

MEによる日本語係り受け解析

太田一樹

東京大学理学部情報科学科 4年

kzk@is.s.u-tokyo.ac.jp

2007/06/07

概要

日本語の構文解析において、係り受け解析は自然言語処理の基本技術の一つとして認識されている。近年では構文情報が付加された大規模コーパスが利用になったことや、使用できるコンピュータの性能が伸びてきたことから、機械学習の手法を用いた係り受け解析手法が数多く提案されている。

今回は特に最大エントロピー法 (ME 法) を用いた日本語係り受け解析手法について調査・実装を行った。ある文節間の係りやすさの確率は、ME によって学習した係り受け確率モデルから計算する事ができる。この確率モデルは、学習データから得られる品詞情報などを元に、2つの文節に係るか係らないかを予測するのに有効な素性を学習することで得られる。

本稿では、ME 法の概要・素性に関する考察・係り受け解析アルゴリズム等について述べる。学習データ・及びテストデータについては、毎日新聞に係り受け情報を付加した京大コーパス [6] を用いた。最終的には京大コーパスを使用した実験で 88.0% という高い精度を出す事に成功した。

1 はじめに

係り受け解析は自然言語処理における基本技術として認識されており、多くの研究が行われている。

初期の研究では、二文節間の係りやすさを決定するルールを人手で作成していたが、網羅性・一貫性という点で問題が多かった。そこで、近年では機械学習アルゴリズムを用いた統計的な解析手法が提案されている。今回はその中でも、主に文献 [3] の手法に従い、最大エントロピー法 (ME) [1] を用いた解析を行った。

ME は学習データ中の素性の頻度から特徴的な素性を学習し、その特徴をもった確率モデルを生成する仕組みである。素性とはデータ内に

観測される特徴のことで、今回は注目する二文節の品詞・活用形・句読点の有無やその組み合わせを利用した。テストの際には学習したデータを元に、テスト文中の二文節間の係る確率を算出する。また、文全体の係り受け確率は、その文中のすべての係り受けの確率の積で求められると仮定する。

日本語の係り受け関係には主に以下の特徴があるとされている。

1. 後方を修飾する。
2. 係り受け関係は交差しない。
3. 係り要素は受け要素を 1 つだけ持つ。

また、各二文節の係り関係は他と独立と仮定

しているが、ある係り受け関係が他の係り受けに影響を及ぼす可能性も有るため、係り関係が決まった所に動的に素性を追加する「動的素性」も導入した。

2 確率モデル

この章では、最大エントロピー法を用いて二文節間の係り受け確率をコーパスから統計的に学習するための方法について説明する。この手法は文献 [3] と同じものである。

2.1 最大エントロピー法 (ME) の概要

ある文脈について、出力値が唯一に決まるような事象を考える。この時、新しい文脈に対して出力値を予想しようと思うと、過去の履歴(文脈に対する出力値)を収集し、履歴から特徴を抽出して、出力値との関係を調べる事である程度予想がつきそうである。

ME では観測された特徴(素性という)と出力値との関係は確率であらわされる。またそれらの確率は、履歴に現れなかったような事象を含めたすべての事象に対する確率が「最も一様であると思われる」ような分布になっている。以下では ME で表される確率分布について説明する。

文脈 $b(\in B)$ で出力値 $a(\in A)$ となるような事象 (a,b) の確率分布 $p(a,b)$ を推定する事を考える。まず、 k 個の素性 $f_i(1 < i \leq k)$ を考える。次に f_k が観測されるような文脈 b の集合を V_{b_j} とし、文脈が $b(\in B)$ でかつ出力値が $a(\in A)$ となる時に 1 を返すような関数 $f_j(a,b)$ を定義すると、以下の様になる。

$$f_j(a,b) = \begin{cases} 1 & (b \in V_{b_j} \text{ かつ } a \in A) \\ 0 & (\text{それ以外}) \end{cases}$$

また、確率分布を推定するにあたって一つ制約を導入する。それは、履歴の中に素性が現れた割合は推定した確率分布の中での割合と同じになるという制約である。履歴中の確率分布

を $p'(a,b)$ とすると、これは以下の制約式で表せる。

$$\sum_{a,b} p(a,b) f_j(a,b) = \sum_{a,b} p'(a,b) f_j(a,b) \quad (j = 1 \dots)$$

これを満たす確率分布 $p(a,b)$ のうち、最も「一様な」分布を推定するのが最大エントロピー法の肝である。一様な分布とはつまり次のエントロピー $H(p)$ を最大にする確率分布である。

$$H(p) = - \sum_{a,b} p(a,b) \log(p(a,b))$$

このような確率分布 p^* は唯一存在し、以下の様に記述される。

$$p^*(a,b) = \pi \prod_{j=1}^k \alpha_{a,j}^{f_j(a,b)}$$

$$(0 \leq \alpha_{a,j} \leq \infty, \pi \text{ は正規化定数})$$

$$\alpha(a,j) = e^{\lambda_{a,j}}$$

$\lambda_{a,j}$ は素性関数 $f_j(a,b)$ のパラメーターで、 $f_j(a,b)$ が推定された確率分布内でどれぐらい重要なのかを表す値である。

