用する手法を提案する．

2 統計的構文解析のためのモデル

本節では，本研究で用いる，形態素・構文解析の曖昧性を解消するための確率モデルについて概説する．

図 1 は形態素・構文解析結果の例である．図 1 において，R は文の構文的な構造を表わす構文木, L は品詞の集合，W は単語の集合を表わす．また，構文木 R は品詞 L を葉とすることを仮定する．本研究では，図 1 のような解析結果 R, L, W に対し，その生成確率を式 (1) に示した確率モデル [9] によって推定し，この値の大きい解析結果を選択することにより曖昧性を解消する．

$$P(R) \cdot P(W|L) \cdot D(W|R) \tag{1}$$

以下，式 (1) を構成する 3 つのサブモデルについて説明する．

第 1 項 $P(R)$ は構文モデルである．このモデルは構文木 R の生成確率を与える確率モデルであり，構文的な統計情報が反映される．構文モデルとしては，R の生成確率を与えるモデルであれば，例えば確率文脈自

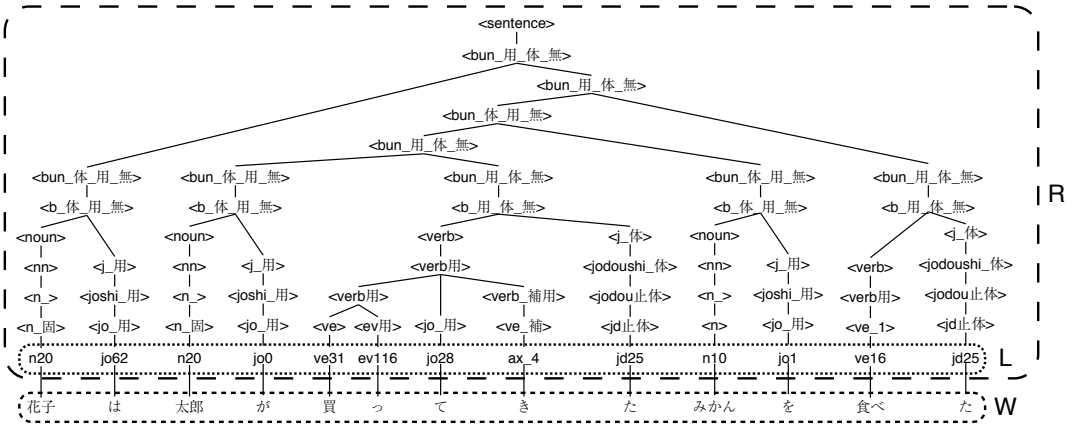

図 1: 解析結果の例

由文法や確率一般化 LR モデル [5] (以下，PGLR モデル) など，任意のモデルを用いることができる．

第 2 項 $P(W|L)$ は単語導出モデルである．このモデルは単語集合 W の生成確率であり，式 (2) に示すように，各品詞 l_i から単語 w_i が文脈自由に導出される確率の積として推定する．

$$P(W|L) \simeq \prod_{w_i \in W} P(w_i|l_i) \tag{2}$$

第 3 項 $D(W|R)$ は従属係数モデルである．このモデルは単語の共起関係を反映した統計量であり，式 (3) に示すように，各単語 w_i と単語生成文脈 c_{ij} の従属係数 $D(w_i|l_i[c_{ij}])$ の積として推定される．

$$D(W|R) = \prod_{c_{ij} \in C} D(w_i|l_i[c_{ij}]) \tag{3}$$

ここで，単語生成文脈 c_{ij} とは，単語 w_i の導出に深く関わりがあると思われる入力文中の他の単語を指す．例えば，単語 w_i が動詞の格要素となる名詞 (図 1 の例では「みかん」) であるとき，単語生成文脈 c_{ij} を表層格と主辞となる動詞 (図 1 の例では「を:食べ (る)」) とすれば，動詞と格要素の共起関係がモデルに反映される．従属係数 $D(w_i|l_i[c_{ij}])$ は，c_{ij} と共起するという条件の下で品詞 l_i から単語 w_i が導出される確率 $P(w_i|l_i[c_{ij}])$ と，そのような制約なしに品詞 l_i から単語 w_i が導出される確率 $P(w_i|l_i)$ の比と定義する (式 (4))[1]．

$$D(w_i|l_i[c_{ij}]) \stackrel{def}{=} \frac{P(w_i|l_i[c_{ij}])}{P(w_i|l_i)} \tag{4}$$

