情 報 学 基 礎 60－5
（2000．9．27）

テキストマイニングのための情報抽出

長野 徹 武田浩一 那須川哲哉
日本アイ・ビー・エム（株）東京基礎研究所
〒242-8502 神奈川県大和市下鶴間 1623-14
e-mail: nagano@trl.ibm.co.jp

テキストマイニングは、膨大なテキストデータからの知識発見を目的としているが、データマイニングと異なり、自由に記述されたテキストが対象となる。そのため、個々のテキストからいかに適切な情報を含むデータを抽出するかが重要な課題である。本研究の対象であるテキストマイニングのシステムでは、単語レベルの情報に加え、係り受け関係の情報や意図表現から得られる情報を付加することで、より多くの情報を含んだデータを抽出することができる。その際、テキストデータからは多種多様な情報が抽出可能であり、抽出される内容の頻度や重要性もまちまちである。そのため、統計的な処理を行うマイニングという観点からは、抽出する語の単位を考慮する必要がある。本稿では、大量のテキストデータから語を取り出す際にどのような単位での抽出が有効であるか示し、実験結果から得られた知見を示す。

Information Extraction for Text Mining

Tohru Nagano, Kohichi Takeda and Tetsuya Nasukawa
IBM Research, Tokyo Research Laboratory, IBM Japan
1623-14, Shimotsuruma, Yamato, Kanagawa, Japan
e-mail: nagano@trl.ibm.co.jp

The extraction of knowledge/information from vast amounts of textual data has been the focus of much research in recent years. Text mining technology aims to find hidden knowledge in textual data. The varied informtaion/term can be extracted from textual data, however all of terms from textual data are not representative for Text Mining. Thus, Text Mining requires the identification of domain specific terms in order to successfully mine linguistic data to extract concept correlations and trends. This paper presents a method for the extraction of significant terms and their use in the representation of textual data specifically for the Text Mining application.

1. はじめに

計算機の高速化・ストレージの大容量化に伴い、大量のデータを高速に処理できるようになった。蓄積されたデータの利用方法として、従来はデータの保存・記録という側面が大きかったが、近年、蓄積された大量のデータから知識を抽出しようする動きが高まっている。

数値データや定型データからの知識発見の手法として Agrawal[4] によって導入されたデータマイニングと呼ばれる手法が発達してきた。データマイニングは大量のデータから属性やデータ間に成り立つ規則を高速に発見することを目的とし、実際に多くの実データに対しての適用が行なわれている。データマイニングと同様に大量のデータからの知識発見を目的とし、対象をテキストデータとしたものがテキストマイニングである。

一方、テキストデータからの知識を抽出するための研究として Information Retrieval(IR) や Information Extraction(IE) の分野での研究も広く行われている。一般に IE や IR はユーザーのクエリーに対して、ドキュメントの集合を返すことを目的としている。言い替えると、IE や IR では、ユーザーの要求する

クエリーが明らかであるのに対し、テキストマイニングでは、IE や IR とは逆にユーザーに対して知識・または知識として有意な情報をを提示することが目的となる。

また、大量のテキストデータからの (広い意味での) 知識獲得の試みは多く行われており、言語知識の獲得[6] を目的としたものや、分野依存の意味情報の抽出を目的としたもの[5] などがある。しかしながら、ほとんどの大量のテキストデータからの知識獲得の技術はテキストを単に単語の集合として扱う (bag-of-word techniques) ものが殆んどである[7]。しかし、実際のテキストの内容を考慮すると、単純な形態素解析の結果得られる単語の集合のみで、実際の内容を表現することは困難である。例えば、テキストデータのクラスタリングの際、多くの場合、テキスト中での問題を表す表現 ("modem is broken") と逆の意味を持つ ("modem is not broken") が同一視されてしまう。

本稿で用いたシステムでは、テキスト中の意図を表す表現や係り受け関係を利用して実際のテキストの内容を考慮した知識獲得を目指している。本稿では、テキストマイニングのための知識抽出の方法を提案し、特に名詞句に含まれる重要語 (複合名詞) を抽出し、より知識として意味のある結果を抽出した。本稿ではコールセンターで得られた 46,000 件のデータに含まれる語に対して重要語を生成し、重要語をマイニングに適用した結果を示す。

