

係り受け整合度を計算する いくつかの統計的手法の比較

江原 暉将

NHK 放送技術研究所

〒 157-8510 東京都世田谷区砧 1-10-11

tel:03-5494-2308 fax:03-5494-2309 email:eharate@strl.nhk.or.jp

あらまし 係り受け整合度を計算する統計的手法のうち、a) 最大エントロピー法、b) 簡易最大エントロピー法、c) 決定木による方法、d) WINNOW 法の4種類の手法について精度を比較する。放送ニュース文を対象にして、「が」格文節の係り先決定に適用した結果、受け文節の種類と文節間距離のみを用いたベースラインシステムの精度は86.2%であったのに比較して、a) 89.0% b) 88.3% c) 86.5% d) 85.9%の精度であった。誤事例について分析したところ、「が」格文節が数詞文節に係る場合に誤ることが多いことなどが明らかになった。

キーワード 係り受け解析、統計的手法、最大エントロピー法、決定木、WINNOW 法

Comparison of several statistical parsing methods for Japanese bunsetsu dependency analysis

Terumasa EHARA

NHK Science and Technical Research Laboratories

1-10-11, Kinuta, Setagaya, Tokyo, 157-8510, Japan

tel:81-3-5494-2308 fax:81-3-5494-2309 email:eharate@strl.nhk.or.jp

Abstract Several statistical methods for Japanese bunsetsu dependency analysis are compared. These methods are a) maximum entropy method, b) simplified maximum entropy method, c) decision tree method and d) WINNOW algorithm method. The base line method uses two features: 1) distance between dependant and head, 2) type of head. The bunsetsu dependency accuracy for these methods are a) 89.0%, b) 88.3%, c) 86.5% and d) 85.9% compared with the base line accuracy 86.2%. These result is obtained by open test for the GA-case-bunsetsu (subject phrase) dependency in TV news articles.

key words Japanese bunsetsu dependency analysis, statistical method, maximum entropy, decision tree, WINNOW algorithm

1 はじめに

係り受け解析は形態素解析と並んで、計算機による日本語解析の基本技術の一つである。初期の係り受け解析システムは、文法情報のみを用いていたが、精度の向上を図るには、意味情報や文脈情報などさまざまな情報を利用しなければならない。近年、確率モデルを用いた係り受け解析や構文解析が提案され、多様な情報を統一的なモデルの下で取り扱う手法が用いられるようになった [3][4][7][9][10][5][11][2]。本文では、このような統計的手法のうち、最大エントロピー法、簡易最大エントロピー法、決定木による方法、WINNOWER 法の 4 種類の手法について比較検討する。実際には、係り文節が「が」格文節である場合に、係り受け整合度を上記各手法を用いて計算し精度を比較する。それと共に、誤事例の分析を行う。

2 文節間係り受け関係

本文で考察する係り受け関係は、文節間のものがあり、以下の条件を仮定する。

- ・ 逆依存はない：文末の文節を除いて、自分より後方の文節に係る¹。
- ・ 交差依存はない：係り受け関係は交差しない²。
- ・ 多重依存はない：文末の文節を除いて、係り先が 1 つ存在し、2 つ以上は存在しない³。

これらの 3 条件を満足する限り、文の係り受け解析結果は、必ず存在する。従って、係り受け解析器の被覆率（再現率ともいう）は 100% である。このことは、他の構文解析と比較した場合、係り受け解析が持つ大きな特長であると言える。規則が被覆し

¹この条件は、倒置を含まない書き言葉では満足される。

²この条件は、

「太郎を 空港に 送って 行った」
のような文の場合、満足されないとする見解が多いが、この場合は、「行く」を補助動詞とみなして、

「太郎を 空港に 送って行った」
とすれば、「送って行った」で 1 文節となるため交差依存はなくなる。「行く」を補助動詞とみなしうる根拠として、

「太郎を 空港に 車で 送って 行った」

は自然であるが、

「太郎を 空港に 送って 車で 行った」

は不自然に感じられることがある。

³この条件は、並列の場合、問題になると言われる。

「A の B と C」

「A が B して C した」

というような表現の場合、A の係り先が B のみの場合、C のみの場合、B と C の双方に係る場合が論理的に考えられる。しかし、上記のような表現で、A が C のみに係る場合は存在しないのではないだろうか。もし、それが事実とすると、A の係り先を C とすることで、B と C の双方に係る場合を表現することが可能であり、この条件の例外とはならない。

