統計数理 (2000)
第 48 巻 第 2 号 271–287

自然言語における統計手法を用いた情報処理

札幌学院大学* 金 明哲

(受付 2000 年 3 月 31 日；改訂 2000 年 9 月 18 日)

要 旨

近年，コンピュータが日本語を自由自在に扱えるようになったことと，機械的に言語処理を行う必要性が高まったことから，言語に関する研究が注目されつつある．特に，最近は言語データをもとにしたデータ主導型アプローチの研究が盛んに行われており，統計的手法による音声認識，音声合成，スペルチェック，形態素解析，機械翻訳，文の生成，テキストデータにおける情報検索及び情報の抽出，文書の自動分類，文章の書き手の推定・判別など列挙できない程の研究事例が報告されている．本稿では，確率モデルによる自然言語の処理と統計手法によるテキスト処理・解析に関する内容を中心として，その研究事例を紹介しながら，統計手法を用いた自然言語処理に関する研究の現状について述べる．

キーワード：統計的手法, 自然言語処理, テキスト処理.

1. はじめに

統計手法による言語の研究は斬新な話題ではない．例えば，広く知られている単語の出現頻度と順位との関係に関するジップの法則 (Zipf (1932)) の発見からすでに半世紀以上の時が流れているし，単語の長さに基づいた書き手の識別に関する試みは一世紀前から行われている (Mendenhall (1887))．ただ，コンピュータが自由自在に言語データを扱えない時期にはこのような研究は膨大な労力を必要としたため，統計手法による言語に関する研究は多くの人々の関心を集めなかった．しかし，80 年代前後からコンピュータが日本語を自由自在に扱えるようになったことと，機械的に言語処理を行う必要性が高まったことから，統計手法による言語に関する研究が多くの分野から注目されるようになった．特に，最近は言語データをもとにしたデータ主導型アプローチの研究が盛んに行われており，統計的手法による音声認識，音声合成，スペルチェック，形態素解析，機械翻訳，文の生成，テキストデータにおける情報検索及び情報の抽出，文書の自動分類，文章の書き手の推定・判別など列挙できない程の研究事例が報告されている．

自然言語に関する研究に用いられている統計数理の手法は多様である．本稿では，そのすべてについて触れることは不可能であるため，確率モデルによる自然言語の処理と統計手法によるテキスト処理に関する内容について，その研究事例を紹介しながら言語と統計との接点及びこれらの研究の現状について述べる．

*社会情報学科：〒069–8555 江別市文京台 11 番地.

2. 言語処理

言語処理では目的によって用いる単位が異なる．例えば，文字スペルチェックでは文字を単位とし，音声認識の場合は音素を単位とし，文法的に言語を処理する場合は単語，あるいは品詞を単位とする．文字，音素，単語，品詞，文節，文などを単位とした場合，単位ごとに記号で表すと言語処理は記号列の処理として見なすことが可能である．記号及び記号の組み合わせに関する統計的性質は記号列の処理に不可欠な情報である．

記号列において, 2 記号, 3 記号 , . . . , n 記号が隣接して出現する共起関係を 2-gram (bi-gram), 3-gram (trigram), . . . , n-gram という．大量の記号列から得られた n-gram の出現頻度に関する統計データはその記号列の処理において非常に重要な情報となる．

例えば，英文の場合アルファベットを単位とした 2-gram では q の後ろにはほとんどの場合 u が続くことから，q の後に続く文字が識別できない場合は u と断定しても間違う確率は非常に小さい．

記号列における n-gram に関する一つの統計量として，シャノンの情報量がある．$n = 1$ のとき，記号 $S = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}$ のシャノンの情報量は

$$SE = -\sum_{i=1}^{n} p(s_i) \log_2 p(s_i) \tag{2.1}$$

により定義され，単位はビット (bit) である．ここで $p(s_i)$ は $\sum_{i=1}^{n} p(s_i) = 1$ を満たす記号 s_i の出現率 (確率) である．SE を記号集合 S における 1 記号当たりのエントロピー (entropy) という．n-gram に関するシャノンの情報量は，n-gram の出現率のばらつきに関する統計量であると見なすことも可能である．出現率のばらつきが大きければ大きいほどその値は小さい．ちなみに英語のアルファベット 26 文字が均一に 1/26 の確率で使用されていると仮定した場合のシャノンの情報量は $SE = 4.7$ である．Shannon が行った英文字に関する統計では約 4.14 ビットである．これは英語のアルファベットの使用率は均一ではなく偏りがあることを意味する．中国語のピンイン (PinYin，中国語のローマ字表記) における 26 アルファベットにおけるエントロピーは 4.11 で，英語より低い．これは中国語のアルファベットの使用は，英語より偏っていることを意味する．$n = 2, 3$ の n-gram のシャノン情報量を比較するとその差はもっと明らかである (金・村上 (1992), 金 他 (1992))．n-gram のシャノン情報量は下記の式により定義される．

$$SE = -\sum_{i_1} \sum_{i_2} \cdots \sum_{i_n} P(s_{i_1}, s_{i_2}, \ldots, s_{i_n}) \log_2 p(s_{i_1}, s_{i_2}, \ldots, s_{i_n}) \tag{2.2}$$

シャノンの情報量は記号列の n-gram の性質に関する最も基本的な統計量であるため，言語計量分析に広く使用されている．

日本語の表記文字 (仮名，漢字) を単位とした場合は文字の種類が多いため，n-gram の統計的性質を見つけ出すことは英語ほど簡単ではないが，コンピュータの恩恵を受け，その計量分析及び統計的性質の応用が着実に進められている．１年間の朝日新聞における日本語の n-gram ($n = 1, 2, 3, 4, 5, 6$) に関する上位 40 位が長尾 (1996) に掲載されている．梅田 (1999) は日本語の苗字の 1〜3 文字組について分析を行っている．

n-gram は自然言語処理に幅広く用いられているが，その利用方法は基本的には同じで，処理対象と密接に関係している大量の言語データから n-gram の統計テーブルを作成し，その情報を用いてパターンマッチングを行う．

多字組 (文字を単位とした n-gram) の頻度表を利用した英単語の誤り訂正に関する早期の試みとしては, Cornew (1968) をはじめ, Harmon (1972), Riseman and Hanson (1974),

Hanson et al. (1976), Hul and Srihari (1982), Shinghal (1983) などがある．Suen (1979) は機械処理に役立つ情報を提供するために，1959 年から 1978 年の間に発表された英語の多字組に関する計量分析の論文を総括している． また，Yannakoudakis and Angelidakis (1988) は英語辞書における単語の n-gram ($n \leq 5$) の統計性質について分析を行った．