2. テキストマイニングのための情報抽出

2.1 重 要 語

語の抽出方法として、Term Weighting による重要語抽出と、固有表現抽出 (Automatic Term Recognition) がある。前者は、統計的に重要である語を抽出する statistically-oriented アプローチであり、後者は言語的な知識を中心とした linguistically-oriented[9] であると言える。統計的なアプローチとしては、Salton[1] の Vector Space Model は多くの情報抽出 (Information Retrieval) 技術に対して応用が行なわれており、単語・ドキュメントのベクタやクラスタの特徴づけのために広く用いられている。

テキストを統計的に分析する上では、テキストの文字列全体を一つの値として扱うことは困難であるため、基本的には、テキストに対して自然言語処理を行い、その結果得られるキーワードを抽出して統計処理を行なうことになる。

例えば、以下の例はあるデータに対して、データマイニングの一手法である相関ルールの導出を行ったもの実験結果の一部であるが、複合語の構成要素が大半を占めている。

- アイテム数：55716
- バスケットあたりの平均アイテム数：10.3
- サポート：120 以上
- コンフィデンス：1%以上
- 出力ルール数：2973
- 導出ルール例：
  "英語版" =⟩ "英語"
  "VOICE" =⟩ "TYPE"
  "アプリケーション" =⟩ "CD"
  "ハードディスク" AND "ROM" =⟩ "CD"
  "ROM" AND "インストール" =⟩ "CD"

テキストマイニングは、大量のテキストデータから傾向やデータの偏りを発見することを目的としている。ただし、定型データと異なり、出力結果はキーワードや文字列として表される。したがって、テキストマイニングのための重要語は統計的に有意で意味的に重要な語でなければ意味がない。

2.2 対象データ

本研究では、PC コールセンターに蓄積された顧客からの問い合わせデータを用い、それを元に語の出現頻度と重要性との関係について調べた。以下の実験では、このPC コールセンターのデータを用いて行なう。

各ドキュメントは ID、日付、タイトル、および定型部分とテキスト部分から構成されている。定型部分には、問い合わせの内容をコールの応対者が判断し、その問い合わせ内容・問題の種類等を選択肢の中から選択された値が入力されている。定型項目には、対象コンポーネント・問題種別・対応種別等複数あり、それぞれ違う観点から成り立っている。各定型項目は 10 ～ 30 の選択肢があり、コールの応対者がその中から 1 つを選択できるようになっている。例えば、定型項目「対象コンポーネント」の場合、"Windows95", "メモリ", "ハードディスク" 等、問い合わせの内容がどのソフトウェアまたはハードウェアに該当するかという観点で分類される。

また、テキスト部分には、コールの応対の記録が応対者によってある程度要約された形でフリーテキストの形式で入力されている。応対の記録は Q&A 形式で保存されているが、本研究では Q のみを用いてある。

対象ドキュメントは 43,378 件で、名詞句に注目すると、各ドキュメントは平均約 14 個の名詞句 (長単位の名詞) を含み、名詞の異なり語数は 31,609 個で、延べ 1,119,039 個の名詞 (短単位の名詞) を含む。したがって、1 名詞句あたり平均 2 個の名詞を含む。

2.3 重要度の指標

重要度の定義として、人手によって付与されている定型項目の値に対しての単語の分散度合を用いた。前

表 1　データサンプル

ID	03240210233
Date	10/21/1999
コンポーネント	Windows95
問題種別	ソフトウェア障害
対応種別	情報提供
タイトル	ＷＩＮＤＯＷＳ９５が起動しなくなった
内容	Ｑ：イーサネットカードを変えたので、ドライバを変えたら、起動しなくなってしまった。 Ａ：新しいイーサネットドライバをインストールしていただくようにご案内。一度ＷＩＮＤＯＷＳ９５をシャットダウンしてから...

述したように、全てのドキュメントは応対者によって、対象コンポーネント・問い合わせ種別・応対種別、などの観点から分類されている。本稿では、この人手でドキュメントを分類する際に判断基準となるような語が重要語であると考えた。この人手による分類を正解として、各ドキュメントを人手による分類と同じように弁別することのできる語を重要性の高い語とする。

このうち、定型項目「対象コンポーネント」は"Windows95", "メモリ", "ハードディスク" など 30 以上の値に分類されている。今回は、この「対象コンポーネント」で分類された値を元に、重要度の計算を行なった。

語 w_i のエントロピーは以下の式：

$$H(w_i) = \Sigma_{t=0}^{N} P(c_t \in w_i) \log_2 \frac{1}{P(c_t \in w_i)}$$

$P(c_t|w_i)$ は分類値 c_t ("Windows95", "メモリ" 等) に分類されたドキュメントに含む語の全ドキュメントに対する出現確率。N は分類の数。で計算される。