ていない文は、そもそも、解析できないからである。そこで、係り受け解析における問題は、いかに適合率の高い解析器を実現させるかということになる。

3 係り受け整合度とその計算方法

係り受け解析の適合率を向上させるために、係り受け整合度を用いる。これは、係り元文節が係り先文節にどれだけ係りやすいかを表現する量である。係り受け整合度が計算できれば、文全体としての最適な整合度を持つ係り受け解析結果が動的計画法を用いて、効率的に求められる [8]。

係り受け整合度に影響する情報としては、品詞などの文法情報はもとより、係り元文節および係り先文節を構成する語と語の意味的な関係や、1 文の範囲を越えたテキスト内文脈情報、テキストには記述されていないテキスト外文脈情報などさまざまなものが考えられる。これらの情報を統一的に扱うために、まず、係り受け関係を素性の列で表現する。本文で用いた素性と素性値の全体については、[2] を参照されたい。

素性の列としての係り受け関係が定義できたので、次に、学習データ（標本）から係り受け関係の正例と負例を作成する。つまり、標本として与えられた各文に対して、形態素解析と文節解析をほどこし、文節列を得る。この文節列を構成する各文節（文末の文節を除く）の係り先を人手によって認定する。このようにして認定された係り元文節と係り先文節の組が係り受け関係の正例となる。また、係り元文節から、文末にいたる文節のうち、真の係り先でない文節と係り元文節の組が負例となる。これら、正例と負例はいずれも素性の列として表現されている。係り受け整合度は、上記の正例と負例に基づいて統計的に計算される。本文では、a) 最大エントロピー法、b) 簡易最大エントロピー法、c) 決定木による方法、d) WINNOWER 法の 4 種類の手法を考察する。

3.1 最大エントロピー法

最大エントロピー法による係り受け整合度計算法は、[2] に述べられているが、概略を説明する。まず、素性の数を次元とする多次元離散分布を用いて確率構造を定義する。つまり、各素性に対して素性値の数だけの値をとる多次元空間を考える。この空間上で正例の確率分布 p と負例の確率分布 q を標本から推定する。そして、係り元文節 w_i が係り先文節 w_j

に係る係り受け整合度 $S(w_i, w_j)$ を

$$S(w_i, w_j) = \frac{p(w_i, w_j)}{q(w_i, w_j)}$$

として計算する。この離散分布の独立な母数の数は、素性の数を K とし、素性 k に対する素性値の数を I_k とすると、

$$I_1 \times I_2 \times \cdots \times I_K - 1$$

となり、実現可能な標本の大きさと比較して多すぎる。一方、もし、全ての素性が独立であると仮定すると、独立な母数の数は、

$$I_1 + I_2 + \cdots + I_K - K$$

となり、著しく減少する。しかし、独立とみなせない素性も多く、独立性の仮定は、適合率の低下をもたらす可能性がある。

そこで、次の2つの方法で整合度計算の精度および統計的推定の精度の両者を確保することを試みる。
(a) 素性の全体を用いず、選択して利用する。
(b) 最大エントロピー法すなわち、対数線形分布を用いる。
(a) の素性選択については、[2] を参照されたい。
(b) について説明する [1]。ここでは、1次および2次の混合した対数線形分布を用いる。このような分布は、

$$p(i_1, i_2, \dots, i_K) = \prod_{k \in D_1} h_k^{(1)}(i_k) \times \prod_{(l, m) \in D_2} h_{l, m}^{(2)}(i_l, i_m) \quad (1)$$

