学術情報センター電子図書館サービス

©(社)情報処理学会

自然言語処理 101−15
（1994. 5. 28）

ARTを利用した多義語の分類とその評価

内田 友幸　　田中 英彦

東京大学 工学部

概要

記号処理的な手法を用いた自然言語理解の研究にはいまだ多くの困難が存在している。そのため現在、新しい角度からのアプローチが望まれている。このような現状に対し、ニューラルネットなどを利用し、新しい考え方でそれらの問題に対し有効な手法を提案することが本研究の目的である。我々が開発したシステムは、入力された自然言語を事象毎に自動分類し、同時にその分類の基準要素を抽出するシステムである。この際、この分類および、分類の基準となる要素は教師なしのニューラルネットワークARTを用いて自動学習させる。本稿では、この分類システムを同一の多義語を含んだ文書に対して適用して評価を行ない、その性能と可能性について述べる。

Clustering Articles including Multi Meaning words by ART Network

Tomoyuki Uchida, Hidehiko Tanaka

Department of Electrical Engineering, Faculty of Engineering,
University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

Abstract

The are many studies about natural language recognition. But it is still difficult because of its ambiguity. So We made new approach to language recognition by neural network ART. This system use only word frequency. So it is simple and objectively. This system put a large quantity of natural language document into some categories, and construct associative relationship between words.

In this report, We make estimation of this clustering faculty by using one multi meaning word. As the result of this experiment, It become clear that this clustering system is useful.

学術情報センター電子図書館サービス

©(社)情報処理学会

1 はじめに

電子計算機の飛躍的発展により、明解なアルゴリズムが存在する良設定問題を解くことは、かなり実現できるようになったといえる。その一方人間が日常的にこなしている、不完全情報下での問題解決などの不良設定問題においてはまだまだ多くの困難が存在している。

このような不良設定問題に対して、人間は日常的に解を得られるので、その人間の思考に即した考え方から、これらの問題に対しより良い手法を考案していきたいと考えている。

人間が、自然言語を理解するには2つのステップがあると考えられる。一つは直観的な連想で概念の候補を提案すること、そしてもう一つが、直観的に提案された候補を論理的に評価し、絞ることである。

論理的に評価する部分は、記号処理的なアプローチでも解決が可能であると考えられるので、その前の段階、直観的な連想の部分に対して本研究では考察を加えていく。

ここで、直観とは何かを考えていくと、人間は過去に類似した経験をしたことが無い事象に対しては戸惑い、類似の経験をしたことがあればすみやかにそのイメージが構築できる。このようなことを考えると、過去に経験した事象の連想とその組合せというものが、人間の直観には大きな意味を持っているものと考えられる。

そこで、本研究ではこの過去に経験した事象を蓄積し、なおかつ新しい入力に対してそれらを連想し、解の可能性のあるものを提案できるようなシステムを目標としてとらえている。この試みの第一段階として自然言語文書を対象とし、その入力された自然言語文書を事象毎に分類し、なおかつそこから文脈に即した単語の連想関係を取得するシステムを開発した

本稿では、多義語を含む文書を分類することによりこのシステムの評価したのでこの評価と、このシステムの可能性を考えた一例について報告する。

また、このような自然言語の分類についてはすでにいくつか行なわれている。Kimotoらは動的なネットワークのシソーラスを使って入力されたキーワードを拡張し、文書データベースの検索効率を高めた [Kim93] 。湯浅らは文書に含まれる名詞の共起関係を調査することで新聞記事を5つのジャンルに分類に成功している [Yua93] 。また、豊浦らはARTを用いてドキュメントの自動分類を行ない [Toy92a] 、同時に単語の専門度について評価を行なっている [Toy92b]。

しかし、定量的な評価に乏しかったり、入力データを人手で加工していたり、パラメータについては経験的に対処している点など、可能性を示唆しただけに留まっている。また、目的が主にデータベースの検索に絞られているため、そこから自然言語理解の可能性を見い出すという視点が乏しいと言える。