これからエントロピーを正規化した冗長度は

$$r(w_i) = 1 - \frac{H(w_i)}{\log_2 r}$$

($log_2 r$ は $H(w_i)$ の上限)

で求められ、$r(w_i)$ は語が偏って出現しているかという指標となる。この値が低ければ、語 w_i はどの分類でも出現する一般的な語であるといえる。

3. 出現頻度と重要度の関係

ここで、文章から抽出される語のうち、どのような語が重要であるかを、全節で述べた指標に基づいて調べた。ここでは、単に単語だけではなく、係り受け関係についての重要度についても調べた。全文章に対して形態素解析を行なった後、Shallow Parser を用いて係り受け解析を行ない、これを基に係り受け関係を抽出した。これらについては後節で延べる。

名詞 vs. 名詞複合語

まず、名詞および名詞複合語について、単語の出現頻度と重要度の関係を調べた。本実験で用いた形態素解析器によって品詞に加えて句の境界が出力されるので、これを用い、形態素解析器が出力した単語と、句の境界で区切ることで生成される単語とを比較する。

名詞句が N 個の短単位の単語から構成されている場合、この名詞句から複合語の候補の数は順列を考慮すると、$N(N+1)/2$ 個の候補が得られる。ここで、形態素解析器の出力する単語を「短単位の単語」と呼び、この複数の短単位の単語から構成される句中の単語を全て結合したものを、「長単位の単語」と呼ぶことにする。したがって、名詞句が短単位の語 1 語から構成される場合、短単位の語と長単位の語は同じとなる。

ここでは、句中に名詞 (一般名詞、固有名詞、一部未知語を含む) を含む句を名詞句と判断し、単語を抽出した。

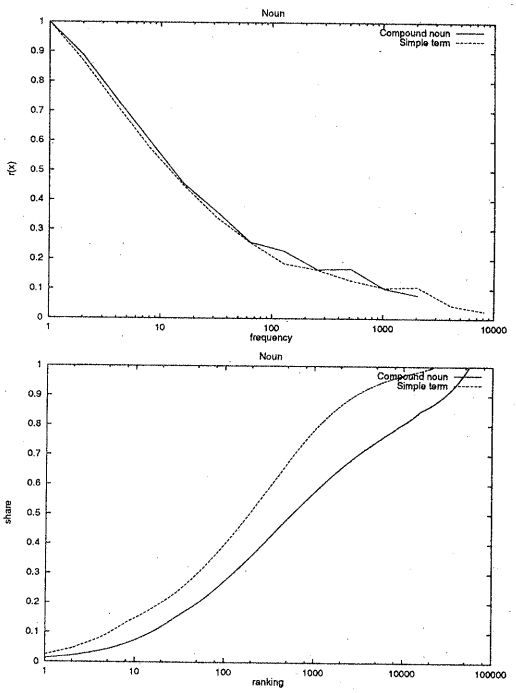

図 1　短単位の名詞 vs. 長単位の名詞

図 1・上に単語の出現頻度と冗長度 $r(S)$ の関係を、短単位の単語と長単位の単語について比較した結果である。図の横軸は単語の出現頻度を示し、対数スケールで示してある。同じ頻度を持つ語は複数存在するため、この図では同じ出現頻度の語は平均をとってある。図 1・下は、出現頻度順に単語をソートした場合、出

現頻度上位 N 語が全体に占める割合を示している。例えば、短単位の単語の場合、上位 100 語は全体の単語数の 約 40% を占める。

この結果、短単位の単語と長単位の単語では、重要度の傾向に差は見られず、出現頻度が同じ語であれば、出現頻度で平均した重要度はほぼ同じとなっている。

動詞語幹 vs. 動詞+意図を表す表現

名詞については名詞句から得られる短単位の語と長単位の語について比較を行なったが、動詞についても同様に、動詞句から語の抽出を行なった。ただし、動詞句については名詞句とは異なり、動詞句中に含まれる意図表現の抽出が重要であると考え、動詞原型と動詞+意図を表す表現、との比較を行なった。