3 モデルの学習

式 (1) の 3 つのサブモデルのうち，構文的な統計情報を反映しているのは構文モデル $P(R)$ である．本研究では，この構文モデルを，教師データとなる構文的な情報が付与されたコーパスを用いずに学習する．一方，$P(W|L)$ は単語の出現頻度を，$D(W|L)$ は単語の共起関係を反映したサブモデルである．これらの語彙的な統計情報は構文モデルとは独立に学習し，構文モデルの学習時に利用する．以下，これらのサブモデルの学習手法について述べる．

3.1 構文モデル $P(R)$ の学習

構文モデルを学習するための訓練コーパスとしては，何も情報が付加されていない平文コーパスを利用する．以下にその手順を示す．

1. 初期構文モデル $P_0(R)$ を学習する．訓練コーパスの文に対して形態素・構文解析を行い，単語数最小法などのヒューリスティクスを用いて，最も正しいと思われる構文木を 1 つだけ取り出す．得られた構文木を教師データとして $P_0(R)$ を学習する．
2. 訓練コーパスの例文の形態素・構文解析を行い，それぞれの例文について，確率モデル $P_i(R)$·$P(W|L)$·$D(W|R)$ の値の大きい上位 T_n 個の構文木を取り出す．これらの構文木を教師データとして，新たな構文モデル $P_{i+1}(R)$ を学習する．
3. i の値を 1 増やす．
4. 訓練コーパス中の各例文について，確率モデルの値の最も大きい構文木が変化しなくなるまで，2.,3. の操作を繰り返す．

2. の操作で用いる確率モデルのうち，単語導出モデル $P(W|L)$ と従属係数モデル $D(W|R)$ は，品詞付きコーパスから事前に学習する．これは，構文モデルとは独立に学習される単語の共起関係や単語の出現頻度といった語彙的な統計情報を曖昧性解消に利用することによ

[1] 式 (4) の対数を取ると，w_i と c_{ij} の相互情報量になる．すなわち，$D(w_i|l_i[c_{ij}])$ は w_i と c_{ij} の共起の強さを表わす統計量である．

り，構文モデル $P_i(R)$ の反復学習の精度を向上させることを目的としている．

3.2 単語導出モデル $P(W|L)$ の学習

単語導出モデル $P(W|L)$ の学習には品詞付きコーパスを用いる．統計情報を学習する対象となる解析システムの品詞体系 L^s と訓練コーパスの品詞体系 L^t が一致している場合には，式 (2) に示した単語導出モデルの各項は式 (5) のように最尤推定すればよい．

$$P(w|l) = \frac{O(w,l)}{\sum_{w \in l} O(w,l)} \tag{5}$$

ここで，$O(w,l)$ は品詞 l を持つ単語 w の出現頻度である．しかしながら，2 つの品詞体系 L^s と L^t が必ずしも一致しているとは限らない．L^s と L^t が異なる場合，ある品詞体系 L^c があり，L^s の品詞から L^c の品詞への対応，ならびに L^t の品詞から L^c の品詞への対応は一意に決まると仮定する．このような条件を満たす品詞体系 L^c は必ず存在するわけではないが，L^c として「名詞」や「助詞」などの粗い品詞体系を仮定すれば，2 つの品詞体系 L^s, L^t の差異を吸収する品詞体系 L^c を決めることは比較的容易であると考えられる．さらに，解析システムの品詞体系 L^s の品詞 l_s から単語 w が導出される確率を式 (6) により推定する．

$$P(w|l_s) = \frac{O(w,l_c) + \gamma}{\sum_{w \in l_s} (O(w,l_c) + \gamma)} \tag{6}$$

ここで，$O(w,l_c)$ は品詞 $l_c (\in L^c)$ を持つ単語 w の訓練コーパスにおける出現頻度であり，γ は確率モデルの平滑化のために全ての単語の出現頻度に足される定数である．