2 システムの構成

本システムは大量の自然言語文書を単語に分解し、その単語の頻度情報を元に適応共鳴理論 ART (Adaptive Resonance Theory)[Car88] を利用して文書のクラスタリングを行なう。また、そのクラスタリングの際に得られる重み値から単語間の連想関係を得る。

ARTの中心部は入力層と内部状態層の2層とその相互の重みつき結合によって構成されている。そして、新しい入力に対して、すでにある内部状態に類似していればその内部状態を更新し、そうでなければ新しい内部状態を作るという動作を繰り返す。このため、従来の多くのニューラルネットとは異なり、誤差ではなく、逆に「ほぼ一致する」ことに基づいて学習を行なことになり、システムの記憶を外界のノイズから守る一方、高速かつ安定な学習が可能になっている。

本システムに入力されるデータは、大量の文書で、それぞれが数個から十数個程度のセンテンスによって構成されている比較的短い文書を想定している。具体的には新聞記事、商用パソコン通信の書き込みを用いた。

これらの文書をそれぞれセンテンスごとに分け、単語に分解する。最終的には、各単語に一つのノードを割り当て、一つの文書内の単語の頻度情報をノードの発火状態に置き換え、ARTに入力できる形にする。

3 新聞記事の分類

入力した新聞記事は1993年7月下旬から8月上旬の約2週間に時事通信によって配信された2075個の記事である。ジャンルは日本経済、世界経済、内政、世界情勢、事件等と多岐に渡っている。

全部で2075個の新聞記事を各記事ごとにセンテンスに分け、さらにそれをパーサーにかけて単語分解をする。単語分解した例を以下に示す。

「こう,し,た,緊急,課題,に,着手,する,ため,、,まず,政治,改革,に,取り組,ま,なけれ,ば,な,ら,ない,。」

このうち、記号だけ、2文字以下の平仮名だけの部分は除外し、以下のような状態にする。

「緊急,課題,着手,政治,改革,取り組,なけれ」

さらに、他の記事には出てこないユニークな単語があれば除外して、単語ごとにノードを割り当てる。最終的に各記事ごとに頻度情報のベクトルにする。今回の実験では8274単語出てきたので、ここでは8274次元ベクトルとなった。

©(社)情報処理学会

これを本システムで分類すると 322 のカテゴリが自動生成し、全記事の 84%の記事がすべてどこかのカテゴリに所属した。このカテゴリのうち1つを以下に示す。記事のタイトル、結合の重みの重い単語を組に順に列挙する。

- ２３日から米加両国を訪問越智運輸相。
- 運輸事務次官に松尾氏が内定運輸相、人事凍結を解除。
- カナダ旅行促進ミッションを派遣運輸省。
- １０月に航空協議開催で合意日加運輸相会談。

「カナダ, 運輸, 運輸省, 会談, 訪問, 越智, 加, 日程, 両国, コルベイユ」

元の記事には文化、事件、経済、政治、外国経済などが区別無く入っているが、上の例では外交関係のさらに細かいテーマにクラスタリングされている様子がわかる。他のカテゴリもこのように、同じジャンルの似たような内容の記事で構成されている。

また、結合の重みの大きい単語はテーマに関係がありそうなものが列挙され、なおかつ、その程度の大きい順に並んでいるのが分かる。各記事に共通に存在し、出現頻度が多いものがより重い結合を持つことから、これらの単語間には何らかの関係があるといえる。そして、これを連想関係の一種と考えることも可能である。

この連想関係は文書の自動分類の結果から自動的に取得でき、なおかつ文脈ごとに取得できるので、従来とは異なった利用法が可能であると考えられる。

4 多義語分類の概要

このクラスタリングに関して多義語の用法に着目することで、定量的な評価を行なった。

類似した文書を一つのカテゴリに分類し、同じ文脈での単語の連想関係を得るには少なくともクラスタリングされた単語の用法が揃っている必要がある。そこで、同じ単語で多くの意味を持つ多義語に着目し、クラスタリングされたカテゴリ内での多義語の用法の揃い具合を調べることでクラスタリングの精度が推定できると考えられる。この揃い具合をクラスタリングの評価とし、諸データの収集、評価を行なった。

