音声言語情報処理 29－32
(1999. 12. 21)

ニュース音声認識のための (n≧4)-gramを併用する言語モデル

加藤直人[*1] 浦谷則好[*2] 江原暉将[*1] 安藤彰男[*1]

*1 NHK放送技術研究所
〒157-8510 東京都世田谷区砧 1-10-11
{katonao, eharate, ando}@strl.nhk.or.jp

*2 ATR音声翻訳通信研究所
〒619-0288 京都府相楽郡精華町光台 2-2
uratani@itl.atr.co.jp

あらまし 音声認識の精度向上には言語制約が強い言語モデルを構成すること必要であり,その一つの方法がタスク適応である.一方で,タスク適応しすぎると頑健性が損なわれるという問題がある.本稿では(n≧4)-gramを利用することによりタスクへの適応をしつつ,2,3-gramも利用することで頑健性もそれほど損なわない言語モデルについて述べる. 提案する言語モデルでは(n≧4)-gramを,従来のn-gramのように宣言的知識として記憶するのではなく,単語出現位置辞書という概念を導入して手続き的知識として記憶することによりそれほどデータ量を増やすことなく利用している.本言語モデルを放送ニュースに応用し,そのperplexityによる評価実験を行ったところ,良好な結果を得た.

キーワード 言語モデル, 音声認識, n-gram, タスク適応, テレビニュース, perplexity

A New Language Model by using (n ≧ 4)-gram for Broadcast News Speech Transcription

Naoto Katoh[*1] Noriyoshi Uratani[*2] Terumasa Ehara[*1] Akio Ando[*1]

*1 NHK Science and Technical Research Laboratories
1-10-11Kinuta, Setagaya-ku, Tokyo 157-8510, Japan
{katonao, eharate, ando}@strl.nhk.or.jp

*2 ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
uratani@itl.atr.co.jp

Abstract Language model adaptation is one of the important methods to construct a speech recognition system for practical use. The conventional adaptation methods adjusted n-gram estimated from various task corpora to ones from a specific task corpus. However the methods are not so effective in some tasks such as TV news, because some of TV news does not use news scripts.
This paper proposes a new language model for Broadcast news speech transcription. Our model can not only adapt to a specific task but also deal with the more tasks by dynamically using (n ≧ 4)-gram and 2,3-gram. The proposed method can reduce amount of (n ≧ 4)-gram data by registering it as procedural knowledge through WPD (Word Position Data). The WPD represents each position of words in a task corpus and is automatically composed of the corpus.
We conducted a series of experiments to evaluate our model and obtained a good result.

key words language model, speech recognition, n-gram, task adaptaion, Broadcast news, perplexity

1　はじめに

聴覚障害者などからテレビ番組，特に「ニュース」の音声を字幕化してほしいという要望が強い．米国では人手でリアルタイムに入力しているが，日本語の場合，漢字に変換する作業等が必要なためリアルタイムで入力するのは熟練した人でも難しい．そこで，このような作業に応用することを目的に音声認識の研究を行っている[1]．

現在，音声認識の言語モデルにはその実現の容易さからn-gramモデルが使われる．実際には記憶量の制約からn = 2,3 (2,3-gram) が使われることが多い．さらに，認識精度を向上させるために**タスク適応**が行われる[2]～[8]．タスク適応は通常，一般の言語モデルをさまざまなタスクから集めた（大量の）コーパスから構築し，あるタスク内で集められた（少量であることが多い）コーパスを用いて確率を補正することにより行われる．実際，放送ニュースの音声認識に関しても，直近の記者原稿を利用することによる効果を確認している[9]．

しかし，一方で，タスク適応は扱うタスクが非常に狭く限定されるような場合には有効であるが，扱うタスクがあまり限定できない場合にはむしろマイナスとなる．

最も望ましい言語モデルは，そのタスク内の言語表現である場合にはタスク用の言語モデルを使い，タスク内に収まらないような言語表現の場合には一般のモデルを使うことである．さらにこのようなことが自動的に行われる必要がある．

