日本語形態素解析におけるニューラルネットの利用

1Q－4

高橋 直人

電子技術総合研究所

1　はじめに

形態素解析時に語彙的曖昧性が生じた場合，文脈と関連の深い単語はそうでない単語よりも正解となる可能性が高いといえる．本稿で述べるニューラルネットは入力された文脈と関連の深い単語を出力するように設計されているので，これを形態素解析システムに組み込めば解析精度を向上させることが可能である．

ニューラルネットには学習能力が備っているので，文脈と単語の関連性を例文に基づいて学習させることができる．またニューラルネットの持つ汎化能力を利用することで，既学習の文脈に対してのみならず未学習の文脈に対しても適切な出力を得ることが可能である．

2　ニューラルネットの構成

実験に使用したニューラルネットは単純な3層フィードフォワード型である．入力層と中間層の間および中間層と出力層との間は全結合されている．入力層と出力層の間を直接結合するリンクは存在しない．入力層および出力層においては1ユニットが1単語を表現するようにした．

入力層には文脈を与える．今回の実験では文脈を「解析しようとする部分の直前8自立語からなる系列」と定義し，位置的に近い自立語（に対応する入力ユニット）ほど大きな入力値を受けるようにした．具体的には，直前の自立語には4.0，その一つ前の自立語には3.5，さらにもう一つ前の自立語には3.0，というように4.0から0.5きざみで0まで減少する値を順に入力した．それ以外の入力ユニットに対する入力はすべて0とした．

各出力ユニットの理想出力値は，そのユニットが表している単語及び入力自立語列に基づき以下のように定義した．

- 解析すべき文字列の先頭部分とマッチし，かつ直前の単語と品詞的に接続可能であるような単語（以下では候補単語と呼ぶ）のうち，入力層に与えられた自立語列に後続する用例の存在する単語の理想出力は1．
- 候補単語ではあるが，入力層に与えられた自立語列に後続する用例のない単語の理想出力は0．
- 候補単語以外の単語は任意の値を出力して構わない．

Applying a Neural Network to Japanese Morphological Analysis
TAKAHASHI Naoto
Electrotechnical Laboratory

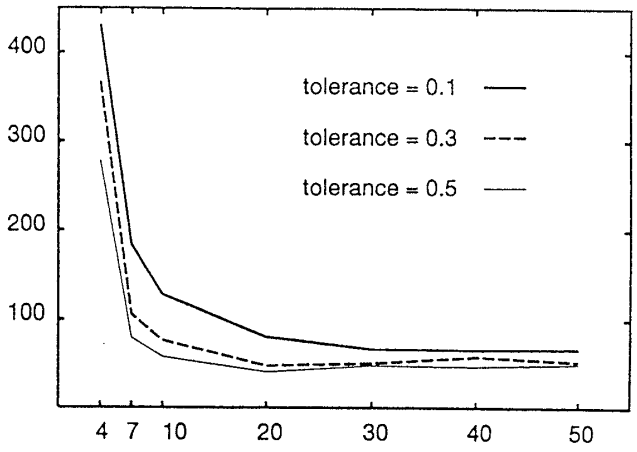

図1: 訓練用3つ組みの学習に必要とされたステップ数

入力自立語列に後続する用例が存在するか否かは，与えられたコーパス（後述）に基づいて判断した．たとえ意味的に後続する可能性があっても，与えられたコーパス中にそういった用例が現れない場合は後続しないものと判断した．

3　実験とその結果

3.1　学習実験

最初の実験ではニューラルネットに訓練用データを与え，正しい後続単語が出力されるように学習させた．訓練用データの基になるコーパスとしては，文献[1]の中から「第12章 市民社会の成長 §2. フランス革命とナポレオン」に出てくる文章を使用した．総文字数は5708，異なり単語数は871，文の数は108である．このコーパスから1）先行自立語列，2）実際の後続単語，3）後続単語以外の候補単語，という3つ組を作成した．さらに同一の先行自立語列を持つ3つ組は1個にマージした．このようにして最終的に得られた訓練用データは1718組になった．

ニューラルネットの学習には標準的なバックプロパゲーションを用いた．学習の係数ηおよび慣性αの値はそれぞれ0.1および0.9に固定した．リンクの重みの初期値は$[-0.1, +0.1]$の間でランダムに設定した．

中間層のユニット数およびtolerance（理想出力と実際の出力との誤差）を変化させたとき，すべての訓練用データを学習し終わるまでに必要とされるステップ数がどう影