音声認識研究では，1970 年代後半から n-gram の統計データが用いられている (Jelinek (1976))．n-gram を用いた日本語における文字列の誤り訂正及び音声認識に関する研究が見られるようになったのは 1980 年代以後であり，早期の代表的な研究事例として池原・白井 (1984), 栗田・相沢 (1984)，鹿野 (1987, 1990) がある．

以下に，日本語を中心とした n-gram を用いた誤り検出及び訂正に関するいくつかの近年の研究事例を紹介する．

新納 (1999) は平仮名 n-gram ($n = 3, 4, 5, 6$) による平仮名列の誤り検出とその訂正の試みを行い，5 年分の新聞データより得られた n-gram のデータを用いて，平仮名列の誤り検出と修正を行う際には $n = 4$ が妥当であると報告している．

竹内・松本（1999）は，文字 3-gram と単語の 3-gram を用いて OCR 誤り訂正システムを構築し，90%の文字読み取り精度を約 92.9%まで，95%の精度を 96.4%まで改善した．また，森 他 (1999) は 2-gram の情報を用いた仮名漢字変換法を提案し，市販されている仮名漢字変換器 Wnn6 の 91.12%の再現率を 95.07%に，91.17%の適合率を 93.94%に改善し，確率モデルによる仮名漢字変換の有効性を実験的に示した．

森・長尾 (1998a) は辞書に登録されていない単語について，n-gram 統計情報を用いてコーパス (corpus) からその単語を抽出し，その単語が属する品詞を推定する方法を提案し，その方法が機械的に辞書を構築するのに有効であることを確認した．また，森・長尾 (1998b) は n-gram を工夫して用いることにより形態素解析の精度を向上させた．

太田 他 (1998) は文字の 2-gram の統計データを用いて，誤りを含む和文テキストにおける全文検索手法を提案した．提案した手法について検索効率の評価実験を行った結果，認識誤りを考慮しなければ 96.01%であった再現率が 99.26%まで改善された．

情報検索では複合語をどのように分割するかが一つの課題である．情報検索システムの索引付けのための複合語の分割に関して，Lee et al. (1999) はいくつかの方法について韓国語を用いて実験を行い，文字を単位とした 2-gram の統計情報を用いた方が単語辞書などを用いるより効果的であるという結果を得た．Nie and Ren (1999) は中国語について実験を行った結果，文字を単位とした 2-gram の統計情報は単語辞書と同等の効果を持ち，単語辞書と同じく重要であると報告している．

甲斐 他 (1999) は，単語の 2-gram の統計情報を用いた音声認識システムにおける未知語・冗長語の処理を試み，文脈自由文法に基づいた方法より有効であると強調している．荒木 他 (1999) は，n-gram を用いて会話文の処理の妨げとなる言い直しの音節列の検出方法を提案し，従来提案されてきた方法と比べ，適合率を約 13 ポイント，再現率を約 20 ポイント向上させた．

また Nagy (2000) は，IEEE-PAMI のレビューのなかで，欧米語における n-gram を用いた文字列の誤り検出と訂正に関する研究をまとめており，参考になる．

n-gram を言語処理に用いるにあたっては，直接 n-gram の統計情報を用いるほかに，問題に適するよう工夫を施すことも重要である．ここでは，最も広く使われている 2-gram を用いて n-gram の近似値を求めるマルコフモデルについて触れておく．

記号列 $s_{i_1}, s_{i_2}, \ldots, s_{i_{n-1}}$ の後に s_{i_n} が続く確率は

$$p(s_{i_n} \mid s_{i_1}, s_{i_2}, \ldots, s_{i_{n-1}}) = \frac{p(s_{i_1}, s_{i_2}, \ldots, s_{i_n})}{p(s_{i_1}, s_{i_2}, \ldots, s_{i_{n-1}})} \tag{2.3}$$

で求めることが可能である．ここで $p(s_{i_1}, s_{i_2}, \ldots, s_{i_n})$ は記号列 $s_{i_1}, s_{i_2}, \ldots, s_{i_{n-1}}, s_{i_n}$ の同時出現確率である．実際に言語処理を行う際には出現率を確率の推定値として用いる．

説明のため英文の品詞付けを例として用いる．一つの単語が複数の品詞属性を持つのは珍しいことではないが，英語の場合は日本語より複数の品詞属性を持つ単語が多い．一つの単語が複数の品詞属性を持っている場合，どのような品詞として解析すべきであるかは自然言語処理において非常に重要な問題である．例えば，英文 “Time flies like an arrow” の “flies” の原型は “fly” で，名詞として「蝿」，動詞として「飛ぶ」の意味を持ち，“like” は形容詞，前置詞，副詞，動詞，接続詞，名詞の属性を持っている．各単語の品詞の解釈が異なると当然それに対応する文の意味も異なる．“flies” を動詞とした場合，“Time flies like an arrow” は「光陰矢のごとし」と解釈できるが，“flies” を名詞とした場合，“Time flies like an arrow” は「時蝿は矢を好む」のように解釈することも可能である．一つの単語が複数の品詞属性を持っているとき，如何にその単語の品詞属性を決めるかに関する問題は，機械翻訳の研究が始まった初期から多くの研究者を悩ませてきた．

ここで入力単語列を $w_1, w_2, \ldots, w_i, \ldots, w_n$，この単語列に対応する品詞列を $h_1, h_2, \ldots, h_i, \ldots, h_n$ とする．この品詞列 $h_1, h_2, \ldots, h_i, \ldots, h_n$ は，下記の条件付確率を最大とする品詞列である．

$$p(h_1, h_2, \ldots, h_i, \ldots, h_n \,|\, w_1, w_2, \ldots, w_i, \ldots, w_n) \tag{2.4}$$

この式はベイズの定理によって

$$\frac{p(h_1, h_2, \ldots, h_i, \ldots, h_n) \times p(w_1, w_2, \ldots, w_i, \ldots, w_n \,|\, h_1, h_2, \ldots, h_i, \ldots, h_n)}{p(w_1, w_2, \ldots, w_i, \ldots, w_n)} \tag{2.5}$$

のように書き直される．分母の $p(w_1, w_2, \ldots, w_i, \ldots, w_n)$ は品詞と無関係である．分子の $p(h_1, h_2, \ldots, h_i, \ldots, h_n)$ は品詞列の n-gram の確率である．理論的には大量の言語データより得られた $p(h_1, h_2, \ldots, h_i, \ldots, h_n)$ の情報を直接用いることが可能であるが，n が大きくなると実用的ではない．品詞における遷移関係は直後の品詞 (2-gram) 間がもっと強い．そこで下記のように連続の 2-gram を用いて近似することが考えられる．

$$p(h_1, h_2, \ldots, h_i, \ldots, h_n) \cong \prod_{i=1}^{n} p(h_i \,|\, h_{i-1}) \tag{2.6}$$

