仕事量基準を用いたコーパスからの定型表現の自動抽出

北 研二[†] 小倉 健太郎[††]
森元 逞[†††] 矢野 米雄[†]

自然言語処理では，形態素レベルにおける曖昧性や統語的な曖昧性などさまざまな曖昧性に対処する必要がある．このような曖昧性を軽減するための実際的な方法の一つに，慣用表現や定型表現等の複合的な表現を一つのまとまりとして処理することがあげられる．近年，世界各地で大規模コーパスの構築が行われており，大量な言語データが容易に手に入るようになってきている．本論文では，頻繁に使用される定型的な表現をコーパスから自動的に抽出する基準として「仕事量」という概念を導入する．仕事量は，いくつかの単語を一まとまりの単位と考えることにより，各単語を別個に処理するよりも，どれだけの処理が削減できるかということを定量的に測る尺度である．また，仕事量基準を用いた定型表現の自動抽出方法について述べ，提案した方法を実際の日本語のコーパスに適用することにより，その有効性を示す．また，コーパスから抽出された定型表現を形態素解析に組み入れることにより，単語区切りや単語誤りをはじめとする形態素レベルの誤りを削減できることを示す．

Automatically Extracting Frozen Patterns from Corpora Using Cost Criteria

KENJI KITA,[†] KENTARO OGURA,[††] TSUYOSHI MORIMOTO[†††] and YONEO YANO[†]

Natural language inherently contains many kinds of ambiguities such as morphological-level ambiguities and syntactic-level ambiguities. One practical and effective method for ambiguity resolution is to process idiomatic patterns and frozen patterns as one unit. By virtue of the development of many large corpora in recent years, a large volume of computer-readable data is now readily available. In this paper, we propose a method for automatically extracting frozen patterns from corpora by introducing cost criteria. By considering frozen patterns as one unit, cost criteria make it possible to measure quantitatively the extent to which processing is reduced. The proposed method is evaluated through experiments using a Japanese corpus. We also show that morphological-level errors are greatly reduced by incorporating frozen patterns into a morphological analysis module.

1. はじめに

機械翻訳をはじめとする自然言語処理において，慣用表現（イディオム）や定型表現は重要な役割を果たす．慣用表現は，複数の単語が連結した結果，個々の単語の意味からは出てこないような意味を全体として持つ．このため，これらの表現は一つのまとまった単位として処理する必要がある．定型表現とは文章中に頻繁に使用される表現であるが，慣用表現とは異なり，基本的には単語単位の処理により全体の意味をつかむことができる．しかし，処理効率の面からは，定型表現も一つのまとまりとして処理することが望ましい．これらの慣用表現や定型表現をどのようにして収集するかというのは，自然言語処理における重要な問題である．

また，最近は用例に基づいた自然言語処理（Example-based NLP）が注目を集めているが，用例ベースの自然言語処理においても定型表現に着目した翻訳方式が一部のシステムで取り入れられている．例えば，古瀬らの TDMT[1),2)] という日英機械翻訳システムでは，日本語の「～したいのですが」という表現が英語の “I would like to～” という表現に対応する等の対訳用例を積極的にシステムの変換知識として取り入れている．

慣用表現や定型表現を人手により網羅的に収集するという試みもいくつかの研究機関でなされているが[3)]，人手による作業では膨大な手間と時間が必要となるし，どれを慣用表現や定型表現とするのかという基準も曖昧にならざるをえない．

† 徳島大学工学部
Faculty of Engineering, Tokushima University
†† NTT 情報通信網研究所
NTT Network Information Systems Laboratories
††† ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

コーパスから単語間の共起関係を自動的に抽出する方法は，これまでにもいくつか提案されている．Church ら[4]は，二単語間の共起の強さをはかる尺度として，情報理論での概念である相互情報量を基に word association ratio と呼ばれるものを導入している．コーパス中から word association ratio の大きな単語対を取り出すことにより，“doctors”，“nurses” のような意味的な関連を持つ単語対や，“set off” 等の語彙・統語論的な関係 (lexico-syntactic relationship) を持つ単語対を自動的に抽出することができる．また，Smadja ら[5),6]は，単語間の共起の強さだけではなく，単語間の相対的距離（二単語間にいくつの単語があるか）等も考慮し，コーパスから連語 (collocation) を抽出する手法を提案している．ほかにも関連する研究として，情報理論でのエントロピーを評価基準に用いて単語あるいは音韻の連鎖を木構造モデルで表現する研究[7),8]や構文情報を利用した連語の抽出[9]，連続語の含有率を使った定型文の抽出[10]等の研究がある．