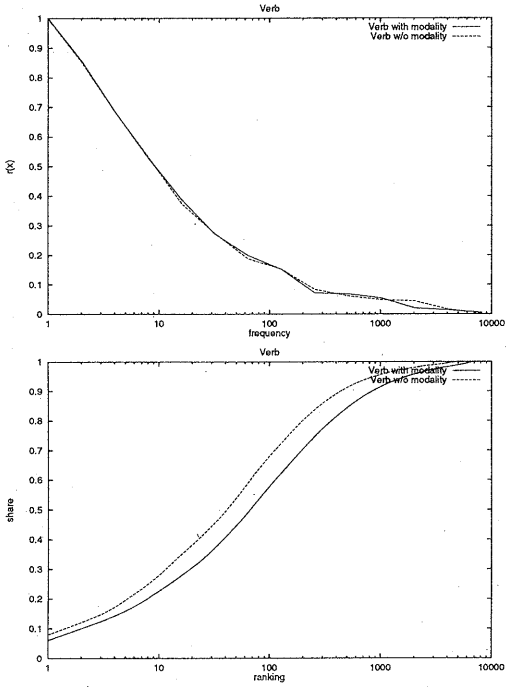

図 2 Verb (stem) vs. Verb (with modality)

意図を表す表現は願望 (~ほしい)、自発意志 (~たい)、否定 (~ない)、等の約 20 種類をパターンマッチで抽出し、動詞句中の動詞語幹に組み合わされて、語を生成する。例えば、「行きたくない」という動詞句からは、「行」(動詞語幹) + { 「たい」(願望) + 「ない」(否定) } という、動詞+意図を表す表現、が抽出され、この場合結果的には元の文字列と同じ「行きたくない」という語として抽出される。ここでは、この語幹のみから得られる「行く」という動詞の集合と、「行きたくない」という動詞+意図を表す表現から得られる語の集合との比較を行なった。

この結果 (図 2・上)、同一頻度で出現する語の平均では、重要度の傾向に差は見られず、出現頻度のみに依存する結果となっている。ただし、1つ1つの語に注目すると、意図表現を含めた場合、単語の出現頻度は各意図表現に散らばるため、低くなり、重要度は上がる。

名詞句 vs. 動詞句

また、名詞と動詞との比較を行なうために、上の2つの比較から名詞長単位と、動詞+意図を表す表現の結果を抜きだし、比較を行なった。その結果、どの頻度においても、名詞の方が動詞に対しては重要度が高いという結果となった。

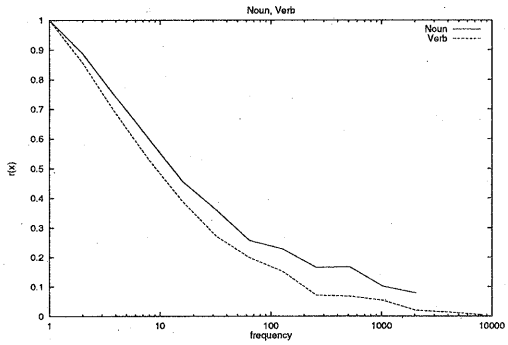

図 3 Noun vs. Verb

係り受け 名詞ー動詞 vs. 名詞ー名詞

また、2語の係り受け関係に関しても重要度を調べた。パーザによってフレーズ間の係り受け関係が得られるが、その中から名詞句と動詞句の係り受け関係と、名詞句と名詞句の係り受け関係について調べた (図 1・上)。この結果、名詞句と動詞句の比較では名詞句の方が重要度は高かったにも拘らず、か借り受け関係に注目すると、名詞ー動詞の係り受け関係の方が、名詞ー名詞の係り受け関係よりも重要度が高かった。

4. 重要語の抽出方法

4.1 中頻度語の抽出

語の重要度は語の頻度に依存し、一般的な傾向として、語の出現頻度が高くなる程、重要度が低くなることがわかった。中頻度語は重要度が高く、ある程度の出現頻度があるため、中頻度語にはテキストマイニングにとって重要語が多く含まれると考えられる。

中頻度語を多く抽出するためには、今まで Stop-Word List 等を用いて省かれていた高頻度語を、より Specific にすることで、中頻度の語にすることを