本稿では一般の言語モデルを使用しつつ，タスク適応も自動的に行う言語モデルについて述べる．提案する言語モデルでは，一般の言語モデルとして2,3-gramを使い，タスク適応として(n≧4)-gramを併用している．ここで，(n≧4)-gramを利用するにあたって問題となるのは記憶量である．これを従来の2,3-gramのように宣言的知識として記憶していたのでは記憶量が爆発してしまう．提案する言語モデルでは(n≧4)-gramを手続き的知識として記憶することによって実現可能な記憶量となる．その際に中心的役割をするのが**単語出現位置辞書**である．単語出現位置辞書はタスク適応するコーパスから自動的に生成される．本言語モデルをperplexityによる評価実験を行った結果についても述べる．

2　アナウンス原稿と記者原稿

ニュース番組でアナウンサーが読む原稿（**アナウンス原稿**）は，一部に割り込みで入ってくるものがあるものの，その多くは記者がある出来事について書いた原稿（**記者原稿**）を，本番直前まで手書きで校正することによって作成される（図1）．また，このような校正はあまり行われない場合も少なくない．

ニュース音声認識の対象としているのはアナウンス原稿であるので，これから言語モデルを構築することが望ましい．しかし，記者原稿はワープロで書かれているので電子化されているが，アナウンス原稿は電子化されていない．そこで，言語モデルの構築には記者原稿を使用している．アナウンス原稿は記者原稿から作成されているので，記者原稿と一致している個所も多く，極端な場合にはまったく一致する文もある．このような特徴を言語モデルに利用すれば，強い言語制約となることが期待できる．

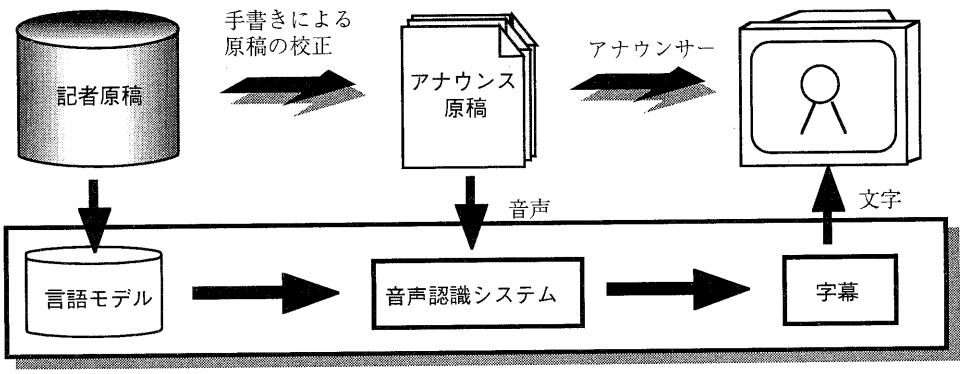


図1　記者原稿と言語モデルの関係

3 (n≧4)-gramを併用する言語モデル

提案する言語モデルでは2,3-gramとともに**(n≧4)-gram**も用いる．2,3-gramによる言語モデルの構築の際には，一般性を持たせるために過去数年分のコーパスと，ある程度のニュースの時事性（弱い適応化）を持たせるために，重みづけした（m倍した）直近のコーパスを用いている．(n≧4)-gramによる言語モデルでは，直近のコーパスから自動的に生成された**単語出現位置辞書**を利用している．

3．1 単語出現位置辞書

単語出現位置辞書とは記者原稿に出現した単語の出現位置（**単語出現位置**）を集めた辞書である．ここで，単語出現位置は記事の分野，その分野における記事の出現位置，その記事における単語の出現位置の3つ情報によって次のように7桁の数で定義している．