本論文では，定型表現をコーパスから抽出する基準として「仕事量」という尺度を導入し，仕事量基準を用いた定型表現の自動抽出方法について論じる．仕事量基準の特徴は，抽出された定型表現を実際に自然言語処理に応用する段階で，「どの程度処理が改善されるのか」ということが考慮されている点にある．従来の研究と異なり，「定型表現を使う立場」に立った抽出基準を用いている点が，本研究の特色であるといえる．

以下，本論文の第2章では，定型表現をコーパスから抽出する基準として「仕事量」を導入する．また，仕事量基準を用いた定型表現の自動抽出方法について論じる．第3章では，第2章で提案した方法を実際のコーパスに適用した結果について述べる．第4章では，本研究の応用として，定型表現を形態素解析に利用した結果について述べる．

2. 定型表現の自動抽出

従来，定型表現は，人手により抽出され分類整理されてきており，コンピュータにより自動的かつ体系的に定型表現を抽出する方法については，あまり研究されてこなかった．本節では，仕事量基準という概念を導入することにより，どういう表現を定型表現として抽出すれば，形態素解析をはじめとする自然言語処理が効率的に行えるかという観点から，定型表現を自動的にコーパスから抽出する方法について提案する．

なお，ここで提案する方法では，自然言語処理の効率化という観点から定型表現を抽出するため，基本的には出現頻度の高い表現が抽出対象となる．したがって，例えば「小骨一本抜かずに」というような慣用的に用いられている表現ではあるが出現頻度の小さな表現は本論文の対象外である．

2.1 基本的な考え方

コーパスから定型表現を自動的に抽出する上での基本的な考えは，コーパス中に頻繁に出現する単語列を抜き出してくるということである．しかし，単に単語列といっても，二つの単語から成るものもあるし，三つの単語から成るものもあり，単純に単語列の出現回数で比較するわけにはいかない．例えば，単語列αの出現回数がnであるとき，αの部分単語列βの出現回数は必ずn以上となるため，単純に出現回数だけからは真のパターンとなる表現を見つけ出すことはできない．

英語を例にとって説明しよう．コーパス中に “in spite” という単語列が110回，“in spite of” という単語列が100回，また “in spite of XYZ” という単語列が10回出現したとする．この場合，“in spite” が最も出現回数が多いからといって，単純に “in spite” を定型表現とするのには問題がある．なぜならば，“in spite” という用いられ方において，ほとんどの場合はその後に “of” を伴って用いられているからである．この例の場合，“in spite” ではなく “in spite of” を真の表現として抽出する必要がある．

したがって，コーパス中から定型表現を抽出する際には，単語列の出現頻度と単語列の部分的な系列（上の例の場合，“in spite”，“in spite of”，“in spite of XYZ” という系列）間の出現頻度の差を同時に考慮する必要がある．

2.2 仕事量基準

コーパスからの定型表現抽出のための基準として，「仕事量基準」と呼ばれるものを導入する．なお以下では，単語列αに対して，次のような表記法を用いる．

$|\alpha|$…単語列αの長さ（語数） (1)

$n(\alpha)$…単語列αのコーパス中での出現回数 (2)

仕事量基準は，簡単にいえば，単語列を一まとめにして一括して処理することにより，全体の仕事量を削減することができるという考えに基づいている．単語列αに対する仕事の削減量$K(\alpha)$を以下で定義する．

$$K(\alpha)=(|\alpha|-1)\times n(\alpha) \quad (3)$$

ここで，$K(\alpha)$ の意味付けを次のように行うことができる．いま，コーパス中に $|\alpha|$ 個の単語から成る定型表現 α があったとする．もし α を $|\alpha|$ 個の単語から成るものとして処理すれば，$|\alpha|$ に比例するだけの仕事量が必要である．しかし，α を一つの表現として処理すれば，仕事量は1でよい．つまり，$(|\alpha|-1)$ 分の仕事量が削減されたことになる．α がコーパス中に $n(\alpha)$ 回出現するならば，

$$(|\alpha|-1)\times n(\alpha) \quad (4)$$

分の仕事量が削減される．すなわち，$K(\alpha)$ は単語列 α を一括処理することにより削減される仕事量を表しており，$K(\alpha)$ の値の大きな単語列ほど一括処理する効果が大きいということになる．