4.1 入力文書

具体的には商用パソコン通信の書き込み 3 カ月分、全体量約 90M バイトのうち「コード」という単語を用いているもの 763 アーティクルを抽出し、これを用いた。

多義語「コード」のそれぞれ違う用法を含むアーティクルの一部を以下に列挙する。

アーティクル数	763
全単語数	88,414
システム入力の全単語数	62,208
ユニークなものを排除した単語数	48,579
単語の種類	8,092
ユニークなものを排除した単語の種類	3,925

表 1: 単語数と種類

- シーケンスソフトは対応と言うことなので問題なく使えます。一般のアプリケーションは制御コードの一部、漢字コードの一部（記号等）が違うと思います。
- VC++ のスケルトンですが、殻のウインドウだけでも生成すると鬼のようにヘッダーだとかソースだとかコードをはいてきます。はじめて見たときはかなりめんくらってしまいました。
- この曲に　メロディをつけますか(笑)?　バッハの平均律１番のプレリュードみたいに(^^)。　でも、コードの進行が、完全に読める（単純すぎる）のでどうでしょうね。　考えてみましょうf^^;　出来たら面白いなぁ(^^)
- あ、質問というのは、このことではなくて、その変換コードを買ってきて、接続したところ、音が片方（L）からしか出なくて、しかも、ものすごく小さいのですが、原因は何でしょう?

4.2 文書内の単語の様子

アーティクルから顔文字のキャラクター、慣例となっている記号を排除し、センテンスを適当に揃える。このようにした文書を新聞記事の場合と同様に ICOT のソフトと EDR を用いて単語に分解を行ない、ベクトル表現に変換する。この結果は表 1 にまとめた。

まず、「コード」を除いた単語が何回登場するかヒストグラムを取った。図 1 にその結果を示す。ただし、縦軸は単語の種類ではなく単語の登場回数を示している。ここで、右側にいくほど多用されている単語という事になるので、右側が一般的な単語、左側が特殊、または専門的な単語と考えられる。単語の登場数が増えるにしたがって急激に減少し、種類で考えればほとんどの単語が 50 回以下程度しか登場していないことが分かる。この場合の単語の平均登場数は 12.43 回であった。

また、同一アーティクル内での重複を無視した場合は若干左側に遷移するものの同様の分布となり、この場合の平均は 9.82 回となった。763 アーティクル中、一つの単語に対し 10 個のアーティクルが関連づけられているとすると、2 つの単語を指

©(社)情報処理学会

定した場合、両方に関連づけられている記事が存在する確率は、

$$1 - \frac{{}_{763}C_{20} \cdot {}_{20}C_{10}}{{}_{763}C_{10}^{2}}$$

このようになり、約0.125の割合となる。前出の新聞記事の場合も同様に計算すれば約0.057の割合であったことから、新聞記事の場合と比べて単語がかなり密な使われ方をしていると言うことができる。

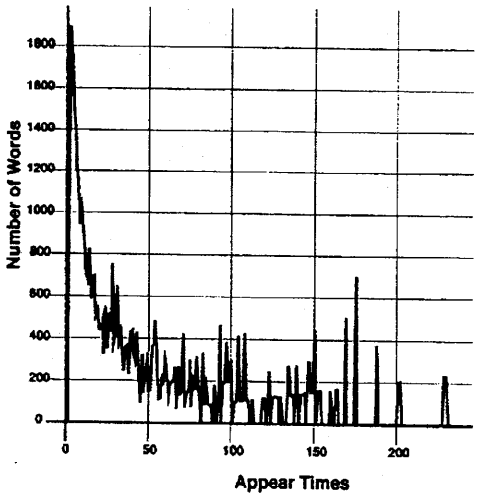

図 1: 単語の登場数とアーティクル全体での使用数

これらのアーティクル内の単語のうち「コード」という単語が全体のどれくらいを占めているかをアーティクルごとに調べると、平均で4%、ほとんどのアーティクルの「コード」の含有率は10%以下であった。