【単語出現位置の定義】

単語出現位置 XYYYZZZ（7桁）

X　（上位1桁目）：分野番号
（1：政治，2：経済，3：社会，4：国際，5：スポーツ，6：ネット，7：首都圏）

YYY（2～4桁目）：その分野における記事位置

ZZZ（5～7桁目）：その記事における単語出現位置

例えば，単語出現位置7002026は，分野が首都圏(7)，記事の出現位置が2番目（002），単語の出現位置26番目（026）ということを表す．

単語出現位置辞書は直近（例えば前日，当日）の記者原稿から自動的に生成される（図3）．記者原稿の各単語に単語出現位置を付与し，単語ごとに集計し，単語を見出しとしてソートする．

3．2 言語モデルでの(n≧4)-gramの利用

言語モデルで，単語出現位置辞書を使った(n≧4)-gramの利用法について説明する．

まず，それまでの各単語で，単語出現位置辞書を引くことによりその単語単語出現位置を求める．今，i番目の単語をw^i，それまでに連続しているn-gramの長さの最大値を*N-1*（すなわち*N*-gramを利用するため），単語出現位置を*c*とおく．次単語$w^{i+1}{}_k$ $(k=1,2,\cdots,k_0)$の生起確率をp'_kを，2,3-gramによる確率p_kに基づいて式（1a），（1b）で動的に再計算することによって行う．

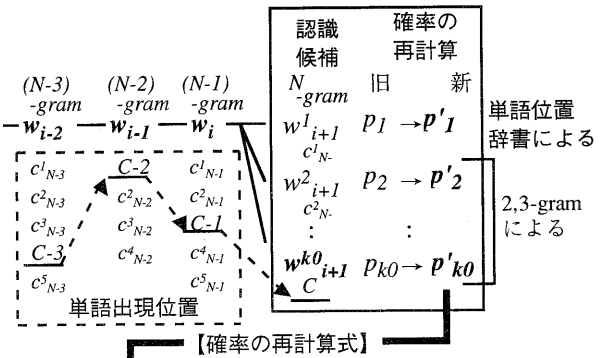
式（1a）は(n≧4)-gramの利用する場合である．前単語の単語出現位置*C*から予測される単語$w^{i+1}{}_1$（$k=1$のときとする）では，その新たな確率p'_1（確率値）を$N\geqq 4$のときに*N*に応じて計算する．ここで，定数λやq_0はコーパスから求めることも可能である．式（1b）は2,3-gramを利用する場合である．2,3-gramから予測される候補$w^{i+1}{}_k$（それを$2\leqq k\leqq k0$のときとする）の場合であり，2,3-gramの確率p_kに応じて残りの$1-p'_1$を分配する．もちろん，式（1b）の単語$w^{i+1}{}_k$（$2\leqq k\leqq k0$）の中に$w^{i+1}{}_1$が含まれているときはそれを除外する．

今回は，(n≧4)-gramを利用するために$N\geqq 4$としているが，$N\geqq N_0$のときのみ上記を実行するようにしておけば，$(n\geqq N_0)$-gramを利用する言語モデルともなる．

4 評価実験

本稿で提案する言語モデルに対してperplexityによる評価実験を行った．言語モデルの構築の際には過去3年分の記者原稿と，前日と当日の直近2日間の記者原稿を用いて，以下の2つのモデルを作成し比較した．

ベースラインモデル
＝3年分の記者原稿　←一般言語モデル
＋m×直近2日の記者原稿　←タスク適応



$$\begin{cases} p'_1 = q_0 \times (1 - exp^{-\lambda N}) & if\ k=1 \quad (1a) \\ p'_k = (1 - p'_1) \times \dfrac{p_k}{\sum_{k=2}^{k0} p_k} & if\ 2\leqq k\leqq k0 \quad (1b) \end{cases}$$

