最大エントロピー法を用いた日本語文節間係り受け整合度の計算

江原 暉将
NHK 放送技術研究所
eharate@strl.nhk.or.jp

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 [2][3][4][6]。本文で述べる方法も確率モデルによる係り受け解析であり、最大エントロピー法を利用している。これによって、対数線形分布が得られ、確率構造が明確に定まるという特長がある。

2 文節間係り受け関係

本文で考察する係り受け関係は、文節間のものであり、以下の条件を仮定する。

- 逆依存はしない：文末の文節を除いて、自分より後方の文節に係る[1]。
- 交差依存はしない：係り受け関係は交差しない[2]。
- 多重依存はしない：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、係り受け解析が他の構文解析と比較して大きな特長であると言える。規則が被覆していない文は、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 [5]。

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

[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しなくなる。「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ということがある。

[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。

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を付録1に示す。

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。

次に、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。

4 対数線形分布の利用と素性の選択

前節で述べた離散分布の母数の数は、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、

$$I_1 \times I_2 \times \cdots \times I_K$$

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、母数の数は、

$$I_1 + I_2 + \cdots + I_K$$

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、本文では次の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) については、後述するとして、まず、(b) について説明する [1]。本文では、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、

$$\log p(i_1, i_2, \cdots, i_K) = \sum_{k \in D_1} \log h_k^{(1)}(i_k) + \sum_{(l,m) \in D_2} \log h_{l,m}^{(2)}(i_l, i_m)$$

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。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、

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

で与える。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ということから選んだ値である。

次に、D_1と D_2に含まれる因子を全素性の中からどのように選択するかを述べる。まず、各素性 kに対して、正例と負例から1次の周辺分布 $p_k^{(1)}(i_k)$ および $q_k^{(1)}(i_k)$ を求める。それらの分布間の距離を、L_1距離つまり、

$$d_1(p_k^{(1)}, q_k^{(1)}) = \sum_{i_k=1}^{I_k} |p_k^{(1)}(i_k) - q_k^{(1)}(i_k)|$$

で表す。この距離は、最大2である。そして、この距離があるしきい値を越える素性を1次の因子として D_1に含める。次に、全ての素性の組合せ (l, m) についても、正例と負例から2次の周辺分布 $p_{l,m}^{(2)}(i_l, i_m)$ および $q_{l,m}^{(2)}(i_l, i_m)$ を求め、1次と同様に距離

$$d_2(p_{l,m}^{(2)}, q_{l,m}^{(2)})$$

$$= \sum_{i_l=1}^{I_l} \sum_{i_m=1}^{I_m} |p_{l,m}^{(2)}(i_l, i_m) - q_{l,m}^{(2)}(i_l, i_m)|$$

を求める。そして、

$$d_2(p_{l,m}^{(2)}, q_{l,m}^{(2)}) - \max(d_1(p_l^{(1)}, q_l^{(1)}), d_1(p_m^{(1)}, q_m^{(1)}))$$

があるしきい値を越える素性の組を2次の因子として、D_2に含める。ただし、D_2に含ませた素性については、D_1から除外する。このようにして、素性の選択を行い、1次と2次の因子群を設定する。

実際には、係り文節の種別によって、係り受けに関する性質が異なるため、これらの因子の設定は、係り文節の種別に依存して行った。また、選択された因子のうち、意味の薄い因子については、経験的に除外したものがある。

5 実験

NHKのニュース原稿を対象にして、実験した。学習データは250文（文末文節を除いて4008文節）、試験データは学習データと異なる240文（文末文節を除いて4032文節）である。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。係り文節種別ごとに利用した因子群を付録2に示す。因子設定のための素性選択で用いたL_1距離でのしきい値は、0.2を用いた。また、2次の因子の最大数は計算時間の関係で8とした。

評価方法は次の基準によった。つまり、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の手法3の欄に示す。表1には、比較のために手法1と手法2も示す。手法1は、「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）因子として用いた結果であり、手法2は、本手法で用いた因子に含まれる全素性を1次の因子として用いた結果である。つまり、手法1は、初期の係り受け解析の模擬であり、手法2は、多数の素性は用いるもの、素性間の独立性を仮定したときの結果である。表1から、素性間の非独立性を考慮した本手法（手法3）が、最も精度が高いことがわかる。さらに、手法2は、手法1とほぼ同程度の精度であり、単に、素性の数を増やすだけでは、精度の向上が望めないことがわかる。

表1: 実験結果

文節種別	試験データ数	正解率 (%)		
		手法1	手法2	手法3
副詞	71	76.1	76.1	74.6
接続詞	47	29.8	25.5	31.9
格助詞	1461	77.6	80.2	80.9
副詞的名詞	222	58.6	54.1	55.0
係助詞	349	53.9	45.8	52.4
並列名詞	90	64.4	63.3	62.2
連体名詞	821	88.6	89.6	89.4
述語連体形	372	82.0	81.2	80.6
述語連用形	410	63.7	64.4	64.4
引用	94	88.3	94.7	94.7
連体詞	90	88.9	88.9	88.9
全文節	4027	75.3	75.6	76.4

6 おわりに

1次と2次の対数線形分布が混合している分布を用いて、係り受け整合度を計算する手法を提案し、実験結果を示した。今後、語事例の分析を通して一層の精度向上を図るとともに、他の手法との比較も行いたい。

参考文献

[1] 江原暉将, 金淵培. 確率モデルによるゼロ主語の補完. 自然言語処理, Vol. 3, No. 4, pp. 67–86, Oct. 1996.

[2] 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.

[3] 春野雅彦, 白井諭, 大山芳史. 決定木を用いた日本語係り受け解析. 自然言語処理シンポジウム資料, 1997.

[4] 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.

