ニューラルネットワークによるインターネット配信ニュースの記事切り出し

畑田 稔†

インターネットの普及により，日々大量のテキストデータをメール配信サービスにより入手できるようになった．大量の情報を高速に処理するためには，深い意味解析に立ちいらず，なるべく個々のデータ形式にとらわれない方法で記事を切り出すことが必要である．本報告では，ニューラルネットワーク技術を適用した記事切り出し方法を提案する．7つの配信サービスに適用し，その有効性を検証した．平均で98.5%という高い認識率が得られた．さらに，ルールで補正することにより，認識率を99.3%にまで高めることができた．また，学習パターンに冗長性があることに着目して，学習の進捗によって冗長パターンを検出し，この学習をスキップすることにより，大幅な高速化（約5倍）を達成した．

Article Segmentation of Internet Delivery News Using Neural Network

MINORU HATADA†

With the spread of the Internet, large volumes of text-based data can be obtained via mail distribution services. Processing large volumes of information at high speeds requires an article segmentation method that does not perform in deep semantic analysis and is not limited to any particular data format. This paper proposes an article segmentation method that employs neural network technology. Its effectiveness was verified by applying it to seven distribution services. On average, a high recognition ratio of 98.5% was obtained. Furthermore, this recognition ratio may be improved to 99.3% through rule-based correction. Because the set of learning patterns contains redundancy, eliminating these redundant learning patterns resulted in a sharp increase in speed (approximately five fold).

1. はじめに

近年，インターネットの普及により，電子メールによるニュース配信サービスが広がっている．2，3年前にスタートしたものが多いが，年とともに充実し，ニュースだけでなく，コラム，広告など内容が多彩となってきた．

データ形式は，現在のところ，文書構造にかかわるタグなどをいっさい含まないテキストデータである．新聞並みに日々目を通すような使い方では特に不便さはない．しかし，受信するメールが増大すると，記事を目で追うのは煩わしい．まず，タイトルに目を通し，読んでみようと思うものがあれば，タイトルをクリックすると，その記事の本文が表示されるといったユーザインタフェースが望まれる．また，検索機能を使って，過去数か月分についてある製品に関する記事をすべて抽出しようという場合も，1つ1つの記事が独立して取り出せないと不便である．これらのニーズに応えるためには，まず，記事の切り出しが必須となる．

記事の切り出し方法としては，構造の特徴をプログラム化する方法がある．これはニュースソース（電子新聞）ごとに異なる多くの条件判定を含むプログラムを開発するものである．このようなニュース配信サービスは発展段階にあるため，構造が変化してゆく．このため，対象とするニュースソースが増えると，条件判定によるプログラム方式は開発および保守の負担が大きくなる．本論文では，文を構成する各行の特徴（属性）を数値表現し，これをニューラルネットワークに学習させることにより，記事の切り出しを行う方法を提案する．

ニューラルネットワークモデルはあらゆるニュースソースに対して共通のものを使用し，学習サンプルのみをニュースソースごととする．このため，ニュー

† 富山県立大学工学部電子情報工学科
Department of Electronics and Informatics, Faculty of Engineering, Toyama Prefectural University

スソースが増えたり，あるニュースソースの構造が変わったりした場合には，新たな学習サンプルを用いて学習することにより，このニュースソース用の結合荷重および出力関数傾きを得ることにより対応できるという利点を有している．事例を用いて，その有効性を評価した．

文章の構造解析に関しては，論説文を対象として，段落間のつながりを助詞，接続語句，時制などから重回帰分析で判定して，分割位置を求める研究がある．文章全体を5段落に分割するケースでは65%程度の適合率が得られている[3]．また，雑誌の目次と各論文の先頭ページをイメージ入力し，タイトル，著者などのブロックの切り出しと文字認識をニューロ・ファジー・システムで行う研究があり，目次および論文先頭ページの切り出しでは，それぞれ99%および94.3%という高い認識精度が得られている[6]．

ニューラルネットワークの学習の高速化に関しては，学習の遅れているパターンを段階に分けて多く提示することにより，学習進捗のばらつきを小さくし，学習時間を短縮する方法が提案されている．手書き数字の認識問題で30%の高速化が報告されている[7]．これに対して，本研究では，記事切り出し問題での学習パターンには大きな冗長性があることに着目して，学習の進捗によって，冗長な学習パターンを検出し，この学習処理をスキップすることにより，およそ5倍の高速化を達成した．

2. 記事切り出し

2.1 ニュース文の特徴

ニュース文の一般的な構造を図1に示す[8),9)]．広告など（広告，お知らせなど）は，存在しないケースもある．また，先頭記事の前および最終記事の後の広告などはそれぞれヘッダおよびフッタに含めて扱う．記事の例を図2に示す．これは擬似的な記事であり，実際の記事では，1行は全角で40字が標準である．