ただし $p(h_1 \,|\, h_0)$ の h_0 は文頭の前の仮想された品詞である．同じく分子の右の条件付確率も次のように近似することができる．

$$p(w_1, w_2, \ldots, w_i, \ldots, w_n \,|\, h_1, h_2, \ldots, h_i, \ldots, h_n) \cong \prod_{i=1}^{n} p(w_i \,|\, h_i) \tag{2.7}$$

したがって，下記のような近似式が得られる．

$$\begin{aligned} &p(h_1, h_2, \ldots, h_i, \ldots, h_n) \times p(w_1, w_2, \ldots, w_i, \ldots, w_n \,|\, h_1, h_2, \ldots, h_i, \ldots, h_n) \\ &\qquad \cong \prod_{i=1}^{n} p(h_i \,|\, h_{i-1}) \times p(w_i \,|\, h_i) \end{aligned} \tag{2.8}$$

このモデルを自然言語処理の分野では隠れマルコフモデル (Hidden Markov Model, 略して HMM) と呼ぶ．$p(h_i \,|\, h_{i-1})$ は品詞の接続関係に関する統計データであり，$p(w_i \,|\, h_i)$ は単語

表 1. 形態素解析の例.

見出し語	読み方	原型	品詞情報
自然	しぜん	自然	副詞
言語	げんご	言語	普通名詞
処理	しょり	処理	サ変名詞
に	に	に	格助詞
は	は	は	副助詞
統計	とうけい	統計	サ変名詞
情報	じょうほう	情報	普通名詞
は	は	は	副助詞
欠かせ	かかせ	欠かせる	動詞，母音動詞，未然形
ない	ない	ない	形容詞性述語接，イ形容詞アウオ，基本形
.	.	.	句点

w_i の品詞 h_i における相対頻度である．この 2 種類のデータが揃えば，最適の品詞列を求めることが可能である．この隠れマルコフモデルによって，長年多くの研究者を悩ませてきた品詞付け問題は非常に高い確率で正解を得るようになった．HMM は言語処理のさまざまな分野で用いられている．

日本文は欧米文と異なり，分ち書きされていないため，品詞づけを行うためには，まず文を単語に切り分けることを行わなければならない．最近，形態素解析システムが多く開発されているが，もっとも広く知られているのは JUMAN (黒橋・長尾 (1998)) と Chasen (茶筅，松本 (1999)) である．JUMAN は連接可能性の辞書を用いたルールベースによる形態素解析システムである．JUMAN を用いて例文「自然言語処理には統計情報は欠かせない.」について形態素解析を行い，その結果を表形式に書き直したものを表 1 に示す．JUMAN による品詞付けの精度は約 94%である (竹内・松本 (1997))．

最近確率モデルによる日本語形態素解析の研究が進んでいる．竹内・松本 (1997) は日本語の特徴に合わせて HMM モデルに工夫を施し，JUMAN より約 2 ポイントの精度向上を実現した．政瀧・匂坂 (1999) は，品詞と可変長形態素列の複合 n-gram を用いた形態素解析手法を提案した．実験の結果，提案された方法の形態素同定率は最高約 99% に上り，JUMAN より品詞付けは約 4 ポイント，読み付けは約 5 ポイント高い正解率を得た．

1980 年代後半から音声認識・理解システムでは確率モデルを用いるのが主流となり，その音声処理部では HMM，言語処理部では 3-gram (trigram) が主に用いられている (Jelinek (1990), 牧野 (2000))．HMM モデルは音声認識だけではなく，テキストの音声合成でもその威力を発揮している (広瀬 (2000)，徳田 (1999)，Donovan and Woodlland (1995))．また，HMM は記号列の誤りの検出と訂正，仮名漢字変換，形態素解析，構文解析など自然言語処理全般に大きく貢献している．

言語と確率モデルに関する専門和書としては中川 (1988)，北 (1999) がある．前者は音声認識における確率モデル，後者では自然言語全般を射程に入れた確率モデルが詳細に述べられている．両者とも確率論の基礎から始まるので，確率論に関する基礎知識が無くても理解可能である．

自然言語処理には，確率モデルのほか決定木モデル，最大尤度，最大エントロピーモデル，EM アルゴリズム，ニューラルネットワークモデルなども用いられている．これらに関する情報は北 (1999)，内元・馬 (1999) から得られる．

3. テキスト処理

近年，インターネットの爆発的な普及によって，コンピュータで直接利用可能な電子化されたテキストデータが急増し，テキストの機械処理に関する研究がますます重要視されている．テキストに関する機械処理としては，テキストの分類（内容，書き手，ジャンルなど，何らかの特徴に基づいて分類），テキストにおける情報の検索と抽出などがあげられる．

3.1 テキストの分類

テキストの分類を行うにあたっては，処理対象となるテキストにおける何らかの特徴に関するデータで構成された，下記のようなデータマトリックス X を用いることが多い．

$$X = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_i \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{cccccc} c_1 & c_2 & \cdots & c_j & \cdots & c_m \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} \end{array}$$

X はテキスト $t_1, t_2, t_3, \ldots, t_i, \ldots, t_n$ における特徴カテゴリー $c_1, c_2, c_3, \ldots, c_j, \ldots, c_m$ の出現頻度に関するデータであり，x_{ij} はテキスト t_i での特徴カテゴリー c_j の出現頻度に関する値である．

コンピュータが今日のように普及していなかった時代では，マトリックス X の次元が高くなると複雑な処理が困難であるため，高次元のマトリックスの煩雑な計算を避け，カテゴリーの平均値，モードのような基本統計量や χ^2 統計量などが文章の分析に多く用いられた．コンピュータの性能の向上と手軽く使えるデータ解析ソフトウエアの普及に伴い，近年は多変量データ解析及び探索的なデータ解析手法が主流となっている．テキストの分析では，主成分分析，因子分析，数量化 III 類・対応分析，多次元尺度法が多く使われ，テキストの分類や文章の書き手の識別では判別分析，クラスター分析が多く使用されている．分類を行う際，学習情報があれば判別分析を用い，学習情報がない場合は，短所があるけれどもクラスター分析手法を用いているのが現状である．

判別分析は，線形判別と非線形判別に大別される．書き手の判別のために判別関数を用いた例としては，Cox and Brandwood (1958), Mosteller and Wallace (1963), 韮沢 (1965), 金 他 (1993b), Burrows and Craig (1994) がある．村田 (1999) は文章をジャンルごとに分類することを通じて，分類にもっとも寄与している論述構造を支える文型をつき止め，日本語の教育に役に立てようと試みている．非線形判別を用いた例としては，距離による判別分析手法を用いた書き手の判別に関する金 (1997) の研究，ニューラルネットワークによる書き手の判別に関する Tweedie et al. (1996) の研究がある．