4.3 アーティクル間の距離の様子

続いてアーティクル間の相関を調査した。この結果を以下に示す。

まず、ランダムにアーティクルを三つ抽出し、各アーティクルとの距離をヒストグラムにした。この結果を図2に示した。ただし、ここでの距離はベクトルの差の平方和とする。

距離が1.3あたりから徐々に増え始め、1.35あたりから1.4のところにピークを持っていることが分かる。

また、最近接アーティクルまでの距離のばらつきが観察される。この3つのアーティクルの間でも白と黒の最小値の間で0.1もの開きが観測できる。

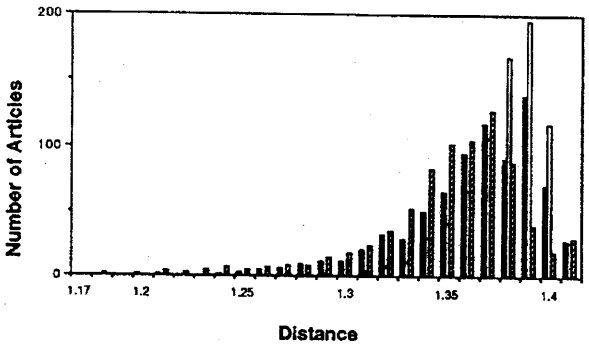

図 2: アーティクル間の距離の様子

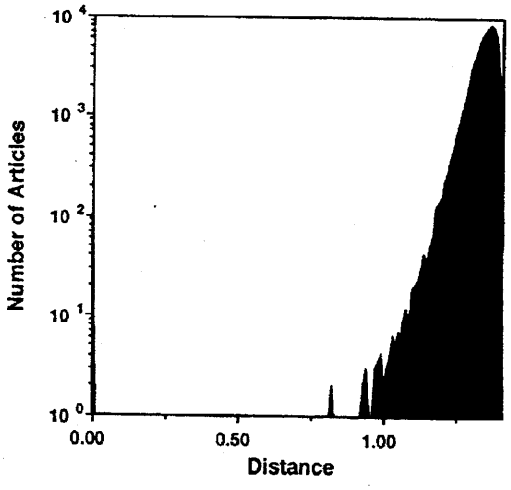

図 3: ランダム抽出した100アーティクルの距離の分布

©(社)情報処理学会

用法	アーティクル数
文字コードなどの符号	453
コンピュータプログラム	196
音楽の和音	54
紐	60
Total	763

表 2: コードの用法とアーティクル数

これらのヒストグラムを 100 アーティクルについて同様に求め、和を取ったものを図 3 に示した。

傾向は図 2と同じで、1.0 まではほとんど無く、徐々に上昇し、1.4 の少し手前でピークが有ることが分かる。

5 多義語を利用した評価

各アーティクルに入っている、「コード」という単語には 4 種類の用法がある。この用法とそれに該当するアーティクルの数を表 2 に示した。

この、カテゴリ内の「コード」の用法の一致する割合をクラスタリングの評価とする。各カテゴリに対してこの割合を調べ、同一用法のアーティクルが９５%以上を占めている場合、ここでは分類成功と考える。そして、分類成功したアーティクルの数の全体のアーティクル数に対する割合を分類成功率と表記する。

5.1 カテゴリ分類

以上のベクトルデータを入力として、ART のシステムにかけ、カテゴリ分類を行なった。閾値を 0.89 から 1.41 まで変化させながらクラスタリングさせた様子を図 4 に示した。閾値を上げていくと徐々にカテゴリサイズが大きくなり、1.3 前後で急激に増加して、最終的には平均のサイズが 100 アーティクルを越えるようになる。このとき、すべてのアーティクルがどこかのカテゴリに所属するようになっていた。