で表現される。ここで、 D_1 は1次のモデルとして選択された因子の全体であり、 D_2 は2次のモデルとして選択された因子の全体である。1次の因子は素性そのものであり、2次の因子は素性の2個組となる。関数 $h_k^{(1)}(i_k)$ は、因子 k に対する1次の周辺分布の i_k での値、 $p_k^{(1)}(i_k)$ そのものであるから、標本から推定するのは容易である。しかし、 $h_{l, m}^{(2)}(i_l, i_m)$ に対しては、因子間で素性が重複して用いられているので、因子 (l, m) に対する2次の周辺分布の (i_l, i_m) での値、 $p_{l, m}^{(2)}(i_l, i_m)$ とは一般に異なるものとなる。この $h^{(2)}$ の計算には、比例反復法を適用する (付録A参照)。

3.2 簡易最大エントロピー法

最大エントロピー法で利用する比例反復法は、計算量が多い。そこで、比例反復法を1回で終了させることを試みる。本方法を簡易最大エントロピ

ー法と呼ぶ。この方法では、付録Aに示すように、 $p(i_1, i_2, \dots, i_K)$ が次のように陽に計算できる。

$$p(i_1, i_2, \dots, i_K) = \prod_{k \in D_1 \cup D_2} p_k^{(1)}(i_k) \times \prod_{(l, m) \in D_2} \frac{p_{l, m}^{(2)}(i_l, i_m)}{p_l^{(1)}(i_l) p_m^{(1)}(i_m)} \quad (2)$$

ここで、 D_2 は D_2 に含まれる素性の全体である。本手法は、最大エントロピー法と比較して、確率値の計算法が異なるのみである。

3.3 決定木による方法

決定木は、正例と負例をできるだけ「きれいに」分割する素性を順次選択して行くものであり、ここでは、[12] で用いたプログラムを若干変更して利用した。変更点は、学習データで決定木を学習した後、入力された試験データに対して、係り受け整合度を計算できるようにした点である。決定木の各節点 v には、学習データのうち、その節点に分類された正例の数 $n_{v, p}$ と負例の数 $n_{v, q}$ が記録されている。これを利用して、係り元文節 w_i が係り先文節 w_j に係る係り受け整合度 $S(w_i, w_j)$ を次のようにして計算する。上記の係り受け関係を表現する素性の列が分類される節点を $v_{(w_i, w_j)}$ とする。このとき、

$$S(w_i, w_j) = \frac{n_{v_{(w_i, w_j)}, p}}{n_{v_{(w_i, w_j)}, q}}$$

である。ただし、 $n_{v, q}$ が0となると、整合度が計算できなくなる。そこで、 $n_{v, p}$ や $n_{v, q}$ が0になる時は、0.5 にフロアリングを行なった。

3.4 WINNOW 法

WINNOW 法では、各因子に対してウェイトが割り当てられており、このウェイトを用いて、整合度を計算するものである。ここでは、正例に対するウェイトと負例に対するウェイトの2種のウェイトを利用する Ballanced WINNOW 法を用いた [13]。用いた因子は、0 と 1 の2値を取るものであり、1次の因子と2次の因子に分けられる。1次の因子 f_{k, i_k} ($k = 1, \dots, K; i_k = 1, \dots, I_k$) は、

$$f_{k, i_k} = \begin{cases} 1 & \text{素性 } k \text{ の素性値が } i_k \text{ である} \\ 0 & \text{その他} \end{cases}$$

で定義され、2次の因子 f_{l, i_l, m, i_m} ($l = 1, \dots, K - 1; i_l = 1, \dots, I_l; m = l + 1, \dots, K; i_m = 1, \dots, I_m$)

は、

$$f_{l,i_l,m,i_m} = \begin{cases} 1 & \text{素性 } l \text{ の素性値が } i_l \text{ であり} \\ & \text{素性 } m \text{ の素性値が } i_m \text{ である} \\ 0 & \text{その他} \end{cases}$$

で定義される。各因子に対するウエイトの値は学習データから判定誤りに基づいて学習される。学習法の詳細は、[13]を参照されたい。因子 f の学習された正のウエイトを w_f^+ とし、負のウエイトを w_f^- とすると、係り元文節 w_i が係り先文節 w_j に係る係り受け整合度 $S(w_i, w_j)$ は、

$$\begin{aligned} S(w_i, w_j) &= \sum_{k=1}^K \sum_{i_k=1}^{I_k} f_{k,i_k} (w_{f_{k,i_k}}^+ - w_{f_{k,i_k}}^-) + \\ &\quad \sum_{l=1}^{K-1} \sum_{i_l=1}^{I_l} \sum_{m=l+1}^K \sum_{i_m=1}^{I_m} \\ &\quad f_{l,i_l,m,i_m} (w_{f_{l,i_l,m,i_m}}^+ - w_{f_{l,i_l,m,i_m}}^-) \end{aligned}$$