近年データマイニングという言葉をよく耳にする．データマイニング法にはさまざまな方法がある．金 (1998a)，Chouchoulas and Shen (1999) はラフ集合理論 (村上・金 (1998)) に基づいて，学習データから書き手の判別ルールを生成し，書き手の識別と文章の分類を試みている．この方法による判別率は通常の判別分析より低いが，書き手の特徴ルールを機械的に抽出するのに有効である．

学習データがない場合，最も多く使用されている分類方法はクラスター分析である．ク

ラスター分析は階層的分析と非階層的分析に大別されるが，テキストの分類には階層的クラスター分析が多く用いられている．階層的クラスター分析を行う際の，テキスト間の類似度の尺度には距離と類似度の 2 種類があるが，どのような尺度を用いるかに関しては明確な基準がなく，データ解析のノウハウに頼っているのが現状である．

テキストの分類では類似度の尺度として内積

$$Sim(t_i, t_k) = \sum_{j=1}^{m} x_{ij} x_{kj} \tag{3.1}$$

コサイン相関値

$$\cos(t_i, t_k) = \frac{\sum_{j=1}^{m} x_{ij} x_{kj}}{\sqrt{\sum_{j=1}^{m} {x_{ij}}^2} \sqrt{\sum_{j=1}^{m} {x_{kj}}^2}} \tag{3.2}$$

が多く使用されている．

距離としては，市外地距離，ミンコフスキー距離，ユークリッド距離，重み付きユークリッド距離，マハラノビス距離が広く知られている．データが相対頻度であれば，Sibson (1969) が Kullback-Leibler 測度をもとに発展させた下記の距離も非常に有効である (McLachlan (1992))．ここではこの距離を K-L-S 距離と呼ぶことにする．

$$d(t_i, t_k) = \frac{1}{2} \sum_{j=1}^{m} \left(x_{ij} \log \frac{2x_{ij}}{x_{ij} + x_{kj}} + x_{kj} \log \frac{2x_{kj}}{x_{ij} + x_{kj}} \right) \tag{3.3}$$

ただし，$x_{ij} = 0$ ならば $x_{ij} \log \frac{2x_{ij}}{x_{ij}+x_{kj}} = 0$, $x_{kj} = 0$ ならば $x_{kj} \log \frac{2x_{kj}}{x_{ij}+x_{kj}} = 0$ とする．

階層的クラスター分析は，類似度や距離を用いてクラスター樹形図を生成し，枝と枝の距離が短いほど似ていると判断する方法である．樹形図生成アルゴリズムは多数提案されている．しかし，それらのアルゴリズムによって生成された樹形図の結果は必ずしも一致しない．どのアルゴリズムにより生成された樹形図がもっとも真の構造を表しているかに関する評価方法は提案されているが，その機能が実装された市販のデータ解析ソフトが少ないため，人文社会系の応用論文におけるクラスター樹形図の選択は，ほとんどの場合解析者の主観にたよっているのが現状である．用いたアルゴリズムが異なった場合, どのアルゴリズムによる樹形図が最も信頼できるかに関する評価の方法の一つとしては，用いた個体間の距離 (あるいは類似度) のマトリックス D と樹形図を生成するアルゴリズムで得られた樹形図における各個体間の距離のマトリックス C とのコーフェン相関係数 (Cophenetic correlation coefficient) を用いる方法がある (西田・佐藤 (1992))．金 他 (1993b), Jin and Murakami (1993) では三種類のクラスター樹形図についてコーフェン相関係数を求め，その値が最も大きい樹形図を用いて議論がなされている．両マトリックスに関する検定統計量を用いてクラスター樹形図を評価することも可能である．その検定方法としては Mantel 検定 (村上・田栗 (1992)) が考えられる．ただし，コーフェン相関係数を用いた結果とは大きく異ならないと考えられる．テキストの分類は一般のパターン認識の特別なケースであり，パターン認識に用いられている汎用の数理的方法はほとんど有効となる．統計的手法によるパターン認識に関する動向については Jian et al. (2000) を参考にすればよい．

テキストの分類に関する処理は，大きく二つの部分，

a. テキストから特徴データを抽出する
b. 適切な数理方法で分類を行う

に分けられる．統計数理研究者は分類の方法に力点を置いているが，用いるデータが分類に与える影響は計り知れない．

3.1.1 書き手ごとの分類

ある作家の作品を好んで読んでいる人は，文章を読むだけでその作家によるものであるか否かを 100%ではないが識別できるそうである．それはその作家の作品を大量に読むと，知らないうちにその作家の作品の独特な共通パターンが脳に焼付けられたためであると考えられる．このような，われわれ人間の脳で行われている文体に関する処理をコンピュータで行うには，どのようにすればよいであろうか？

このような処理を機械的に行うためには，文章から書き手の特徴と思われる情報を抽出しなければならない．書き手の特徴は文章を構成するさまざまな要素及びその要素の組み合わせに現れると考えられる．文章を構成する要素は記号，文字，単語，文，段落などに分けることが可能である．

文章を構成する最小の要素は記号と文字である．ここでいう記号は句点，読点，疑問符などを指す．記号の使用頻度，使用形態，文章中での配置位置とバランスなどは書き手の特徴情報になる場合がある．

読点の使用形態についてみよう．読点は，文章を書くとき，切れ・続きを明らかにするために，文の中の意味の切れ目につける符号である．そのなかで，文中の並立する要素の間に打つ読点は，どの書き手においてもそう変わるものではないが，それ以外の読点の打ち方に関しては，明確な基準がないため，書き手によって異なることが考えられる．

例えば，副助詞「は」の後ろで必ず読点を打つ人もいれば，場合によって打ったり打たなかったりする人もいるようである．本稿の読者の中でも多くの方は，文のある位置に読点を打つか否かについて迷った経験があるであろう．これは読点の打ち方に明確な基準がないからである．明確な基準がないと書き手の特徴が出やすい．

文章に現れている読点に関して書き手の特徴情報をいかに抽出するかについては，読点をどの文字の後に打つか，読点を打つ間隔，読点をどの品詞の後に打つかなどが考えられる．このような読点の打ち方について計量分析を行った結果，読点をどの文字・単語の後に打つかに関する情報に書き手の特徴が最も明確に現れ，かつ得られた情報が最も安定していることがわかった (Jin and Murakami (1993)，金 (1994))．また，読点の打ち方に書き手の特徴が明確に見られるのは文学作品だけではなく，論文においても書き手の特徴が明確に現れることが実証された (金 他 (1993b))．吉岡（1999）は現代新書 40 冊について計量分析の結果，読点は書き手の特徴として有力な情報になるということを実証した．読点の打ち方には書き手の特徴が明確に現れ，文学作品に限らず，論文の著者の識別にも有効である．ただし，文章のなかで，使用された読点が少ない，例えば日記のような極端に短い文章の書き手を識別するのは困難であろう．記号「！」,「……」,「？」,「ー」などの使用頻度にも書き手の特徴が現れる可能性もあるが，問題はそれに関するデータが文章の内容と関係なく，書き手において安定性を持っているか否かである．