今回はこのパラメーターを求めるために、Amis¹ という ME 用の学習機を用いた。Amis は文脈 b とそこで観測された素性 $f_j(a,b)$ を与えると、 $\lambda_{a,j}$ もしくは $\alpha_{a,j}$ を推測するツールである。

2.2 ME の係り受け解析への適用

この節では前節で述べた ME のモデルを係り受け解析に適用する方法について述べる。まず、二文節が係り受け関係になるか無いかを出力値に取る k 個の素性 f_k を考える。出力値 a は 0 か 1 かのどちらかである。このとき、文中の任意の 2 つの文節 s'_i, s'_j に着目する。この 2 つの文節に現れる素性群を文脈 b とすると、文節間が係り関係にある確率 $p^*(1|b)$ は、次のように表される。

$$p^*(1|b) = \frac{p^*(1,b)}{p^*(1,b) + p^*(0,b)}$$

¹<http://www-tsuji.is.s.u-tokyo.ac.jp/amis/>

学習コーパスから ME を用いて事前に $p^*(a, b)$ を求めておくと、テストコーパス中のある2文節間が係り受け関係にある確率 $p^*(1|b)$ が計算できる。

素性については前文節、後文節、二文節間それぞれの持つ特徴を組み合わせたものを用いた。詳しくは次章で述べる。

3 素性の選択

ME では人手で素性を明示的に加える必要がある。今回の学習に用いた基本素性を表 1 に示す。若干の差異があるものの、[3, 4, 5] 等で一般的に用いられている素性である。

表 1: 使用した基本素性

前/後文節	主辞見出し, 主辞品詞, 主辞品詞細分類, 主辞活用, 主辞活用細分類, 語形見出し, 語形品詞, 語形品詞細分類, 語形活用, 語形活用細分類, 語形活用型, 語形活用形, 括弧の有無, 句読点の有無, 文節の位置 (文頭 or 文末)
文節間	距離 (1,2-5,6 以上), 括弧の有無, 句読点の有無, ”は”の有無

また、これらを組み合わせた素性も加えた。文献 [3] を参考にし、有効と思われる素性を人手で追加した。結果的に 2~ 5 個の基本素性を組み合わせたものが追加されている。文中に 1 度しか現れないような素性も用いた。

また、後述するように係り関係の情報を素性として与える「動的素性」も一部用いている。

4 解析手法

この節では、係り受けの解析アルゴリズムを説明する。基本的には文献 [3] で紹介されている BeamSearch アルゴリズムを用いた。幅 k による精度の変化は実験結果の章に掲載した。また、解析する過程において 2 点工夫を施した。

4.1 交差判定

日本語の係り受け関係には交差判定が無い事を利用し、交差するような係り受け関係を排除しながら BeamSearch を行った。これにより解析時間は長くなるが、有意な精度向上が見られた。

4.2 動的素性

BeamSearch をしていく過程では、係り受け関係は後ろの方から決定されていく。文献 [3] では全ての係り受け関係は独立であると見なし解析を行っているが、文献 [4] においてそれとは反する結果が出ている。

具体的には動的素性と呼ばれるものを導入し、精度の向上が行われている。これは既に係り受け関係が決定した場所について、動的に素性を追加しながら解析を行う手法である。文献 [4] では以下の 3 タイプの動的素性を考慮している。

- 着目している係り先に係る文節 (A)
- 着目している係り元に係る文節 (B)
- 着目している係り先が係る文節 (C)

動的素性によって有意に精度が上がるという事実は、係り受け関係同士が影響し合っている事を意味している。この動的素性を BeamSearch に取り込む試みを行った。BeamSearch の場合、係り受け関係は文脈の後ろから決定していくため、(A)(B) のケースは導入できないので、(C) のケースのみを取り入れた。

ただし、これについても解析時間が増えてしまう。動的素性としては主辞品詞・主辞活用・語形品詞・語形活用を用いた。動的素性の効果については後述する。

5 実験結果と考察

京大コーパス (Version 3.0)[6] を以下の3つに分割して実験を行った。

- 学習データ: 一般記事 1月 1,3-8 日
- テストデータ: 一般記事 1月 9 日
- ディベロップメントデータ: 一般記事 1月 10 日

これは文献 [2] で紹介されている物と同じテストセットである。すべての実験は Core 2 Duo 2.4Ghz, 主記憶 8Gbyte の Linux 上で行った。

5.1 実験結果

今回の結果と、文献 [2] に掲載されている他の手法での正解率を表 2 にまとめる。ただし、係り受け正解率とは文末の一文節を除くすべての文節に対して、正しく係り先が同定出来たものの割合を示す。

表 2: 手法毎の正解率

手法	正解率 (%)
本稿 (BeamSearch)	88.02
S04(SVMs)	89.56
S04(linear SVM)	87.36
KM02(Chunking)	89.29
USI98(BeamSearch)	87.14

S04 = Sassano 2004[2], KM02 = 工藤, 松本 2002[4], USI99 = 内元 1998[3].

