係り受け周辺確率に基づく文節間距離

海野 裕也　　坪井 祐太
日本アイ・ビー・エム株式会社 東京基礎研究所
{yunno, yutat}@jp.ibm.com

1　はじめに

構文解析結果を検索や情報抽出に応用する試みは多いが [9, 10]，解析誤りによる抽出漏れを引き起こす可能性がある．実験性能で優れても，現実の応用では少数の抽出漏れが深刻となることは多い．そのため，文字列一致や単語共起（Bag of words）を採用せざるを得ず，精度よく単語関係を抽出できなくなることもある．

本研究の目的は，検出漏れを防ぎつつ，意味的に重要な 2 単語の共起を見つけることにある．単語一致をベースに考え，見つかった単語を含む文節対を，解析誤りに頑健な距離関数で並べ替えること目指す．そこで 1-best の解析結果ではなく，全解析候補に対する確率分布そのものを使う．係り受け木上で文節間の道の距離を設計し，その期待値を距離関数とした．正確な計算は計算量が大き過ぎるので，周辺確率の積で確率を近似することで，多項式時間で計算する方法を示す．

新聞記事を使った 2 つの検証実験を行った．直接の文節係り受け関係の検索では，1-best の構文解析結果から探すよりも周辺確率で順位付けしたほうが精度の高い検索ができた．一方，間接的な係り受けの検索実験では，上記距離関数を使ったときと 1-best の構文解析を使ったときとで大きな違いは得られなかった．しかし，個別に見ると，特に 1-best で解析誤りがあったときには提案した距離関数の方が頑健なスコアを与えていることが観察された．

2　係り受け条件付き対数線形モデル

$\mathbf{x}$ を長さ n の入力文節列，$\mathbf{y}$ を出力の係り受け構造とする．$\mathbf{y}$ は長さ n のベクトル $\mathbf{y} = (y_1, \ldots, y_n)$ で，各要素 y_i は i 番目の文節が係る先の文節インデックスを示す．たとえば，3 番目の文節が 5 番目の文節に係っていることを $y_3 = 5$ とあらわす．$\mathcal{Y}(\mathbf{x})$ は文節列 $\mathbf{x}$ に対して，ありうる全ての係り構造候補の集合とする．

文節列と係り構造に対する確率変数をそれぞれ X, Y とし，係り受け条件付対数線形モデルを

$$\mathrm{P}(Y = \mathbf{y}|X = \mathbf{x}) = \frac{1}{Z}\exp(\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}, \mathbf{y})) \qquad (1)$$

と定義する．ただし，$Z = \sum_{\tilde{\mathbf{y}} \in \mathcal{Y}(\mathbf{x})} \exp(\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}, \tilde{\mathbf{y}}))$ は分配関数である．また，$\mathbf{\Phi}(\mathbf{x}, \mathbf{y})$ は素性関数で，$\mathbf{x}$ と $\mathbf{y}$ のペアをベクトルに変換する関数である．以降，確率変数は略記する．$\mathbf{\Phi}$ は個々の係り関係ごとに分解できるとする．すなわち，入力文節列 $\mathbf{x}$，文節インデックス i，その係り先 y_i を引数とする関数 $\mathbf{f}$ を使って，

$$\mathbf{\Phi}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}, i, y_i)$$

と分解できるものとする．つまり，複数の文節の係り先に依存した素性を使うことはできない．

パラメタ $\mathbf{w}$ は学習データ $T = \{(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})\}_{k=1}^{m}$ に対する対数尤度 $\mathcal{L}(\mathbf{w}) = \sum_{(\mathbf{x},\mathbf{y}) \in T} \mathrm{P}(\mathbf{y}|\mathbf{x})$ を最大化するように選ばれる．$\mathcal{L}$ は $\mathbf{w}$ に対して凸なので，勾配法や擬似ニュートン法で大域最適解を求めることができる．この際，対数尤度 $\mathcal{L}$ の勾配を求める必要があるが，$\mathbf{y}$ の候補集合 $\mathcal{Y}(\mathbf{x})$ が文長に対して指数爆発するため，簡単には求まらない．しかし，素性森のパラメタ学習 [4] と同様，内側・外側アルゴリズムと同種の動的計画法により，文長に対して多項式時間で計算できる．係り受け解析の生成モデルに対する内側・外側アルゴリズムは Lee & Chi の研究 [2] で詳しく述べられている．また，この計算過程で，文節 i が文節 j に係る事象，すなわち $Y_i = j$ の周辺確率 $\mathrm{P}(Y_i = j|\mathbf{x})$ も，内側確率と外側確率の積で求めることができる．