日本文を構成する主な要素は漢字と仮名であるため，漢字と仮名が文中に占める割合も書き手の特徴情報となる可能性がある．一般的に，漢字の素養が高ければ漢字を多く使用すると思われがちである．確かに，漢字の使用率に書き手の特徴が現れるケースも少なくない (金 他 (1993a))．しかし，同じ書き手でも，異なる作品のなかで使用されている漢字の比率は，必ずしも安定しているとは限らない．それは文中に用いる漢字・仮名の使用率は文章の内容に依存するからである．

文章を構成する要素の単位を「単語」とした場合，単語における書き手の特徴情報は，好

んで使用されている単語，単語の組み合わせなどに現れると考えられる．好んで使用されている単語に関してはどのように計量するかが問題である．文体分析に最も早く用いられた方法は，文中に現れた単語の長さによる分析である．Mendenhall (1887) が単語の長さによる書き手の識別・推定に関する論文を有名なジャーナル Science に発表してからすでに 110 年以上過ぎている．単語の長さの情報は，欧米では，書き手の識別などによく用いられている (Fucks (1952)，Williams (1956, 1975)，Herdan (1958), Brinegar (1963))．日本文の文体研究における単語の長さに関する実証的な研究は，最近ようやく行われつつある．日本語では何を単位として単語の長さを計量するかが問題であるが，もっとも簡単なのは表記文字を単位とした計量方法である．表記文字を単位として計量分析を行った結果，日本文でも単語の長さに書き手の特徴が現れることが明らかにされた (金 (1995，1996))．

また，より有効な書き手の特徴を抽出するためには，単語を品詞ごとに分けて処理した方がよいという提案もある (金 (1995，1996))．これは，すべての単語を用いると，文章の内容に依存性が高い単語が書き手の特徴を弱めてしまうことが考えられるからである．

文章の中から安定した書き手の特徴を抽出するためには，なるべく文章の内容と関連性が薄い要素を用いるべきである．品詞を分析の単位とした場合，文章の内容と最も依存性が低いのは助詞である．助詞の出現率は品詞の中で約 30%〜40%を占め，その使用率は最も高く，かつ通常頻繁に使われている助詞は約 20 種類前後であるため，計量分析にも都合がよい．助詞は文章の内容に関する依存性が低く，また読点と異なり，文・文章の作成に必ず使用しなければならないため，日記のような短い文章の場合でも有効である．金 (1997) は日記の文章に関して，助詞に関する情報のみで判別分析を試み，約 95%の判別率で書き手が判別できるという結果を得た．この分析で用いた日記の長さは，短いのは約 200〜300 単語である．

村上・今西 (1999) は源氏物語について助動詞に焦点をあてて分析を行い，宇治十帖における助動詞の使用形態がその前の 44 巻との間に差が見られるということを明らかにした．

また，書き手が好んで使用している単語の使用頻度の情報を用いることも多い (Mosteller and Wallace (1964)，Morton (1965)，Burrows (1987), Burrows and Hassal (1988)，伊藤・村上 (1992))．そのほかに，語彙量 (Yule (1944)，Efron and Thisted (1976), Thisted and Efron (1987)，Tweedie and Baayen (1998))，品詞の使用率 (樺島・寿丘 (1965)，村上 (1999))，文の長さ (Yule (1939)，Williams (1940)，Wake (1957)，Morton (1965))，段落の長さ (樺島・寿丘 (1965))，会話の比率 (樺島・寿丘 (1965))，色彩語・比喩語などの使用率 (安本 (1958, 1981, 1994，2000))，文頭・文末のパターン (Cox and Brandwood (1958), Milic (1967))，音韻の特徴 (Fucks (1952, 1954))，統語パターン (Potter (1989)) なども書き手の特徴情報として用いられている．

このような文体の計量分析に関するより詳細な情報は，欧米文に関しては参考文献 Holmes (1994), 日本語現代文に関しては参考文献 金 (1998b, 1999) を参照してほしい．

文章に見られる書き手の特徴は，書き手によって現れる形態が異なるため，書き手の識別・判別などを行う際，特定の要素についてこのような分析手法を用いるべきであるというような公式がない．また書き手の特徴は，記述する内容や執筆時の環境，情緒，経年に伴い変化することも十分考えられる (金 (2000))．

3.1.2 内容ごとの分類

文書の内容・テーマごとの機械的な分類は，情報検索や電子化されている大量の文書の管理・利用の基礎となる．

文書の自動分類も学習データを用いた分類と学習データを用いない分類に分けられる．学

習データを用いた分類とは，文書をあらかじめ決められたグループ（例えば，新聞記事をあらかじめ分けておいた「経済」,「娯楽」,「社会」,「政治」などのグループ）に，分類することである．

Iwayama and Tokunaga (1994) は確率モデル (SVMV) に基づいた文書の自動分類を試み，0.63 の適合率を得ている．間瀬 他 (1998)，福本・鈴木 (1999)，Marc (1995) は文書間の類似度による分類を行っている．近年の研究動向をみると，類似度による分類が主流である．間瀬 他 (1998) は学習データのキーワードベクトルと分類すべき文書におけるキーワードベクトルとの重みつき類似度を用いて，特許を分野ごとに分類する方法を提案した．重みはキーワードをいくつかの項目に分け，項目ごとに配分している．福本・鈴木 (1999) は「語の重み付け学習」の手法による文書の自動分類手法を提案した．ここでいう「語の重みつけ学習」とは，事前に分類されている学習用データを用いて分類を行い，分類が誤った文書について正しく分類されるよう用いた語 (キーワード) に重みをつける方法である．重みをつけるべきである語の選定は，χ^2 統計量を工夫して用いている．また重みは，正しく分類できるように繰り返し学習を行う方法で定めている．そして Iwayama and Tokunaga (1994) と同じデータを用いてテストを行った結果，Iwayama and Tokunaga の確率モデル (SVMV) より高い適合率 (0.75) を得た．

河合 (1992) は意味属性に関する情報を用いて文書の分類を行い，湯浅 他 (1995) は単語の共起関係の情報を用いて文書の分類を試みている．

金・宮本 (1999) は作文の文章の中に使用されているすべての名詞を抽出し，シソーラス辞書を用いて作成した意味的な同義語グループをカテゴリーとしたデータセット X を作成し，距離による判別分析で (110 の個体を 10 グループに分類) 内容・テーマごとに文章の分類を試み，90%台の判別率を得ている．