広告などと通常の記事ではスタイルが異なっているのが普通である．本研究の目的はこのようなニュース文から記事1，記事2，記事Nを切り出すことである．

記事の始まりと見出し，記事の終わりの主な特徴は次のとおりである．

記事の始まりと見出しの特徴

- その前が空行であることが多い．
- 最初が見出し（1行とは限らない）であり，本文との間に空行があることが多い．
- 記事の先頭の行に，分類語が置かれるケースがある．

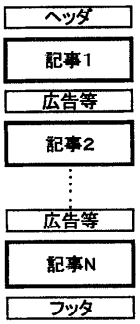

図1 ニュース文の構造
Fig. 1 Structure of a news article.

[ニュース]

○ A社、ノートブックを大幅値下げ

米A社は2日(米国時間)、ノートブックラインの大幅値下げを発表、一部のモデルの価格を999ドルに引き下げた。
今回の値下げは、ローエンドからハイエンドまでの全モデルを対象としている。
なお、値下げに加えてRAM、CD-ROMドライブの無料オプションも発表された。
(InternetNews)

図2 記事サンプル
Fig. 2 Sample of article.

- 見出しの頭には，記号文字（○，◇など）が付されることがよくある．
- 見出しは通常行の幅よりも短く，末尾に句点がつかない．
- 分類語は通常見出しよりもさらに短く，末尾に句点がつかない．
- 記事全体あるいは見出しは罫線（----------など同一文字の繰返し）で囲まれることがある．

記事の終りの特徴

- その後が空行であることが多い．
- 「複写禁止」とか「Reported by ...」といった決まり文句で終わることが多い．
- 最後の行に出典を示し，これは【】,() などの括弧で括られていることが多い．

もちろん，記事本文中にも空行が多く，また，箇条書きなどでは先頭に記号文字が付き，行末に句点が付かないことが多い．したがって，1つの行の特徴では，記事の始まり，見出し，あるいは記事の終わりかどうかを決められず，前後の数行の状態を見て判断する必要がある．

2.2 階層型ニューラルネットワーク

ニューラルネットワークの代表的なアルゴリズムの1つとしてバックプロパゲーション（BP）法がある[1),5)]．

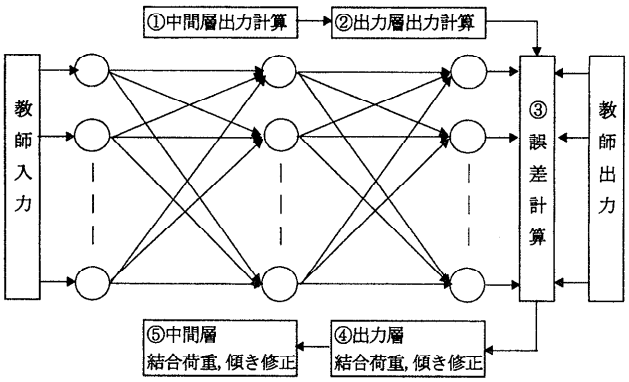

図 3 階層型ニューラルネットワーク
Fig. 3 Multi-layered neural network.

BP 法は，図 3 に示すような階層型ニューラルネットワークを対象として，二乗誤差和を評価関数として誤差を逆伝搬させながら結合荷重を修正することにより学習を進める．本研究では，これを発展させた拡張 BP 法[2] を適用する．拡張 BP 法の概略は次のとおりである．

i 番目の入力ユニットの出力信号を x_i，j 番目の中間ユニットの入力および出力信号をそれぞれ u_j および y_j，k 番目の出力ユニットの入力および出力信号をそれぞれ v_k および z_k とすると，ニューラルネットワークの状態は

$$u_j = \sum_{i=0}^{N^I} \omega_{ij}^M x_i \tag{1}$$

$$v_k = \sum_{j=0}^{N^M} \omega_{jk}^O y_j \tag{2}$$

$$y_j = \frac{1}{1+\exp(-\beta_j^M u_j)} \tag{3}$$

$$z_k = \frac{1}{1+\exp(-\beta_k^O v_k)} \tag{4}$$

で記述される．ここで，N^I，N^M および N^O はそれぞれ入力層，中間層および出力層のユニット数である．ω_{ij}^M および ω_{jk}^O はそれぞれ，入力–中間層および中間–出力層の各ユニット間の結合荷重である．入力層，中間層の 0 番目のユニットは，しきい値を結合荷重に含めて統一的に扱うための仮想的なユニットである．$x_0 = 1$，$y_0 = 1$ であり，$-\omega_{0j}^M$ および $-\omega_{0k}^O$ は，それぞれ j 番目の中間ユニットおよび k 番目の出力ユニットのしきい値にあたる．