で与えられる。ここで、 f の値は、係り受け関係 (w_i, w_j) を表現する素性列から決まる因子の値である。

4 実験

NHK のニュース原稿を対象にして、実験した。今回の実験では、係り文節種別が、「が」格文節のもののみを用いた。実験データは 490 文（「が」格文節数 573 文節）からなり、これを (A)250 文（「が」格文節数 256 文節）と (B)240 文（「が」格文節数 317 文節）に分け、交叉確認法によって精度を評価した。このコーパスは、平均 17.8 文節／文と比較的長文のものである。

評価方法は次の基準によった。試験データの各係り元文節と正例、負例に含まれる係り先文節の組に対して、システムが計算した整合度が正例において最大となる場合を正解、そうでない場合を不正解として正解率を計算した。実験結果の正解率を表 1 に示す。各手法は以下の通りである。

- ・手法 a：最大エントロピー法
- ・手法 b：簡易最大エントロピー法
- ・手法 c：決定木による方法
- ・手法 d：WINNOW 法
- ・手法 e：「受け文節の種別」と「係元と係先の間の文節数」の 2 個の素性のみを（1 次）の因子として用いる手法（ベースライン）
- ・手法 f：全素性を 1 次の因子として用いる手法

表 1: 実験結果

手法	正解率 (%)		
	学習データ (A)	学習データ (B)	全体
手法 a	87.9	90.0	89.0
手法 b	87.1	89.2	88.3
手法 c	83.6	88.6	86.5
手法 d	82.4	88.6	85.9
手法 e	84.4	87.7	86.2
手法 f	86.7	87.7	87.3

表 2: 学習に要した時間

手法	時間 (秒)
手法 a	221
手法 b	113
手法 c	42
手法 d	851
手法 e	12
手法 f	13

手法 e は、初期の係り受け解析の模擬であり、手法 f は、多数の素性は用いるもの、素性間の独立性を仮定したものである。

また、データ (A) について、各手法が学習に要した時間を表 2 に示す。手法によって、使用している因子の数も違うし繰り返しの終了条件も異なるので、表 2 はあくまで参考データである。

表 1 の結果、最大エントロピー法（手法 a）が最も精度が高く、簡易最大エントロピー法（手法 b）が次に精度が高かった。さらに、全素性を 1 次の因子として用いる手法（手法 f）が続いており、それ以外の手法は同じくらいの精度であった。ただし、すべての手法について、精度の差はそれほど小さくなく、統計的に有意な差とはいえない（1%で 5.73 データ）。

5 誤事例の分析

最も精度が高かった最大エントロピー法の誤事例は、573 データ中、63 例である。これらを「係元と係先の間の文節数」で分類すると表 3 のようになる。表 3 から、間の文節数が 4 以上になると正解率が極端に悪くなることがわかるが、誤事例の絶対数としては、むしろ、間の文節数が 0 や 1 の場合の方が多い。誤りの原因を分析すると以下のような現象が見られた。

数詞文節に係らない。間の文節数が 0 の誤事例を見ると、例文 1、2 のように受文節が数詞文節になっている場合が 14 事例あり、「が」格で数詞文節に係

表 3: 係元と係先の間の文節数によるデータの分類

間の文節数	総 数	正解数	誤り数	正解率
0	361	339	22	93.9
1	90	76	14	84.4
2	53	49	4	92.5
3	34	31	3	91.2
4	15	10	5	66.7
5 以上	20	5	15	25.0
合 計	573	510	63	89.0

ることを表す情報が不足していることが分かる。(以下では、係り元文節を” (”と”) ”で、正しい係り先文節を” < ”と” > ”で、システムが計算した整合度が最も大きい文節を” [”と”] ”で囲んで表す)