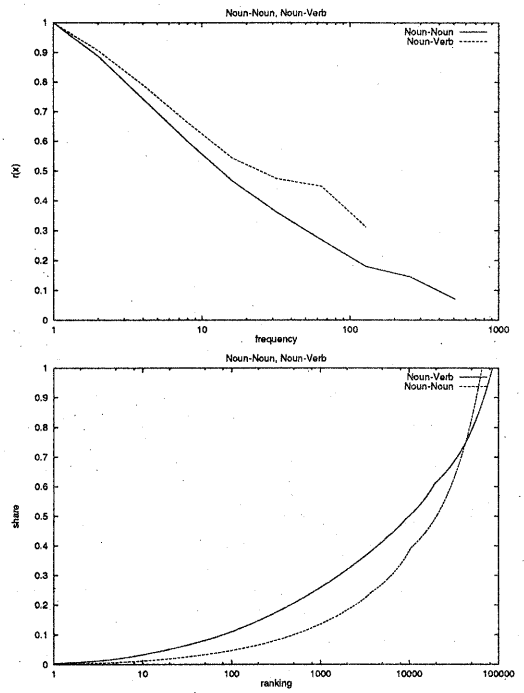

図 4 Noun-Verb vs Noun-Noun

考える。

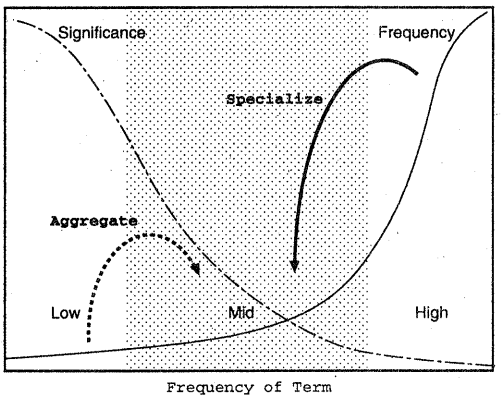

図 5 コンセプト

また、低頻度語は類義語辞書等によって１つの表現にまとめることで、中頻度の語にすることが必要となる。ただし、低頻度語には膨大な数の複合語があり、その為の複合語辞書を作るのは現実的ではない。また、低頻度語の多くの部分はミススペルや拗音撥音等の違いによるものである。したがって、一般的な高頻度語を組み合わせて、中頻度の複合語を作ることを考える。例えば、本稿で用いたPCコールセンターのデータの場合、短単位の単語「リカバリー」と「ＣＤ」はそれぞれ 9070 回、8756 回の高頻度で出現する。ただし、各単語自体には (PC コールセンターにとって) の意味は薄く、これらの短単位の単語から組み合わせて生成される語 (例えば、「リカバリーＣＤ」「アプリケーションＣＤ」) が重要である。

検索のような情報検索・情報抽出を目的とした場合、複合名詞「リカバリーＣＤ」が名詞句中にあれば、考えられる複合語「リカバリーＣＤ」「リカバリー」「ＣＤ」全てに対してインデックスを作っておけば、それぞれ目的の語を検索することができる。これに対してテキストマイニングの場合、このように全ての語を用いると、結果として「リカバリー」と「ＣＤ」には強い相関関係がある、といった結果が多く出力され、有用な結果が埋もれてしまうことが懸念される。

したがって、対象データから得られる単語集合の中から意味のある語を抽出することが重要である。

4.2 アルゴリズム

複合語は、句中の短単位の語の組み合わせによって得られるが、もちろん、全ての組み合わせに複合語としての意味があるわけではない。重要度が高く、意味のある複合語のみを抽出するために、以下のようにして重要と考えられる複合語を作成した。

(1) 全ての句に含まれる bi-gram (w_k, w_j) の数をカウントする。

(2) T_{ulim} を越える頻度で出現する bi-gram を取り出す。

(3) 閾値 T_{com} を越える頻度で出現する (w_k, w_j) を結合し、新しく複合語の候補として登録する。

(4) この複合語の重要度 $r(w_k w_j)$ を再計算し、この複合名詞の重要度が各要素の重要度 $r(w_k), r(w_j)$ を越えれば、複合名詞として登録する。

(5) 以上のプロセスを繰り返し行うことで、短単位の単語から複合語を生成する。

本来閾値の検定は行うべきであるが、今回は行っていない。

従来の重要語の抽出手法では、テキストマイニングのような統計的な処理を目的にしていないため、一般的な高頻度語は不要語として捨てられがちであるが、本手法を用いると、文中の全ての単語を利用することができる。これは、テキストマイニングのような統計処理にとっては重要なことであり、単に頻度が高いという理由で不要語としてしまうことはできない。

5. 実験と評価

5.1 自然言語処理部

まず、本研究で用いた自然言語処理について説明す

る。自然言語処理のモジュールは、形態素解析器・係り受け解析器(Shallow Parser)・意図表現抽出器(Modality Extractor)・複合語生成器(Term Extractor)から成り立っている。