拡張 BP 法は，評価関数

$$E = \frac{1}{2}\sum_{k=1}^{N^O}(T_k - z_k)^2 \tag{5}$$

をもとに，出力誤差を逆伝搬させながら結合荷重，出力関数傾きを修正することにより学習を行う．ここで，T_k は k 番目の出力ユニットへの教師信号を表す．

出力層では，結合荷重および出力関数のパラメータの修正量は

$$\Delta\omega_{jk}^O = \eta\delta_k^O\beta_k^O y_j \tag{6}$$

$$\Delta\beta_k^O = \varepsilon\delta_k^O v_k \tag{7}$$

となる．ここで

$$\delta_k^O = (T_k - z_k)z_k(1-z_k) \tag{8}$$

である．

中間層に対しては，

$$\Delta\omega_{ij}^M = \eta\delta_j^M\beta_j^O x_i \tag{9}$$

$$\Delta\beta_j^M = \varepsilon\delta_j^M u_j \tag{10}$$

となる．ここで

$$\delta_j^M = y_j(1-y_j)\sum_{k=1}^{N^O}\delta_k^O\beta_k^O\omega_{jk}^O \tag{11}$$

である．

実際のプログラムでは，振動を減らし，学習の収束を早めるために，式 (6) の代わりに

$$\Delta\omega_{jk}^O(t+1) = \eta\delta_k^O\beta_k^O y_j + \alpha\Delta\omega_{jk}^O(t) \tag{12}$$

を用いる[5]．ここで，α は 1 より小さい正の定数，t は修正の回数を表す．式 (7)，式 (9) および式 (10) についても同様である．

2.3 記事切り出しのニューラルネットワークモデル

本節では，記事切り出し問題をどのようにモデル化するかについて述べる．まず，文の基本単位を行とする．行は複数の属性値を持つ．ここで，属性は，2.1 節であげた記事の始まり，見出し，および記事の終わりに現れやすい特徴を 2 値 (0/1) 化したものである．

3 章で述べる評価実験で用いた行属性を以下に示す．ここで，行の長さは英数字など半角文字数で表し，漢字など全角文字は 2 文字として扱う．

(1) 空行属性：空行のとき1，そうでないとき0.

(2) 出典・分類語属性：空行でなく，行の長さが20文字未満で，かつ，行末が句点でないとき1，そうでないとき0．記事の末尾に出典が記載されることが多いが20文字前後のときが多い．分類語は単語1つのときは10文字前後が多い．このため，上限を20文字とした．

(3) 見出し語属性：行の長さが20文字以上，60文字未満で，かつ，行末が句点でないとき1，そうでないとき0．実際の見出しの8割程度が60字未満であることから，この上限を選定した．この値を大きくしすぎると，記事本文中にこの条件を満たすものが増加し，認識率の悪化を招く．

(4) 見出し記号属性：行の先頭文字が記号文字（○，◇など）のとき1，そうでないとき0.

(5) 括弧記号属性：行の先頭文字が括弧記号文字（【，[，(など）のとき1，そうでないとき0.

(6) 罫線属性：行が「----------」,「☆★☆★☆★」など同じパターンの繰返し☆のとき1，そうでないとき0.

(7) 著者属性：「Reported by ...」といった著者を表す決まり文句を含むとき1，そうでないとき0.

(8) 著作権属性：「複写禁止」といった著作権を表す決まり文句を含むとき1，そうでないとき0.

行の属性の数を N_A で表す．ニューラルネットワークに与えるデータは，対象行を中心として前後の N_L 行の属性も含める．全文の先頭部あるいは末尾部で，前後の行が N_L 行に満たないときは，仮想的に空行が存在するものとする．一方，出力層のデータは対象行が「記事の始まり」,「記事の終わり」，および「見出し」（実際のデータでは，まれに見出しが2行以上のケースがあるが，その場合，先頭の行のみを見出しとして扱う）であるか否かを表す3次元の2値データとする．したがって，第 n 行の属性ベクトルを ATT(n)，出力データを TAG(n) で表すと，記事切り出しにおけるニューラルネットワークは，**図4**のように表現でき，入力ユニット数が $(2N_L+1)\times N_A$ で，出力ユニット数が3である．

本論文では，ニューラルネットワークに教師データとして与えるものを学習サンプル，学習結果を用いて認識精度を評価するものを評価サンプルと呼ぶ．学習サンプルおよび評価サンプルには，あらかじめ「記事の始まり」,「記事の終わり」および「見出し」の位置にタグ <art>, </art> および <ti> を挿入しておく．