（ここで，q_0，λは定数，$0\leqq q_0\leqq 1$，$\lambda > 0$）

図2　単語出現位置辞書による確率の再計算

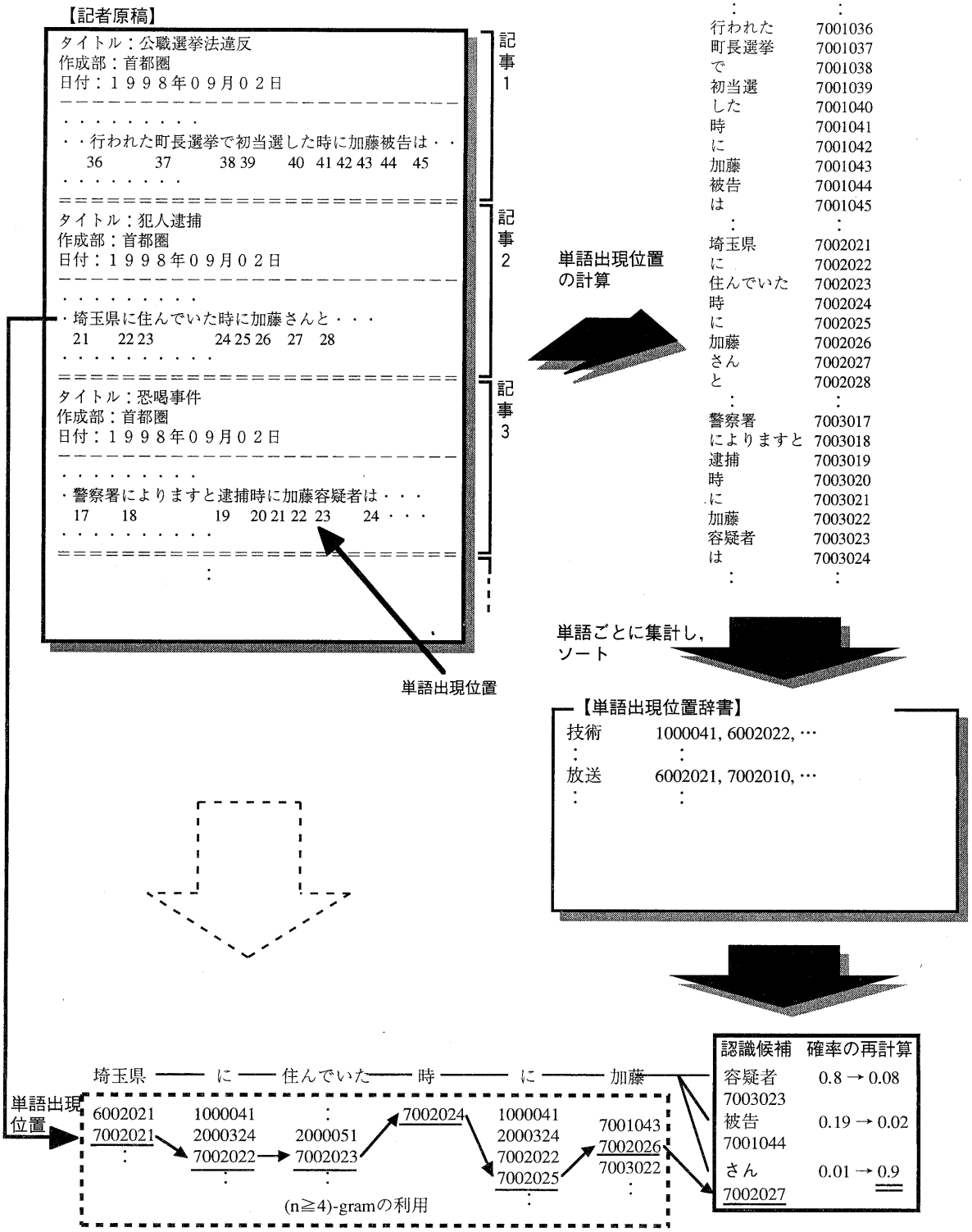


図３　単語出現位置辞書の自動生成と音声認識への応用

提案する言語モデル

＝３年分の記者原稿　　←一般モデル

＋m×直近２日の記者原稿　←弱いタスク適応

＋(n ≧ 4)-gram　　←強いタスク適応

また，パラメータである λ や q_0 は今回は $\lambda = 2,3$, $q_0 = 0.9$ としている．

評価データには１９９８年９月３０日のニュースのアナウンス原稿*1を，以下の３つの場合に分けて使用した．

評価データＡ：元となる記者原稿とほぼ同じ（３６記事）

評価データＢ：元となる記者原稿とある程度同じ（２０記事）

評価データＣ：元となる記者原稿なし（１４記事）

ただし，これらの評価データの分け方は，m= 100 におけるベースラインモデルの perplexity の値 perplexity0 が

評価データＡ：perplexity0 ＜ 20 のとき

評価データＢ：20 ≦ perplexity0 ＜ 100 のとき

評価データＣ：100 ＜ perplexity0 のとき