この方法は，品詞の粒度の違いは考慮されているが，単語分割の基準の違いについては考慮されていないという問題点も残されている．

3.3 従属係数モデル $D(W|R)$

従属係数モデルの各項 (式 (4)) の推定について説明する．式 (4) の分母は文脈自由な単語の導出確率であるので，式 (6) によって推定する．また，式 (4) の分子も，式 (6) と同様に，式 (7) のように推定する．

$$P(w|l_s[c]) = \frac{O(w,l_c,c) + \gamma}{\sum_{w \in l_s} (O(w,l_c,c) + \gamma)} \tag{7}$$

$O(w,l_c,c)$ は，訓練コーパスにおいて，品詞 l_c を持つ単語 w が単語生成文脈 c と共起する頻度を表わす．

例えば，表層格 p の格要素である名詞 w と，その主辞となる動詞 v の共起関係をモデルに反映させる場合には，単語生成文脈 c を「p:v」とし，$D(w|N_s[p{:}v])$ という従属係数を学習すればよい．この従属係数を学習するためには，共起頻度 $O(w,l_c,p{:}v)$ を獲得しなければならない．この共起頻度は，例えば「名詞 + 格助詞 + 動詞」という品詞が連続して並んでいればこれを共起事例として数え上げるといったヒューリスティクスを使えば，品詞付きコーパスからでも学習することが可能である．他にも，格と動詞の共起関係や，修飾・被修飾関係にある名詞間の共起関係なども，同様にモデルに反映させることができる．

4 予備実験

本節では，3 節で提案した手法を評価するために行った予備実験について述べる．

形態素・構文解析を行うシステムとして，現在我々が公開している MSLR パーザ[2]を使用した．また，解析用の文法と辞書として，MSLR パーザに付属のもの (以下，それぞれ MSLR 文法，MSLR 辞書と呼ぶ) を使用した．MSLR 文法は，図 1 のような文節のまとまりと文節間の係り受け関係を表わす構文木を生成する．MSLR 文法の規則数は 1,498，非終端記号数は 220，前終端記号数 (品詞数) は 556 である．MSLR 辞書は，EDR 日本語単語辞書 [7] を元に作成された辞書であり，241,189 個の単語が登録されている．

まず，品詞タグが付与されている RWC コーパス [3] から単語の出現頻度 $O(w,l_c)$ を求め，単語導出モデル $P(W|L)$ を推定した．RWC コーパスの品詞体系と MSLR 文法の品詞体系が異なるため，両者の差異を吸収する品詞体系 L^c として,「名詞」「助詞」などの粗い品詞体系を設定した．また，活用語については，RWC コーパスが語幹と語尾をひとつの単語としていたのに対し，MSLR 文法の品詞体系では語幹と語尾は異なる単語としていたため，RWC コーパス中の活用語を全て語幹と語尾に自動的に分割し，単語の出現頻度を求めた．

式 (6) の推定には表記と品詞体系 L^c による品詞が一致した単語のみを使用する．RWC コーパス中の単語数を表 1 の「総単語数」に，式 (6) の推定に実際に利用した単語数を「使用単語数」に，その割合を「使用率」にそれぞれ示す．異り数での使用率が約 15%であるのに対し，のべ数での使用率が約 86%であることから，RWC コーパスの品詞体系と MSLR 文法の品詞体系の品詞間の対応は，高頻度語については比較的よく取れていることが推察できる．

次に，従属係数モデル $D(W|L)$ を推定した．今回の

[2] http://tanaka-www.cs.titech.ac.jp/pub/mslr

表 1: 単語導出モデルの推定に使用したデータ

	総単語数	使用単語数	使用率
のべ数	142,555,149	123,425,604	86.58 %
異り数	440,216	67,342	15.30 %

実験では，モデルに反映させる単語の共起関係として，格要素となる名詞とその係り先の動詞との間の共起関係のみを考慮した．RWC コーパスと EDR コーパスから，名詞 n が動詞 v の表層格 (助詞)p の格要素となる事例 (n, p, v) をのべ 7,864,359 組抽出し，これをもとに式 (7) を推定した．

最後に，構文モデル $P(R)$ を学習した．構文モデル $P(R)$ を与える確率モデルとしては PGLR モデルを使用した．構文モデルを学習するための訓練テキストとして，EDR コーパスの例文 1 万文を用意し，3.1 節で示したように，これらの文の解析と構文モデルの学習を 5 回繰り返した．初期構文モデル $P_0(R)$ の学習は，単語数最小法と文節数最小法のヒューリスティクスなどにより，訓練テキストに対する構文解析結果を一意に決め，それらをもとに学習した．