☆「====解説==========」のように，文字を含むケースも罫線として扱う．

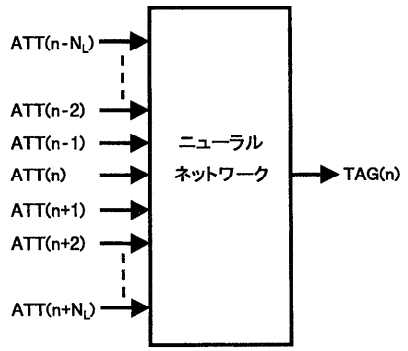

図4 記事切り出しニューラルネットワーク
Fig. 4 Neural network for article segmentation.

```
<art>                                                    1
[ニュース]                                               2
---------------------------------------------------      3
<ti>○ A社、ノートブックを大幅値下げ                      4
---------------------------------------------------      5
                                                         6
  米A社は2日(米国時間)、ノートブックラインの大幅値下げを発表、一部のモデルの価格を999ドルに引き下げた。   7 8 9
  今回の値下げは、ローエンドからハイエンドまでの全モデルを対象としている。   10 11
  なお、値下げに加えてRAM、CD-ROMドライブの無料オプションも発表された。   12 13
(InternetNews)                                           14
</art>                                                   15
```

図5 タグの挿入
Fig. 5 Insertion of tags.

図2の記事例に対してタグを挿入した結果を**図5**に示す．ここで，行末の数字は行番号を表す．たとえば，第4行では ATT(4) = [0,0,1,1,0,...]，TAG(4) = [0,0,1] となる．また，第6行では ATT(6) = [1,0,0,0,0,...]，TAG(6) = [0,0,0] となる．

学習アルゴリズム

[ステップ1]

学習サンプルを読み込み，各行の属性ベクトルを求め，配列に格納する．同様にタグ情報も行単位の配列に格納する．これらをそれぞれ属性配列，タグ配列と呼ぶ．

学習サンプルにはあらかじめタグが挿入されているが，属性ベクトルはこれらのタグの存在を除去して算出される．

[ステップ2]

階層型ニューラルネットワークの入力–中間層間の結合荷重を −0.5～0.5 の乱数で，中間 出力層間の結合荷重を −0.1～0.1 の乱数で初期化する．また，出力関数の傾きをすべて1で初期化する．

[ステップ 3]

学習サンプルの全行数に等しい数の学習パターンが存在する．スキップ条件（後述）を満たさない学習パターンに対して，拡張BP法を用いて結合荷重と出力関数の傾きを修正する．このとき，出力ユニットの誤差の最大値 $\max_k\{|T_k - z_k|\}$ を行単位の誤差配列に格納する．また，全体を通じての最大誤差も求める．

[ステップ 4]

ステップ 3 全体での最大誤差が 0.05 未満のとき，学習を完了する．また，この最大誤差が 0.1 未満であり，かつ前回の学習サイクルにおける最大誤差との差が 0.01 未満の状態が 10 回連続したときも学習を完了する．

学習を継続するときは，ステップ 3 に戻る．

スキップ機能

記事の切り出しで大きな役割を果たすのは，主として記事の切れ目の前後の行である．記事の中央部分から取り出された学習パターンは一般にあまり意味を持っていない．これは，誤差がすぐさまほぼゼロになる学習パターンが多数あることを意味する．このような学習パターンについて，毎回同じように，誤差を求めて，結合荷重と傾きを補正するのは無駄が多い．ステップ 3 の誤差配列は，このような無駄を除去するためのものである．この誤差配列に格納された誤差の値が 0.025 よりも小さく，ステップ 3 全体での最大誤差の前回との差が 0.05 よりも小さければスキップの対象とする．しかし，このような学習パターンのすべてを学習対象から外すと全体としての学習精度が劣化するため，$1/N_s$ の割合で学習対象にする．具体的には，スキップが各学習サイクルでほぼ均等になるように，(学習サイクル + 学習パターン) mod N_s の値が 0 に等しいときは学習対象にし，そうでないときは，学習をスキップする☆．

認識アルゴリズム

[ステップ 1]

記事切り出しを行うニュース文を読み込み，属性ベクトルを算出して，属性配列に格納する．

[ステップ 2]

学習によって得たニュースソース別の結合荷重，出力関数の傾きデータを用いて，中間層の計算を経て，出力ユニットの出力の値を求める．ニューラルネットワーク自体での認識では，出力の値が 0.5 未満のとき論理値 0，そうでなければ論理値 1 とする．

☆ スキップするか否かを乱数を使ってランダムに決める方法もテストしたが，誤差が振動しやすく，良い結果が得られなかった．

[ステップ 3]

後述するアルゴリズムにより，ニューラルネットワークの認識結果を補正する．

ニューラルネットワークの認識結果の補正

記事の始まりと見出しは，分類語とか罫線が存在しても，比較的近接している．これに対して，記事の終わりと次の記事の始まりの間隔は，広告などの存在の有無により，ばらつきが大きい．また，記事の最終行には出典とか記者名があったり，なかったりで，必ずしも形式が一定していない．これらのことから，認識誤りは記事の終わりで発生しやすく，記事の始まりとか見出しでは認識誤りが比較的少ない．