解析では $\mathrm{P}(\mathbf{y}|\mathbf{x})$ を最大にする $\mathbf{y}$ を探索する．これは CKY アルゴリズムと同種の手法で効率的に実行でき，空間計算量が $O(n^2)$，時間計算量が $O(n^3)$ である．

3　周辺確率を用いた文節間距離関数の設計

3.1　係り受け木上での文節間距離

係り受け構造に対する 2 文節間の距離 l を，係り受け木上での道の重みつき距離として定義する．

係り受け構造 $\mathbf{y}$ と一対一対応する有向グラフ $G = (V, E)$ を以下のように構築する．まず，頂点集合 $V = \{v_1, \ldots, v_n\}$ の要素数は文節数 n に一致する．そして，$\mathbf{y}$ 中で i 番目の文節が j 番目の文節に係るときに限り，v_i から v_j への有向辺が存在する．つまり，辺集合を $E = \{(v_i \rightarrow v_{y_i}) | i \in [1, n]\}$ とする．ここで，交差・非交差に関わらず，$\mathbf{y}$ が有効な係り受け構造であるとき，G は木をなす．これを係り受け木と呼ぶ．

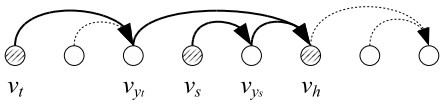

図 1: 係り受け木と頂点間の道の例

係り受け木 G 上で，任意の 2 頂点 v_t, v_s に対して辺の向きを無視すれば必ずただ 1 つの道が存在する．各頂点の出次数が 1 であることから，この道の中に頂点 v_h が存在し，辺の向きに沿って v_t から v_h へと，v_s から v_h への道に分けられる．実際の例を図 1 に示した．v_t, v_s 間の道に含まれる辺は実線で示している．

任意の有向辺 $(v_i \to v_j)$ に対して 2 つの重み関数 $e_1(i,j), e_2(i,j)$ を与える．そして，v_t から v_s までの距離 $l(t,s)$ を，v_t から v_h までの e_1 の総和と，v_s から v_h までの e_2 の総和として定義する．これは v_t から v_s までの道をたどるとき，辺の向きに沿うときは e_1 を，逆らうときは e_2 を足した総和である．

$$l(t,s) = \sum_{(i,j)\in \mathrm{path}(t,h)} e_1(i,j) + \sum_{(i,j)\in \mathrm{path}(s,h)} e_2(i,j)$$

$l(t,s)$ の計算を効率的に行うことを考える．ここで，後方係り受けに限定することで，簡単な再帰式で l を計算できることを示す．後方係り受けに限定すると，v_h は v_t, v_s のいずれよりも右側になくてはならない．つまり，$t \le h$ かつ $s \le h$ が成り立つ．仮に $t < s$ とすると，$t < s \le h$ より必ず $t \ne h$ である．したがって，辺 $(v_t \to v_{y_t})$ は必ず v_t, v_s 間の道に含まれる．$t > s$ の場合も同様である．以上をまとめると，

$$l(t,s) = \begin{cases} e_1(t,y_t) + l(y_t,s) & (t < s) \\ 0 & (t = s) \\ e_2(s,y_s) + l(t,y_s) & (t > s) \end{cases} \quad (2)$$

となることがわかる．この関数は，各文節を高々 1 回ずつたどるので，$O(n)$ で計算することができる．

3.2 距離関数の近似期待値

解析誤りに対して文節間距離 l を頑健にするため，決定的な解析結果を使わず，識別モデル $\mathrm{P}(\mathbf{y}|\mathbf{x})$ に対する期待値を計算する．正確に計算することは難しいので，周辺分布の積で近似した分布に対する期待値を計算する方法を示し，これを距離関数 L_e として定義する．

(1) 式で定義した分布 $\mathrm{P}(\mathbf{y}|\mathbf{x})$ 上での l の期待値 $\mathrm{E}[l(t,s)] = \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \mathrm{P}(\mathbf{y}|\mathbf{x}) l(t,s)$ を求めたい．$\mathcal{Y}(\mathbf{x})$ の要素数は文長に対して指数爆発するため，全列挙はできない．そこで，同時確率を周辺確率の積で近似した分布 $\mathrm{P}(\mathbf{y}|\mathbf{x}) \simeq \prod_{i=1}^{n} \mathrm{P}(y_i|\mathbf{x})$ での期待値を考え，$L_e(t,s)$ とおく．(2) 式を代入して，まず $t < s$ に関して考える．簡単のため $\mathrm{P}(y_i|\mathbf{x})$ を p_i と略記すると，

$$\begin{aligned} L_e(t,s) &= \sum_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \left(\prod_{i=1}^{n} p_i\right) l(t,s) \\ &= \sum_{y_1} p_1 \cdots \sum_{y_i} p_i \cdots \sum_{y_n} p_n \left\{e_1(t,y_t) + l(y_t,s)\right\} \\ &= \sum_{y_t} p_t e_1(t,y_t) + \sum_{y_1} p_1 \cdots \sum_{y_n} p_n l(y_t,s) \end{aligned}$$