基本的には，単語列 α に対して $K(\alpha)$ の値を計算し，値の大きな単語列を定型表現として抽出すればよい．しかし，ある単語列が別の単語列を部分列として含む場合を考慮する必要がある．いま，単語列 α が単語列 β の部分単語列であったとする．このとき，当然，

$$n(\alpha)\geq n(\beta) \quad (5)$$

が成り立つ．α と β の両方を慣用表現としてテキストを処理する場合を考える．β はテキスト処理の際に $n(\beta)$ 回参照されるが，α について考えると，α の出現回数 $n(\alpha)$ のうち $n(\beta)$ 回については β が参照されるので，純粋に α が参照されるのは，

$$n(\alpha)-n(\beta) \quad (6)$$

回だけである．したがって，α と β の両方を定型表現として採用する場合には，

$$K(\alpha)=(|\alpha|-1)\times(n(\alpha)-n(\beta)) \quad (7)$$

となる．

一般に，共通の部分単語列を持つ単語列の集合

α

$\alpha, \beta_i (i=1,\cdots,m)$

$\gamma_j, \alpha (j=1,\cdots,n)$

を同時に定型表現と考える場合，共通の部分単語列 α に対しては，

$$K(\alpha)=(|\alpha|-1)\times\Big(n(\alpha)-\sum_i n(\alpha,\beta_i) -\sum_j n(\gamma_j,\alpha)+\sum_i\sum_j n(\gamma_j,\alpha,\beta_i)\Big) \quad (8)$$

となる．

2.3 定型表現の現実的な抽出方法

コーパスから定型表現を抽出するためには，まずコーパス中を走査し，各単語列の出現回数を調べ，次に同一の系列に属する単語列（部分単語列の関係が成り立つような単語列）の K の値を動的に変更しながら，K の値の大きなほうから単語列を順次取ってくる必要がある．しかし，後者の操作を厳密に実現しようとすると，計算量が爆発的に増大する．なぜならば，単語列を一つ取ってくるたびに，その時点までに取られた単語列のすべてに対して同一の系列に属するか否かを調べて，K の値を再計算しなおさなければならないからである．

		K の値			
単語列	出現回数	最初	再計算1	再計算2	再計算3…
AB	100	100	70	60	40
ABCD	30	90	90	90	90
ABC	40	80	20	20	20
ABE	20	40	40	40	40
…	…	…	…	…	…

図1 K の再計算の例
Fig. 1 Example of re-computation of K values.

簡単な例で説明する．いま，単語A，B，C，D，E…に対して，単語列の出現回数が**図1**のようになったとする．このとき，まず K の値の最も大きい AB ($K=100$) が取られる．次に二番目に K の値の大きい ABCD ($K=90$) が取られるが，すでに取られている AB は，ABCD の部分単語列となっているので，AB に対する K の値は 70 に修正される（再計算1）．このあと，ABC が取られ，再度 AB に対し再計算が行われる（再計算2）．同様に，ABE が取られた時点でも，AB に対して再計算が行われる（再計算3）．上の例は比較的単純な例であるが，実際には単語列は一方向だけではなく，両方向（前後）に伸長するため，より複雑な計算が必要とされる．

われわれは，計算量の増加をおさえるために，再計算の操作を近似的に行う方法を考案した．この方法について，以下で説明する．

●再計算の近似方法1

K の値の再計算は単語系列中で隣り合った単語列どうしに限る．なお，二つの単語列が隣り合っているというのは，一つの単語列の左端あるいは右端に単語を一つ連接させることによって，もう一つの単語列が得られることをいう．上の例の場合，単語列 AB と ABCD は単語系列中で隣り合っていないので，再計算1は行わない．