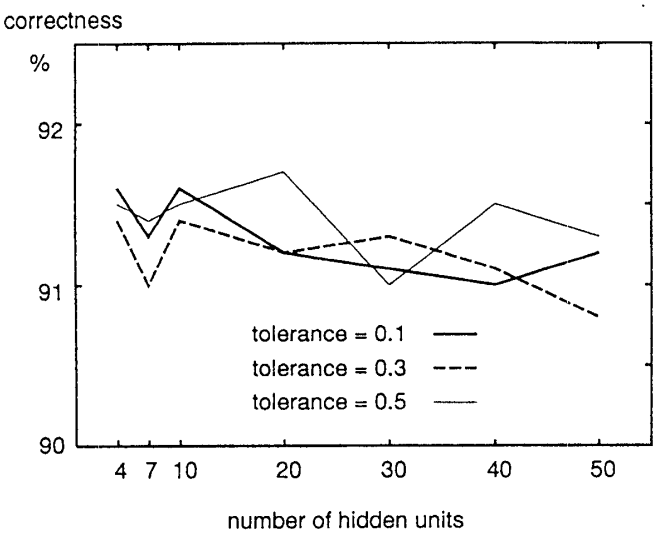

図 2: 漢字仮名交じり文に対する語彙的曖昧性解消テストの正解率

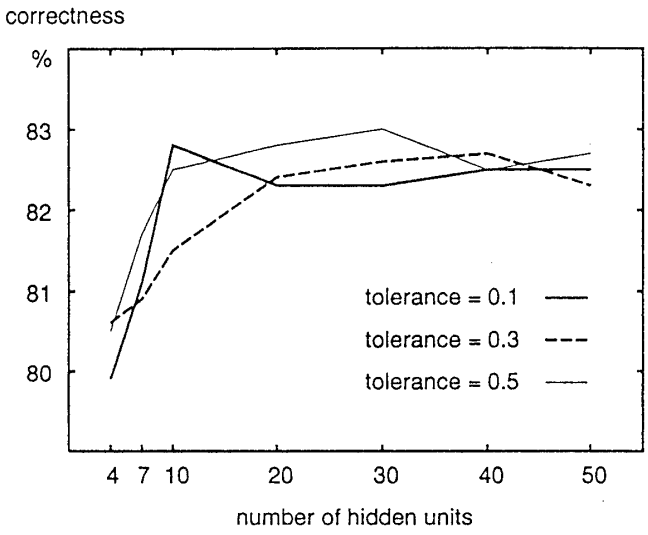

図 3: 平仮名べた書き文に対する語彙的曖昧性解消テストの正解率

響を受けたかを図1に示す．なお，中間層のユニット数が4未満になると，すべての訓練用データを学習することはできなくなった．

3.2 汎化実験

次に学習の完了したニューラルネットがどの程度の汎化能力を持つかを調べる実験を行った．訓練用コーパスと同一の分野でしかも字面の異なる文章として，文献[2]の中から「フランス革命」「ナポレオン(1世)」「ウィーン会議」の3項目の解説文を採用した．総文字数は 1902，異なり単語数は 382，文の数は 33 である．

このコーパスから学習時と同様の方法で3つ組を作成し，その中の先行自立語列を学習済みのニューラルネットに順に入力した．このとき，もし各候補単語を表している出力ユニットの中で，実際の後続単語を表している出力ユニットからの出力値が最大になっていれば正解，そうでなければ不正解という判断基準で語彙的曖昧性解消のテストを行った．

テスト用コーパスを原文のままの漢字仮名交じり文としてテストした場合の正解率を図2に，またテスト用コーパスを平仮名べた書きに変換してテストした場合の正解率を図3にそれぞれ示す．入力が漢字仮名交じり文の場合，中間ユニット数が少ないほど正答率が高くなる傾向にある．これに対して入力が平仮名べた書きの場合は，中間ユニット数が多いほど正答率が高なる傾向が見られる．

4 おわりに

ニューラルネットを用いて文脈に沿った単語を選択するための方法と，シミュレーションによる実験結果を示した．本稿で述べたニューラルネットを日本語形態素解析システムに組み込むことで，解析精度の向上が期待できる．

今回の実験では単語の表現に局所表現を用いたが，一般的には局所表現よりも分散表現の方が汎化の点で有利であると言われている．単語の分散表現を自動的に生成するための手段としては文献[3]が興味深い．ただし，今回の枠組の中で分散表現を用いる際には，1) 複数の単語を，2) 同時に，3) 順位付けで，出力するための方法を工夫する必要がある．今後はこの点を考慮しつつ，解析の対象をより広範囲に広げてゆきたい．

参考文献

[1] 村川堅太郎，江上波夫，山本達郎，林健太郎：詳説世界史 (再訂版)，pp. 221 − 231，山川出版社 (1982).

[2] 歴史教育研究会編：世界史事典，旺文社 (1992).

[3] Miikkulainen, R. and Dyer, M. G. : Natural Language Processing With Modular PDP Networks and Distributed Lexicon, Cognitive Science, 15, pp. 343 − 399 (1991).

[4] 高橋直人：後続部分予測機能を持つ日本語解析システム，信学技報 NCL92 − 40 (1992).

[5] 上原龍也，野上宏康，齋藤佳美，相原義弘，天野真家: かな漢字変換における共起情報の適用方式の拡張，情報処理学会第 44 回全国大会論文集，Vol.3, pp. 187 − 188 (1992).

[6] 山本喜大，久保田淳市：共起グループを用いたかな漢字変換，情報処理学会第 44 回全国大会論文集 Vol.3, pp. 189 − 190 (1992).