となる．ただし，$\sum_{y_i}$ は $\sum_{y_i=i+1}^{n}$ のことである．$l(t,s)$ は $\min(t,s)$ より左の文節の構造に依存しない．また，必ず $t < y_t$ なので，$l(y_t,s)$ の計算に t とそれより左の構造は依存しない．そのため，

$$\begin{aligned} &\sum_{y_1} p_1 \cdots \sum_{y_n} p_n l(y_t,s) \\ &= \sum_{y_t} p_t \left\{ \sum_{y_{t+1}} p_{t+1} \cdots \sum_{y_n} p_n l(y_t,s) \right\} \\ &= \sum_{y_t} p_t L_e(y_t,s) \end{aligned}$$

としてよい．$t > s$ のときも同様に計算され，以下の再帰式が得られる．

$$L_e(t,s) = \begin{cases} \sum_{y_t} \mathrm{P}(y_t|\mathbf{x}) \left(e_1(t,y_t) + L_e(y_t,s)\right) & (t < s) \\ 0 & (t = s) \\ \sum_{y_s} \mathrm{P}(y_s|\mathbf{x}) \left(e_2(s,y_s) + L_e(t,y_s)\right) & (t > s) \end{cases}$$

L_e は再帰的に計算されるので，動的計画法を使うことができる．2 引数で，引数の上限は n なので，空間計算量は $O(n^2)$ である．それぞれの計算で最大 n 回のループがあるので，時間計算量は $O(n^3)$ である．これはいずれも CKY アルゴリズムと同じである．したがって，1-best の解析を行ってから文節間距離 l を求めた場合と同等のコストで計算することができる．

4 実験

2 種類の係り受け関係の検索実験を行った．ひとつは直接的な係り受け関係の検索で，係り受け周辺確率によって検索する．もう一方は間接的な係り受け関係の検索で，期待距離の距離関数によって検索を行う．

4.1 パラメタ学習

(1) のモデルパラメタ $\mathbf{w}$ を，MAP 推定法で学習する．パラメタの事前分布として 0 中心，分散 $\sigma = 0.1$ の正規分布を用いた．これは L2 正則化を行うのと同じである．学習データとして京大コーパス ver. 4.0 の 1 月 1 から 8 日までを使った．内元ら [11] と同じ素性関数を用いた．最適化には L-BFGS 法 [3] を使用した．

参考のため 1 月 9 日のデータで 1-best 解の精度を調べたところ，文節正解率は 87.5% であった．なお，積極的な素性選択とパラメタ調整は行っていない．