$\alpha = 0.2, \rho = 1.18, \theta = 0.0001$ の条件で 763 記事をクラスタリングした結果、118 のカテゴリが生成し、291 記事以外はすべてどこかのカテゴリに所属した。分類された記事の割合は全記事の 62%である。このようにして分類したカテゴリのうち 4 個を例として以下に示す。この例はすべて同一の用法がクラスタリングされた例なので分類に成功した例といえる。

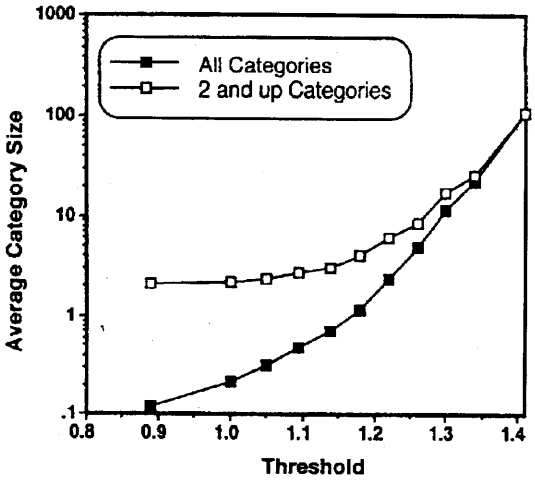

図 4: 平均のカテゴリサイズ

カテゴリはそれぞれアーティクルのタイトルの列挙と重みの高い単語の列挙の 2 つで表している。タイトルの冒頭の数字は「コード」の用法の番号、単語の部分はそれぞれ、単語の内容、その単語への結合重みで構成されている。

1. 1_ＰＯＰＣＡＬＣバージョンアップ
   1_ダイナでＪＥＤＶに成功
   1_ＢＡＳＩＣのポインタ
   1_文字列入力をＥＳＣで中止するには
   1_文豪ドライバ
   1_題名忘れです

   バイト...0.703178　　込...0.118292
   文字...0.409215　　思...0.099473
   漢字...0.371992　　書...0.096601
   コード...0.342490　　知...0.077781
   処理...0.131934　　試...0.061340

2. 2_ＷＩＮとアセンブラ
   2_Ｃ言語 VS アセンブラその 1
   2_Ｃ言語 VS アセンブラその 3
   2_ＡＳＭライブラリ
   2_ライブラリ作り
   2_Ｃの引数渡し
   2_戦略ミス

   アセンブラ.0.4775　　ライブラリ...0.1983
   使...0.4698　　プログラム...0.1746
   私...0.3023　　話...0.1649

学術情報センター電子図書館サービス

©(社)情報処理学会

倍...0.2427	差...0.1439
事...0.2261	速...0.1363

3. 3_すばらしい日聴きました
   3_５番聴いたよまとめてドン第６回
   3_いとしいあのこ
   3_感想大切なあなた
   3_DriveOff、かっこいい(^^)
   3_Re:久の感想でした
   3_どもども(*^^*)海老蟹さん
   3_負けないで、聴きました

曲...0.8771	進行...0.1013
聴...0.3236	データ...0.0877
コード...0.2072	雰囲気...0.0788
作...0.1773	他...0.0669
音...0.1121	でしょ...0.0592

4. 4_Re:実験結果
   4_Responseto#1577

時...0.3001	ハング...0.2192
配置...0.3001	干渉...0.2192
必要...0.2765	パソコン...0.1956
使...0.2529	WX2...0.1956
本体...0.2192	DOS5...0.1382

入力したアーティクルには多彩な話題が入っているが、上のカテゴリ内には多義語の用法の同じものが集まっている。重みの大きな単語からも分かるように、1では文字コード、漢字コードの話題、2ではアセンブラ関係のプログラミング、3では曲データの感想の話題など、それぞれ似たような趣旨のアーティクルが集まっている様子がわかる。他の多くのカテゴリもこのように、同じ用法で構成された、似たような内容のアーティクルで構成されている。ただし、4の紐の用法には先の例のようなカテゴリしか観察されず、綺麗に分類された例が見受けられなかった。これは紐という用法のサンプル数が少なかったことと多岐に渡るコンテキストで用いられたためと考えられる。