正しい事象の発生は，記事の始まり，見出し，記事の終わりの繰返しであるが，ニューラルネットワークでは，この順序性は明示的には学習していない．そこで，ニューラルネットワークの認識結果について，この事象の発生順序をチェックし，記事の終わりで認識誤りが起こりやすいことを配慮して以下に述べるアルゴリズムにより補正を行った．このような補正によって，かえって誤りが増えるケースもありうるが，全般的には認識精度の向上が得られた．

(1) 記事の始まりを待つとき，
- 記事の終わりが現れた場合，この記事の終わりは認識誤りと見なす．

(2) 見出しを待つとき，
- 記事の始まりが現れた場合，この記事の始まりは認識誤りと見なす．
- 記事の終わりが現れた場合，この記事の終わりは認識誤りと見なす．

(3) 記事の終わりを待つとき，
- ニュース文の終わりに達した場合，これまでの行で，記事の終わりに対応する出力が一番大きい行を記事の終わりと見なす．
- 記事の始まりが現れた場合，これまでの行で，記事の終わりに対応する出力が一番大きい行を記事の終わりと見なす．

3. 実験結果

3.1 実験データ

評価実験では，7 つのニュースソースについて，それぞれ連続した 20 日分の記事を用い，日付の古い 10 日分を学習サンプル，日付の新しい 10 日分を評価サンプルとした．一部の実験では，学習サンプルを増やした．ここで用いたタグ `<art>`，`</art>` および `<ti>` は元のニュース文にはいっさい存在しない．もし，学習サンプルのニュース文自体に，これらのタグが存在

表 1 学習サンプルニュース文の諸元
Table 1 Characteristics of learning samples.

ニュースソース	1	2	3	4	5	6	7
平均ヘッダ行数	31	5	30	22	49	60	46
平均フッタ行数	21	6	20	19	41	42	38
平均記事行数	37	14	46	25	49	59	52
記事件数	100	20	115	145	116	107	114
広告	あり	なし	なし	あり	あり	あり	あり
総行数	4324	414	6017	4745	6903	7824	7244

表 2 評価実験結果（学習サンプル数=10）
Table 2 Evaluation test results (the number of learning samples of 10).

ニュースソース	1	2	3	4	5	6	7	平均
学習回数	19	21	20	39	21	18	25	
実質学習回数	4.4	13.3	4.0	8.7	3.8	3.3	3.9	
再現率 N	100	100	99.7	99.5	99.4	98.4	86.3	97.4
適合率 N	100	100	99.7	100	99.7	99.4	99.2	99.7
再現率 M	100	100	100	99.8	100	98.8	97.1	99.3
適合率 M	100	100	100	99.8	100	98.8	97.4	99.3
認識率 N	100	100	99.7	99.8	99.6	98.9	92.8	98.5
認識率 M	100	100	100	99.8	100	98.8	97.2	99.3
記事数	100	20	115	145	116	107	114	

していた場合，別の綴りに代えておく必要がある．評価サンプルに事前にこれらのタグを挿入したのは，認識精度をプログラムで計測するためであり，実際の応用では，タグは関与しない．したがって，記事の切り出しを行うニュース文にこれらのタグがあっても差し支えない．

学習サンプルとして用いたニュース文の諸元を**表 1** に示す．

3.2 評価実験結果

評価指標として以下に示す再現率と適合率を用いる[3),4)]．

$$\text{再現率} = \frac{\text{抽出した正しいタグの個数}}{\text{評価サンプルにあるタグの個数}}$$

$$\text{適合率} = \frac{\text{抽出した正しいタグの個数}}{\text{抽出したすべてのタグの個数}}$$

再現率と適合率では分母が異なっているが，100%に近いときは両分母の差は小さい．そこで，総合的に 1 つの指標を使うときは，便宜上，再現率と適合率の平均値を用い，これを認識率と呼ぶことにする．

評価実験結果を**表 2** に示す．再現率，適合率，認識率に付加した N および M はそれぞれニューラルネットワーク自体の認識結果および補正後の認識結果を表す．また，平均は記事数を加味した加重平均である．

評価実験に用いた各種パラメータの値は，$\alpha = 0.9$, $\eta = 0.2$, $\varepsilon = 0.01$, $N^M = 6$, $N_L = 10$, $N_S = 20$ である．

ニュースソース 7 を除くと，ニューラルネットワークの再現率は 98.4%以上，適合率は 99.4%以上という良好な結果が得られた．広告は罫線で囲まれるなど目立った特徴を有しているために，広告の存在による認識精度の低下は見られない．ニュースソース 7 の再現率はやや悪く，86.3%であるが，補正アルゴリズムにより，92.8%に向上した．

