

## 相互結合型ニューラルネットワークによる日本語の係り受け解析

4 F - 7

高橋 直人 板橋 秀一

筑波大学

## 1. はじめに

日本語の構文構造は、文節間の係り受け関係という形で表現することができる。1つの文に対する可能な係り受け関係は一般に複数個存在し、それら複数の候補の中から最適と思われる係り受け関係を決定することは、ある条件の下での最適化問題に帰着することができる。本稿では係り受け解析をこのような最適化問題としてとらえ、Hopfield型ネットワークで最適解を得るための試みについて述べる。ネットワークの各ユニットには成立する可能性のある文節間係り受け関係を割り当てる。また制約条件として文法的な制限だけでなく、係り元の文節と係り先の文節の意味的整合性を反映させることも可能である。

## 2. 重み・しきい値の決定

本ネットワークの各ユニットに割り当てられるのは文節ではなく、成立する可能性のある文節間の係り受け関係である。各ユニットは0または1の値をとり、最終的に値が1であるユニットが表す係り受け関係のみが残るものとする。例えば、

「日曜日に地震で壊れた建物を見た。」

という文は図1のように5つの文節からなっており、可能な係り受け関係は図中の矢印で示したように全部で6つ存在する。目標はこの6関係の中から必要十分な数の互いに矛盾しない係り受け関係を選び出すことである。そのための制約項として以下の4項を考える。

## 1) 係り受け関係の総数による制約

最後の文節以外のすべての文節は、それよりも後方にある文節のどれか1つにかかる。したがって最終的に1となるべきユニットの数は(文節数)-1である。可能な係り受け関係の総数を $n$ 、最終的に値1となるべきユニットの数を $m$ とすると、値1を持つユニット数がちょうど $m$ 個のとき最小値を与える項は、

$$E_1 = \left( \sum_{i=1}^n u_i - m \right)^2$$

と表せる。

## 2) 相互排他性による制約

ある1つの文節に係るのは1つの文節だけである。したがって同一の文節に係り元とする係り受け関係が複数存在する場合、それらのうちではただ1

つだけが成立することになる。また係り受けの非交差条件から、2つの係り受け関係が互いに交差するときにはそれらの両方が成立することはあり得ないことがわかる。

このような係り受け関係間の相互排他性を表すために、相互排他行列 $X$ を導入する。 $X$ は $n \times n$ 行列( $n$ は成立する可能性のある係り受け関係の総数)であり、その $(i, j)$ 要素の値は

$$X_{ij} = \begin{cases} 1 & \text{係り受け関係 } i \text{ と係り受け関係 } j \text{ が} \\ & \text{相互排他的である場合} \\ 0 & \text{otherwise} \end{cases}$$

である。相互排他行列は明らかに対称行列になる。例えば図1の文に対する相互排他行列は

$$X = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

となる。

この相互排他行列 $X$ を用いて、値1をとるユニット中のどの2つも互いに排他的でない場合に最小値をとる項を表すと

$$E_2 = \sum_{i=1}^n \sum_{j=1}^n X_{ij} u_i u_j$$

となる。

## 3) 係り受け関係の尤度による制約

ある係り受け関係がどの程度成立しやすいかは、係り元の文節と係り先の文節の意味的関係によって決定される。今 $y(i)$ を「係り受け関係 $i$ の意味的尤度が高いほど小さい値をとる関数」とし、 $Y$ を第 $i$ 要素が $y(i)$ であるような尤度ベクトルとする。図1の例では「地震で見た」という係り受けよりも「地震で壊れた」という係り受けの方が意味的尤度が高いと考えられるので、 $y(3) < y(4)$ となるであろう。