また、重みの大きな単語は、例えば、2の用法には「アセンブラ」、「プログラム」など、関連の深いものが集まっている様子が読みとれる。

5.2 結果の定量的評価

続いて、クラスタリングした結果の定量的評価を行なった。

カテゴリ内で一番多い用法の、カテゴリ内の占有率をそれぞれのカテゴリについて計算した。これを占有率ごとにプロットしたものを図5に示す。比較として、全くランダムにカテゴリを1000回生成させた時の平均の結果も合わせて示した。

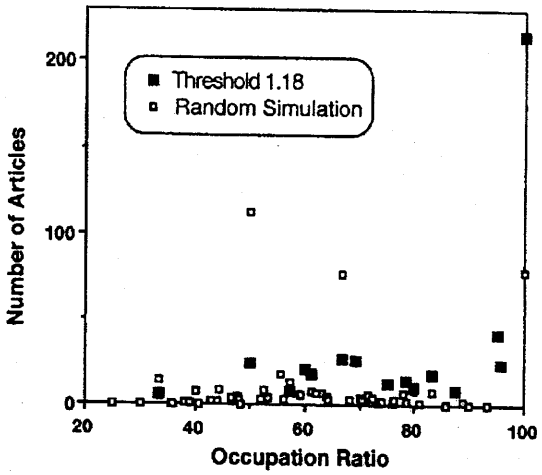

図 5: カテゴリ内の同一用法の占有率

これを見るとランダムに分類した時に比べて明らかに同一用法の占有率が上がっているのが分かる。例えば100%の部分ではランダムが78.2アーティクルなのに対して、本システムでは215アーティクルと2.75倍も多い値を示している。また、50%以下の分類に失敗したと考えられる部分も本システムの方が圧倒的に少ないのが分かる。

また、この際、成功したと判断される占有率が95%以上のアーティクル数の割合は全アーティクルに対しては36.8%、カテゴリ分類された全アーティクルに関しては59.8%の割合となっている。

次に、パラメータを変えた時に分類成功率がどのように変化するのかをα, ρ, θについて調べ、グラフを取ってみた。また、この際カテゴリのサイズが2の場合、成功率は50%か100%にしかならないため、評価が実際より高く見積もられてしまうことが考えられるため、対象とするカテゴリのサイズが2以上と3以上の2つのケースについてそれぞれプロットを行なった。以下にその結果を示す。

図6を見ると、閾値をあげていくと徐々に成功の割合は上がってきて、1.18あたりでピークに達し、その後は低下していく様子がわかる。この傾向はカテゴリサイズが、2以上の合計と、3以上の合計の双方とも同様の結果が得られた。

閾値が小さい場合、カテゴリが小さく、クラスタリングされるアーティクル数自体が小さいため成功率は小さくなる。また、閾値が大きくなるとカ

©(社)情報処理学会

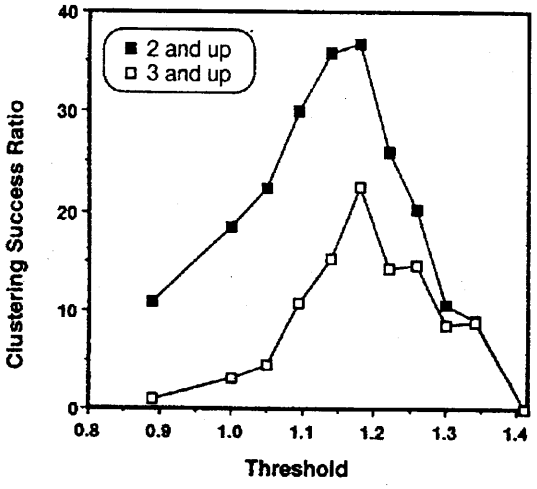

図 6: 閾値 (ρ) と成功率 ($\alpha = 0.2, \theta = 0.0001$)