例文 1 東京都の平成三年度の普通会計決算は、収入にあたる歳入総額が六兆八千百七億円で、前年度を三点六パーセント上回り、また支出にあたる(歳出総額が) < 六兆七千二百六十億円と > 前年度を三点二パーセント [上回りました。]

例文 2 ルーキーとして初受賞になる日本ハムの片岡選手は、十九試合に出場して、(打率が) < 三割八分五厘、 > ホームラン六本、打点二十を [マークし、] 打率と打点の二つの部門でトップの成績を挙げました。

補助動詞に係らない。 例文 3 のように補助動詞が係り先の場合、意味的整合度が低くなり、誤る場合がある。2 文節間ではなく、3 文節間での整合度を用いる必要がある。

例文 3 (日本が) 苦手と< していた > クラスを制した点でも値千金の金メダルと [言えます。] ⁴

例文 4 この (大会が) [素晴らしい] 試合と世界の相互理解の機会と< なることを > 期待します。

品詞の不整合 動詞「ある」の否定形が形容詞「ない」となるため、整合度が低くなったり、断定の助動詞の「で」を格助詞の「で」と誤るために整合度が低くなる。

例文 5 … 警視庁ではフライデーが最近、掲載した記事をめぐる (トラブルが) 原因ではないかと > [みて] 捜査しています。⁵

例文 6 (お金が) < 問題では > [ないのだ。]

⁴ 「日本が」の係り先が「制した」である解釈も可能であるが、ここでは、作業者の解釈を正解とした。

⁵ 「トラブルが」の係り先が「原因では」である解釈も可能であるが、ここでは、作業者の解釈を正解とした。

並列構造の認定誤り 並列構造の認定誤りとして例文 7 がある。このような場合は、構文全体を見ることで、解決できる可能性がある [6]。

例文 7 中型で強い台風十号は、発達しながら南大東島の南の海上を北西に進んでおり、今夜には (南大東島が、) [また] あすの朝には沖縄本島付近が暴風域に< 巻き込まれる > 恐れがあります。

連体形に係る誤り 例文 8 のように受け文節が連体形であり誤る場合がある。連体形文節の係り先の文節も加えた 3 文節間の関係を調べる必要がある。

例文 8 … (ロシアが) これまでのイデオロギーに [偏った] 歴史教育を見直す作業を< 進めているなかでの > 判断だとしています。

6 おわりに

統計的に係り受け整合度を計算するいくつかの手法について、テレビニュース文を対象にして整合度の計算実験を行ない、精度を比較した。また、誤事例の分析を通して、不足している情報について考察した。今後、素性の数を増やすなどの改良を行ない、さらに精度を向上させたい。

参考文献

- [1] 江原暉将, 金淵埤. 確率モデルによるゼロ主語の補充. 自然言語処理, Vol. 3, No. 4, pp. 67-86, Oct. 1996.
- [2] 江原暉将. 最大エントロピー法を用いた日本語係り受け整合度の計算. 言語処理学会第 4 回年次大会発表論文集, pp. 382-385, 1998.
- [3] 藤尾正和, 松本裕治. 統計的手法を用いた係り受け解析. 自然言語処理研究会資料 NL-117-12, 情報処理学会, 1997.
- [4] 春野雅彦, 白井諭, 大山芳史. 決定木の混合を利用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. 217-220, 1998.
- [5] 柏岡秀紀, 河田康裕, 金城由美子, Andrew Finch, Ezra Black. 確率付き決定木を用いた日本語構文解析. 言語処理学会第 4 回年次大会発表論文集, pp. 213-216, 1998.
- [6] 黒橋禎夫, 長尾真. 並列構造の検出に基づく長い日本語文の構文解析. 自然言語処理, Vol. 1, No. 1, pp. 35-57, Oct. 1994.
- [7] 森信介, 長尾真. 係り受けを用いた確率的言語モデル. 自然言語処理研究会資料 NL-122-6, 情報処理学会, 1997.