意味的尤度の高い係り受け関係が多く選ばれるほど小さい値をとるような項は尤度ベクトル $Y$ を用いて

$$E_3 = \sum_{i=1}^n Y_i u_i$$

と表すことができる。

#### 4) 文節間の距離による制約

日本語の係り受けの原則の1つに、「係り先を2つ以上持つ文節は、(意味的制約を考えなければ) 位置的に近いものにほど係りやすい」というものがある。係り受け関係  $i$  の係り元を  $k_i$  番目の文節、係り先を  $l_i$  番目の文節とすると、 $Z_i = l_i - k_i$  であるようなベクトル  $Z$  を考えると、文節間の距離が短い係り受け関係が選ばれるほど小さな値を取る項は

$$E_4 = \sum_{i=1}^n Z_i u_i$$

と表すことができる。

以上4項の1次結合  $aE_1 + bE_2 + cE_3 + dE_4$  ( $a, b, c, d$  は定数) と、ユニット  $u_i$  の状態変化規則が

$$\sum_{j=1}^n w_{ij} u_j - \theta_i \begin{cases} < 0 & \text{then } u_i = 0 \\ = 0 & \text{then } u_i \text{ は変化しない} \\ > 0 & \text{then } u_i = 1 \end{cases}$$

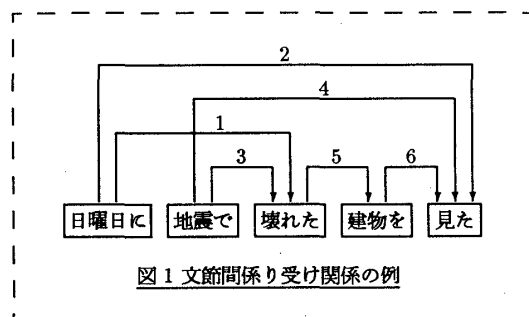
であるようなネットワークのエネルギーの式

$$E = -\frac{1}{2} \sum_{i=1}^n \sum_{j \neq i} w_{ij} u_i u_j + \sum_{i=1}^n \theta_i u_i$$

とを比較することにより、重み  $w_{ij}$  およびしきい値  $\theta_i$  は

$$\begin{cases} w_{ij} = -2(a + bX_{ij}) \\ \theta_i = (1 - 2m)a + cY_i + dZ_i \end{cases}$$

となることがわかる。



### 3. 実験・考察

以上をもとに簡単な実験を行なった。尤度ベクトルとしては各係り受け関係の係り元と係り先の文節に含まれる自立語の単語共起頻度を用いた。また各ユニットの初期値は、同一の係り元を有する係り受け関係が2つ以上ある場合、係り元と係り先の文節の位置的距離がもっとも近いものを1、それ以外のものを0とした。図1のネットワークで言うと、ユニット1, 3, 5, 6の初期値が1、ユニット2, 4の初期値が0となる。また、各ユニットのしきい値  $\theta_i$  を時間と共に変化させ、Boltzmann Machine に類似した動作を行なわせるようにした。

約200の例文から単語共起頻度を採取し、それら例文中で用いられている語からなる文を文節に分解したものを入力データとして係り受け関係の解析を行なったところ、 $E_i$  の係数  $a, b, c, d$  を適切に選べばほぼ期待どおりの解析結果が得られることがわかった。またネットワークが安定状態に収束するのに必要とされるステップ数は、ユニット数にほぼ比例することがわかった。

今回作成したネットワークは厳密な Boltzmann Machine ではないが、好ましくない安定状態におちいることが少ない。これは各ユニットの初期値が最適解の比較的よい近似になっているためと考えられる。

今回は尤度ベクトル  $Y$  として単語共起頻度を用いたため、そのままでは解析すべき文が変化すれば  $Y_i$  の最大値・最小値も変化してしまい、 $E_i$  の係数の決定が比較的難しくなる。これを回避するための手段として、入力文ごとに  $Y$  を正規化するようにした。

### 4. おわりに

各ユニットに文節ではなく係り受け関係を割り当てたため、重み・しきい値の計算が簡単になったが、一方で相互排除行列を計算する必要が生じた。相互排除行列の計算量はユニット数の2乗に比例するので、これを効率良く計算する必要がある。

今回はあらかじめ分割された文節間の係り受け関係のみの解析であるが、今後は形態素解析の段階での disambiguation にもニューラルネットを利用する予定である。

### 参考文献

- [1] Hopfield, J. J. and Tank, D. W. 1985. Neural computation of decisions in optimization problems. *Biol. Cybernetics* 52, pp.141-152.
- [2] 麻生英樹 1988. ニューラルネットワーク情報処理. 産業図書.