ボルツマンマシンによる日本語係り受け解析

三浦　康秀, 高橋　治久
電気通信大学電気通信学研究科電子情報学専攻

本稿はボルツマンマシンによる係り受け解析の手法を提案し, その汎化性能の検証を行う. ボルツマンマシンによる係り受け解析は長田・吉田 [1] らがすでに行っているが, 本研究では文節から観測される特徴である素性の概念を導入し, その改良を提唱する. さらに手法の有効性の検証には, 汎化性能が重要という観点から, 上記論文で行われていなかった未知の文に対しての解析を行った. 京大コーパスを用いた実験を行ったところ, このモデルは上記論文のモデルより汎化性能を大幅に向上させることを示した.

Japanese Dependency Structure Analysis based on Boltzmann Machine

Yasuhide Miura, Haruhisa Takahashi
Department of Communications and Systems, The Univercity of Electro-Communications

This paper describes an analysis of Japanese dependency-based structure using Boltzmann Machine, which was originally proposed by Nagata and Yoshida[1]. The performance of Boltzmann-Machine-based methods highly depends on the construction of energy function. We introduce a new construction method of the energy function introducing a grammatical characteristic principle which enables us to improve the performance. Further, we propose a new performance evaluation indicator as well as an evaluation method based on the concept of generalization performance. Experimental results based on Kyoto University corpus show that our model has higher performance compared to the former model.

1　はじめに

近年のIT（Information Technology）産業の成長は我々に日常的に大量の情報を得られる機会をもたらした. このため, 大量の文書から目的の情報を探し出す, 文書の内容を要約する, 文書を翻訳する, などの高度で大規模な自然言語処理技術の必要性が高まってきている. 本研究で取り挙げる, 構文解析およびその部分問題である文節係り受け解析はこうした自然言語処理技術の基礎となる.

従来, 係り受け解析は人手で作成した規則を用いて行われてきた. しかし, 係り受け解析に有効だとされる素性の数は膨大であり, また競合することがあるため, 人手による規則の作成には網羅性, 一貫性という点で問題がある. 近年, 係り受け情報などが付与された例文集であるテキストコーパスの大規模なもの（京大コーパス,EDR コーパス, など）が利用可能になっている. そこで, コーパスから統計を収集し, そのデータを一種の言語知識として考え, 係り受け関係を求める手法が提案されている.

統計を用いての係り受け解析の一つの手法として相互結合型のニューラルネットを用いる手法が考えられる. 実際に, ボルツマンマシンを用いた例として長田・吉田 [1] の研究がある. しかし, この文献では統計を取った文による解析しか行っておらず, ボルツマンマシンを用いた手法の汎化性能の検証がなされていない. そのため, 実際の解析にこの手法が出来るか否かは明らかでない. そこで本研究では, 最大エントロピー法を用いて係り受け解析を行った内元・関根・井佐原 [2] の手法を参考にデータから観測される特徴である素性（この場合においては品詞などの文節情報）の統計を加え長田・吉田 [1] による手法の改良を行った. さらに未知の文に対しての汎化性能の検証を行い, 提唱手法の有効性について検討した.

2　係り受け解析

係り受け解析とは文節単位に区切られた日本語文に対し正しい係り受け構造を特定する問題である. 係り受け構造を特定することは文節間の関係を特定することになり, 日本語文の解析を行う上で非常に有用なものとなる.

本研究ではこの係り受け解析を相互結合型ニューラルネットワークの一種であるボルツマンマシンを用いて行う.

3　素性

本研究で統計の収集に用いる, 文節または文節間から観測される素性について説明する. 素性とは一文中の二つの文節に着目したとき, それぞれの文節（前文節, 後文節）が持ち得る属性または二文節間に現れる属性である. 本研究では内元・関根・井佐原 [2] の手法, 京大コーパスのマニュアルを参考に, 表記, 読み, 主辞[1]見出し, 主辞品詞, 主辞品詞細分類, 主辞活用型, 主辞活用形, 語形[2]見出し, 語形品詞, 語形品詞細分類, 句読点の有無, の素性を考えた.

4　ニューラルネットワーク

4.1　ホップフィールドネットワーク

ホップフィールドネットワークとは相互結合型のニューラルネットワークであり, 素子 i から素子 j への結合の強さを w_{ij} で表すと,$w_{ij} = w_{ji}$ が成り立つ. 素子の閾値を θ_i とすると, 状態 u_i の素子 i が受け取る入力の総和 C_i は,

$$C_i = \sum_{j=1}^{n} w_{ij} u_j + \theta_i$$

と表され, 状態遷移は,

$$u_i = \begin{cases} 0 & : C_i < 0 \\ no\,change & : C_i = 0 \\ 1 & : C_i > 0 \end{cases}$$

のように定義される. このときネットワークのエネルギー関数を

$$E = -\sum_{i=1}^{n} \sum_{j>i} w_{ij} u_i u_j - \sum_{i=1}^{n} \theta_i u_i$$

のように定義すると, エネルギー関数は時間と共に単調減少し, 定常状態に収束することが証明されている.