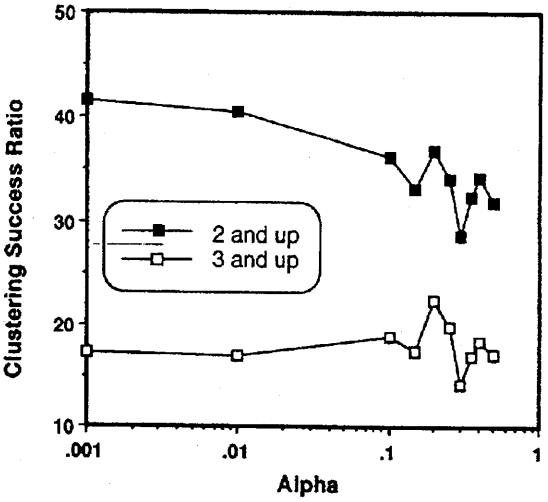

図 8: αと成功率 ($\rho = 1.18, \theta = 0.0001$)

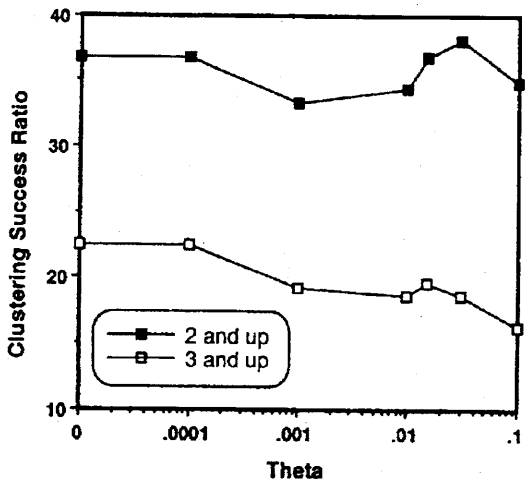

図 7: θと成功率 ($\alpha = 0.2, \rho = 1.18$)

テゴリサイズが大きくなり、違う用法のカテゴリ同士であったもの同士が同じカテゴリに入ってしまうため成功率が落ちてしまうと考えられる。以上の結果からこの例では閾値は 1.18 が適当であるということがいえる。

次にθのグラフを考えてみる。θはノイズ除去の閾値を意味しているので、小さくなるとノイズ除去をあまり行なわなくなる。

図 7 を見るとカテゴリサイズが 2 の時を含めるかどうかで傾向が変わってきている。カテゴリサイズが 2 の時を含めると 0.3 あたりにピークが見られる。しかし、含めないと 0.15 あたりに小さなピークがあるものの 0 に近い方が成功率は高くなる傾向がある。

これはカテゴリのサイズによるものと考えられる。カテゴリサイズが 2 であるカテゴリ数は 0.0001 の時 66 個しか無いのに対し、0.3 では 88 個と 33% も多い。カテゴリサイズが 2 の時は 100%になりやすいのでこの効果が見かけ上の成功率を高めていると考えられる。

もともとこのノイズ除去は微細なノイズが入力に混じっている時に有効なものであり、単語の頻度情報のような離散的に大きな値が並んでいるような入力に関してはあまり効力がないものではないかと考えられる。

そのため、このパラメータは 0.0001 以下の小さな値にしておく方が良いようである。

最後にαについて考えてみる。αは学習の度合を意味しているので、小さくなると学習をあまり行なわなくなる。

学術情報センター電子図書館サービス

©(社)情報処理学会

図8を見るとこの場合もカテゴリサイズが2の時を含めるかどうかで傾向が変わってきている。カテゴリサイズが2の時を含めると0に近い方が成功率は高くなるが、含めないと0.2あたりにピークができる。

この場合もカテゴリサイズが小さくなって見かけ上の成功率が上がっている効果があるものと考えられる。カテゴリサイズが2のカテゴリ数は0.2が66個なのに対し、0.001では111個にもなり、68%も増加しているのが分かる。

