日本語形態素解析用ニューラルネットワークについて

1C−2 高橋直人 板橋秀一 平井有三

筑波大学

1 はじめに

筆者らは相互結合型ニューラルネットワークを用いた日本語解析に関する研究を行なってきた[1][2][3]が、従来の方法では係り受け解析に比べると形態素解析の効率が悪いという問題があった。本稿では、以前の形態素解析ネットワークについてどこが悪いのか考察し、より効率の良いネットワークの提案を行なう。また、新しいネットワークを用いて実際に解析を行なった結果を示し、従来の方法での結果と比較を行なう。

2 従来の方法とその問題点

従来の形態素解析ネットワークでは、各ユニット u_i は与えられた文に含まれる可能性のある単語を表すものとし、エネルギー関数としては以下の4関数の一次結合 $E = pE_L + qE_O + rE_G + sE_I$ を採用していた。

1. 選択された各単語の文字数の合計と、もとの文の文字数が等しいときに最小値をとる関数

$$E_L = \left(\sum_{i=1}^{n} L_i u_i - N \right)^2 \tag{1}$$

ただし L_i はユニット u_i で表される単語の文字数で N はもとの文の文字数を示す

2. 選択されたどの2単語も、同一の部分文字列を共有していない場合（単語同士の間で文字の重なりがない場合）に最小値をとる関数

$$E_O = \sum_{i=1}^{n} \sum_{j \neq i} O_{ij} u_i u_j \tag{2}$$

ただし

$$O_{ij} = \begin{cases} 1 & u_i \text{ と } u_j \text{ で表される単語} \\ & \text{が重なりあっている場合} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

3. 隣接する単語の組合せがすべて文法的に接続可能な場合に最小値をとる関数

$$E_G = \sum_{i=1}^{n} \sum_{j \neq i} G_{ij} u_i u_j \tag{4}$$

ただし

$$G_{ij} = \begin{cases} 1 & u_i \text{ と } u_j \text{ の接続が文法的に} \\ & \text{正しくない場合} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

4. 選択された単語列に含まれる自立語の数が少ないほど小さな値をとる関数

$$E_I = \sum_{i=1}^{n} I_i u_i \tag{6}$$

ただし

$$I_i = \begin{cases} 1 & \text{ユニット } u_i \text{ に対応する単語が} \\ & \text{自立語の場合} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

従ってユニット間の結合の強さ w_{ij} およびユニットのしきい値 θ_i は

$$\begin{cases} w_{ij} &= -2(pL_iL_j + qO_{ij} + rG_{ij}) \\ \theta_i &= p(L_i^2 - 2NL_i) + sI_i \end{cases} \tag{8}$$

となる。

また、E_L, E_O, E_G, E_I の係数 p, q, r, s、および焼きなましのスケジュールは以下の通りであった。

$$p = 200, \quad q = 1800, \quad r = 1500, \quad s = 300 \tag{9}$$

$$T(t) = \frac{T_0}{1 + t/\tau} \tag{10}$$

ただし

$$T_0 = 5000, \quad \tau = 500 \tag{11}$$

この方法の第一の問題点として、ユニット間に好ましくないリンクが存在するということがあげられる。

式(8)からわかるように、ユニット間リンクは L_i, O_{ij}, G_{ij} から作られる。このうち O_{ij} は重なりあっている単語のみに影響を与え、G_{ij} は隣接した単語にのみ影響を与えるので、文の構造が一次元であるという事実を反映しているといえる。しかし、L_i は単にユニット u_i が表す単語の文字数であるから、w_{ij} が L_i と L_j から構成されるということは文の上で遠く離れた2単語間にもリンクが張られることを意味し、従って文の一次元構造を無視しているといえる。このようなリンクの存在は好ましくない。

文を単語列に分解するための制約としては、近隣の単語間のみに相互作用を許すような関数を考えるべきであろう。

第二の問題点としては、各制約関数の係数 p, q, r, s、ネットワークの初期温度 T_0、時定数 τ がいずれも大きいということである。

制約関数の係数が大きくなると、ユニットの値の変化に伴うネットワークのエネルギーの変化量が増大する。これはエネルギーの極小値のくぼみが深くなることを意味してお

On Neural Networks for Japanese Morphological Analysis
TAKAHASHI Naoto, Shuichi ITAHASHI and Yuzo HIRAI
University of Tsukuba

り、従って一度極小値につかまるとなかなか脱出できなくなることになる。エネルギーの深いくぼみから脱出できるようにするためにはネットワークの温度を高くする必要がある。ネットワークを高い温度から急速に冷やすとエネルギー最小値以外の状態に収束する可能性が高くなるが、高い温度からゆっくりと冷やすのではネットワークが収束するまでに長い時間がかかる。