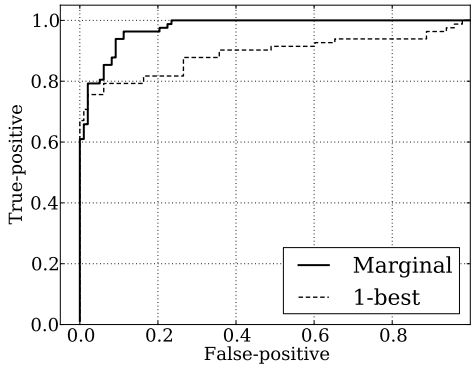

図 2: 直接的係り受け関係の検索における ROC 曲線

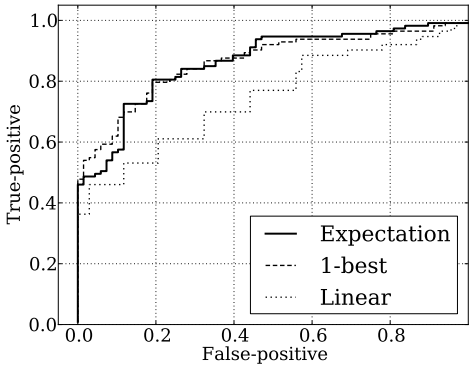

図 4: 間接的係り受け関係の検索における ROC 曲線

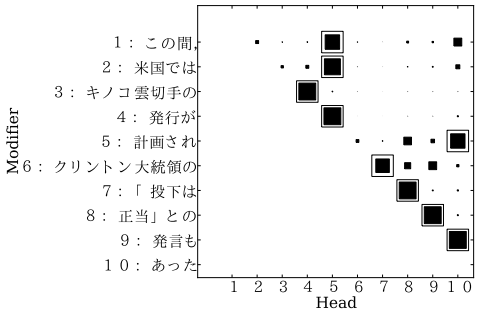

図 3: 係り受け周辺確率の例

4.2 実験データ

解析対象として CD-毎日新聞データ集'95 年版を使用した．単純に句点で文区切りを行うと，全部で 1,038,665 文あった．各文は MeCab [7] と CaboCha [8] を用いて文節列に変換してから，各実験を行った．L_e における枝の重み関数 e_1, e_2 は常に 1 を返すものとした．

4.3 周辺確率による直接係り受けの検索

係り受け周辺確率の有効性を示すため，直接係り関係にある文節対を検索する実験を行った．クエリとして単語対を与え，まず両方を含む文を検索する．そして，それぞれを含む文節間の係り受け周辺確率 $\mathrm{P}(y_i|\mathbf{x})$ の順に文を並び替えた．ベースライン手法では，まず 1-best の解析結果で係り関係にあるか調べ，次に文節インデックスの差が小さい順に並べる．これも同じときは正解を優先的に選び，精度の上限を出した．

図 2 は「大統領」と「発言」という単語対に対して上記 2 種類の順位付けを行った際の ROC 曲線 [1] である．このクエリに対して，全部で 180 件の文が見つかった．“Marginal” が周辺確率による順位付けを行った場合，“1-best” が 1-best 解によるベースライン手法である．正解か否かは人手で与えた．それぞれの AUC は，0.974, 0.889 と顕著に差が出た．ROC 曲線も， True-positive 0.8 あたりに 1-best で検出できる限界があり，極端に False-positive が増えている．

例として，図 3 に「この間，/米国では/キノコ雲切手の/発行が/計画され/クリントン大統領の/『投下は/正当』との/発言も/あった」という文の全文節間係り受け周辺確率を示した．Modifier が係り元，Head が係り先文節を示す．周辺確率に比例した面積の正方形をプロットし，1-best 解の係り先を四角で囲った．1-best の構文解析結果は「クリントン大統領の」の係り先を間違えているが，正解の「発言も」に対してある程度周辺確率が振られている（Modifier = 6 に対する正解は 7 ではなく 9）．また,「この間」,「米国では」といった，係り先が非自明な文節の周辺確率が全体に散っており，いずれも期待通りの結果といえる．

4.4 意味的つながりをもつ間接的な係り受けの検索

間接的な係り受けの検索で期待距離 L_e を実験した．クエリとして単語対を与えると，前節同様まず両方を含む文を検索する．それぞれの単語を含む文節に対して，各距離尺度を計算し，近い順に文を並び替える．先と同様,「大統領」「発言」というクエリに対して上記実験を行った結果の ROC 曲線が，図 4 である．なお，正解の基準として文節が直接係り関係にあるものに加え,「大統領は/ 発言を/した」のように主述関係のあるもの,「大統領は/発言を/ 否定した」のように述語の目的語になるものを含めた．これらの文は，前節のような直接の係り関係だけでは抽出できない．比較した 3 手法は，“Expectation” が L_e で定義される距離期待値，“1-best” が 1-best 解における距離 l，“Linear” が文節列上の距離（文節インデックスの差）である．それぞれの AUC は，順に 0.862, 0.862, 0.747 であった．

まず，“Linear” に比べて他の 2 手法は，いずれの点でも精度の高い検索ができている．これは，単純な文内の出現位置よりも係り受け木上での距離を設計したほうがよいことをあらわしている．一方，“Expectation” と “1-best” では前節で顕著に現れていた差がほとんど

表 1: 1-best と比べて距離指標が大きく変わった文例

Text	Expect	1-best	Correct
1: マスコミでも/（略）/冷淡な/評価が/多く、/クリントン**大統領**も/回想録を/受け/「自分が/ 反戦活動を/した/こと/は/正しかった」と/記者会見で/**発言**した	3.28	1	1
2: 今年/春、/米**大統領**の/「原爆投下は/正当だった」と/する/**発言**が/日本国内の/ 反発を/買った	2.94	1	1
3: 最も/率直に/厳しい/見解を/述べたのは/エリツィン**大統領**で、/八日の/会見で/ 「核廃絶を/（略）/あるのか」と/**発言**した	2.90	1	1
4: 米国の/ヒラリー**大統領**夫人は/五日に/北京入りの/予定で、/政府間会議で/**発言**する/ほか性と/（略）/（ＮＧＯ）フォーラムでも/発言する/予定	4.74	8	5
5: 米**大統領**が/戦勝国・敗戦国との/関係で/結ばれた/条約に/ついて、/ 自国に/不利な/**発言**を/するのは/前例が/なく、/米国内でも/反響を/呼びそうだ	5.62	7	2
6: （略）/クリントン米**大統領**は/三日の/定例ラジオ講話で、/（略）/限定する」と/述べ、/国連防護軍の/再編・強化支援を/示した/新政策から/後退する/**発言**を/行った	3.95	5	3