SVM を利用する手法に比べると、1%程度正解率が低いという結果になった。計算量に関しては S04 は $O(n)$ 、それ以外については $O(n^2)$ となっている。

5.2 BeamSearch の幅毎の比較

次に、BeamSearch 時の幅を変えながら文正解率の変化をプロットした。その結果は図 1 のようになった。

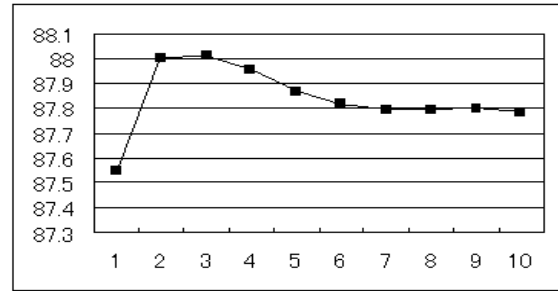


図 1: BeamSearch の幅と正解率 (%) の関係

幅を大きくすると必ずしも正解率が向上するとは限らない事が分かる。以下では、正解率の一番高かった幅 3 (88.0116%) で実験を行う。

今回実装した手法の元論文である USI98[3] に、多少工夫を施したことで正解率が微増していることが分かる。しかし、Support Vector Machine を利用した手法と比べると 1%程度の正解率の差が出た結果となった。

5.3 交差判定の効果

交差判定を抜いて精度を測ると、正解率は 87.3113% (-0.70023%) になった。日本語の係り受け解析においては、非交差条件を考慮する事で精度が有意に上がる事を示せた。

5.4 動的素性の効果

動的素性を抜いて精度を測ったところ、正解率は 87.8661% (-0.1455%) という結果になった。減少率はかなり少ないので、他の特定の素性との兼ね合いでたまたま精度が向上している可能性もある。

また今回は文献 [4] に沿って主辞品詞と主辞活用を動的素性として選んだが、他にも適した物があるかもしれない。他にも「着目している係り先に係る文節」の素性情報も動的に加える余地が有りそうである。動的素性についてはもう少し詳しい検証が必要だ。

5.5 係り先距離毎の比較

係り先の距離毎に係り受け精度を算出したところ、表3のようになった。

表3: 距離毎の正解率

距離	正解率 (%)	正解文節数/総文節数
1	97.7894	6901/7057
2-3	70.0939	1493/2130
4-5	78.1088	603/772
6-7	59.2326	247/417
8-9	69.5489	185/266
10~	70.1705	247/352

これを見ると距離2以上の正解率が距離1に比べると低い水準になってしまっている。ただ、距離が少ないほど数が多くなる傾向にあるので全体としての精度はそれほど変わらないものとなっていると思われる。

解析の過程で長距離の係り受けを推定すると、それより前の文節の係り受け自由度が交差判定によって下がるためにこのような結果になっているものと思われる。これを解決するためには、BeamSearchのような近似的な方法ではなく係り受け関係を有効グラフと見た時に、選ばれた辺の確率の積を最大にするような全域有向木を決定的に求めるような手法が必要と思われる。

6 まとめ

MEを用いて日本語係り受け解析を行い、機械学習と自然言語処理の理論と実際に触れた。

今回実装した手法では人手による素性選択に依ってしまうところが大きく、そこが演習前の機械学習に対するイメージと最も異なる点で

あった。また、素性を加えるにつれて精度も上がるが、それに比例して当然学習に要する時間も増大する。なので実際の問題に機械学習を適用させるとなると、少ない素性で精度が出たり計算量が少ないようなモデルのほうが適しているように思えた。

参考文献

- [1] Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra: A maximum entropy approach to natural language processing, Computational Linguistics, Volume 22, Issue1 (March 1996), Pages: 39 - 71 (1996).
- [2] Manabu Sassano: Linear-Time Dependency Analysis for Japanese, Proceedings of the 20th international conference on Computational Linguistics, Article No. 8 (2004).
- [3] 内元清貴, 関根聡, 井佐原均: MEによる日本語係り受け解析, 情報処理学会研究報告. 自然言語処理研究会報告, Vol.98, No.99 pp. 31-38 (1998).
- [4] 工藤 拓, 松本 裕治: チャンキングの段階適用による係り受け解析, 情報処理学会論文誌, Vol 43, No. 6 pp. 1834-1842 (2002).
- [5] 工藤 拓, 松本 裕治: 相対的な係りやすさを考慮した日本語係り受け解析 SIGNL-162 (2004).
- [6] 黒橋禎夫 and 長尾眞, 京都大学テキストコーパス・プロジェクト. In 言語処理学会第3回年次大会, pages 115-118, 1997.