3.2 情報検索

情報検索は，文献検索あるいは文・テキスト検索と同義である (Ingwersen (1992))．情報検索に関する研究は約半世紀前から行われているが，初期段階の情報検索はキーワード，著者の氏名，出版社，発刊の年月日などの情報に基づいて，必要となる図書，資料，論文などを検索するのが主であった．今日の情報検索は，従来のものと比べ検索範囲が広くなり，検索端末から全文を閲覧することも前提としている．

最もシンプルな情報検索の方法は，検索質問としてキーワードを入力し，入力されたキーワードが検索対象の中に含まれているかをチェックし，含まれていればその文書を検索対象として出力する方法である．この方法をさらに発展させたものとして，検索の索引語を用いる方法がある．索引語とは文書を特徴づける単語 (キーワード) に，その文書の中における重要度に関する情報を付加した単語ベクトルである．重要度の計算に用いられている最も基本となる計算式は

$$w_{ik} = f_{ik} \log \left(\frac{N}{n_k} \right) \tag{3.4}$$

である．f_{ik} は検索質問ベクトルの中のカテゴリー c_k の文書 t_i における出現頻度，N は検索対象の総文書数，n_k は c_k が含まれている文書の数を示す．索引ベクトルの重みを求める式は数多く提案されている．Lee et al. (1999) は 128 種の重みの計算式について韓国語を用いてテストを行い，下記の式が最も効果的だったと報告している．

$$w_{ik} = \frac{\left(0.5 + 0.5\frac{f_{ik}}{\max f}\right)\ln\frac{N}{n_k}}{\sum\left[\left(0.5 + 0.5\frac{f_{ik}}{\max f}\right)\ln\frac{N}{n_k}\right]^2} \tag{3.5}$$

入力された検索質問について自動的に索引付けを行う研究もさまざまな角度から行われている．藤田 (1999) は日本語の場合は，単語または語句ベースの索引語を利用した方が，より進んだ問題解決能力につながるシステムを実現できると考えている．第 2 章で取り上げた Lee et al. (1999), Nie and Ren (1999) も自動索引付けに関する研究である．

表記のマッチングによる検索では，意味的には同じであっても表記が異なる場合は検索の対象にはならない．そこで，大井 他 (1997) は意味的類似度と多義解消を用いた類似検索方法を提案した．

検索質問が文・文書である場合は，文・文書を意味的な単位に分割 (形態素解析) し，検索質問の文・文書の意味的単語ベクトルと検索対象となる文書との類似度を算出する方法で検索候補を選び出すことが可能である．これは前節で述べた金・宮本 (1999) の文書の自動分類とも大きく関係する．

3.3 自由回答文の分類と解析

アンケート調査の結果をコンピュータで分析するのはごくあたり前のことであるが，自由回答文に関しては，分析者が回答文を読んで分析・解析するのが一般的である．しかし，大量のアンケート調査の場合には，一つ一つの回答文を読んで分析・解析するのは至難である．そこで，自由回答文をコンピュータで分析・解析する試みが行われている．

川端 (1999) は霊能者の世界観を明らかにするために，ある宗教団体の霊能者に対するアンケート調査の自由回答文のコンピュータによる分析・解析を試みた．この研究では，自由回答文 (平均の長さはおよそ 100 文字) を電子化し，意味のある文字列 490 コート (パターン) について相互間の関連性を算出し，その中から関連性が高い文字列を用いて分析を試みている．

アンケート調査の自由回答文に何らかの傾向を見ることは，自由回答文を内容ごとに分類し，分類されたグループごとの特徴をつかむことと等価であると考えてもよいであろう．アンケート調査の自由回答文の機械的解析に関する研究としては，川端 (1999) 以外に乾 他 (1998)，大隈 (2000) がある．

4. むすびにかえて

本稿では，統計手法と自然言語における情報処理の接点について，研究事例を用いて概述した．統計手法を用いた自然言語における情報処理・解析の応用事例をもれなくまとめるのは筆者の能力では不可能に近い．またすべてを包含しようとすると焦点がぼやけるので，本稿では，確率モデルによる自然言語の処理と，統計手法によるテキスト処理・解析に関する研究の現状に限った．

情報過多・情報洪水と言われている現在，蓄積されているデータの約 8 割がテキスト形式であるという説もある．さまざまな情報が付加され電子化されたコーパス・言語データの急速な増加に伴い，コーパスに関する計量的な研究もまた注目を集めている．興味をもたれた方には専門書として斎藤 他 (1998) を薦める．本稿のテキスト処理に関する内容をコーパスの利用に関する研究とみなしても何の差支えもないであろう．

膨大なテキストデータから必要となる情報を掘り出す意味で，テキストマイニング (Min-

ing)，情報マイニングという言葉をよく耳にする．テキストマイニング, 情報マイニングはデータマイニングが語源で (那須川 他 (1999))，テキストデータから必要となる情報を抽出することを指す．第 3 章で述べた文書の分類，情報検索，自由回答文の機械処理などもテキストマイニングの一部分である．

本稿では構文解析，機械翻訳，文の生成など多くの自然言語処理範疇内の話題に触れていない．興味のある方には Manning and Schütze (2000)，長尾 (1996) を薦める．前者は統計的手法による自然言語処理に関する専門書であり，後者は広義の自然言語処理に関する専門書である．

謝　辞

本稿を丁寧に読んで，ご助言をくださった査読者と宮崎大学の藤井良宜先生に感謝いたします．

参 考 文 献

荒木哲郎，池原　悟，三品尚登 (1999)．N-gram を用いた対話文の言い直し表現の検出，自然言語処理，**6**(3)，23–41．

Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass letter, *J. Amer. Statist. Assoc.*, **58**, 85–96.

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Clarendon, Oxford.

Burrows, J. F. and Craig, D. H. (1994). Lyrical drama and the "Turbid Mountebanks": Styles of dialogues in Romantic and Renaissance tragedy, *Computers and Humanities*, **28**, 63–86.

Burrows, J. F. and Hassal, A. J. (1988). Anna Boleyn and the authenticity of fielding's feminine narrative, *Eighteenth-Century Studies*, **21**, 427–453.

Chouchoulas, A. and Shen, Q. (1999). Arugh set-based approach to text classification, *Lecture Notes in Artificial Intelligence*, **1711**, 118–127.

Cornew, R. W. (1968). A statistical method of spelling correction, *Inform. and Control*, **12**, 79–93.

Cox, D. R. and Brandwood, L. (1958). On a discriminatory problem connected with the words of Plato, *J. Roy. Statist. Soc. Ser. B*, **21**, 195–200.

Donovan, R. and Woodland, P. (1995). Automatic speech synthesizer parameter estimation using HMMs, *Proc. ICASSP*, **95**, 640–643.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know?, *Biometrika*, **63**(3), 435–447.