ニュースソース 7 には，箇条書きの「売り上げランキング」，1 つの記事が複数の小さな記事からなる「ダイジェストニュース」など，通常のニュース記事とは形式がきわめて異なる記事が含まれるため，他のニュースソースに比べて認識率が悪くなっている．

全ニュースソースの平均では，ニューラルネットワーク自体の認識率は 98.5%（エラー率 1.5%），補正後の認識率は 99.3%（エラー率 0.7%）であり，補正アルゴリズムによりエラー率が半減した．

評価実験にはパソコン（Pentium Pro☆ 200 MHz, 64 MB メモリ，Windows NT☆☆ 4.0）を使用した．表 2 には学習回数を示したが，各ニュースソースの学習時間の合計値は 67 秒である．なお，プログラムは C++ 言語で開発した．

学習サンプル数と認識率との関係

ニュースソース 7 の認識精度が若干低いのは，学習サンプルが 10 日分では不足している可能性があると考え，学習サンプル数と認識率との関係を評価した．

☆ Pentium Pro は米国 Intel Corp. の商標である．
☆☆ Windows NT は米国 Microsoft Corp. の登録商標である．

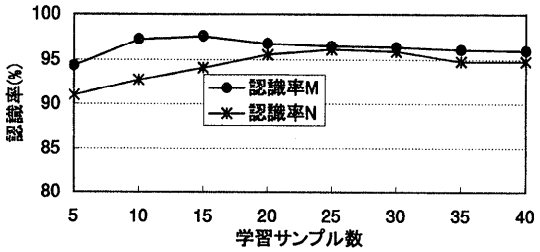

図 6 学習サンプル数と認識率との関係（ニュースソース 7）
Fig. 6 Relationship between the number of learning samples and recognition ratios for news source 7.

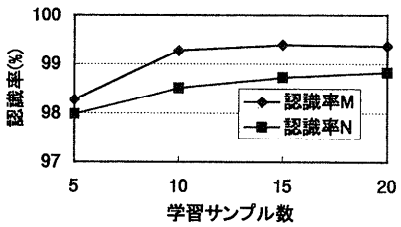

図 7 学習サンプル数と認識率との関係（全ニュースソース平均）
Fig. 7 Relationship between the number of learning samples and recognition ratios (average of all news sources).

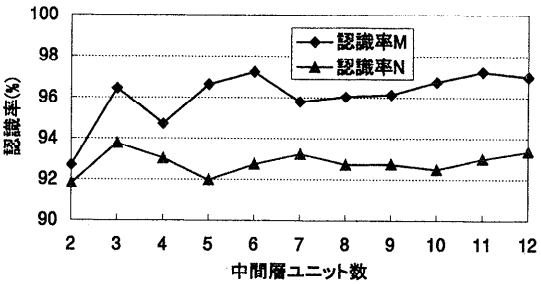

図 8 中間層ユニット数と認識率との関係（ニュースソース 7）
Fig. 8 Relationship between the number of intermediate layer units and recognition ratios for news source 7.

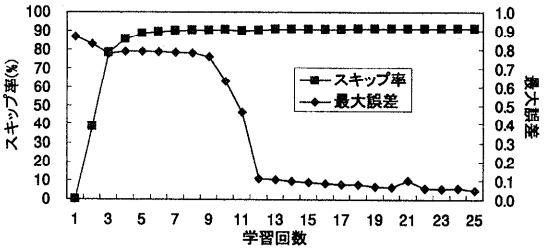

図 9 学習曲線（ニュースソース 7）
Fig. 9 Learning curve for news source 7.

その結果を図 6 に示す．ニューラルネットワーク自体の認識率は学習サンプル数 25 のとき最大値 96.2%に達する．一方，補正後の認識率は学習サンプル数 15 のとき最大値 96.8%に達する．

同様に，ニュースソース 6 の場合にも，学習サンプル数を増大すると，認識率が向上し，学習サンプル数 15 でネットワーク自体および補正後の認識率が最大となり，それぞれ 99.4%および 98.9%となった．全ニュースソースの学習サンプル数を変更したときの結果を図 7 に示す．

補正後の認識率は学習サンプル数が 15 のとき最大で，99.4%となっている．学習サンプル数が 10 以上で認識率が向上するのは，ニュースソース 6 と 7 の認識率が向上するためであり，その他のニュースソースについては，学習サンプル数が 10 の段階で 100%または 100%にきわめて近い値に達している．このため，図 7 には示されていないが，図 6 から分かるように，ニューラルネットワーク自体の認識率は学習サンプル数が 25 のときピークに達する．