このことから学習の度合は0.2程度の適切な値を探し、その値を採用することが重要であることが推察される。

6 抄録の作成について

自然言語の分類について今まで述べてきた訳であるが、最後にこのシステムの可能性の一例として抄録の作成について述べる。

本システムは分類と同時に分類の基準となる重みつきの単語の組合せを得ることができる。この重みをアーティクル中のセンテンスの重要度の指標の一つとして使ってみると複数の複数のセンテンスの中から重要と思われるセンテンスを抜き出すことができる。以下にその例を示す。

まず、ある欧州の通貨関係の記事が集まったカテゴリ内の主な結合状態は以下のようになっている。

通貨...0.6424	ＥＲＭ...0.1799
欧州...0.5361	市場...0.1356
変動...0.2657	仏...0.1214
幅...0.2430	フラン...0.0937
拡大...0.2128	為替...0.0909

カテゴリ内のある記事「欧州通貨は実質的な変動相場制。」内のセンテンスに含まれる単語のこれらの結合重みの単純な和をとってやると以下のようになる。

2.50 田中誠士第一勧銀総合研究所金融市場調査部長の話、欧州通貨制度ＥＭＳ内の為替の変動レート幅を拡大することで名目的にＥＭＳは維持されたが、実質的には変動相場制になったとみてよいのではないか。

1.86 今後欧州通貨は各国の経済のファンダメンタルズ基礎的諸条件を反映した水準を探る局面になり、通貨混乱は終息していくだろう。

0.81 そうなれば一時的に円に逃避していた資金も欧州市場に戻っていくとみている。

0.09 ドル・円相場については、急激な円高は景気の先行きの懸念材料となり、日本の長期金利は低下する可能性がある。

0.02 反対に、米国の金利は１０、１２月に利上げがあるのではないか。

0.11 このため、今後のドル・円相場は、日本の貿易黒字削減を目指した日米包括経済協議の進展具合など予断を許さない条件もあるが、投機的な動きに一時的に円高が進むことがあっても、長続きはしないだろう。

これをみると、一番初めのセンテンスが一番高い値を示している。このセンテンスがこの記事内で一番重要かどうかは議論の余地があるが、一つの指標にできる可能性はあると考えられる。

7 考察

本システムに利用したARTは教師なし学習を行なうので、新聞記事の自動分類では、単語の頻度情報を本システムにかけるだけで自動分類が行なわれた。このような外部からの情報が必要のない自律的な処理は、莫大な量の自然言語情報を処理する上では重要な性質ではないかと考えられる。

また、多義語の分類では、その用法を調べることでクラスタリングの妥当性を評価できるようになった。その結果、クラスタリングされたアーティクルの約3分の2が分類に成功した。

また、現在は多義語の用法を評価の基準としてとらえているが、これを逆に用いることで多義語の意味の同定に利用できる可能性も示唆される。

さらに、分類されたカテゴリ先の単語の結合重みを利用すれば目的とする文書の特徴的な単語を取得できるので、抄録の作成等にも有効であろうと考えられる。

しかし、単語の頻度情報しか利用していないため、複雑な概念を取り扱えない、文書単位でしか操作を施していないため、一つの文書内に複数の概念トピックが存在している状況に対処できないなどの問題点も考えられる。

参考文献

[Car88] Carpenter, G. A. and S. Grossberg: The ART of Adaptive Pattern Recognition by a Self-organizing Neural Network, *IEEE Computer*, Vol. 21, No. 3, pp. 77–88, 1988.

[Kim93] Kimoto, H. and T. Iwadera: Associated Information Retrieval System(AIRS), *IEICE TRANS. INF. & SYST.*, Vol. E76-D, No. 2, 1993.

[Toy92a] 豊浦 潤：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, 情報処理学会, 研究会報告, 自然言語処理, 88-6, 1992.

[Toy92b] 豊浦 潤：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, 情処第45回全国大会, 3-247, 1992.

[Yua93] 湯浅, 上田, 外川：大量の文書データから自動抽出した名詞間共起関係による文書の自動分類, 情報処理学会, 自然言語処理, 98-11, 1993.