25 個の評価用例文について形態素・構文解析を行い，学習された 3 つのサブモデル $P_i(R), P(W|L), D(W|R)$ による曖昧性解消の精度を調べた．まず，3.1 項の 2. の操作において，上位 T_n 個[3]の構文木を選ぶモデルとして，構文モデル $P_i(R)$ だけを用いて学習を繰り返した場合，正解となる構文木を付与できた文の数は，$i = 0$ のときに 3 個，$i = 5$ のときに 6 個であった．次に，単語導出モデル・従属係数モデルを組み合わせたモデル $P_i(R)P(W|L)D(W|R)$ を用いたところ，正解となる構文木を付与できた文の数は変わらなかった．

現段階では，我々は MSLR 文法による構文木が付与された十分な量の評価用データを持っていないため，今回の実験では 25 個の例文についてしか評価を行わなかった．そのため，語彙的な統計情報を利用したことによる改善は認められなかった．今後は十分な量の評価用データを作成し，単語の共起関係，単語の出現頻度といった語彙的な統計情報を構文的な統計情報の学習に利用することの有効性を検証する予定である．

5 おわりに

本研究では，品詞付きコーパスから事前に学習された単語の共起関係や単語の出現頻度といった語彙的な統計情報を利用して，構文的な統計情報の教師なし学習を行う手法について述べた．ここで提案した手法では，構文的な統計情報を学習する際に，語彙的な統計情報は利用しているが，構文的な知識については何も使わないということを前提としていた．しかしながら，たとえ部分的であっても構文的な知識を与えることができれば，学習の精度も向上すると考えられる [8]．句構造文法の場合には，文法体系の異なる構文木が付与されたコーパスの情報を利用することは難しいが，今回の実験のような文節の係り受け解析を前提にした文法を用いる場合には，例えば文節の係り受け解析の情報が付与されたコーパスの利用を考えるべきである．文節切りの基準が異なるために，コーパスにある全ての情報をそのまま利用することはできないにしても，文節境界が一致する例文に付与された情報は利用できるだろう．今後は，異なる文法や品詞体系を基準とする構文的な情報が付与された既存のコーパスを構文的な統計情報の学習に利用することも検討していきたい．

[3] 今回の実験では $T_n = 10$ とした．

参考文献

[1] E. Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the National Conference on Artificial Intelligence*, 1997.

[2] 藤尾正和, 松本裕治. 語の共起確率に基づく係り受け解析とその評価. 情報処理学会論文誌, Vol. 40, No. 12, pp. 4201–4212, 1999.

[3] K. Hasida, H. Isahara, T. Tokunaga, M. Hashimoto, S. Ogino, W. Kashino, J. Toyoura, and H. Takahashi. The RWC text databases. In *Proceedings of the the first International Conference on Language Resources and Evaluation*, pp. 457–462, 1998.

[4] 乾健太郎, 脇川浩和. 品詞タグつきコーパスにおける品詞体系の変換. 情報処理学会情報処理学会自然言語処理研究会 (NL-132-12), Vol. 99, No. 62, pp. 87–94, 1999.

[5] K. Inui, V. Sornlertlamvanich, H. Tanaka, and T. Tokunaga. Probabilistic GLR parsing: A new formalization and its impact on parsing performance. 自然言語処理, Vol. 5, No. 3, pp. 33–52, 1998.

[6] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 人工知能学会全国大会論文集, pp. 58–61, 1997.

[7] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. Technical Report TR–045, 1995.

[8] 白井清昭, 徳永健伸, 田中穂積. PGLR 法を用いた構文木付きコーパスの自動獲得. 情報処理学会第 57 回全国大会講演論文集, pp. 213–214, 1998.

[9] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 統計的構文解析における構文的統計情報と語彙的統計情報の統合について. 自然言語処理, Vol. 5, No. 3, pp. 85–106, 1998.

[10] M. Ueki, T. Tokunaga, and H. Tanaka. Sharing syntactic structures. In *Proceedings of the Machine Translation Summit VII*, pp. 543–546, 1999.

[11] 内元清貴, 関根聡, 井佐原均. 最大エントロピー法に基づくモデルを用いた日本語係り受け解析. 情報処理学会論文誌, Vol. 40, No. 9, pp. 3397–3407, 1999.