中間層ユニット数と認識率との関係

中間層ユニット数と認識率との関係を図 8 に示す．中間層ユニット数が 3 から 12 までの範囲では，ニューラルネットワーク自体の認識率は 92.0%から 93.8%の範囲にあり，補正後の認識率は 94.7%から 97.2%の範囲にある．中間層ユニット数が 2 のときは $\eta = 0.2$ ではローカルミニマムに陥るため，$\eta = 0.02$ としてゆっくり学習させるようにした．しかし，最大誤差が 0.05 未満とはならないので，学習回数 300 回（中間層ユニット数が 3 から 12 までのときの学習回数に比べると約 10 倍の値）で打ち切った．このときの最大誤差は 0.2 である．

図 8 では，中間層ユニット数が 3 のときニューラルネットワーク自体の認識率は最良であり，補正後の認識率もそう悪くない．しかし，上で述べたように，中間層ユニット数が小さいと学習に時間がかかり，容易に収束しないケースもあるため，中間層ユニット数を 6 とした．ニュースソース 7 では，ニューラルネットワーク自体の認識率も平均的な値にあり，補正後の認識率は最良値を示している．

なお，中間層ユニット数が 1 のときは，いずれのニュースソースでも，二乗誤差は小さくならず，学習不能であった．

スキップ機能の効果

スキップ定数 N_S の値を 20 としたとき，ニュースソース 7 に対する学習曲線を図 9 に示す．各種パラメータの値は表 2 のときとまったく同じである．学習回数 3 回あたりから大幅なスキップが起こり，6 回目から 90%，13 回目から 91%に達している．スタート時ではスキップがないため，全体平均ではスキップ率は 84%となる．

表 3 スキップ定数 N_S と実質学習回数との関係
Table 3 Relationship between skip constant N_S and the number of real learning.

スキップ定数 N_S	1	5	10	15	20	25	30	35	40	45	50
再現率 N	97.4	97.6	97.7	97.7	97.6	97.6	97.8	97.8	97.6	97.8	97.7
適合率 N	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.7	99.6
再現率 M	99.2	99.3	99.3	99.4	99.4	99.3	99.4	99.3	99.4	99.3	99.4
適合率 M	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3	99.3
認識率 N	98.6	98.6	98.7	98.7	98.7	98.7	98.7	98.7	98.6	98.7	98.7
認識率 M	99.3	99.3	99.3	99.4	99.4	99.3	99.3	99.3	99.3	99.3	99.4
実質学習回数	21.3	6.8	5.3	5.1	4.6	4.4	4.4	4.2	4.1	4.5	3.9

表 4 学習時間と認識率
Table 4 Learning times and recognition ratios.

	除去なし	最大 20 行	最大 15 行	最大 10 行
総行数	7,244	4,019	3,552	3,020
学習時間（秒）	11.1	8.8	5.4	6.6
認識率 N	92.8	93.1	92.2	89.1
認識率 M	97.2	96.6	95.3	89.2

スキップ定数 N_S の値により，認識精度および実質学習回数がどう変わるかを**表 3** に示す．再現率，適合率，認識率は 7 つのニュースソースに対する単純平均であり，実質学習回数は総 BP 学習回数を学習サンプルの総行数で除算したものである．認識精度はスキップにより若干改善されている．これは学習サンプルから冗長性が取り除かれたためと思われる．実質学習回数は，スキップを大きくするほど低下する傾向があるが，学習の収束に手間取り，かえって増大するケースもある．$N_S = 20$ のとき，5 倍弱の高速化が達成できている．

記事本文の中央部は記事切り出しへの関与が少ないことは容易に推測される．中央部を除去して，学習サンプルの記事本文の行数を最大 20 行，15 行，10 行としたときの学習時間と認識率を**表 4** に示している．最大 20 行の場合，認識率はほぼ同じで，学習時間は 2 割短縮した．スキップ操作を除いたときの学習時間は 37.9 秒であり，認識率 N および認識率 M はそれぞれ 93.2%および 96.3%である．したがって，記事本文の中央部を除去するよりも，本研究で用いたスキップ操作の方が学習の高速化への寄与度が大きい．両者を併用すれば，ニュースソース 7 に対して，学習速度は約 6 倍となる．除去する行数をさらに増やして，本文の行数を最大 15 行にすれば，学習時間は短縮されるが，認識率が若干低下する．除去行数をさらに増やして，本文の行数を最大 10 行にすると，認識率の低下が顕著となり，学習時間も逆に増大している．これは，学習が進みにくくなり，学習の繰返し回数が大きくなるためである．なお，平均記事行数の小さいニュースソース 2，4 などでは，記事本文中央部除去による学習時間の大幅な短縮は期待できない．

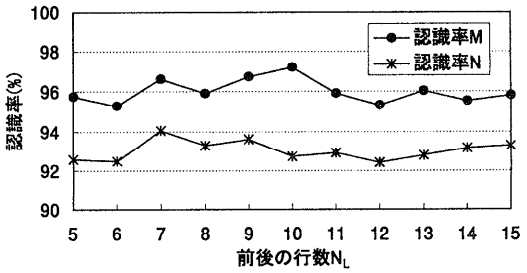

図 10 入力層に与える前後の行数 N_L と認識率との関係（ニュースソース 7）
Fig. 10 Relationship between the number of lines allocated to input layer and recognition ratios for news source 7.