[5] 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.

[6] Satoshi Sekine, Kiyotaka Uchimoto, and Hitoshi Isahara. Corpus-based Japanese parser using context information. In *Proc. of NLPRS'97*, pp. 161–166, 1997.

付録1　係り受け整合度計算のための素性

係り元文節に関する素性

1 係文節の種別：副詞, 接続詞, 格助詞, 係助詞, 並列名詞, 連体名詞, 述語終止形, 述語連体形, 述語連用形, 引用, 連体詞, その他
2 係文節の位置：文頭, 文頭から2文節目, 文頭から3文節目以降
3 係文節の,「が」格パターンの有無：無, 有
4 係文節の,「を」格パターンの有無：無, 有
5 係文節の,「に」格パターンの有無：無, 有
6 係文節の,「で」格パターンの有無：無, 有
7 係文節の,「から」格パターンの有無：無, 有
8 係文節の,「と」格パターンの有無：無, 有
9 係文節の,「へ」格パターンの有無：無, 有
10 名詞化付属語の有無：無, こと, の, 他
11 格助詞の有無：無, から, が, で, と, に, へ, を, 他
12 連体格助詞の有無：無、の, 他
13 並立助詞の有無：無, と, や, 他
14 接続助詞の有無：無, が, て, で, に, 他
15 係助詞の有無：無, は, 他
16 引用助詞の有無：無, と, 他
17 読点の有無：無, 有
18 係文節の名詞種別：名詞無, 人名, 数詞, 組織名, 代名詞, 地名, 転成名詞, 関係名詞, 形容名詞, 普通名詞

係り先文節に関する素性

19 受文節の種別：名詞, 動詞（連体形以外）, 動詞連体形, 形容詞・形容動詞, その他
20 受文節の位置：文末, 文末以外
21 受文節の,「が」格パターンの有無：無, 有
22 受文節の,「を」格パターンの有無：無, 有
23 受文節の,「に」格パターンの有無：無, 有
24 受文節の,「で」格パターンの有無：無, 有
25 受文節の,「から」格パターンの有無：無, 有
26 受文節の,「と」格パターンの有無：無, 有
27 受文節の,「へ」格パターンの有無：無, 有
28 使役表現の有無：無, 有
29 受身表現の有無：無, 有
30 可能表現の有無：無, 有
31 希望表現の有無：無, 有
32 てある表現の有無：無, 有
33 打消表現の有無：無, 有
34 受文節の名詞種別：名詞無, 人名, 数詞, 組織名, 代名詞, 地名, 転成名詞, 関係名詞, 形容名詞, 普通名詞
35 受文節の述語種別：述語無, 形式述語, 格助詞相当述語, 特殊述語, 普通述語

係り受け関係に関する素性

36 係元と係先の引用範囲の別：引用範囲外, 引用範囲の最後, 引用範囲で最後以外
37 係元と係先の接頭語的記号の数の一致の別：一致で最後, 一致で途中, 不一致で最後, 不一致で途中

文節間の距離に関する素性

38 係元と係先の間の文節数：0,1,2,3,4,5 以上
39 係元と係先の間の係元と同一の文節数：0,1,2 以上
40 係元と係先の間の係先と同一の文節数：0,1,2 以上
41 係元と係先の間の読点数：0,1,2 以上
42 係元と係先の間の係助詞「は」の数：0,1,2 以上
43 係元と係先の間の格助詞「が」の数：0,1,2 以上

係り文節と受け文節の意味的整合性に関する素性[4]

44 係元から係先への「が」格整合度：0,1,2,3,4,5
45 係元から係先への「を」格整合度：0,1,2,3,4,5
46 係元から係先への「に」格整合度：0,1,2,3,4,5
47 係元から係先への「で」格整合度：0,1,2,3,4,5
48 係元から係先への「の」格整合度：0,1,2,3,4,5
49 係先から係元への「が」格整合度：0,1,2,3,4,5
50 係先から係元への「を」格整合度：0,1,2,3,4,5
51 係先から係元への「に」格整合度：0,1,2,3,4,5
52 係先から係元への「で」格整合度：0,1,2,3,4,5
53 係先から係元への「の」格整合度：0,1,2,3,4,5
54 係元と係先の意味的類似度：0,1,2,3

付録2　利用した因子群

副詞:35,38 19,38 35,40 34,36 34,38; 20 39 41 42 43
接続詞:2,20; 19 34 35 36 40
格助詞:34,39 39,41 19,38 35,39; 21 22 23 24 25 27 36 40 42 43 44 45 46
副詞的名詞:20,38; 19 18 40 41 42
係助詞:20,23 20,38 22,23 20,22; 19 21 24 25 26 27 34 35 36 39 40 41 42 43 44 46 47
並列名詞:34,39 34,41; 19 36 38 40 42 43 54
連体名詞:19,43 34,43 19,41 19,40; 36 38 39 42 48
述語連体形:19,40 34,40 19,39 39,43; 36 38 41 42 49
述語連用形:; 19 34 35 36 20 38 39 40 41 43 54
引用:19,40 35,40 22,40 34,42 42,43 21,42 35,42 34,40; 20 23 24 25 26 36 38 41 54
連体詞:42,43 19,42 34,42 19,43 34,43 19,41 19,40 41,43; 36 38 39

[4]意味的整合度は、「田中康仁（著）語と語の関係データベース」の頻度統計から相互情報量と $t-score$ を用いて計算した。その際、「類語国語辞典」の分類番号を用いてデータの過疎性に対処した。意味的類似度は、「類語国語辞典」の分類番号を用いて計算した。いずれも、値の大きい方が、整合性や類似性が高い。データの利用を許諾していただいた兵庫大学 田中康仁教授および角川書店に感謝する。