単語の連想関係に基づく意味マップによるテキスト表現の試み*

2C−7

豊浦潤　有田英一†
三菱電機（株）中央研究所‡

1　はじめに

近年社会の情報化に伴い、情報源の複数化によるテキスト間の冗長性の増大、情報サイクルの短縮による言葉の意味の多義化といった問題が生じている。そのため情報の受け取り手側に情報獲得のボトルネックが生じている。筆者等はこうした問題を解決するために情報獲得を自動的・自己組織的に実行する情報システムの研究を進めており、その第一歩として教師なしニューラルネットワークによるテキストの自動分類機構を提案している[1]。本稿では、[1]から得られる単語の連想関係を利用して作成するマップ形式のテキスト表現を提案する。

2　テキスト表現

2.1　アプローチ

要約作成に対するアプローチとしてテキストをボトムアップに解析する手法を採る場合、莫大な辞書が必要となる。この困難を避けるためには予め幾つかの典型的なシナリオを用意し、各テキストに特徴的なタームを後から追加するトップダウン手法が有効であるが、シナリオの形式とその作成方法が問題となる。以下では一つの試みとして、[1]で自動抽出される共起関係の強いワードの集合をシナリオとして利用すること検討する。

2.2　テキストのパタン表現

情報検索システムに対するテキストの入力形式としては、キーワードの総数(m)を次元とし各要素はキーワードの出現頻度に比例する、ベクトル表現が用いられる。ベクトル表現は、語の順序関係や接続関係を省略した斜め読み的なテキスト表現と言える。以下、i番目のテキストに対するベクトルを $\vec{t_i} = (t_{i1}, t_{i2}, \ldots, t_{im})$ と記す。

図1に、[1]で提案した機構の概略を示す。図中のニューラルネットは $\vec{t_i}$ を入力層：$\vec{a}$ に対する入力として分類を

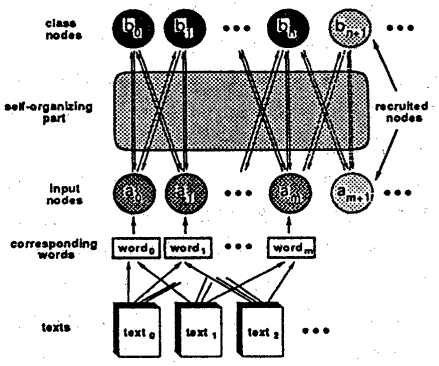

図 1: テキストの自動分類モデル

実行すると同時に、同一クラスに分類されたテキスト中のワードの共起関係の強さを重み付きパタンの形式でリンク上に保存する。このパタンの内で重みの大きいノードに対応するキーワードはそのクラスを代表するワードであると解釈される。以下これをテンプレートと呼び、j番目のテンプレートを $\vec{c_j}$ などと書く。

3　マッピング

3.1　連想性

図1の各ノードは入力パタンに対し競合的に動作し、$\vec{t_i}$ がクラス j に分類されたとすると入力ノードの活性度は

$$a_k = \frac{t_{ik} + B\,c_{jk}}{t_{ik} + A}$$

に収束する**。$\vec{a}$ の収束値は入力パタンにこれまでの学習結果より得た連想を加えたものと解釈できる。

3.2　専門性

一般に用語には専門用語と一般用語の範疇があり、テキストの内容表現には前者が適していると言われてい

*Semantic Map for Text Representation using Association between Words

†Jun Toyoura, Hidekazu Arita
‡Mitsubishi Electric Corp.

**$1 \simeq A \gg B > 0$ はパラメータで B は連想の強度を表す

る。専門用語の定義に関しては幾つかの見解があるが[2]、ここでは「専門用語とは一般的でない用語である」という立場で、無作為に抽出されたワードの専門性を判定する方法について考察する。

図1では入力ノードに対応するキーワードとして、入力テキストから字種の変化などの接続条件のみを用いて抽出したワードを用いており、予めキーワードテーブルを用意するなどの制約条件は極力排除している。これは新語の登場など入力テキストのダイナミズムに対応するためであるが、その結果生成されたテンプレート中には専門用語と一般用語が混在している。

今、テキストが分類された各々のクラスがそれぞれ一つの専門分野を表しているとすれば、一般的な用語とは分類クラスに依らずにテキストに出現する用語であるから、図2のようにテンプレート同士を抑制的に結合したネットワーク上では、汎用性の高い一般用語は活性度が抑えられる。具体的にはテンプレート j 上のワードの専門度:sを下式で定義した*。

$$s_k = C c_{jk} - D \sum_{i \neq j} c_{ik}$$

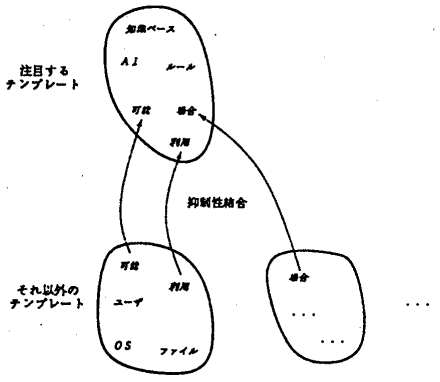

図 2: テンプレートからの専門度評価モデル

4 実験

[1] で行なった 369 のテキストの自動分類実験の結果生成された約 30 のテンプレートを用いて、各テキストの出現ワードから連想性・専門性を計算しそれぞれを X, Y 軸とし平面上にマッピングした。その一例を図3に示す。

[1] の実験ではテキストとしてＡＩ関係の技術記事を用いたため、図のマップでは一般的尺度では専門性が高

*1 $\simeq C \gg D > 0$ はパラメータで D は一般性の抑制度を表す

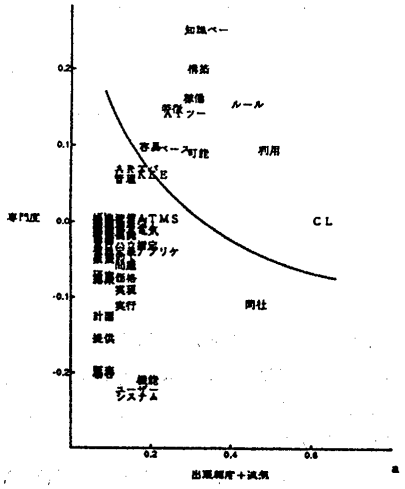

図 3: マップ形式のテキスト表現

い、「システム」、「ユーザ」などの専門度は低いと判定さる。マップの右上に分布するワード、例えば図の破線の右側の数ワードは連想度・専門度が高く、テキストの内容を代表するのに特に適切なキーワードと考えられる。そして、これらの中でも「知識ベース」のように左上部に位置するワードは、このテキストに限って言えば出現頻度・連想度は余り高くないが、分類されたクラスに固有のキーワードであること、「ＣＬ」のように右下に位置するワードは、分類されたクラスに対する専門性は低く、このテキスト固有のキーワードであることが推察される。実際にこのテキストは、「ルールベースを利用したＡＩツール：ＣＬ」に関する記事であり、マップは上に述べた内容を良く反映している。

今後はこのマップの、検索時のテキスト選別や、テキストを代表するキーワードの付与に対する有効性を具体的に確証していく予定である。

参考文献

[1] 豊浦 他, 自己組織型ニューラルネットワークによるドキュメントの自動分類, 情処ＮＬ研資, *92*, 21, pp.41-48, 1992.

[2] 石井, 専門用語を抜き出す試み, 専門用語研究, *3*, pp.32-37, 1991.