4.2　ボルツマンマシン

ボルツマンマシンはホップフィールドネットワークの動作を熱力学的な確率動作としたものである. 具体的には, 素子 i が状態 0 から状態 1 へ状態遷移する際に状態 1 を受容する確率を, ネットワークの温度 T と共に次のように定める.

$$P_{i,T}\{accept\ 1\} = \frac{1}{1 + exp\left(\frac{-C_i}{T}\right)}$$

これにより T が高いときは C_i に依存しない状態遷移が行われる.

5　係り受け解析の手法

本稿では解析する文中において考えられるすべての係り受け関係をボルツマンマシンの素子に割り当て, 一般的な日本語の係り受け関係の特徴にコーパスから得た文節の表記, 素性の統計を考慮した以下のようなエネルギーを構成して係り受け解析を行う.

5.1　エネルギー関数の構成

5.1.1　非交差性

係り受け関係を考えるとき, 一般的に入れ子構造になっておりそれぞれの係り関係が交差することはない.

$$E_1 = \sum_{i}^{n} \sum_{j>i} X_{ij} u_i u_j$$

$$X_{ij} = \begin{cases} 1 & \text{if 素子 } u_i \text{と素子 } u_j \text{が交差している} \\ 0 & \text{その他} \end{cases}$$

5.1.2　係り先専有性

係り受け関係では, 文末の文節を除けばそれぞれの文節はそれより後ろの一つの文節のみに係る.

$$E_2 = \sum_{k=1}^{m} \left(\sum_{i=1}^{n} Y_{ki} u_i - 1 \right)^2 - 1$$

$$Y_{ki} = \begin{cases} 1 & \text{if 素子 } u_i \text{の係り元が文節 } k \\ 0 & \text{その他} \end{cases}$$

[1]各文節内で, 品詞の大分類が特殊, 助詞, 接尾辞となるものを除き, 最も文末に近い形態素.

[2]各文節内で, 特殊を除き最も文末に近い形態素. もしそれが助詞, 接尾辞以外の形態素で活用型, 活用形をもつものである場合はその活用部分とする.

5.1.3 卑近接続性

距離的に近い文節ほど係り受け関係が成立しやすい.

$$E_3 = \sum_{i=1}^{n} Z_i u_i$$
$$Z_i = l_i - k_i$$

l_i：係り先文節番号,k_i：係り元文節番号.

5.1.4 頻度との関連性（表記）

表記に関する統計において文節同士係り受け頻度が高いものほど係り受け関係が成立しやすい.

$$E_4 = \sum_{i=1}^{n} Q_i u_i$$
$$Q_i = -\frac{h_i}{h_{ki-max}}$$

h_i：表記に関する統計において求められた係り受け u_i の頻度.$h_{ki-max} = max\{h_j | for\ k_j = k_i\}$.

5.1.5 頻度との関連性（素性）

素性に関する統計において文節同士係り受け頻度が高いものほど係り受け関係が成立しやすい.

$$E_5 = \sum_{i=1}^{n} S_i u_i$$
$$S_i = -\frac{h_i}{h_{ki-max}}$$

h_i 素性に関する統計において求められた係り受け u_i の頻度.$h_{ki-max} = max\{h_j | for\ k_j = k_i\}$.

5.2 ネットワークのエネルギー

以上の５つのエネルギーによりネットワークのエネルギーを次のように定める.

$$E = aE_1 + bE_2 + cE_3 + dE_4 + eE_5$$

a, b, c, d, e は定数である. そしてこれよりネットワークの重み, 閾値を以下のようになる.

$$w_{ij} = -2aX_{ij} - 2b\sum_{k=1}^{m} Y_{ki}Y_{kj}$$
$$\theta_i = b\sum_{k=1}^{m} Y_{ki} + cZ_i + dQ_i + eS_i$$

6 実験

6.1 アニーリングスケジュール

ボルツマンマシンのネットワークの温度を徐々に下げていくときに, シミュレーテッドアニーリングという手法をとる. 本研究ではアニーリングスケジュールを予備実験[3]を参考に決定した. これは, 初期温度 T_0 を定め, ステップ毎に $T_{k+1} := \alpha T_k$ のようにして値を小さくしていき, 1 ステップあたり L 個の素子について計算をランダムに行い, やがて K ステップ後において状態遷移が起こっていなければ計算を終了する, という手法である. 本研究ではパラメータは,$T_0 = \sum(|w| + |\theta|)$,$\alpha = 0.95$,$L = 25$,$K = 5$ に設定した.