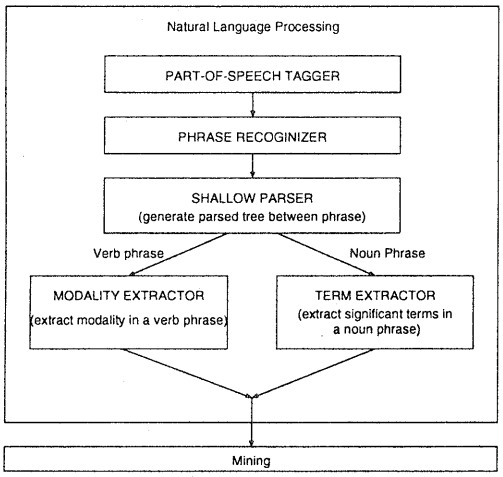

図 6 概要

形態素解析器は JMA (Japanese Morphological Analyzer)[2] を用いた。JMA は品詞に加えて句の境界を出力できるので、この出力で得られる句の境界を元に文を句に分割し、パージングを行い、各名詞句と動詞句に対して、意図表現の抽出または、複合語の生成を行う。

Shallow Parser

句構造の係り受け関係を生成するために、Shallow Parser を作成し、これを用いて係り受けを生成した。Strzalkowski[3] らは検索の Keyword Expansion のために句構造の係り受け構造を統計的に用いているが、本研究では、「何が」－「どうした」(ガ格には限らない) という関係が文の内容を表すのに有効であると考え、係り受け関係を抽出した。

Full Parser を用いた理由の 1 つは処理すべきデータ量が膨大 (今回の実験に用いたデータは約 21 万文) であることと、文の長さがそれほど長くないということである。今回用いたコールセンターのようなデータは新聞記事のように文が長くなく、比較的簡潔に記述されているからである。また、コールセンターのデータは、顧客との会話を元に構成されていることや、ログの蓄積目的として機械処理が前提とされていないため、文法的に正しくない文章が少なくない。このような理由から、重要語抽出においては Shallow Parser でも十分な精度が得られると考えた。

Parsing はパターンベースで行なわれ、以下のように高速で解析が行なわれるようになっている。

まず、前に述べたように JMA は句境界を含んだ、品詞タグつきの単語列を出力する。句境界から、文 T はフレーズ列 $p(i)$ $(0 \leq i \leq N_T)$ を出力し、各フレーズは単語列 $t_i(j)(0 \leq j \leq N_{Ti})$ から成り立っている。T の長さを N_T とし、フレーズ pi 中の単語列の長さを N_{Pi} とする。

ルールは品詞列から成り、例えば、ルール r は $\mathtt{r} = \{(\mathtt{prop}, \mathtt{conj}_1), (\mathtt{prop}, \mathtt{punct})\}$ (prop : 固有名詞, conj : 接続助詞 punct : 区点・終端記号) である。() 内の品詞の組は、それぞれフレーズ内の (自立語, 付属語) の組になっている。例えば、文 $\mathtt{T} = \{(\mathtt{prop}, \mathtt{conj}_1), (\mathtt{noun}, \mathtt{conj}_5), (\mathtt{prop}, \mathtt{punct})\}$ は上記のルールにマッチし、係り受け関係が抽出される。

Modality Extractor

動詞句の場合、句中にある特定の意図表現を抽出して、動詞の語幹に付与する。このモジュールは動詞句に含まれる (主に動詞句の付属語に含まれる) 意図表現を抽出して、動詞に意図表現を付与する。意図表現には、質問・要望・自発意志・否定、などを含み、他のモジュールと同様に簡単なルールで記述される。簡単な例では、動詞句に含まれる "?" は質問の意図を含むことから、質問の意図表現とされる。

Term Extractor

名詞句の場合、分野依存の固有名詞等を句中から抽出する。このモジュールでは前述のアルゴリズムを用いて、短単位の名詞から複合名詞を生成する。また、自明な複合名詞を作るためのルール、例えば、姓と名を結合するためのルールは品詞列 $\mathtt{r} = (\mathtt{noun}_{\mathtt{firstname}}, \mathtt{noun}_{\mathtt{lastname}})$ で与えられる。

5.2 マイニングアプリケーション

テキストマイニングの目的は、テキストを含むデータベースから、効果的に知識を抽出し、それらの知識を利用者に提示することである。そこで、我々は、上記の NLP を用いた知識発見の為のアプリケーションを開発した。このクライアントは時系列分析やキーワード間の相関関係と調べるといったマイニングの機能と、キーワード、およびフルテキストの検索といった基本的な IR の機能を備えている。