なくなった．これを詳細に調べるために，距離指標が大きく上がった・下がった正例を 3 つずつ調べたのが表 1 である．それぞれの距離指標とあわせて，正しい構文木上での距離を “Correct” に記した．まず，期待距離の方が遠くなった 3 例は，直接の係り関係にあるが，文節が離れており推定が難しい文であった．前節で見たとおり，係り先の曖昧な文節は周辺確率が散らばりやすく，これは予想される結果である．一方，期待距離の方が小さくなった例では，いずれも直接係らず，かつ構文解析を間違えている．1-best で間違えても期待距離では近くなっているため，これも予想通りの結果である．しかし，十分に指標が小さくならなかったため上位に上がらなかったと考えられる．ひとつの理由として，係り受け木上の距離 l に問題がある．辺の重みを全て 1 に固定したため，正しい解析ができたとしてもこれらの文例では文節間の距離はかなり大きい．したがって期待距離によって解析誤りが緩和されても正しく検出することは難しい．距離 l の設計と解析誤りに対する頑健さは別の問題として扱うべきであり，より適切な l を設計した上で比較する必要がある．

5 関連研究

工藤 [6] は決定的な形態素解析の代わりに，形態素の出現の周辺確率を使って曖昧な単語分割を行っている．また，岡野原ら [5] は単語分割確率を各単語境界での周辺確率で近似することで，検索に必要な確率値の圧縮を行っている．これらの確率的な単語分割と本手法を組み合わせることで，形態素解析・構文解析双方の誤りに頑健な検索・抽出システムの実現が期待できる．

6 結論

従来の決定的な構文解析結果を処理するのではなく，全構文解析候補に対する確率分布全体を活用した文節間距離尺度を提案した．まず，係り受け木上で頂点間の道の重み和として距離を定義した．そして，周辺分布の積で近似した分布における距離の期待値を効率的に計算する方法を示し，距離関数とした．これは，決定的な構文解析と同じ計算量で計算可能である．応用の例として，周辺確率を使った単語の係り関係の検索と，距離関数を使った単語対の検索を行った．前者の実験では決定的な手法より高い性能を示したが，後者では性能差が現れなかった．しかし，特に構文解析に間違えた文に対してより適切な距離尺度を与えていた．

今回，係り受け木上の辺の重みは定数としたが，検索の目的にあわせて適切に決める必要がある．また，双方向の係り受け木上での距離と，その期待値の効率的な計算方法が課題として残されている．

参考文献

[1] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, Vol. 30, No. 7, pp. 1145–1159, 1997.

[2] S. Lee and K.S. Choi. Reestimation and Best-First Parsing Algorithm for Probabilistic Dependency Grammar. In *Proceedings of the Fifth Workshop on Very Large Corpora*, pp. 41–55, 1997.

[3] D.C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, Vol. 45, No. 1, pp. 503–528, 1989.

[4] Y. Miyao and J. Tsujii. Maximum entropy estimation for feature forests. In *Proceedings of the second international conference on Human Language Technology Research*, pp. 292–297, 2002.

[5] 岡野原大輔, 工藤拓, 森信介. 形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用. Technical report, 第 1 回 NLP 若手の会, 2006.

[6] 工藤拓. 形態素周辺確率を用いた分かち書きの一般化とその応用. 言語処理学会第 11 回年次大会発表論文集, pp. 592–595, 2005.

[7] 工藤拓, 山本薫, 松本裕治. Conditional Random Fields を用いた日本語形態素解析. 情報処理学会自然言語処理研究会 SIGNL-161, pp. 89–96, 2004.

[8] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[9] 池田和幸, 高須淳宏, 安達淳. 単語間の係受け情報を用いた文献検索手法. 学術情報センター紀要, Vol. 9, pp. 143–159, 1997.

[10] 竹内淳平, 辻井潤一. 係り受け関係と言い換え関係を用いた柔軟な日本語検索. 言語処理学会第 11 回年次大会発表論文集, pp. 568–571, 2005.

[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.