行数と認識率との関係

ニュースソース 7 に対する入力層に与える前後の行数 N_L と認識率との関係を**図 10** に示す．ニューラルネットワーク自体の認識率は $N_L = 7$ のとき最大値 94.1%となり，補正後の認識率は $N_L = 10$ のとき最大値 97.2%となる．

属性値決定に関する上限値と認識率との関係

ニュースソース 7 に対して，分類語属性での文字数の上限値を 10 文字および 30 文字としたときの補正後の認識率はそれぞれ 96.8%，97.0%となり，20 文字のときの認識率 97.2%より若干悪い．また，見出し語属性での文字数の上限値を 50 文字および 70 文字としたときの補正後の認識率はともに 96.6%となり，60 文字のときの認識率 97.2%より若干悪い．

学習終了条件と認識率との関係

ニュースソース 7 に対して，学習終了とする最大誤差の値を 0.02，0.05，0.1，0.2 としたときの認識率 N および認識率 M の値はそれぞれ 92.8%，92.8%，93.1%，93.0%および 97.0%，97.2%，97.1%，96.4%である．大差はないが，学習終了とする最大誤差の値が 0.05 のとき認識率 M は最良である．

4. おわりに

インターネットの普及により，新聞の電子化が進み，日々大量のテキストデータが簡単に入手できるようになった．これら大量のテキストデータから所望の情報を効率良く得るには，まず，個々の記事の切り出しが必要である．本研究では，ニューラルネットワーク技術を適用することにより，記事の切り出しを行う方法を提案した．7つのニュースソースについて評価実験を行い，平均で98.5%という高い認識精度が得られることを明らかにした．最も悪いケースでは，ニューラルネットワーク自体の認識率は92.8%であるが，補正アルゴリズムにより認識率を97.2%にまで向上させることができた．平均的にはこの補正アルゴリズムにより認識誤りを半減でき，認識率98.5%を99.3%に高めることができた．

大量のデータを人手をかけずに処理するためには，認識率100%を目指さなければならない．今回の研究では，文の意味解析の世界にまったく入っていないが，さらに認識率を向上させるには何らかの形で意味処理を取り込むなどの工夫が必要である．

参考文献

1) 堤 一義：ニューラルネットワーク研究の最新動向，システム/制御/情報，Vol.36, No.10, pp.619–624 (1992).
2) Sperduti, A. and Starita, A.: Speed Up Learning and Network Optimization with Extended Back Propagation, *Neural Networks*, Vol.6, pp.365–383 (1993).
3) 田村直良，和田啓二：セグメントの分割と統合による文章の構造解析，自然言語処理，Vol.5, No.1, pp.59–78 (1998).
4) 江里口善生，木谷 強：冨田一般化LRパーザを用いた情報抽出，情報処理学会論文誌，Vol.38, No.1, pp.44–54 (1997).
5) 麻生英樹：ニューラルネットワーク情報処理，産業図書 (1989).
6) Guadarrama, R.S., Dimitriadis, Y.A., Palmero, G.I.S., Izquierdo, J.M.C. and Coronado, J.L.: *Building Digital Libraries from Paper Documents, Using ART Based Neuro-fuzzy Systems*, Lecture Notes in Computer Science, Vol.1240, pp.294–303, Springer (1997).
7) 小原和博，中村行宏：バックプロパゲーション・ニューラルネットへの学習セットの選択的提示法，電気学会論文誌C，Vol.117, No.9, pp.1281–1290 (1997).
8) IDGコミュニケーションズ：ComputerWorld Today, http://www.idg.co.jp/cwt/ (1998).
9) インプレス：Internet Watch, http://www.watch.impress.co.jp/internet/ (1998).

(平成10年5月29日受付)
(平成11年1月 8日採録)

畑田 稔（正会員）

1972年京都大学大学院工学研究科博士課程修了．同年日立製作所入社，日立研究所，マイクロエレクトロニクス機器開発研究所，システム開発研究所に勤務．1998年11月より富山県立大学教授．制御系の安定問題，大規模システムの解析と評価，複合マイクロコンピュータシステム等の研究を経て，インターネット・イントラネット技術を活用した情報サービスシステムの研究開発に従事．1971年電気学会論文賞受賞．京都大学工学博士．システム制御情報学会会員．