5.3 マイニング結果

このシステムの 1 つの機能である相関関係の抽出機能を用いて、本手法の評価を行った。データの対象分野の専門辞書が用意されていれば、ハードウェアとソフトウェアとの相関であったり、ハードウェアと動詞

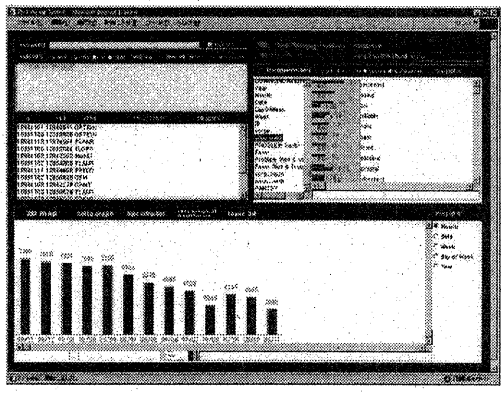

図 7 アプリケーション

の関係から、どのような箇所に問題があるのかを示すことができる。ここでは、名詞と動詞の共起関係を調べた。

前節で書いたように語の抽出にはいろいろな方法がある。例えば、名詞の場合、短単位の単語・名詞句全体 (長単位の単語)、また組合せ方によって多くの語が抽出できる。また動詞の場合、原型動詞や、意図表現を含めた動詞が抽出される。

これらのうちの抽出された語を用いてマイニング処理を行うのであるが、マイニング結果が正しい意味を持たなければ、マイニングとしては価値がない。例えば、「暑い」「暑くない」という語は原型のみを取り出せば同じく「暑い」になってしまうが、意味的には全く逆になってしまう。したがって、このうち動詞は意図表現が付随した語を固定して用いた。よって以下の実験では、名詞に関して、短単位の名詞、長単位の名詞、および本手法を用いた方法との比較を行った。

短単位の単語

表 2 に短単位の名詞と動詞 (形容詞・形容動詞を含む) との相関関係を相関の高い方からリストした。相関の度合は単に頻度の高いものから表示している訳ではなく、以下の条件を満たしたもののみ表示してある:

(1) 全ドキュメントの中で 10 回以上共起しているもの。

(2) 相対頻度が 10 倍以上の確率で出現しているもの

相対頻度は、以下のように計算される。例えば、名詞 "WINDOWS" は "起動する" や "シャットダウンする" のような多くの動詞と共起する。もし、"起動する" が "シャットダウンする" の 2 倍以上共起していれば、"WINDOWS" に関して "スタートアップする" の相対頻度は $(1+2)/2 = 1.5$ と計算される。

以上の条件を満たしたものを頻度上位からソートして示してある。順位を Rank で示している。語が複数の短単位の単語から構成されている場合は "-" で区切って表示してある。この場合、名詞は短単位の単語を用いたので、名詞には "-" で区切られた語は無い。動詞に関しては、意図を表す表現が含まれているので、"-" で区切られた語を含む。

表 2 短単位の語を用いた相関関係

Rank	Freq.	Noun	Verb
0	327	お願い	致す
1	229	キー	押す
2	79	転送	願う
3	74	メール	受信する
4	73	スキャンディスク	かかる
5	73	コンセント	抜く
6	33	ランプ	点灯する
8	65	HD	増設する
10	62	発信	聞こえ-ない
11	58	ネットワーク	出来-ない

特徴的に共起する語の組合せがあるが、短単位の単語の場合、多くの組合せは一般的な呼応関係を示すだけで、なにか問題として現象を表すものではない。

長単位の単語

次に、長単位の名詞を用いた結果を示す。それ以外の設定は短単位の単語と同じである。

表 3 長単位の語を用いた相関関係

Rank	Freq.	Noun	Verb
0	260	お願い	致す
3	58	発信-音	聞こえ-ない
5	41	画像	取り込む
8	30	対応-お願い	致す
10	25	引き上げ	願う
12	24	実行	指定する
13	23	電源-コンセント	抜く
15	22	左-上	点滅する
17	21	折り返し-ご-連絡	致す
19	17	購入-時期	送信する

中頻度語

表 4 に本手法を用いた結果を示す。表 2 のように単に、一般的な係り受けの関係だけではなく、より詳しく、有意な共起関係が得られている。

図 8 は以上の 3 つの方法で得られた強い相関関係を持つ語の共起がどのような頻度で出現しているかを示したものである。X 軸は、表 2 3 4 の Rank 、Y 軸はその語の共起の頻度を示している。