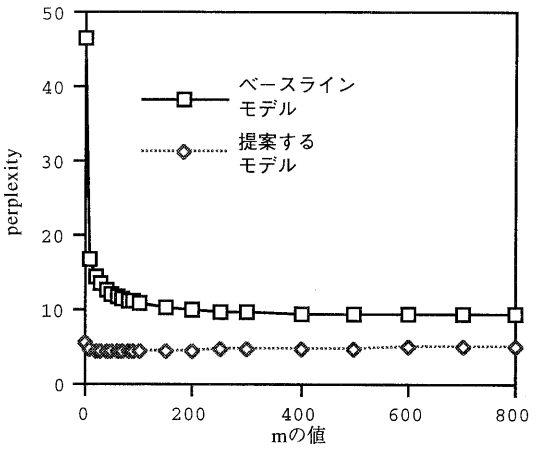
とした．この境界値は実際の perplexity の結果をいくつ見て決めた．

さまざまな m で実験した．結果を図４に示す．図４を見ると，いずれの場合も提案する言語モデルがベースラインモデルより優れていることがわかる．評価データＡは現在我々が音声認識の対象として考えているものであるが，ベースラインモデルではm＝600のとき，提案する言語モデルではm＝50 のときに最小となっている．すなわち，提案するモデルでは直近の２日の記者原稿を重みづけが小さく，(n ≧ 4)-gram が有効に働いているのが

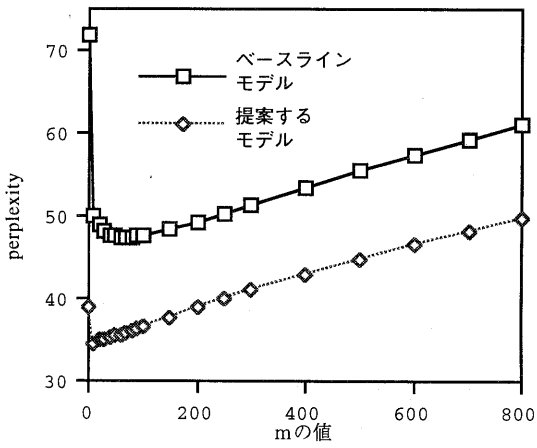
表１　評価データＡでの perplexity が最小となる場合

評価データ	評価データＡ	評価データＢ	評価データＣ
ベースラインモデル（α）	９．６	５７．４	３８３．５
提案するモデル（β）	４．５	３５．４	２４９．０
削減率（$(\alpha-\beta)/\alpha$）	５３．１％	３８．３％	３５．１％

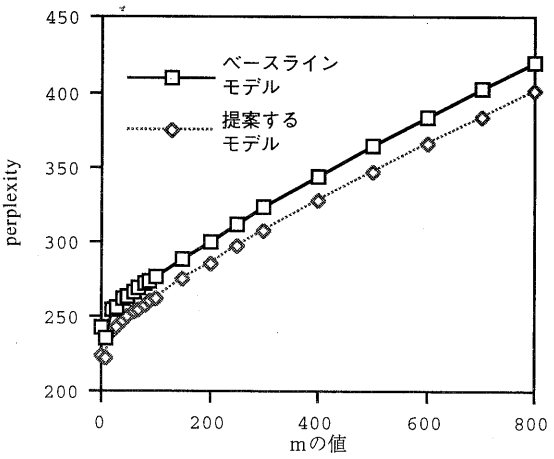
＊１正確にはアナウンス原稿ではなく，アナウンサーが話したものを書き起こした原稿である．したがって，不要語等が含まれる．