同一の単語系列中で単語列の長さが2以上離れているものについての再計算は，隣り合った単語列どうしの再計算に帰着される．したがって，同

一の単語系列中の再計算は原則的には隣り合った単語列どうしで行えばよいが，Kの値の大きい順に取ってきたときに必ずしも隣り合った単語列が順に取ってこられるとは限らない．しかし，順に取ってくるかぎり，いずれは単語系列中で隣り合った単語列が現れ，その際に再計算をすれば十分であるという理由に基づいている．

●再計算の近似方法2

再計算は一度だけに限る．上の例の場合，ABCが取ってこられた時点で，ABの再計算が行われているので，ABEが取ってこられてもABの再計算は行わない．つまり，再計算3は行われない．これは，単語列をKの値の大きい順に取ってきているため，あとからされる再計算ほどKの値に与える影響が少ないという理由に基づいている．

3. 定型表現の抽出実験

前章で説明した方法を実際のコーパスに適用し，定型表現抽出の実験を行った．

3.1 言語データ

ATR自動翻訳電話研究所で作成された対話データベースADD[11),12)]から抜き出してきたデータを用いて実験を行った．対話データベースADDは，主に電話またはキーボードを介した目的指向型の対話に基づいて作成された話し言葉ないしは疑似話し言葉に関する大規模なデータベースであり，ADD中の各単語には形態素解析をはじめとする様々な事前分析により，表記，ひらがなによる読み，標準表現，品詞，活用型，活用形，音便等の属性が付与されている．ADDには，「国際会議」，「旅行」，「ホテル予約」などに関する対話が含まれているが，今回の実験には「国際会議」に関する対話を用いた．また，書き言葉に関する実験を行うために，新聞記事から抽出したデータを用いた．定型表現抽出実験に用いたデータのサイズを**表1**に示す．表で「異り単語列数」とあるのは，10単語連鎖までの単語列のうち異なった単語列の数を示している．

表1 定型表現抽出実験に用いたデータのサイズ
Table 1 Size of text data used for frozen pattern extraction experiments.

メディア	総文数	延べ単語数	異り単語列数
キーボード会話	1,197	12,669	57,445
電話会話	2,758	32,399	156,598
新聞記事	626	16,095	109,434

3.2 定型表現抽出結果

今回行った定型表現抽出実験では，単語列として10単語連鎖までを扱った．また，抽出された単語列のうちKの値の大きいもの0.5%についてのみ，Kの再計算を行った．**図2**に，それぞれのメディア（キーボー

順位	キーボード会話	電話会話	新聞記事
1	でしょ，う，か	でしょ，う，か	て，いる
2	です，か	ん，でしょ，う，か	し，て，いる
3	し，たい，の，です，が	ん，です，けれども	し，た
4	の，です，が	の，方	て，い，た
5	わかり，まし，た	あ，そう，です，か	と，いう
6	の，でしょ，う，か	そう，です，か	で，ある
7	はい，わかり，まし，た	はい，わかり，まし，た	で，は
8	の，です，か	と，いう，こと	し，て
9	ます，か	ん，です	に，は
10	どうも，ありがとう，ございまし，た	です，ね	か，も，しれ，ない
11	し，て	ます，でしょ，う，か	さ，れ，た
12	の，です	ああ，そう，です，か	さ，れ，て，い，る
13	お，願い，し，ます	し，て	し，て，いる，の，で，は，ない，か，と
14	て，おり，ます	と，思い，ます	で，は，ない
15	ませ，ん	な，ん，です，けれども	た，の
16	そう，です，か	の，方，に	に，よっ，て
17	たい，の，です，が	て，おり，ます	だろ，う
18	し，ます	の，方，は	し，て，い，た
19	と，思い，ます	わかり，まし，た	だっ，た
20	まし，た	に，なっ，て，おり，ます	に，も

図2 各メディアからの定型表現の抽出結果
Fig. 2 Examples of extracted frozen patterns.

