神経回路網モデルを用いた日本語単語の多義性解消

菊井 真*，橘 公一*，古家 美佳*，村田 真樹**，馬 青*
* 龍谷大学理工学部
** 通信総合研究所けいはんな情報通信融合研究センター
qma@math.ryukoku.ac.jp

1 はじめに

日本語単語の多義性解消は自然言語処理の多くの場面で必要となる基礎技術である．例えば，

①「太郎が壁にかける．」
②「太郎が花子にかける．」

という二つの文がコンピュータに与えられたとする．コンピュータに①の文が「太郎が壁に（絵など）を掛ける」，②の文が「太郎が花子に（期待など）をかける」という意味の解を正しく導き出させるには，まず，「かける」という単語の意味をそのまわりの文脈情報を用いて正しく解析しておかなければならない．実際，辞書で「かける」という単語を調べると，「掛ける」「欠ける」「賭ける」「駆ける」など数十個の同じ読みの単語が得られる．このように，「かける」という単語には複数の意味があるので，その意味上の曖昧性を解消しておかないと，①の文，②の文が誤って解釈される可能性がある．そうすると，例えば自動英訳をしようとしたら，翻訳結果が誤ってしまう．そこで，多くの自然言語処理課題においては，まず単語の多義性を解消し，コンピュータに①の文，②の文の「かける」をそれぞれ正しく認識させる必要がある．また，意味の少ない単語の場合はまだましであるが，高頻度に用いられる平易な単語ほど語義の数が多く，この問題は深刻となる．以上のことから多義性解消の重要性がわかる．

日本語単語の多義性解消に関する研究は以前よりサポートベクターマシンなど様々な手法を用いて行われてきた[1]．しかしながら，神経回路網は汎化能力の高い学習機械であるにもかかわらず，我々の知る範囲ではそれを用いた，日本語単語の多義性解消の研究はこれまで非常に少なかった．この数少ない研究の中，高橋の研究を挙げることができる[2]. しかし, 高橋の研究では，実験に用いたデータの規模が小さかった上，すべての単語を同一の神経回路網で取り扱っていた．そのため，神経回路網の収束性の問題から，取り扱うデータの大規模化が困難と思われる．また，多義性解消に用いた文脈情報も改善の余地が残っている．

本稿では，神経回路網モデルを用いた単語多義性解消の新しい手法を提案する．提案手法では, 1 つの多義語に 1 つの神経回路網を割り当て，取り扱うデータの大規模化を可能とした．また，多義性解消に必要な情報としては，形態素情報や構文情報など，最近の多義性解消の研究によく用いられているものを用いた．さらに，実験には SENSEVAL-2 日本語辞書タスク[3]のデータを使用し，これまで提案されてきた既存手法との比較を容易にした．

2　問題設定

本研究では CRL により加工された SENSEVAL-2 日本語辞書タスクのデータを使用した．ここで，まず，SENSEVAL-2 日本語辞書タスクの概要について述べる．この日本語辞書タスクには，評価用データとして 100 単語（名詞 50 個，動詞 50 個）についてそれぞれ 100 事例，合計 10000 事例が与えられている．学習用データとしては RWC コーパスが与えられている．このコーパスは毎日新聞（1994 年）の 3000 個の記事中の単語に，岩波国語辞典に基づいて定義された語義を付与したコーパスである．このタスクの目的は，この語義をその単語のまわりの情報などを用いて推定することである．

我々はこのタスクの評価用データ 10 個（名詞 5 個，動詞 5 個）についてのみの小規模な多義性解消の実験を行った．また，精度は SENSEVAL-2 のホームページより取得できる scorer2 という評価用プログラムは使わず，我々自身で作成した評価用プログラムによって算出した．

3　素性（解析に用いる情報）

他の機械学習手法と同様，我々の提案する手法も素性（解析に用いる情報）を定義しなければ用いることができない．本節では我々が実験で使った 14 個の素性について定義する．

- 文字列素性
  - ― 解析する形態素自身の文字列
- RWC 形態素素性
  - ― 解析する形態素自身の RWC コーパスの品詞情報，品詞細分類情報
  - ― 解析する形態素の直前の形態素の単語の分類語彙表（国立国語研究所 1964）の 5 桁，品詞情報
  - ― 解析する形態素の直後の形態素の単語の分類語彙表の 5 桁，品詞情報
- JUMAN 形態素素性
  コーパスを JUMAN（黒橋 長尾 1998）で形態素解析し，その結果を素性として利用する．
  - ― 解析する形態素自身の JUMAN の解析結果の品詞情報，品詞細分類情報
  - ― 解析する形態素の直前の形態素の単語の分類語彙表の 5 桁，品詞情報
  - ― 解析する形態素の直後の形態素の単語の分類語彙表の 5 桁，品詞情報
- 同一文内共起素性
  コーパスを JUMAN で形態素解析し，その解析結果の形態素列を素性として利用する．
  - ― 同一文中の各形態素（単語）の分類語彙表の 5 桁