- [8] 尾関和彦. 最適文節列を選択するための多段決定アルゴリズム. 音声研究会資料 SP-86-32, 電子情報通信学会, 1986.
- [9] Satoshi Sekine, Kiyotaka Uchimoto, and Hitoshi Isahara. Corpus-based Japanese parser using context information. In *Proc. of NLPRS'97*, pp. 161–166, 1997.
- [10] Satoshi Sekine. Corpus-based parsing and sublanguage studies. Ph.d. thesis, New York University, 1998.
- [11] 白井清昭. 統計情報を利用した統合的自然言語解析. 東京工業大学テクニカルレポート TR98-0004, 東京工業大学, 1998.
- [12] 田中英輝. 日英機械翻訳システムにおける基本動詞の曖昧性解消に関する研究. 学位論文, 九州大学, 1995.
- [13] Takefumi Yamazaki and Ido Dagan. Mistake-driven learning with thesaurus for text categorization. In *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, pp. 369–374, 1997.

A 比例反復法による 2 次対数線形分布の母数の推定

2 次対数線形分布の母数である

$h_{l,m}^{(2)}(i_l, i_m)$ ($(l, m) \in D_2$; $i_l = 1, \dots, I_l$; $i_m = 1, \dots, I_m$) は以下の比例反復法によって求めることができる。2 次対数線形分布の制約式は

$$\begin{aligned}
 p_{l,m}^{(2)}(i_l, i_m) &= \sum_{i_1=1}^{I'_1} \cdots \sum_{i_{l-1}=1}^{I'_{l-1}} \sum_{i_{l+1}=1}^{I'_{l+1}} \cdots \sum_{i_{m-1}=1}^{I'_{m-1}} \sum_{i_{m+1}=1}^{I'_{m+1}} \cdots \sum_{i_K=1}^{I'_K} \\
 &\quad \prod_{k' \in D_1} h_{k'}^{(1)}(i_{k'}) \times \prod_{(l', m') \in D_2} h_{l', m'}^{(2)}(i_{l'}, i_{m'})
 \end{aligned} \tag{3}$$

ここで

$$I'_k = \begin{cases} I_k & k \in D_1 \cup D'_2 \\ 1 & \text{その他} \end{cases}$$

である。そこで、 $h_{l,m}^{(2)}(i_l, i_m)$ の r 回目の推定値を $h_{l,m}^{(2),(r)}(i_l, i_m)$ とし、式 (3) の右辺に $h_{l,m}^{(2),(r)}(i_l, i_m)$ を代入して求めた 2 次周辺分布を $p_{l,m}^{(2),(r)}(i_l, i_m)$ とするとき、 $r+1$ 回目の推定値を

$$h_{l,m}^{(2),(r+1)}(i_l, i_m) = \frac{p_{l,m}^{(2)}(i_l, i_m)}{p_{l,m}^{(2),(r)}(i_l, i_m)} h_{l,m}^{(2),(r)}(i_l, i_m)$$

のようにして求める。この反復を、推定値が収束するまで繰り返し行う。そのときの初期値は、2 次の

全因子 D_2 を構成する各素性 k の (1 次) の周辺分布の推定値を $p_k^{(1)}$ とし、 D_2 での当該素性の出現回数を n_k とするとき、

$$h_{l,m}^{(2),(0)}(i_l, i_m) = (p_l^{(1)}(i_l))^{\frac{1}{n_l}} (p_m^{(1)}(i_m))^{\frac{1}{n_m}}$$

で与えた。これは、もし、 D_2 に含まれる各素性が独立であれば、推定値が初期値に一致することから選んだ値である。初期値の定義から 0 回目の 2 次周辺分布の推定値は、

$$p_{l,m}^{(2),(0)}(i_l, i_m) = p_l^{(1)}(i_l) p_m^{(1)}(i_m)$$

となる。そこで、1 回目の $h^{(2)}$ の推定値は、

$$\begin{aligned}
 h_{l,m}^{(2),(1)}(i_l, i_m) &= \frac{p_{l,m}^{(2)}(i_l, i_m)}{p_l^{(1)}(i_l) p_m^{(1)}(i_m)} \times (p_l^{(1)}(i_l))^{\frac{1}{n_l}} (p_m^{(1)}(i_m))^{\frac{1}{n_m}}
 \end{aligned} \tag{4}$$

である。そこで、式 (4) の $h^{(2),(1)}$ と $p^{(1)}$ を本文の式 (1) に代入することにより、簡易最大エントロピー法の確率計算式 (2) を得る。