ド会話，電話会話，新聞記事）から抽出された定型表現のうち，Kの値が上位20個のものを示す．比較的妥当と思える定型表現が抽出されており，われわれの提案した方法の有効性を示していると考えられる．

3.3 メディアによる定型表現の違い

有田[13]らは，メディアに依存する会話の性質をとらえるために，キーボード会話と電話会話を比較対照し，品詞の相対的出現頻度，埋め込み表現の深さ，指示詞と指示対象との距離，一発話内の単語数，談話構造などの観点からそれぞれの特徴を分析している．ここでは，キーボード会話と電話会話から抽出された定型表現の違いについて分析を行ったので，その結果について述べる．

電話会話では，「んでしょうか」，「んですけれども」，「んです」，「なんですけれども」のような準体助詞の「ん」になっているのに対して，キーボード会話では「のですが」，「のでしょうか」，「のです」のように準体助詞の「の」が含まれるややかたい表現になっている．ただし，「の」と「ん」の違いを除けば同じである定型表現が多い．

また，電話会話では，接続助詞の「けれども」や終助詞の「ね」が付いた表現が多くみられる．婉曲的な表現を表す接続助詞「が」が付いたものはキーボード会話と電話会話に共通にみられる．これらの表現は，新聞記事からの抽出結果にはみられないものであり，この現象はキーボード会話が話し言葉と書き言葉の中間的なところに位置していることを示唆している．

また，電話会話の表現の「の方」，「の方に」，「の方は」などは，「の」の前のものを直接指すことを避けた婉曲的な表現であり，話し言葉一般に共通する定型的な表現といえる．

今回行った実験では，対象領域がともに「国際会議」に関するものであったために，語彙の違いによる定型表現の大きな違いは見受けられず，同じ定型表現が多く抽出された．

4. 形態素解析への応用

コンピュータにより自然言語を処理する際には，まず形態素解析を行い，文章中に現れる単語を認識/同定し，各単語の文法的な属性を把握する必要がある．しかし，日本語では文章を文字で表記しようとするとき，英語のように単語ごとに分かち書きされることはなく，漢字と仮名の混ざったべた書きとなるため，形態素解析はそれほど単純な処理ではない．

一般に日本語の形態素解析においては，単語の区切りが一意に定まらず複数の可能性が生じる場合がある．このような形態素解析の曖昧性を軽減させる方法として，文字列として最も長い単語を優先的に採用する最長一致法や文中の文節数を最小にする候補を優先する文節数最小法などが考案されている．しかし，これらの方法単独では単語区切りの誤りを激的に軽減させることが難しいため，字種に関するヒューリスティクスや単語の共起頻度を利用する方法などが考えられている．

形態素解析の際に，定型表現を優先的に取り扱うことにより，形態素解析の誤りを削減することができる．前節で抽出された定型表現が，形態素解析において実際に有効に働くかを確認するため，定型表現を利用した形態素解析実験を行った．以下で，この実験について述べる[14]．

4.1 実験概要

今回の実験では，広く一般的に使われている最長一致法に単語の連接条件（例えば，自立語と付属語の接続条件等）を組み入れた方法を基準として，定型表現を利用し優先的にそれを取り扱った場合と，定型表現をまったく利用しない場合を比較した．

定型表現を利用した形態素解析では，定型表現は処理効率を上げるための一まとまりの処理単位と考えることができるが，形態素解析の結果得られた定型表現からその表現に含まれる形態素情報を取り出せなくてはならない．このため，各定型表現に対し，その表現に含まれる単語の種々の属性を記述している．例として，「わかりました」に対する形態素情報記述を**図3**に示す．

形態素解析の際に用いる定型表現を抽出するために用いたデータは約1万語，テスト・データは定型表現抽出に用いたデータとは異なる約3,000語のテキスト・データを用いた．形態素解析に使用した辞書の大きさは約14,000語である．

出現表記：	わかりました
単語分割：	わかり/まし/た
ひらがな表記：	わかり/まし/た
標準表現：	分かる/ます/た
品詞・活用型・活用形：	(本動詞，五段活用，連用形)/
	(助動詞，　，連用形)/
	(助動詞，　，終止形)

図3 定型表現の形態素情報記述の例
Fig. 3 Description of morphological information for a frozen pattern.