4　神経回路網による多義性解消

神経回路網として３層パーセプトロンを用いた．提案手法では，今後の取り扱うデータの大規模化を念頭に，１つの多義語の解析に１つの３層パーセプトロンを割当てることにした．

解析データは３節にも述べたように，文字情報であるので，３層パーセプトロンの入力とするためには符号化（数値要素を有

すベクトルに変換）する必要がある．そこで，各素性をそのとりうる値の種類数を次元とした，0-1 の 2 値サブベクトルに変換する．そして，各単語の解析に用いた 14 個の素性に対応する 14 個のサブベクトルから構成されたものをその単語の入力データとする．また，各単語の教師データはその単語がもちうる語義の数を次元とした 0-1 の 2 値ベクトルである．

5　実験

5．1　実験データとパラメータ

本実験で使用した単語 10 個と各単語の学習用データの数，評価用データの数，語義の数は表１に示す．但し，使用した素性の数が少なかったため，学習データに，入力データが同じにもかかわらず教師データが異なる「衝突」が生じた．このような衝突データが多いと，３層パーセプトロンの学習ができなくなる恐れがある．今回の実験ではこのような衝突データを予め取り除いた．

用いた３層パーセプトロンの各パラメータの値は以下のように決めた．学習率と慣性率は 0.9 と 0.1 にそれぞれ設定した．目標誤差値は 0.01，学習回数は 100 万回までとした．入力層のユニット数は各単語の入力データの次元数，出力層のユニット数は各単語の教師データの次元数とした．そして，パーセプトロンの汎化能力にもっとも影響を与える中間層のユニット数はクロスバリデーションで選定した．我々は各単語の学習用データを先頭から「9：1」「1：9」「4：1：5」「6：1：3」「2：1：7」に分割し，9 割を学習用データ，1 割を評価用データとして用いた（5 種類のデータセットができる）．中間層の値（ユニット数）を「（入力データの次元数）×1」「（入力データの次元数）×0.5」「（入力データの次元数）×0.25」「（入力データの次元数）×0.125」「（入力データの次元数）×2」（5 種類）に変えながら，この分割したデータセット毎に３層パーセプトロンで学習とテストの予備実験（5×5=25 回）をし，最も精度の良い中間層のユニット数を本実験で使用した．各単語の入力層-中間層-出力層のユニット数は表２に示す．

表１ 各単語とデータの数

単語	学習用データの数	評価用データの数	語義の数
場合	206	100	2
近く	169	100	3
関係	426	100	3
問題	802	100	4
市民	136	100	3
図る	96	100	2
決める	285	100	3
超える	123	100	5
認める	263	100	4
狙う	76	100	2

5．2　実験結果

実験の結果は表３に示す．我々の実験では単語「市民」を除けば高い精度を得ることができた．「市民」を含んだ全単語の平均解析精度は 0.85 に達した．この結果は SENSEVAL-2 のコンテストでの最も良い実験結果よりもよかった．但し，我々の実験は 10 単語のみの小規模なものなので，本当に我々の用いた手法の方が有効であるとは断言できない．

表２：各単語に用いる 3 層パーセプトロンの各ユニット数

単語	入力層のユニット数	中間層のユニット数	出力層のユニット数
場合	77	77	2
近く	148	37	3
関係	286	71	3
問題	382	47	4
市民	107	107	3
図る	67	8	2
決める	99	12	3
超える	105	13	5
認める	124	16	4
狙う	64	8	2

表３：各単語の精度

単語	精度
場合	0.79
近く	0.88
関係	0.84
問題	0.97
市民	0.50
図る	0.91
決める	0.91
超える	0.82
認める	0.89
狙う	0.99
	平均　0.85

６　おわりに

我々の実験は小規模ではあったが，神経回路網が日本語単語の多義性解消に有効な手段として利用できることを示した．今後はまず，大規模な実験で提案手法の有効性を検証する予定である．そして，他の手法との融合を図り，高性能な多義性解消システムの構築を目指す.

参考文献

[1] 村田真樹，内山 将夫，内元清貴，馬青，井佐原均：SENSEVAL2J 辞書タスクの CRL の取り組み－日本語単語の多義性解消における種々の機械学習手法と素性の比較－，自然言語処理，Vol. 10, No.3, pp. 115-133, 2003

[2] 高橋直人：階層型ニューラルネットによる語彙的曖昧性の解消，情報処理学会論文誌，Vol. 36, No. 9, pp.2102-2112, 1995.

[3] 白井清昭：SENSEVAL-2 日本語辞書タスク，自然言語処理 Vol.10, No. 3, pp. 3-24, 2003.