（ａ）評価データＡに関する実験結果



（ｂ）評価データＢに関する実験結果



（ｂ）評価データＣに関する実験結果

図４　それぞれの評価データにおける perplexity

わかる．このときの評価データB，Cの場合のperplexityも合わせて表1に示す．表1を見ると，提案する言語モデルでは評価データB，Cでも削減率が大きいことがわかる．これはベースラインモデルが評価データAに関してタスク適応しているために，特に評価データCでのperplexityが大きい．一般モデルはベースラインモデルのm＝1の場合であり，perplexity＝242.5である．これに対して提案するモデルでもperplexity＝249.0であるので一般モデルとほとんど変わらなく，一般言語モデルとしての特性も失っていないことがわかる．

5　おわりに

n≧4であるn-gramを利用した言語モデルについて述べ，その有効性を評価実験により確認した．今後は実際に計算機上にインプリメントし，認識実験によって本言語モデルの検証を行う必要があろう．

今回は単語出現位置が連続している場合のみを用いたが以下のような使い方や，音声認識の改良も考えられる単語出現位置辞書はまた次のような応用も考えられる．

・単語出現位置の連続していなくとも値が近い単語候補を優先することにより構文構造の変化に対応．
・上位1桁目を使うことにより単語の分野制約．
・単語出現位置辞書により記者原稿での出現位置に句読点を挿入．
・実験結果のようにperplexityがここまで小さくなるので，デコーダー上での第2パスの削除や最良優先探索の導入による音声認識の高速化．

これらの詳細について機会をあらためて報告したい．

謝辞

本研究を進めるにあたって適切な助言をいただいた当研究所音声認識・言語情報処理グループに感謝する．

【参考文献】

[1]今井　亨，小林彰夫，尾上和穂，安藤彰男，“ニュース番組自動字幕化のための音声認識システム，” 情報処理学会研究会，HI80-11(SLP-23-11), pp.59-64, 1998.

[2]R. Kneser and V. Steinbiss, “On the dynamic adaptaion of stochastic language models,” Porc. IVASSP’93, Vol.2, pp.585-588, 1993.

[3]P.Clarkson and A. Robinson, “Language model adaptation using mixtures and exponential decaying cache,” Poc. ICASSP’97, Vol2, pp.799-802, 1997.

[4]S. Matsunaga, T. Yamada, and K. Shikano, “Task adaptaion in stochastic language models for continuous speech recognition,” Proc. ICASSP’92, Vol.1, pp.165-168, 1992.

[5]政瀧浩和，匂坂芳典，久木和也，河原達也，“最大事後確率推定によるN-gram言語モデルのタスク適応，” 信学論 (D-II), Vol.J81-D-II, No.11, pp.2519-2525, 1998.

[6]伊藤彰則，好田正紀，“対話音声認識のための事前タスク適応の検討，” 電子情報通信学会研究会報告，SP96-81, pp.91-98, 1996.

[7]赤松裕隆，甲斐充彦，中川聖一，“新聞・ニュース文の大語彙連続音声認識，” 情報処理学会研究会報告，SLP98-21-11, pp.97-104, 1998.

[8]政瀧浩和，松永昭一，匂坂芳典，“品詞および可変長単語列の複合N-gramの自動生成，” 信学論 (D-II), Vol.J81-D-II, No.9, pp.1929-1936, 1998.

[9]小林彰夫，今井　亨，安藤彰男，中林克己，“ニュース音声認識のための時期依存言語モデル，” 情報処理学会論文誌，Vol.40, No.4, pp.1421-1429, 1998.