4.2 実験結果

形態素解析により得られた結果を，単語区切りの誤り率，単語の誤り率，項目の誤り率の三つの基準で評価した．なお，項目とは，形態素解析の結果得られる単語，ひらがな表記，品詞，活用型，活用形，音便の六つの項目のことである．項目誤り率を評価基準とした理由は，間違った形態素解析の結果を人手で修正する際に，項目誤り率が実質的な修正作業量を把握するための尺度と考えられるためである．それぞれの誤り率は，次のように定義される．

$$\text{単語区切りの誤り率}=\frac{\text{単語区切り誤り総数}}{\text{正解の単語総数}}\times 100 \quad (9)$$

$$\text{単語誤り率}=\frac{\text{単語誤り総数}}{\text{正解の単語総数}}\times 100 \quad (10)$$

$$\text{項目誤り率}=\frac{\text{項目誤り総数}}{\text{正解の項目総数}}\times 100 \quad (11)$$

単語区切りの誤りは，単語の分割が正しく行われたか否かで判定する．すなわち，単語の分割さえ合っていれば，単語の各形態素情報に誤りがあっても正解とする．単語誤りでは，単語のどれか一つの項目でも誤りがある場合は，その単語を誤ったと判定する．項目誤りは，単語の各項目に対し，どのくらい誤り項目数があったかで計算する．なお，評価に用いたテキスト・データには，あらかじめ人手によって，正解の形態素情報が付けられており，これと形態素解析の結果を比較することにより，各誤り率を計算した．

結果を**表2**に示す．単語区切りの誤り率，単語の誤り率，項目の誤り率のいずれに対しても，定型表現を利用した場合のほうがよい結果を示した．定型表現を利用する場合には，処理単位が一まとまりの単語列となっているために，特に単語区切りの誤りに対して有効であることが読みとれる．

今回の形態素解析実験では，定型表現を利用した場合，「おうかがい」を「おう/か/が/い」（正解は「お/うかがい」）と誤るような接頭辞に起因する誤りや，「本動詞＋接続助詞＋補助動詞」を「本動詞＋接続詞＋本動詞」としてしまう誤り（例：「していただく」）に特に有効であることがわかった．

表2 形態素解析実験の結果
Table 2 Experimental results of a morphological analysis with and without frozen patterns.

定型表現の利用	メディア	項目誤り率	単語誤り率	区切り誤り率
なし	キーボード会話	14.9	29.2	11.0
	電話会話	13.5	28.8	8.1
あり	キーボード会話	8.6	20.8	2.7
	電話会話	9.0	22.3	2.8

5. おわりに

本論文では，頻繁に使用される定型的な表現をコーパスから自動的に抽出する基準として「仕事量」というものを提案した．仕事量は，いくつかの単語を一まとまりの単位と考えることにより，各単語を別個に処理するよりも，どれだけの処理が削減できるかということを定量的に測ることを可能にする．また，仕事量基準を用いた定型表現の自動抽出方法について述べた．この方法をATR対話データベース中のキーボード会話と電話会話，および新聞記事に適用し，提案した方法の有効性を示した．また，コーパスから抽出された定型表現を形態素解析に組み入れることにより，単語区切りや単語誤りなどを削減できることを示した．

現在の方法では，単語の連続的な連鎖から成る定型表現しか扱うことができないが，定型表現にはギャップを持つもの（「まるで〜のようだ」等）もある．今後の課題として，ギャップを持つ定型表現を扱えるように，本研究を拡張したいと思っている．また，本論文で提案した方法を，相互情報量等の情報理論的な尺度を用いた方法と比較することも今後の課題として上げられる．

謝辞 本研究の大部分はATR自動翻訳電話研究所で行われた．研究の機会を与えてくださったATR自動翻訳電話研究所 樽松明社長（現在 電気通信大学教授），ならびに有益な助言をいただいたATR自動翻訳電話研究所の皆様に感謝いたします．

参考文献

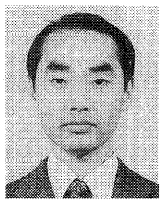
1) 古瀬 蔵，飯田 仁：変換と解析の協調的処理による翻訳手法—変換主導型翻訳手法—，情報処理学会自然言語処理研究会報告 87-4, pp. 27-34 (1992).
2) Furuse, O. and Iida, H.: Cooperation between Transfer and Analysis in Exampled-based Framework, *Proc. of COLING-92*, pp. 645-651 (1992).
3) 首藤公昭，吉村賢治，武内美津乃，津田健蔵：日本語の慣用的表現について，情報処理学会自然言語処理研究会報告 66-1 (1988).
4) Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol. 16,