短単位の名詞を用いた場合の共起関係は、高頻度の組合せが多く、長単位の名詞を用いた場合の共起関係は低頻度の組合せが多い。本手法を用いることで、中

表 4　本手法で抽出した語を用いた相関関係

Rank	Freq.	Noun	Verb
0	260	お願い	致す
2	79	転送	願う
3	58	発信-音	聞こえ-ない
4	56	回避-方法	教える
8	36	電源-ランプ	点灯する
12	27	実行	指定する
14	25	修正-F D	届く
15	25	引き上げ	願う
19	23	左-上	点滅する
20	23	左-上	起動し-ない

頻度の相関関係を抽出することができた。

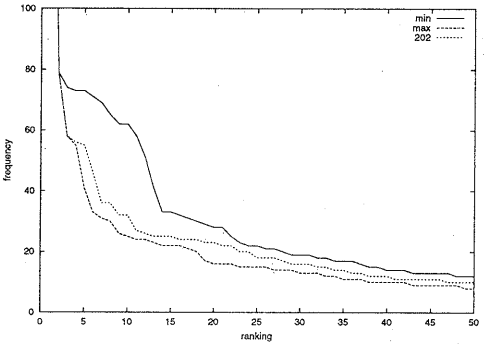

図 8　Ranking

6. おわりに

テキストマイニングは、知識を自動的に抽出することだけが目的ではなく、膨大なテキストデータがどのような性質を持っているかをユーザーに提示できる仕組みであると考える。自由に記述されたテキストデータは、様々な表現を含むため、単なる単語の集合を自然言語処理によっていかに知識として有用な語にすることが課題である。

また、今回用いたPC コールセンターのデータのように、予め記述されている内容が明らかでないようなデータから知識を得ようとする場合、得られる結果が知識としての価値があり、かつ自明な知識は省かれなければ、有用な知識を得ることは難しい。

本稿では、係り受け関係や意図表現を抽出することで意味のある単位での語の抽出を行い、また単語の重要度をエントロピーによって定義し、従来、不要語として扱われることの多かった高頻度語を複数組み合わせることで、意味のある中頻度語を得ることができた。

中頻度語が重要であるという知見自体は、従来の情報検索、情報抽出研究によって既に得られているものであり、目新しいものではない。ただテキストマイニングという応用を考えたとき、本稿で述べたように、単なる単語の集合を、いかに意味のある語にするかということが重要な課題である。

参考文献

1) Gerard Salton. *"SMART and SIRE Experimental Retrieval Systems."* 1983. McGraw-Hill.
2) Hiroshi Maruyama, et al. *"Japanese Morphological Analysis Based on Regular Grammar"* Transactions of Information Processing Society of Japan (IPSJ), Vol.35, No.7, pages 1293-1299. 1994.
3) Tomek Strzalkowski, et al. *"Information Retrieval using Robust Natural Language Processing."* Proceedings of the Association for Computational Linguistics (ACL1992). 1992.
4) Rakesh Agrawal. *"Mining Association Rules between Sets of Items in Large Databases."* Proceedings of the 1993 ACM SIGMOD, pages 207-216. 1993.
5) Hahn U. et al. *"Deep Knowledge Discovery from Natural Language Texts."* Proceedings of KDD-97 pages 175-178. 1997.
6) Knight M. *"Mining Online Text."* Communications of the ACM, Vol42, Number11, pages 58-61. 1999.
7) Mladenic D. et al. *"Text-Learning and Related Intelligent Agent: A Survey."* IEEE Intelligent Systems. Volume14, Number4, pages 44-54. 1999.
8) Marti A Hearst. *"Untangling Text Data Mining."* Proceedings of ACL-99, pages 3-10. 1999.
9) Kyo Kageura. *"Methods of Automatic Term Recognition - A Review -"* Tech rep., Terminology 3. 1996.
10) Philip Resnik. *"Using Information Content to Evaluate Semantic Similarity in a Taxonomy."* Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI95). 1995.
11) Tohru Hisamitsu, et al. *"Measuring Representativeness of Terms."* Proceedings of the SIGNL133-16 Information Processing Society of Japan. pages 115-122. 1999.
12) Slava M Katz. *"Distribution of Content Words and Phrases in Text and Language Modeling."* Natural Language Engineering 2(1), pages 15-59. Cambridge University Press 1996. 1995.
13) 那須川哲哉, 他 "テキストマイニング －膨大な文書データの自動分析による知識発見－" pp358-364, Vol.40, No.4, 情報処理. 1999.