6.2 評価尺度

実験の評価尺度について述べておく. 本実験では解析結果の評価尺度として以下のような正解率と誤解率というのを考える.

$$\text{正解率} = \frac{\text{発火した正しい係り受けの素子の数}}{\text{文に存在する係り受けの数}}$$
$$\text{誤解率} = \frac{\text{発火した誤った係り受けの素子の数}}{\text{発火した素子の数}}$$

6.3 統計の収集

実験するに先立って, 京大コーパス（Version3.0）より,1 月 1 日〜1 月 9 日までの全記事,1 月〜6 月までの社説記事, 計 18461 文より統計を収集した. 統計の収集は前文節, 後文節両方に関して, 文献 [2] を参考に, 次頁の表 1 の素性の組み合わせで行った.

6.4 パラメータ a, b, c, d, e の設定

エネルギー関数のパラメータ a, b, c, d, e の値をどのように定めればよいかは実験的に求めるしかない. そこで本研究では未知の文を用いて, まずパラメータ a, b, c, e の４つをそれぞれ 1,5,10,15,20 に変化させ予備実験を行こない, そして結果の良かったパラメータ a, b, c, e を用いてさらにパラメータ d を

[3]ボルツマンマシンによる日本語形態素解析の試み. 電気通信大学卒業論文 1993.

1	表記, 読み
2	主辞品詞, 主辞活用形 語形品詞
3	主辞品詞, 主辞活用形 語形品詞, 句読点の有無
4	主辞品詞, 主辞品詞細分類, 主辞活用形 語形品詞, 語形品詞細分類
5	主辞品詞, 主辞品詞細分類, 主辞活用形 語形品詞, 語形品詞細分類, 句読点の有無
6	主辞品詞, 主辞品詞細分類, 主辞活用型 主辞活用形, 語形品詞, 語形品詞細分類
7	主辞品詞, 主辞品詞細分類, 主辞活用型 主辞活用形, 語形品詞, 語形品詞細分類 句読点の有無
8	主辞品詞, 主辞品詞細分類, 主辞活用型 主辞活用形, 語形見出し, 語形品詞 語形品詞細分類
9	主辞品詞, 主辞品詞細分類, 主辞活用型 主辞活用形, 語形見出し, 語形品詞 語形品詞細分類, 句読点の有無

表 1: 収集した統計

10,20,40,60,80,100,120,140,160,180,200 の値に変化させ予備実験を行い, 最終的に最も結果の良かったパラメータ a, b, c, d, e を用いて実験を行っている.

6.5 予備実験

実験に用いる素性統計とパラメータ a, b, c, e を求めるため, 表 1 の素性の組み合わせ２～９, それぞれに対して, 京大コーパスより未知の文 50 文を用いて予備実験を行った. その結果, 素性統計の組み合わせ５を用いてパラメータ $a = 10, b = 20, c = 15, e = 20$ のときに最も良い結果が得られた. そこで, 素性統計の組み合わせ５にパラメータの組み合わせ $a = 10, b = 20, c = 15, e = 20$ を用いパラメータ d を 10,20,40,60,80,100,120,140,160,180,200 に変化させ, 京大コーパスより統計を取った文 100 文と未知の文 100 文を用いてさらに予備実験を行った. その結果,$d = 140$ のときに未知の文に対して最も良い結果が得られた.

6.6 実験

素性の組み合わせ５を用いて, パラメータを $a = 10, b = 20, c = 15, d = 140, e = 20$ として, 京大コーパスより統計を取った文 1000 文, 未知の文 1000 文を用いて実験を行った. 結果は以下のようになった.

文の種類	正解率	誤解率
統計を取った文 1000 文	96.8 %	5.5 %
未知の文 1000 文	60.9 %	37.5 %

表 2: 実験の結果

なお, 同じ文を長田・吉田 [1] の手法で解析した場合, 統計を取った文で, 正解率 92.0 %, 誤解率 11.1 %, 未知の文で, 正解率 43.3 %, 誤解率 57.3 %, が得られた. 結果として本手法は, 長田・吉田 [1] の手法と比べて, 統計を取った文で正解率 4.8 %, 誤解率 5.6 %, 未知の文で正解率 16.8 %, 誤解率 19.8 %の精度の向上が得られた.

7 まとめ

本研究の手法で未知の文に対して係り受け解析を行った結果, 長田・吉田 [1] の手法に比べて大幅な精度の向上が得られた. しかし, 本研究の解析精度（正解率 60.9 %, 誤解率 37.5 %）も実用上十分とは言えない. 本研究では統計を取った文と, 未知の文に対する解析精度に大きな差が見られた. これは統計を用いて解析を行う以上, 仕方のないところもあるが, 結果として収集した統計の量が不十分であったと考えられる. また, パラメータの決定も, まず a, b, c, e を 1,5,10,15,20 の組み合わせで決定し, その後 d を決定する, といった手法を取ったがこれが最良のパラメータという保証はない. さらなる解析精度の向上には, より多くの統計を収集する, パラメータをより高精度で決定する, などが考えられる.

参考文献

[1] 長田靖, 吉田敬一. BNN を用いた日本語文の係り受け解析. 情報処理学会 自然言語処理研究会 NL142-13,pp91-96,2001.

[2] 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 自然言語処理研究会 NL128-5,pp31-38,1998.