Fucks, W. (1952). On mathematical analysis of style, *Biometrika,* **39**, 122–129.

Fucks, W. (1954). On Nahordnung and Fernordnung in samples of literary texts, *Biometrika*, **41**, 116–132.

藤田澄男 (1999)．自然言語処理を利用した情報検索・分類へのアプローチ，*JPSJ Magazine*, **40**, 352–357.

福本文代, 鈴木良弥 (1999)．語の重み付け学習を用いた文書の自動分類，情報処理学会論文誌，**40**, 1782–1791.

Hanson, A. R., Riseman, E. M. and Fisher, E. (1976). Context in word recognition, *Pattern Recognition*, **8**, 35–45.

Harmon, L. D. (1972). Automatic recognition of print and script, *Proc. IEEE*, **60**, 1165–1176.
Herdan, G. (1958). The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics, *Biometrika*, **45**, 222–228.
広瀬啓吉 (2000). 21 世紀に向けての音声合成の技術展望，情報処理，**41**，277–281.
Holmes, D. I. (1994). Authorship attribution, *Computer and the Humanities*, **28**, 87–106.
Hull, J. J. and Srihari, S. N. (1982). Experiments in text recognition with binary n-gram and Viterbi algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-4**, 520–530.
池原 悟，白井 諭 (1984). 単語解析プログラムによる日本文誤字の自動検索と二次マルコフモデルによる訂正候補の抽出, 情報処理学会論文誌, **25**, 298–305.
Ingwersen, P. (1992). *Information Retrieval Interaction*, Taylor Graham, London (藤原鎮男監訳 (1995).『情報検索研究』，トッパン，東京).
乾 伸雄，内元清貴，村田真樹，井佐原均 (1998). 文末表現に着目した自由回答アンケートの分類，情報処理学会研究報告，98-NL-128, 181–188.
伊藤瑞叡, 村上征勝 (1992). 三大秘法禀承の計量文献学的新研究, 大崎学報，**148**，1–52.
Iwayama, M. and Tokunaga, T. (1994). A probabilistic model for text categorization: Based on a single random variable with multiple values, *Proceedings of 4th Conference on Applied Natural Language Processing*, 162–167.
Jelinek, F. (1976). Continuous speech recognition by statistical methods, *Proc. IEEE*, **64**(4), 532–556.
Jelinek, F. (1990). Self-organized language modeling for speech recognition, *Tradings in Speech Recognition* (eds. A. Waible and K. Lee), 450–506, Morgan Kaufman.
Jian, A. K., Duin, R. P. W. and Mao, J. C. (2000). Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 4–37.
金 明哲 (1994). 読点の打ち方と著者の文体特徴，計量国語学，**19**，317–330.
金 明哲 (1995). 動詞の長さの分布に基づいた文書の分類と和語および合成語の比率，自然言語処理，**2**(1), 57–75.
金 明哲 (1996). 動詞の長さの分布と文書の書き手，社会情報，**5**(2), 13–22.
金 明哲 (1997). 助詞の分布に基づいた日記の書き手の認識，計量国語学，**20**, 357–367.
金 明哲 (1998a). 助詞分布における書き手の識別ルールの抽出，言語処理学会第 4 回年次大会予稿集， 676–679.
金 明哲 (1998b). 日本語における計量文体学の近年の進展，情報，**1**(2), 57–64.
金 明哲 (1999). 日本現代文における書き手の特徴情報，人文学と情報処理, **20**, 64–71.
金 明哲 (2000). 文体の変化，*ESTRELA*, 5 月号, 77–80.
金 明哲，宮本加奈子 (1999). ラフな意味情報に基づいた文章の自動分類，言語処理学会第 5 回年次大会発表論文集，235–238.
金 明哲，村上征勝 (1992). 中国語高頻度単語のピン音表記の統計的特性, 統計数理, **40**, 131–151.
Jin, M. and Murakami, M. (1993). Authors' characteristic writing styles as seen through their use of commas, *Behaviormetrika*, **20**, 63–76.
金 明哲，徳田尚之，村上征勝，田中栄一 (1992). 音声学的観点からの中国語高頻度単語の計量分析, 行動計量学, **19**, 49–65.
金 明哲，樺島忠夫，村上征勝 (1993a). 手書きとワープロによる文章の計量分析, 計量国語学, **19**, 133–145.
金 明哲，樺島忠夫，村上征勝 (1993b). 読点と書き手の個性，計量国語学，**18**，382–391.
樺島忠夫，寿岳章子 (1965).『文体の科学』，綜芸舎，京都.
甲斐充彦，広瀬良文，中川聖一 (1999). 単語 N-gram 言語モデルを用いた音声認識システムにおける未知語, 冗長語の処理，情報処理学会論文誌，**40**, 1383–1395.

川端 亮 (1999). 非定型データのコーディング・システムとその利用，文部省科学研究費補助金課題番号 08551003 研究成果報告書.
河合敦夫 (1992). 意味属性の学習結果に基づく文書の自動分類方式，情報処理学会論文誌，**33**，1114–1122.
北 研二 (1999).『確率的言語モデル』, 東京大学出版会，東京.
栗田泰一郎, 相沢輝照 (1984). 日本語に適した単語の誤入力訂正法とその大語彙単語音声認識への応用, 情報処理学会論文誌, **25**, 831–841.
黒橋禎夫，長尾 眞 (1998). 日本語素解析システム JUMAN version 3.5, http://www.lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html.
Lee J. H., Cho, H. Y. and Park, H. R. (1999). N-gram-based indexing for Korean text retrieval, *Information Processing and Management*, **35**, 427–441.
牧野正三 (2000). 21 世紀に向けての音声認識，情報処理，**41**，273–276.
Manning, C. D. and Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.
Marc, D. (1995). Gauging similarity with n-grams: Language-independent categorization of text, *Science*, **267**, 843–848.
間瀬久雄，辻 洋，絹川博之，石原正博 (1998). 特許テーマ分類方式の提案とその評価実験，情報処理学会論文誌，**39**, 2207–2216.
松本祐二 (1999). 日本語形態素解析システム茶筅 (Chasen)，http://cactus.aist-nara.ac.jp/index.html.
McLachlan，G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Interscience, New York.
Mendenhall, T. C. (1887). The characteristic curves of composition, *Science*, **IX**, 237–249.
Milic, L. T. (1967). *A Quantitative Approach to the Style of Jonathan Swift*, Mouton, The Hague.
森 信介，長尾 眞 (1998a). N グラム統計によるコーパスからの未知語抽出，情報処理学会論文誌，**39**, 2093–2100.
森 信介，長尾 眞 (1998b). 形態素クラスタリングによる形態素解析精度の向上，自然言語処理，**5**, 75–99.
森 信介，土屋雅稔，山地 治，長尾 眞 (1999). 確率的モデルによる仮名漢字変換，情報処理学会論文誌，**40**, 2946–2953.
Morton, A. Q. (1965). The authorship of Greek prose, *J. Roy. Statist. Soc. Ser. A*, **128**(2), 169–233.
Mosteller, F. and Wallace, D. L. (1964). Inference in an authorship problem, *J. Amer. Statist. Assoc.*, **58**, 275–309.
村上征勝 (1999). 源氏物語の計量分析，人文学と情報処理，**20**，81–86.
村上征勝，今西祐一郎 (1999). 源氏物語の助動詞の計量分析，情報処理学会論文誌，**40**，774–782.
村上征勝，金 明哲 (1998).『人文科学とコンピュータ講座，第５巻 数量的分析編』，尚学社，東京.
村上正康，田栗正章 (1992).『多変量データ解析基礎』，培風館，東京.
村田 年 (1999). 論述構造を支える文型の基礎的研究── 多変量解析によるジャンル判別に有効な文型抽出──，専門日本語教育研究，**1**，32–39.
長尾 眞 (1996).『自然言語処理』，岩波書店，東京.
Nagy, G. (2000). Twenty years of document image analysis in PAMI, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 38–62.
中川聖一 (1988).『確率モデルによる音声認識』，電子情報通信学会，東京.
那須川哲哉，諸橋正幸，長野 徹 (1999). テキストマイニング，*JPSJ Magazine*, **40**, 358–373.
Nie, J.-Y. and Ren, F. (1999). Chinese information retrieval: Using characters or words?, *Information Processing and Management*, **35**, 443–462.