従って制約関数の係数はできるだけ小さい方が望ましいといえる。

3 新しいネットワーク

今回の実験では、前節で述べた問題点を改善したネットワークとして以下のようなものを採用した。

1. 制約関数の変更
   以前のネットワークでは、もとの文を洩れも重なりもなく分解するために E_L と E_O という2つの関数を用いていたが、新しいネットワークではこれらを廃止し、代わりに以下の関数を採用する。

$$E_C = \sum_{k=1}^{N} \left(\sum_{i=1}^{n} C_{ki} u_i - 1 \right)^2 \tag{12}$$

ただし n はユニットの総数、N はもとの文の文字数である。また C は $n \times N$ 行列で、

$$C_{ki} = \begin{cases} 1 & \text{ユニット } u_i \text{ で表される} \\ & \text{単語がもとの文中で } k \\ & \text{番目の文字を含む場合} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

と定義される。

式(12)は、もとの文に含まれる各文字がちょうど一回ずつ現れるような単語の組み合わせが選択された時に最小値0をとる。この関数は k 番目の文字を共有しているユニットの間にのみ相互作用を許すので、E_L を用いた時のような問題は生じない。

関数 E_G および E_I は以前のものをそのまま用いる。

2. 各種定数
   前節で述べたように、制約関数の係数は小さく抑える方が望ましい。係数の実際の値は経験的に良いものを採用するが、その際に値が大きくなり過ぎないように注意する。すなわち、ある関数による制約が弱いという結果が出た場合には、その関数の係数を大きくするよりは他の関数の係数を小さくする方向で対処する。

このようにして作られた新しいネットワークにおけるユニット間リンクの重み w_{ij} およびユニットのしきい値 θ_i は以下のようになる。

$$\begin{cases} w_{ij} &= -2 \left(p \sum_{k=1}^{N} C_{ki} C_{kj} + r G_{ij} \right) \\ \theta_i &= -p \sum_{k=1}^{N} C_{ki} + s I_i \end{cases} \tag{14}$$

となる。

また、制約関数の係数および焼きなましのスケジュールは以下のように決定した。

$$p = 4, \quad r = 2, \quad s = 3 \tag{15}$$

$$T(t) = \frac{T_0}{1 + t/\tau} \tag{16}$$

ただし

$$T_0 = 5, \quad \tau = 10 \tag{17}$$

ユニットとしては2値ユニットを用いた。

この新しいネットワークを用いて以前の実験で用いたのと同一の文を解析した結果を Figure 1 に示す。新しいネットワークでの解析成功率は約95%で、以前よりも約5%向上した。またネットワークが収束するまでのステップ数は以前の1/10から1/100で済んだ。

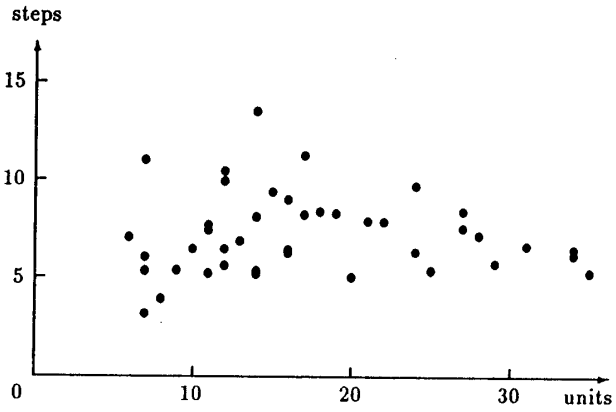

Figure 1: ネットワークが収束するまでのステップ数とそのときのユニット数との関係。ただしすべてのユニットが平均して1回ずつ発火するのに要する時間を1ステップとする。

4 おわりに

日本語形態素解析を相互結合型ニューラルネットワークを用いて行なう方法とその実験結果について述べた。以前のネットワークにおける制約関数を一部変更し、さらにパラメータを調節した結果、従来の1/10～1/100のステップ数で約95%の解析成功率を得ることができるようになった。

References

[1] 高橋直人, 板橋秀一：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, 情報処理学会第40回全国大会 4F-7, pp.464-465, 1990.

[2] Takahashi, N. and Itahashi, S. : Japanese Sentence Analysis Utilizing Mutually Connected Neural Network, *Proceedings of PRICAI'90*, pp.257-262, 1990.

[3] 高橋直人, 板橋秀一： ニューラルネットによる日本語形態素・係り受け解析, 情報処理学会研究報告, 90-NL-80, 1990.

[4] Hopfield, J. J. and Tank, D. W. : Neural Computation of Decisions in Optimization Problems, *Biological Cybernetics*, 52, pp.141-152, 1985.