No. 1, pp. 22-29 (1990).
5) Smadja, F. A. and McKeown, K. R.: Automatically Extracting and Representing Collocations for Language Generation, *Proc. of ACL-90*, pp. 252-259 (1990).
6) Smadja, F. A.: Fron N-grams to Collocations: An Evaluation of Xtract, *Proc. of ACL-91*, pp. 279-284 (1991).
7) Bahl, L. R., Brown, P. F., de Souza, P. V. and Mercer, R. L.: A Tree-Based Statistical Language Model for Natural Language Speech Recognition, *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP 37, No. 7, pp. 1001-1008 (1989).
8) 田本真詞, 伊藤克亘, 田中穂積: 木構造を用いた音韻連鎖統計モデル, 電子情報通信学会音声研究会報告 SP 92-23, pp. 77-84 (1992).
9) Basili, R., Pazienza, T. and Velardi, P.: A Shallow Syntactic Analyser to Extract Word Associations from Corpora, *Literary and Linguistic Computing*, Vol. 7, No. 2, pp. 113-123 (1992).
10) 加藤真人, 相沢輝昭: 外電ニュースの定型文抽出とその英日機械翻訳, 情報処理学会自然言語処理研究会報告 93-2, pp. 7-14 (1993).
11) 江原暉将, 小倉健太郎, 篠崎直子, 森元 逞, 榑松 明: 電話またはキーボードを介した対話に基づく対話データベース ADD の構築, 情報処理学会論文誌, Vol. 33, No. 4, pp. 448-456 (1992).
12) 匂坂芳典, 浦谷則好: ATR 音声・言語データベース, 日本音響学会誌, Vol. 48, No. 12, pp. 878-882 (1992).
13) 有田英一, 小暮 潔, 野垣内出, 前田広幸, 飯田仁: メディアに依存する会話の様式—電話会話とキーボード会話の比較—, 情報処理学会自然言語処理研究会報告 61-5 (1987).
14) 小倉健太郎, 森元 逞: 形態素解析への慣用表現と頻度情報適用による定量的効果, 情報処理学会自然言語処理研究会報告 90-6 (1990).

(平成5年3月18日受付)
(平成5年7月 8 日採録)

北 研二（正会員）

1957年生. 1981年早稲田大学理工学部数学科卒業. 1983年から1992年まで沖電気工業(株)勤務. この間, 1987年から1992年まで ATR 自動翻訳電話研究所に出向. 1992年9月から徳島大学工学部勤務. 現在, 同助教授. 工学博士. 確率・統計的自然言語処理, 音声認識, 音声言語統合方式の研究に従事. 電子情報通信学会, 日本音響学会, ACL 各会員.



小倉健太郎（正会員）

1954年生. 1978年慶應義塾大学工学部管理工学科卒業. 1980年同大学院修士課程修了. 同年日本電信電話公社入社. 1986年 ATR 自動翻訳電話研究所に出向. 1990年日本電信電話(株)に復帰. 現在, 情報通信網研究所知識処理研究部主任研究員. 機械翻訳の研究に従事. 電子情報通信学会, 計量国語学会各会員.

森元 逞（正会員）

1946年生. 1968年九州大学工学部電子工学科卒業. 1970年同大大学院修士課程修了. 同年, 日本電信電話公社に入社. 以来, 同社電気通信研究所にて, オペレーティングシステム等の研究開発に従事. 1987年より ATR 自動翻訳電話研究所へ出向. 音声言語翻訳システム, 特に, 音声言語統合方式, 音声言語翻訳方式の研究を行っている. 現在, ATR 音声翻訳通信研究所第4研究室室長. 電子情報通信学会, 人工知能学会各会員.

矢野 米雄（正会員）

1945年生. 1969年大阪大学工学部通信工学科卒業. 1974年大阪大学大学院工学研究科博士課程修了. 工学博士. 同年徳島大学工学部助手. 1975年同講師. 1981年同助教授. 1990年同教授. 1979～1980年米国イリノイ大学 Computer-based Education Research Laboratory 客員研究員. 環境型知的 CAI システム, 人脈活用システム, ゲーム環境の研究に従事. CAI 学会理事, 同関西支部副支部長. 日本教育工学協会理事. 日本教育工学会評議員. 電子情報通信学会, 米国 IEEE 各会員. 「実例パソコン分子モデリング」(講談社), 「ソフトエデュケーション」(監修, くもん出版).