韮沢　正 (1965). 由良物語の著者の統計的判別, 計量国語学，**33**，21–28.
西田英朗，佐藤　嗣 (1992).『実例クラスター分析』，内田老鶴圃，東京.
大井耕三，隅田英一郎，飯田　仁 (1997). 意味的類似性と多義解消を用いた文書検索方法，自然言語処理，**4**(3), 51–69.
大隅　昇 (2000). 定性情報のマイニング ── 自由回答データ解析 ──, *ESTRELA*, **5**, 14–26.
太田　学，高須淳宏, 安達　淳 (1998). 認識誤りを含む和文テキストにおける全文検索手法，情報処理学会論文誌，**39**, 625–635.
Potter, R. G. (1989). *Literary Computing and Literary Criticism*, University of Pennsylvania Press, Philadelphia.
Riseman, E. M. and Hanson, A. R. (1974). A contextual postprocessing system for error correction using binary n-grams, *IEEE Trans. Comput.*, **C-23**, 480–493.
斎藤俊雄，中村純作，赤野一郎 (1998).『英語コーパス言語学』，研究社，東京.
政瀧浩和，匂坂芳典 (1999). 品詞及び可変長形態素列の複合 N-gram を用いた日本形態素解析，自然言語処理，**6**(2), 41–57.
Shannon, C. E. (1951). Prediction and entropy of printed English, **BSTJ**, 55–64.
Shinghal, R. (1983). A hybrid algorithm for contextual text recognition, *Pattern Recognition*, **16**, 261–267.
Sibson, R. (1969). Information radius, *Z. Wahrsch. Verw. Gebiete*, **14**, 149–160.
鹿野清宏 (1987). Trigram Model による単語音声認識結果の改善, 信学技報, SP87–23.
鹿野清宏 (1990). 統計手法による音声認識, 電子情報通信学会誌, **73**, 1276–1285.
新納浩幸 (1999). 平仮名 N-gram による平仮名列の誤り検出とその修正，情報処理学会論文誌，**40**, 2690–2698.
Suen, C. Y. (1979). n-gram statistics for natural language understanding and text processing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-1**(2), 164–172.
竹内孔一，松本裕治 (1997). 隠れマルコフモデルによる日本語形態素解析システムのパラメータ推定，情報処理学会論文誌，**38**, 500–509.
竹内孔一，松本裕治 (1999). 統計的言語モデルを用いた OCR 誤り訂正システムの構築，情報処理学会論文誌，**40**, 2679–2689.
Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly-discovered poem?, *Biometrika*, **74**, 445–455.
徳田恵一 (1999). 隠れマルコフモデルの音声合成への応用，電子情報通信学会技術研究報告，SP99-61, 47–54.
徳永健伸 (1999).『情報検索と言語処理』，東京大学出版会，東京.
Tweedie, F. J. and Baayen, R. H. (1998). How variable ma a constant be? Measures of lexical richness in perspective, *Computers and the Humanities*, **32**, 323–352.
Tweedie, F. J., Singh, S. and Holmes, D. I. (1996). Neural network application in stylometry: The federalist papers, *Computer and the Humanities*, **30**, 1–10.
内元清貴，馬　青 (1999). 形態素構文解析，人文学と情報処理，**21**, 13–29.
梅田三千雄 (1999). 日本の苗字の計量分析，情報処理学会論文誌，**40**，796–804.
Wake, W. C. (1957). Sentence-length distribution of Greek authors, *J. Roy. Statist. Soc. Ser. A*, **20**, 331–346.
Williams, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style, *Biometrika*, **31**, 356–361.
Williams, C. B. (1956). Studies in the history of probability and statistics, *Biometrika*, **43**, 248–256.
Williams, C. B. (1975). Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, *Biometrika*, **62**, 207–211.

Yannakoudakis, E. J. and Angelidakis, G. (1988). An insight into the entropy and redundancy of the English dictionary, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **10**, 960–969.
安本美典 (1958). 著者の推定──源氏物語, 宇治十帖の著者について──, 心理学評論, **2**, 147–156.
安本美典 (1981). 『因子分析法』, 培風館, 東京.
安本美典 (1994). 文体を決める三つの因子, 言語, **23**(2), 22–29.
安本美典 (2000). 因子分析による現代作家の統計的分析, *ESTRELA*, 5 月号, 2–6.
湯浅夏樹, 上田 徹, 外川文雄 (1995). 大量文書データ中の単語間共起を利用した文書分類, 情報処理学会論文誌, **36**, 1819–1827.
吉岡亮衛 (1999). 新書の数量的分析, 人文学と情報処理, **20**, 51–56.
Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship, *Biometrika*, **30**, 363–390.
Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge.
Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequencies of Language*, Harvard University Press, Cambridge, Massachusetts.

On Natural Language Statistical Information Processing

Mingzhe Jin

(Department of Social Information, Sapporo Gakuin University)

This paper describes the work done on statistical natural language processing. It is organized into two parts. Part One explains symbol and word-centered work in language processing with n-gram and Marcov models, etc. Part Two describes text processing: text classification (or categorization), authorship attribution (or stylometry) and information retrieval (or extraction), etc.

Key words: Statistical method, natural language processing, text processing.